Ungváry Rudolf– Orbán Éva

OSZTÁLYOZÁS ÉS INFORMÁCIÓKERESÉS

Kommentált szöveggyűjtemény

Második kötet:
Az információkeresés és elmélete

Ungváry Rudolf– Orbán Éva

OSZTÁLYOZÁS ÉS INFORMÁCIÓKERESÉS

Kommentált szöveggyűjtemény

Második kötet:
Az információkeresés és elmélete

Országos Széchényi Könyvtár
Budapest
2001

Ez a könyv a Művelődési és Közoktatási Minisztérium által a felsőfokú könyvtáros–informatikus-képzés korszerűsítésére kiírt tankönyvpályázat nyertes munkája.

UNGVÁRY Rudolf (1936–)
Osztályozás és információkeresés : kommentált szöveggyűjtemény / Ungváry Rudolf, Orbán Éva ; [közread. az] Országos Széchényi Könyvtár. – Budapest : OSZK, 2001. – 2 db ; 24 cm
ISBN 963 200 424 8 ö
Tft.: Orbán Éva (1957–). – Mt.: Országos Széchényi Könyvtár (Budapest) (közread.)

1. köt., Az osztályozás és elmélete. – 543 p.
ISBN 963 200 425 6
2. köt., Az információkeresés és elmélete. – 535 p.
ISBN 963 200 426 4

ISBN 963 200 424 8 ö
ISBN 963 200 426 4

Kiadja
az Országos Széchényi Könyvtár
Felelős kiadó: Dr. Monok István főigazgató
Nyomdai kivitelezés: AKAPRINT Nyomdaipari Kft.
Felelős vezető: Freier László
Munkaszám: 24759

Szerkesztették:
Ungváry Rudolf és Orbán Éva

A bevezetőket és kommentárokat írta:
Ungváry Rudolf

Lektorálta:
Horváth Tibor

Fedélterv:
Hangay Gabriella és Markó Natália

Technikai szerkesztő:
Korpás István

Szakfordítók:

Bendl János	Pesthy Mónika
Czékli Béla	Somi Kovács Mária
Csuhaj Erzsébet	Strelisky János
Enyedi Ágnes	Szőnyi Katalin
Gergely Júlia	Tamás Gáspár Miklós
Gőri Bella	Ungváry Rudolf
Horváth Magda	Válas György
Medve Éva	Varga Ildikó
Németh Erzsébet	Zsadon Béla

Orbán Éva

Szakmai lektorok:

Csát József	Pap Mária
Fogarassy Miklós	Sárdy Péter
Gerő Vera	Sipos Márta
Hegyközi Ilona	Ungvári Gyula
Horváth Tibor	Ungváry Rudolf
Hoványi János	Vadász Ágnes
Orbán Éva	Varga Dénes
Pálvölgyi Tibor	Zöldi Péter

Nyelvi lektorok:

Fogarassy Miklós	Szász Anna
Sipos Márta	Ungváry Rudolf

Korrektorok:
Burány Tamásné és Korpásné Balczer Gabriella

ÁTTEKINTÉS

I. RÉSZ
A GÉPESÍTÉS KEZDETEI

II. RÉSZ
ON-LINE ÉS INTERNET

TARTALOMJEGYZÉK

I. RÉSZ
A GÉPESÍTÉS KEZDETEI

II. RÉSZ
ON-LINE ÉS INTERNET

„...and in such indexes, although small pricks to their subsequent volumes, there is seen the baby figure of the giant mass of things to come."

„...s az ilyen kis mutatók (bár pontok csak az általuk jelölt fóliánsokhoz képest) gyermeki alakban már mutatják az eljövendő dolgok irdatlan méretét."

William Shakespeare: Troilus és Cressida, I. felv. 3. szín, 343.

„...For, by the way, I'll sort occaion, as index to the story we late talked of, to part the Qquen's poud kindred from the king."

„...mert útközben, a megbeszélt ügyek kezdőjeléül, a királyné családját a herceg mellől majd eltávolítom."

William Shakespeare: III. Richárd, II. felv. 2. szín, 149.

„...an index, and obscure prologue to the history of lust."

„...címlap, a gyönyör történetének kétes hírű bevezetője."

William Shakespeare: Othello, I. felv. 1 szín, 263.

„...the flattering index of a direful pageout."

„...víg címlapja egy szörnyű színdarabnak..."

William Shakespeare: III. Richárd, IV. felv. 4. szín, 85.

„Hey! What act, that roars so luod and thunders in the index?"

„Haj! Miféle tett az, mely már a címlapon így mennydörög."

William Shakespeare: Hamlet, III. felv. 4. szín, 52.*

---

* *Az idézetek forrása:*
Schmidt, Alexander: Shakespeare-lexicon. A completes dictionary of all the English words, phrases and constructions in the works of the poet. Vollständiger englischer Sprachschatz… – Rev. and enl. by Gregor Sarrazin. Vol. 1–2. – 5. ed. – Berlin: de Gruyter, 1962.

I. RÉSZ
A GÉPESÍTÉS KEZDETEI

AZ INFORMÁCIÓKERESŐ GONDOLKODÁS KEZDETEI

„Az emberi agy nem így működik. Asszociációkat követ. Megragad valamit és már kapcsol is tovább...”

Vannevar Bush: As we may think [Úgy, ahogy gondolkodunk]
In: Atlantic Monthly, 1945 július.

A szakirodalom mennyiségének növekedésével párhuzamosan a század első évtizedeiben egyre nyilvánvalóbbá vált, hogy mind az Egyetemes Tizedes Osztályozás, mind pedig a hagyományos – Cutter-féle – tárgyszavas osztályozási eljárások túlságosan nehézkesen használhatók dokumentációs célokra. Az utóbbiakról – mivel természetes nyelven alapultak – a dokumentátorok jelentős részének mégis volt olyan érzése, hogy elvileg alkalmasabb lehetne a dokumentációs mikrofeltárás céljaira, mint az ETO és a hozzá hasonló mesterséges nyelven alapuló rendszerek, s eme ösztönös várakozásuk következtében annál nagyobb lett idővel a kiábrándulás. A könyvek keresésére készült tárgyi (tárgyszavas) katalógusok a részletes és gyors információk szolgáltatására törekvő szakemberek szemében a harmincas–negyvenes évekre az alkalmatlanság jelképeivé váltak.

A két szakterület fokozódó eltávolodásának egyik első jele volt, hogy 1909-ben az Amerikai Könyvtáros Társaságból (American Library Association; ALA) kivált a Szakkönyvtárak Társasága (Special Library Association; SLA) és egyesült az Amerikai Dokumentációs Intézettel (American Documentation Institute). A könyvtárosok és dokumentátorok közötti ellentéteket jellemezve írta *Jesse H. Shera* egyik 1977-ben megjelent történeti visszapillantásában: „számos könyvtáros – nevüket borítsa jótékony homály – a dokumentációt »amatőrök által eltorzított könyvtárosságnak« tekintette”.[1]

1 Shera, J. H.: Könyvtárosság, dokumentáció, tájékoztatástudomány. In: Könyvtári Figyelő, 1970, 16. évf., 3. sz., p. 222–229.

Részletes történeti áttekintés található magyarul:

Saracevic, T.: Az információ eredete, fejlődése és kapcsolatai. In: Tudományos és Műszaki Tájékoztatás, 1994, 41. évf., 7–8. sz., p. 313–320.,

Meadow, Ch. T.: Vissza a jövőbe: az adatbázisipar kronológiája. In: Tudományos és Műszaki Tájékoztatás, 1989, 36. évf., 6. sz., p. 266–271. továbbá angolul:

Lilley, Doroty B, Trice, Ronald W.: History of information science, 1945–1985. – San Diego… : Academic Press, 1989. 181 p. – (Library and Information Science).

A második világháború után felgyorsult a dokumentáció fejlődése. Az 1948–1968 közötti korszakot az „innováció, a szakmai kommunikáció ugrásszerű bővülése, a növekedés és a kutatások elmélyülése”[2] jellemezte. A dokumentáció fokozatosan átalakult, kibővült információtudománnyá (tájékoztatástudománnyá). Ebben az időszakban alkotta meg *Claude Shannon* (1916) és *Warren Weawer* a matematikai információelméletet, *Norbert Wiener* a kibernetikát, kezdték építeni az első kereskedelmileg forgalmazott második generációs nagyszámítógépeket.

Fontos szerepet játszott, hogy 1950-ben fogtak hozzá az Egyesült Államokban a tudományszervező *Vannevar Bush* kezdeményezésére a National Science Foundation égisze alatt a nem-konvencionális információs rendszerek fejlesztéséhez. (Ebből a szempontból a helyzet emlékeztetett azokra az első kötetük bevezető kommentárjában említett 1850-es évekre, amikor az amerikai könyvtárügy ugrásszerű fejlődését ugyancsak nagyszabású állami kulturális programnak köszönhette.)

Bush a háború alatt kb. hatezer vezető amerikai szakember munkáját koordinálta annak érdekében, hogy a tudományos kutatásokat a hadviselés szolgálatában felhasználhassák; ő irányította az atombomba előállításának tudományos programját. Még a háború befejezése előtt, 1945 áprilisában tanulmányt írt „Úgy, ahogy gondolkodunk” címmel[3], melyben az emberi észjáráshoz közelebb álló, a mai hypertext-kapcsolódásokra emlékeztető világméretű információkereső rendszer vízióját fogalmazta meg, összekapcsolt számítógépek hálózatával; olyan – Memex nevű – „költői gépekkel”, melyek az analógia és az asszociációk útján működnek, megragadva és reprodukálva a képzelet anarchikus jellegét. Később könyvben is kiadott[4] elképzelései nemcsak az Egyesült Államok háború utáni informatikai fejlesztési stratégiáját határozták meg, hanem évtizedek múlva is erősen hatottak, például *Douglas Engelbertre*, az egér és az ablaktechnika, *Theodor Holm Nelsonra,* a hypertext föltalálójára, és a számítógépes technológia számos más úttörőjére. *Bush* a problémák gyökerét a „szelekcióban”, vagyis az osztályozásban és indexelésben látta:

„Főként az indexelő rendszerek természetellenessége az oka, hogy képtelenek vagyunk elérni a rögzített adatokat. Amikor valamilyen adatot tárolnak, alfabetikusan vagy numerikusan iktatják, az információ pedig alosztályról alosztályra követve található meg (ha ugyan megtalálható). Csak egy bizo-

2 Lilley, Doroty B, Trice, Ronald W.: History of information science, 1945–1985. p. 41.

3 As we may think. In: Atlantic Monthly, 1945, July, No. 176, p. 101–108). Magyarul: Út az új gondolkodás felé. In: Hypertext+multimédia. Oktatási segédanyag. A szöveget vál. … Sugár János; szerk. Klaniczay Júlia. – Budapest: Artpool, 1996. p. 3–14.

4 Endless horizons. – Washington: Public Affairs Press, 1946.

nyos helyen lehet, hacsak nem készítünk másolatokat; szabályokra van szükség, hogy megtudjuk, milyen úton juthatunk el az információhoz, a szabályok pedig fárasztóak."

A dokumentációs–információs szolgáltatások fejlesztését meghatározóan támogatták az ebben az időszakban alakult védelmi szervezetek és hadiiparilag fontos intézmények is, mint a Nemzeti Repülési és Űrkutatási Hatóság (National Aeronautik and Space Administration; NASA), a Légierők Tudományos Kutatási Hivatala (Air Force Office of Scientific Research), az Amerikai Hadászati Műszaki Információs Központ (Armed Services Technical Information Agency; ASTIA). Mindezzel szoros szellemi és anyagi kölcsönhatásban született meg a korszak elején az információkeresés (information retrieval) szakterülete is.

A nyugati világ, elsősorban az Amerikai Egyesült Államok ezekben az ötvenes és hatvanas éekben – az ún. hidegháború idején – alapozta meg technológiai fölényét a Szovjetunióval szemben. Az információtudomány (és ezen belül az információkeresés) kialakulása és fejlődése elválaszthatatlan ettől a technológiai fejlődéstől.

Az űrkutatásban elért kezdeti szovjet sikerek a hatvanas évek elején egy ideig még eltakarták ennek a technológiai fölénynek a körvonalait. Az évtized végére azonban már megjelentek olyan publikációk a tudomány kelet-európai és szovjet művelőitől, melyekben óvatosan, de utaltak a fokozódó lemaradásra, elsősorban az információs technológiák területén. A hetvenes évek közepétől pedig szovjet irányítással megindult egy kelet-európai „nemzetközi" információs rendszer – az NTMIR – fejlesztési programja, mely a szocialista világ felbomlásával a nyolcvanas évek végére hamvába hullt.[5]

Az információkeresés új szakterületéhez az út a hagyományos osztályozásról való leváláson át vezetett. A negyvenes évek végén eleinte még *Samuel Clement Bradford* (1878–1948), az „angol dokumentáció atyjának" kezdeményezésére az ETO korszerűsítésében keresték a dokumentációs tartalmi feltáráshoz a megoldást. A munkák koordinálására a Nemzetközi Dokumentációs Szövetségben (FID) három bizottság alakult. Az egyik konkrétan az ETO fejlesztésével foglalkozott. Az osztályozáskutatási bizottságot *Ranganathan* vezette, *Eric de Grolier* támogatásával (utóbbi e tevékenységének keretében írta meg az első kötetben bemuta-

5 Magyarországon a nyelvész Petőfi S. János volt egyike az elsőknek, aki 1969-ben a tezaruszkészítésről írt könyvével felhívta a figyelmet az információtudományban bekövetkezett nyugati fejlődésre. „A tezaurusz-kérdés jelenlegi helyzete különös tekintettel a tudományos, műszaki–gazdasági tájékoztatásra" című műve az egyik első hazai dokumentuma az információkeresés szakirodalmának (Budapest: Országos Műszaki Könyvtár és Dokumentációs Központ, 1969. 168 p.) (Lásd még az *Alan Gilchristhez* fűzött kommentárt az első kötetben).

tott monográfiáját az osztályozási rendszerekben használt általános kategóriákról). Az osztályozási rendszerek összehasonlító vizsgálatára alakult, *Douglas J. Foskett* vezette harmadik bizottságot évtizedekig az angol osztályozáskutató társaság (CRG) uralta, és az ötvenes–hatvanas években a fazettás osztályozás fejlesztésének szentelték a legnagyobb figyelmet.

Ennek a bizottságnak a kezdeményezésére került sor 1954-ben az első nemzetközi osztályozási konferencia megrendezésére az angliai Dorkingban; ez volt egyben az első olyan nemzetközi konferencia is, melyet az információkeresésnek szenteltek. Az osztályozás és az információkeresés konferenciáinak története azonban ettől kezdve szétvált (az osztályozási konferenciák további történetével első kötetünk elején foglalkozunk).

Az Egyesült Államokban ugyanebben az időszakban minden jel szerint hamarabb és gyakorlatiasabb formában kellett használható tartalomfeltárási és keresési megoldásokhoz jutni. Az úttörő szerepet az eredetileg filozófiai végzettségű *Mortimer Taube* (1910–1965) játszotta 1952-ben[6], amikor az elvileg a tárgyszavakon, s így a természetes nyelven alapuló osztályozást újította meg radikálisan: Uniterm-rendszerében a hagyományosnál sokkal egyszerűbb alapszavak kizárólag a nyelvi–szemantikai egységek szerepét játszották. A tartalmi feltárás eredményeként a szavakat – indexkifejezéseket – egymással teljesen egyenrangúan, mechanikusan egymás mellé kellett rendelni. Ez volt a koordinált indexelés. Az összetett tárgyszavak hagyományos permutációjának helyébe az egyszerű szavak puszta kombinációja került.

Például a „Rézből készült csővezetékek védőbevonatai" tárgyat a következő Uniterm-lánc reprezentálja: „Réz", „Csővezeték", „Védelem", „Bevonat". (A hagyományos tárgyszórendszerben az összetett tárgyszó – melyet a katalógusban a kereshetőség céljából permutáltan kell szerepeltetni – a következő, ***egyetlen egység*** lett volna: „Védőbevonat, rézcsövek, csővezeték". A többi alakjáról „lásd" utalások készültek: „Rézcsövek, Védőbevonat, Csővezeték", „Csővezeték, Rézcsövek, Védőbevonat". Egy dokumentumot elvileg egyetlen ilyen összetett tárgyszó reprezentált.) A dokumentum tartalmát leíró „unitermek" nem alkotnak egyetlen összetett tárgyszót, hanem mindegyik önállóan reprezentálja a dokumentum tartalmát. A keresés célja, hogy a szóba kerülő egyszerű szavak alapján a releváns dokumentumokat mintegy „összefésüljék". A releváns találat akkor keletkezik, ha a kereséskor használt kifejezések kombinációihoz ugyanaz a dokumentumazonosító tartozik.

6 Taube, M.: Specifity in subject headigs and coordinate indexing. In: Library Trends, 1952, Okt, p. 219–223.

Ez a ma természetesnek tűnő eljárás a maga idejében forradalmian hatott, amiből talán az is megérthető, milyen más, hierarchikus elveken nyugodott a hagyományos tárgyszavas katalogizálás. (És talán az is érthetőbb, miért tűnt a természetes nyelven alapuló, hagyományos tárgyszavas osztályozás elvileg mégis alkalmasabbnak a korai dokumentátorok egy részének szemében a folyóiratcikkek feltárására: ők a tárgyszavakban ösztönösen nem a tárgyi katalógusok általában több szóval megnevezett „osztályait", hanem a mikrofeltárásra alkalmas „nyelvi leíró" eszközöket szerettek volna látni. *Taube* intuíciójára és elfogulatlanságára volt szükség, hogy ez a vágy megvalósuljon.)[7]

Ugyanakkor az Uniterm túlságosan is radikális javaslat volt, és a maga eredeti formájában, mint annyi más felfedezés, nem gyökeresedett meg. A felhasználás tapasztalatai alapján világossá vált: a kereséskor figyelembe kell venni elvont szemantikai (analitikus) összefüggéseket (mi minek a fajtája, része, oka, tulajdonsága stb.)[8], hogy az eljárás hatékony legyen. *Fred Jonker* fogalmazta meg 1959-ben[9], hogy az egyes szavak specifikus, konkrét kiválasztása valójában csak az egyik valószínű választás bekövetkeztét jelenti a lehetséges szóválasztások spektrumából. Ennek alapján alkotta meg *Jonker* az információkeresés egyik klasszikus alapfogalmát, a deszkriptív kontinuumot.

A *Mortimer Taube* által 1952-ben útnak indított koordinált indexelés kiszabadította az indexelést (a „mutatókészítést") a hagyományos, intellektuális tárgyszavas eljárás szemléletéből és ezzel az információkereső gondolkodás szellemét is kiszabadította a palackból. Ugyanakkor felhívta a figyelmet az indexelés/osztályozás nyelvi aspektusaira, a szintaktikai és szemantikai problémákra.

A fejlődés útja innen egyrészt szintaktikai relációk kidolgozásának kísérleteihez vezetett, párhuzamosan azzal, hogy a *Ranganathan* eszméiből született fazettás osztályozás is tovább fejlődött a brit osztályozáskutatási bizottság (Classification Research Group; CRG) tagjainak tevékenysége révén (ezekkel a kérdésekkel az első kötetben *Jason Farradane* és a CRG kapcsán foglalkozunk).

---

7 A tárgyszórendszerek rejtett szerepére utal *Fredrick W. Lancaster* is az első kötetben, melyet az *Alan Gilchristhez* fűzött bevezetőben olvashatunk (a szerk.).

8 Az angol szakirodalomban inkább az „analitikus relációk" kifejezés fordul elő, olykor „báziskapcsolatok". A nyelvészetben „paradigmatikus összefüggésekről" beszélnek. A 60-as évek orosz szakirodalmában az „értelmi összefüggések" kifejezést is használták. Minden esetben szövegfüggetlen, a fogalmak, ill. a szavak között eleve meglévő összefüggésekről van szó.

9 Jonker, F.: The descriptive continuum. A „generalized" theory of indexing. In: Proceeding of the Internation Conference on Scientific Information, Washington, D. C., November 16–21, 1958. 2 vols. – Washington, D. C.: National Academy of Sciences, 1959. p. 1291–1312.

A koordinált indexelés a maga eredeti formájában egy másik, az információ kereshetőségét javító felismerést is inspirált, mely első lépésben a különféle mechanikusan előállítható mutatókban (indexekben) öltött testet. Az eredetileg textilmérnök *Hans Peter Luhn* 1956–58-ban[10] vezette be a folyóiratcikkek címszavait kereshetővé tevő permutált indexet, s ennek nyomán vált általánosan használttá a kulcsszó fogalma a dokumentumok szövegéből kiválasztott szavakra. Különösen a referáló folyóiratok kiadói voltak elragadtatva az eljárástól, mely egyszerre volt olcsó, intellektuális beavatkozás nélkül készíthető keresőeszköz, ugyanakkor nagy mennyiségű információtételt lehetett vele kereshetően feldolgozni. 1953 és 1963 között az eredetileg vegyész *Eugen Garfield* (1925) kidolgozta a hivatkozási indexelés módszerét, melyet később a tudományos alkotótevékenység mérésére is használtak.

Az Uniterm nyelvekkel és a gépi indexekkel párhuzamosan kezdték használni az ötvenes évek második felétől az információkeresés (information retrieval) és az információkereső rendszer fogalmát, terjedt el – igazából még a számítógépek számottevő megjelenése előtt, az ötvenes évek végétől, hatvanas évek elejétől, a manuális rendszerekben – a rekord, a fájl, az elérés, az invertálás és az adathordozó, az indexelő/információkereső nyelv és az indexelés fogalma.

Az új szakterület keresztapja ugyancsak *Mortimer Taube*, aki egyik 1950-ben írott tanulmányában használta először az angol vadászati nyelvben honos „retrieval" kifejezést: „…searching and **retrieval** of information from storage encoding to specification by subject." […a tárgy pontos meghatározása érdekében kódolt információ keresése, **becserkészése és visszanyerése** a tárolóból]. Magyarul a leginkább még a „becserkészés és visszanyerés" szóösszetétel áll a legközelebb a szó eredeti angol jelentéséhez, ehelyett az információkeresés, „információ-visszakeresés"[11] kifejezések terjedtek el, megkülönböztetésül a szűken értelmezett, egyszerű kereséstől (searching) és szinonim, de legfeljebb még szűkebb értelmű lekérdezéstől, sávos letapogatástól (scanning). Információkeresésen (retrieval, information retrieval) a kereséssel összefüggő teljes folyamatot értjük, kezdve a keresőkérdés elemzésével, a keresőprofil és a keresési stratégia és taktika kialakításán és a tárolóban végzett keresőműveleteken át a találatok képzéséig és kiadásáig.

---

10 Luhn, H. P.: Keyword-in-context index for technical literature (KWIC index). – Yorktown Heights, N. Y.: IBM, 1959.

11 Átfogó – „information retrieval" – értelemben mindig az „információkeresés" (és nem az ugyancsak elterjedt „információ-visszakeresés") kifejezést használjuk. Kivétel, ha a „visszakeresés" kifejezés korábban magyarul megjelent mű címében szerepel. A „vissza…" használata egyrészt felesleges, mert minden tárolt információ megkeresése „visszakeresés", másrészt hibás, mert a keresés lehet eredménytelen is, de attól még ugyanúgy keresés játszódott le.

Vickery így jellemzi ezt a korai időszakot:

„Az információkeresés különféle irányzatainak és eljárásainak egységesítésére az első lépéseket 1947-ben *J. E. Holmstron* tette, amikor az egyik FID-konferencián az »osztályozások osztályozási rendszerét« mutatta be. A bibliográfia szervezésével foglalkozó 1950-ben rendezett chicagói konferencián már mind az osztályozásról, mind pedig a gépi kereső- és tárolóeszközökről hangzottak el előadások… A személyes találkozásoknak az volt az eredménye, hogy az információkereséssel foglalkozó kutatók elkezdték végre olvasni egymás közleményeit, lassan megértették egymás szóhasználatát és gondolatait, és megtanulták, hogyan értelmezzék a nyilvánosságra hozott elképzeléseket."[12]

Az 1958-ban Washingtonban a Légierők Tudományos Kutatási Hivatala támogatásával rendezett információkeresési konferencián *Taube* meg *Luhn* (és velük az Uniterm és az indexek) játszották a főszerepet.
Az 1966-os cranfieldi konferenciának a tárgya *Cyrill W. Cleverdon* cranfieldi vizsgálatai és általában az információkeresés értékelése volt. Ezt követően kisebb-nagyobb rendszerességgel rendeztek konferenciákat „International Conference on Mechanized Information Storage and Retrieval Systems" [A gépi információtároló és -kereső rendszerek nemzetközi konferenciája] címen, napjainkban pedig olykor évente több információkereséssel foglalkozó nemzetközi konferencia szervezésére kerül sor.[13]
1966-ban jelentette meg az eredetileg pszichológus *Carlos A. Cuadra* az információtudomány évkönyvének első számát[14], 1968-ban az eredetileg vegyész *Allen Kent* (1921) útnak indította a könyvtári és információtudományi enciklopédiáját[15], melynek 1997-ben adta ki az 59. kötetét.
A negyvenes évek végétől kezdtek néhányan az osztályozás modern matematikai alapjaival is foglalkozni. Az osztályozás algebrájának, s így a későbbi automatikus osztályozásnak immár klasszikus előfutára az első kötetben *Alan Gilrichst* kapcsán tárgyalt *Calvin Mooers* mellett *Robert A. Fairthorne*, a CRG tagja. Számtalan korai tanulmányában – melyeket 1961-

12 Vickery, B. C.: On retrieval system theory. 2nd ed. – London: Butterworths, 1965. p. 2.

13 Az egyik legismertebb az 1963-ban indult, az illinoisi egyetemen az adatfeldolgozás könyvtári alkalmazásának intézete által évenként rendezett konferencia (Annual Meeting of the Clinic in Library Applications of Data Processing), és az 1976-ban indult, minden év végén Londonban megrendezett „On-line information Meeting", melyet az On-line Review, a The Electronic Library, és még több más szakfolyóirat kiadója, a Learned Information (Europe) Ltd. szervez.

14 Annual Review of Information Science and Technology. Ed. C. A. Cuadra. 1966–.

15 Encyclopedia of Library and Information Science / Ed. by Allen Kent. – Vol. 1–59. – New York : Dekker Inc., 1968–

ben kötetben is kiadott –, feltérképezte az osztályozás matematikai–logikai jellegzetességeit. Az elsők között vetette föl, hogy az automatizált osztályozás jól alkalmazható dokumentumkeresési célokra. Amikor ezeket a javaslatait megfogalmazta, gyakorlatilag még nem léteztek kereskedelmi forgalomban kapható számítógépek és csak sok évvel később került sor az első komoly kísérletekre a dokumentumklaszterálás terén.

Végül egy terminológiai megjegyzés.

Az információkeresés félreérthető fogalom. Nemcsak azért, mert az egyszerű keresést, lekérdezést is érthetik rajta, hanem mert az „információ" szó is félreérthető benne. Ez egyrészt nem azonos az információelmélet egzakt, mérhető információ-fogalmával, másrészt tisztázatlan, hogy tényadatokról (elsődleges információkról) vagy hivatkozási adatokról (másodlagos információkról) van szó. A problémára rávilágít *Cornelis van Rijsbergen* kissé egyoldalú értelmezése:

„Ami azt illeti, sok esetben az információkeresés megfelelő módon leírható úgy is, hogy az »információ« szót egyszerűen a »dokumentum« szóval helyettesítjük. …a legtökéletesebben szókimondó definíciót *Lancaster* adta meg: »Az információkeresés az az általánosan elfogadott, de némileg pontatlanul használt kifejezés, amelyet a könyvemben tárgyalt tevékenység megjelölésére alkalmaztak. Egy információkereső rendszer nem ad információt a használónak kérdése témájáról, azaz nem változtatja meg ismereteit. Csupán arról ad tájékoztatást, hogy van-e, vagy nincs, s ha van, hol van olyan dokumentum, amely választ ad kérdésére.« Ez a meghatározás kizárja… a kérdés–felelet rendszereket. Ugyancsak kizárja azokat az adatkereső rendszereket, amelyek például a tőzsdei árfolyamok on-line szolgáltatására valók."[16]

Van Rijsbergen az elsődleges, tényadatok keresésére nem az információ-, hanem az adatkeresés kifejezést javasolja. Valójában nem igaz, hogy a hivatkozási adatokat/információkat szolgáltató információkereső rendszer nem változtatja meg használójának ismereteit; megváltoztatja, csak éppen nem az elsődleges, hanem a másodlagos információk (tehát a hivatkozások) szintjén. Hiszen a használat előtt nem tudta (vagy rosszul tudta), milyen dokumentumot és hol keressen, utána – szerencsés esetben – viszont tudni fogja.

Az információkeresés kontra adatkeresés terminológiával pedig az a baj, hogy a tényadatok/faktografikus adatok keresői és az ilyen rendszerek kezelői nem feltétlenül tudnak erről a finom distinkcióról, és többnyire a maguk rendszereit is információkereső rendszereknek nevezik. Ezért helyesebb, ha

16 Van Rijsbergen: Információ-visszakeresés. [közr. az] Országos Széchényi Könyvtár Könyvtártudományi és Módszertani Központ. – Budapest: Múzsák Közművelődési Kiadó, 1987. p. 7.

az információkereső rendszert ebből a szempontból mind a másodlagos (a dokumentumtételek, reprezentációk, hivatkozások), mind az elsődleges információk keresőrendszerének generikus fölérendelt fogalmának tekintjük, és ha szükséges pontosítani, faktografikus, illetve hivatkozási információkeresésről beszélünk. Kötetünkben a dokumentációs és könyvtári/bibiográfiai, tehát hivatkozási információk keresésével foglalkozunk.

ROBERT A. FAIRTHORNE (1905)

1905-ben született a nagy-britanniai Abingdonban. Matematikus, több számítógép-tudományi, informatikai és osztályozási társaság tagja. Egyike volt az elsőknek, akik már a 40-es évek végén az osztályozás formális, matematikai–logikai elméletével, a bibliográfiai osztályozások struktúrájával, a tartalomelemzéssel foglalkoztak, mintegy az automatikus indexelés előfutáraként. 1961-ben, a közölt részleteket is tartalmazó monográfiája első kiadása idején a brit légierő matematikai osztályának (Mathematics Department of Royal Aircraft Establishment) tudományos főmunkatársa és az ohioi (USA) egyetem könyvtártudományi iskolájának (School of Library Science, Western Reserve University) vendégprofesszora volt.
Fairthorne azokat az osztályozás matematikájával foglalkozó szakembereket képviseli kötetünkben, akik az absztrakt algebra és a formális logika ún. extenzionális értelmezése alapján közelítik meg az automatikus indexelés és osztályozás kérdését. Létezik az ún. intenzionális megközelítés is, ezt kötetünkben *Jurij A. Šrejder* képviseli. A kétfajta megközelítéssel részletesen *Šrejder* szemelvényének bemutatásánál foglalkozunk.
Fairthorne itt közölt tanulmányai még az 50-es években jelentek meg folyóiratokban, 1961-ben gyűjtötte össze őket kötetbe.

Az információkeresés tudománya felé[17]

Az automaták és az információ[18]

1713-ban egy Humphrey Potter nevű ember vezetékkel kapcsolta össze a gőzszivattyú szelepeit és így maga a szivattyú szabályozta a gőz áramlását. Ő készítette el az első „munka” automatát vagy robotot. Egészen addig csak arra használták a gőzt, hogy mozgassa a gépet, ki-bekapcsolnia az embernek

17 Fairthorne, R. A.: Towards information retrieval. London : Butterworths, 1961. 162 p.
18 Automata and information. In: Towards information retrieval, p. 11–21. [Eredetileg megjelent 1952-ben.]

magának kellett, ugyanúgy, ahogy egy lóval meg kell értetnie, hogy mikor induljon el vagy álljon meg. Az automaták *addig* arisztokratikus *szerkezetek* voltak; megbámulni való kiállítási tárgyak.

Amit Humphrey elkezdett, mások tökéletesítették. A gőzgép egyre inkább önirányítóvá vált. A kazánokra biztonsági szelepeket szereltek, így amikor a nyomás kritikus értékű lett, a szelep magától kinyílt és a gőzt kiengedte. A szivattyús gépek valójában nagyon lassúak voltak. A gyorsabb gépekre „vezérműveket" szereltek, amelyek csökkentették a gőzellátást, amikor a gép túl gyorsan működött, és növelték az adagot, ha nagyon lelassult.

Humphrey vezetékével, sőt a „vezérművel" összehasonlítva az olyan bonyolult dolgok mint az automatikus közlekedésirányító lámpák, a távirányított repülőgépek, az „elektronikus agyak" és még sok ezer kevésbé reklámozott, de hasonlóképpen hasznos, gépeket és folyamatokat irányító szerkezet varázslatosnak tűnik. Az emberek azt kérdezik: „Gondolkodnak-e a számítógépek?", de ezek a *szerkezetek* csak bonyolultságukban különböznek a vezetékektől. Sokkal több dolog található a belsejükben és több dolgot tudnak gyorsabban csinálni. Ez talán a pszichológiai tevékenység egyetlen követelménye. Ha így van, bizonyára a nem gondolkodó mechanizmusok a ritkák, nem a gondolkodók.

Akár gondolkodnak ezek az önirányító szerkezetek, akár nem, viselkedésüknek ugyanazt a modellt kell követnie amelyet a legegyszerűbb szabályozóknak. Divatos antropomorf kifejezéssel élve: tudniuk kell azokról az eseményekről, amelyek működésüket befolyásolják (de semmi másról), tudniuk kell, hogy adott helyzetben mit tegyenek; képesnek és hajlandónak kell lenniük ezen lépések megtételére. *Sem* gondolkodásra, sem öntudatra nincs szükség ahhoz, hogy a számítógépekben, irányító szerkezetekben és hasonló konstrukciókban a megkívánt működési, viselkedési mintát kialakíthassuk, még akkor sem, ha képesek a tapasztalatokból tanulni és kiszámíthatatlan cselekedeteket produkálni. E működési, viselkedési minta legegyszerűbb példája az egyszerű vezérmű. Az ilyen rendszerekkel foglalkozó tudományágat „kibernetikának" nevezik, a görög „kormányos" szó után, amely a latin nyelv közvetítésével került be az angolba, hogy az „irányítástechnika" legáltalánosabb fogalmát képviselje. A kifejezés modern használatát néhány amerikai matematikusnak és ideggyógyásznak köszönhetjük, de a „cybernetique" szót már *Ampere* is alapvetően ugyanebben az értelemben használta.

Minden irányító rendszerben a bemeneti egységek válaszolnak a releváns történésekre, a reakciókat egy jelátalakítónak továbbítják, amely azokat azután kimenő jelekre fordítja. Ezek a különböző végrehajtó szervekben megindítják a megfelelő tevékenységet, *melynek* nyomán megváltozik a megválaszolt és tevékenységet kiváltó helyzet, és újra kezdődik a folyamat. A jelátalakítót – ha úgy tetszik – agynak is nevezhetjük, ám ahhoz, hogy megfelelő szerkezeteket készítsünk, elegendő olyan műszerként kezelni, amely a jeleket bizonyos szabályok

szerint változtatja meg. A fontosabb kérdések a következők: melyek az ilyen berendezés működésének a korlátai? Milyen szabályok szükségesek, hogy ezeken a korlátokon belül a kívánt viselkedést, működést elérjük?

A gyakorlatban ezek a szabályok és jelek rendkívül változatosak. Lehetnek vezérművekkel és emelőkkel előidézett egyszerű mozgásváltoztatások. Ilyen például az a folyamat, amelyben az irányító karjainak mozgatása a megkívánt szelephelyzetet előállítja. Az olyasféle gépeken, mint a számítógépek, a jelek távirati jellegűek és betűket vagy számokat képviselnek. A bejövő jeleket olyan módon változtatják meg, hogy a kimenő jelek megfeleljenek az utasításokban megadott aritmetikai műveletek eredményének. Az utasítások lehetnek a bejövő jel részei, tartalmazhatja őket a vezérmű, vagy származhatnak e két forrásból egyszerre. A vezérmű belső feltételei kialakíthatók úgy, hogy a korábbi események függvényében módosuljanak, és hogy a kimenő jel ugyanúgy függjön ezektől a változásoktól, mint a bejövő jeltől és az ebből származó utasítástól. Így a vezérmű képes saját utasításait megváltoztatni. Lehetséges, de nem kívánatos az ilyen konverterek teljes hierarchiájának megalkotása, amelyben mindegyik megváltoztatja a szabályokat azért, hogy a szabályokat megváltoztassa … és így tovább. Szerencsére a gép véges és véges az élettartama is, így az ilyen olcsó „platonizmust" egyszerűen elkerülhetjük, ha szigorúan dolgokban és eseményekben gondolkodunk. Bármilyen nyelven írjuk is le a műveleteket, alapvető tény, hogy a továbbított jelet teljes mértékben a kapott jel és a vezérmű mindenkori állapota határozza meg. Ha valaki rendelkezik az állapotok és a jelek összes lehetséges kombinációinak és az ezekre adható korrekt válaszoknak és új gépi állapotoknak a listájával, az üzemképes gép viselkedése, működése máris pontosan meghatározható. Az ilyen listák *azonban* nem szolgáltatnak információt arról, hogy a viselkedés hogyan és miért jön létre. A viselkedés bármilyen teljes megfigyelése önmagában nem adhat választ ezekre a kérdésekre. Egy egyszerű pénztárgépben például a gép állapotát a belső számlálón nyilvántartott összeg mutatja és a bejövő jel az éppen beütött tétel árát jelzi. A megfelelő válasz az ár kijelzése a képernyőn és a belső állapot megváltoztatása olyan módon, hogy az már a beütött ár és a korábbi összeg együttes összegét képviselje. Ebben az esetben a belső állapot nem befolyásolja a kimenő jelet, bár a bonyolultabb számítógépekre ez nem jellemző. Befolyásolja vagy sem, a részösszegek és az egyedi értékek összes lehetséges kombinációjának a listája a megfelelő válaszokkal helytálló lehet, de rendkívül terjedelmes. A legegyszerűbb gépek esetében is billió tételt és összeget jelentene.

Minden ilyen viselkedési rendszer logikailag megfelel az úgynevezett Thue rendszernek. Ez annak a jól ismert játéknak az általánosított formája, amelyben egy bizonyos szóból úgy kell másikat csinálni, hogy egyszerre csak egy betűt változtatunk meg. A behelyettesítés szabályai a szóval együtt szisztematikus vagy önkényes törvény szerint változnak. Úgy tűnik, hogy minden

realizálható mechanizmus viselkedését ilyen rendszerek irányítják. Milyen korlátozást jelent ez? Az ilyen rendszer sohasem képes saját működésének szabályait teljes részletességgel kikövetkeztetni, mivel ez azt jelentené, hogy saját magával írná le saját magát…

Mivel a gép működési szabályait magából a gépből nem következtethetjük ki, azokat előre kell a tervezőnek megalkotnia. A tervező lehet élő ember, egy magasabb rendű gép, mindkettő vagy egyik sem. Az önműködő gép ebből következően a tervező eszköze, amelynek nincs se több, se kevesebb filozófiai jelentősége, mint bármely más eszköznek. Az automata a csavarhúzótól voltaképpen csak bonyolultságában, árában és abban különbözik, hogy a használó az összes szükséges utasítást egyszerre adhatja meg és aztán elmehet, ahelyett, hogy végig a gép mellett állna és irányítaná. Igaz, hogy egy sakkautomata, amelyet a jó sakkjáték szabályait leírni tudó ember alkotott meg, jobb sakkozó, mint a tervezője. Minden eszköz jobban végzi a dolgát, mint a tervező. Ezért készítjük őket. Egy szerszám a készségek javítására szolgál, vagyis a munka hatékonyságát növeli. Nem szükségszerű viszont, hogy az eszköz csökkentse a munka teljes mennyiségét.

A szerszámot összegyűjtött és előre megmunkált anyagokból kell elkészíteni. Nem biztos, hogy a jövőben is kifizetődő a munka, amelyet térben és időben megtakarítanak vele és hogy nem megy azok rovására, akiket a megtakarítás nem érint közvetlenül. Az olyan gépek például, amelyek a folyóiratokat túl szorosan kötik össze, vagy a feloldás nélküli, illetve kideríthetetlen jelentésű rövidítések használata elrettentő példái az olyan eszközöknek, amelyek a munkát nem megtakarítják, hanem szaporítják. Az ilyen hátráltató eszközök tipikus példái a fotónyilvántartások. Megfelelő körültekintés és előkészítés nélkül egy ilyen alkalmatlan formában rögzített adathalmaz annyira felduzzadhat, hogy jóval nehezebb és drágább egy tételt megtalálni és leolvasni, mint az információt újonnan kikeresni az eredeti forrásból.

Ez utóbbi példa érzékelteti ama főbb akadályok egyikét, amelyek az automaták sikeres könyvtári alkalmazásának útját állják. A gép komplexitása bizonyos fokig függ működési szabályainak komplexitásától. Ha egy könyvtári tevékenység szabályait már megfogalmazták, az alapvető tevékenységre vonatkozó előírások nagyon egyszerűek lehetnek; például a beérkező alfabetikus információ sokszorosítása és újrarendezése. Azok a szabályok *viszont,* amelyek szerint a bejövő ingereket feldolgozásra alkalmas jellé, majd ezeket megfelelő output formára fordítják, rendkívül bonyolultak lehetnek, ha ismertek egyáltalán. Az input ingerek különböző méretű és eltérő fizikai jellemzőkkel rendelkező, papírlapokon megjelenő, különböző és azonos alakú jelek. A jelek értelmezését a kontextus határozza meg és a legkülönfélébb formáknak (gépírás, kézírás és hangjegyek) megegyező az interpretációja, amely szintén a szövegkörnyezettől függően változik. Hogyha egységes formájú inputtal rendelkezünk lyukkártyán vagy -szalagon, mágnesszalagon stb. a működés szabályait korlá-

tozhatjuk az alapvető kiválasztásra, elrendezésre és automatikus átírásra stb. A gépi utasításokat ilyen módon csökkenthetjük, de a dolgozókra és a használókra vonatkozó szabályok biztosan megsokszorozódnak és szigorúbbakká válnak.

Egy könyvtár olyan széles körű információkat szerez be, amilyenre csak képes. Nem korlátozódik bizonyos nyelvekre, betűtípusra vagy formátumra. Így a jelenleg rendelkezésünkre álló információs automaták csak a könyvtárak által, a könyvtárak számára speciálisan előkészített információt képesek haszonnal feldolgozni. A gépírás általános elterjedése azt jelenti, hogy ha a kereskedelemben hozzá lehetne jutni gépírást leolvasó gépekhez, az automaták alkalmazási köre kibővülne: a leolvasókat nemcsak automata írógépekhez lehetne csatlakoztatni, hanem kártya- és szalaglyukasztókhoz is. Túl sok ügyviteli munkát fecsérelünk ugyanazon információ különféle megjelenéseinek elkészítésére. Például arra, hogy kártyákról jegyzékekre írjuk át az adatokat, pedig ezt a lyukszalag irányította gépek el tudják végezni. A gépírást leolvasó gép iránti igény ilyenkor nagyon is valóságos.

Bármely információs berendezés eredményeit végül az emberek számára megfelelő formában kell kibocsátani. Az információt legalább térben, időben, szimbolikában nem szabad túlzottan sűríteni. Az automatákat – megfelelő információátalakítók segítségével – a folyamat többi részéhez kell „illeszteni". Ezek érzékelhető formájúra alakítják az információt. Például a mikrofilmhez nagyítóra van szükség, a múló látványokat fényképen kell megörökíteni, a numerikus táblázatokat érdemes grafikusan is megjeleníteni. Ha erről nem gondoskodunk és időnként – különösen tudományos kísérletekben és számításokban – nem tesszük ezt meg, a nagy sebességű berendezések használata annak a hintalóüzemnek a technológiájára fog hasonlítani, amelyben a mintákat kézzel festik, de a farkakat olyan nagy teljesítményű farokberakóval teszik fel, amely másodpercenként több »megafarok« sebességgel működik.

Mindezek eredményeképpen a meglevő automaták általában nemigen alkalmasak közvetlen könyvtári használatra: nincs elég hozzáférhető tároló kapacitásuk vagy pszeudomemóriájuk; túl gyorsan dolgoznak; inputjaikhoz és időnként outputjaikhoz is átírásra van szükség. Bármely könyvtári automatáról legyen is szó, hatása a jellegzetes könyvtári tevékenységekre nagymértékben csak közvetett. Van azonban igény „szamárgépekre", amelyekkel a felesleges műveleteket felszámolhatjuk, a segédmunkát, a fennmaradó szállítási és anyagmozgatási feladatokat elvégezhetjük. Magyarán: egyetlen nagy sebességű gép sem versenyezhet azzal, ha a munkát egyáltalán nem kell elvégeznünk, de mindig marad sok olyan elkerülhetetlen fizikai munka, amely számos apró és eltérő mozzanatból tevődik össze. Kis robotokra van szükségünk, amelyek jó „izommemóriával", és a számoló őrültekénél is kevesebb „intellektuális" bátorsággal rendelkeznek, viszont erősek vagy ügyesek. Egy ilyen csudabogár az a kameraállvány, amelyet olyan filmjelenetek megalkotására találtak ki, melyekben az élő és a modell felvételeket kell egymásra kopírozni… Az ilyen

megbízható utánzás, amely például a szállításra, mozgatásra vagy ugyanazon anyag többszöri legépelésére és hasonló mechanikus tevékenységek segítésére szolgál, azonnal alkalmazható a könyvtárakban. Ha például bizonyos szállítási tevékenység mindig egyazon úton történik, a szállítandó mennyiség általában indokolttá teszi mechanikus szállítási és anyagmozgatási berendezések használatát. Az irányított szállítók – targoncák és ehhez hasonlók – bármely úton végigmehetnek, de emberi irányítást igényelnek. A javasolt szerkezet, nevezzük „utánzónak", pontosan azt a tevékenységet fogja ismételgetni, amelyen egy emberi oktató végigvezette, és ezt teszi egészen addig, amíg egy új, csak egyszer bemutatott tevékenység a működését nem módosítja. Egy automatikus villanyírógép valójában ugyanilyen „utánzó", de szokásait jóval nehezebb megváltoztatni, mint a javasolt szerkezetét, mivel tevékenységét előre lyukasztott kártyák vagy szalagok segítségével végzi, a lyukakat pedig nem lehet megszüntetni. A mágneses adathordozók legnagyobb előnye, hogy rendkívül tartósak, de ugyanakkor tökéletesen kitörölhetők, akár egy-egy helyen, akár teljes terjedelmükben.

Mint bármely eszköz terméke, ezen betűk vagy tevékenységek sorozatának értéke is a használó képességeitől függ. Értelmetlen dolog, ha az automatának vagy a másik embernek azt mondjuk: „Tudod, hogyan gondolom". A jó eszközök arra alkalmasak, hogy ugyanazzal az erőfeszítéssel, segítséggel többet érhessünk el, nem pedig arra, hogy kevesebb erőfeszítéssel ugyanannyit. Ezt tudja mindenki, csak azok nem, akiknek ezekre az eszközökre elő kellene teremteniük a pénzt.

Az említett utánzó–másoló automaták elkészítése csupán gazdasági probléma. Megszerkeszthetők olyan módon, hogy hozzáférhető mágneses adatrögzítőket és szabályozó berendezéseket csatlakoztatunk a szóban forgó munkák elvégzésére alkalmas motorizált targoncákhoz, írógépekhez vagy akár gépzongorához.

Bár az automatákat nem lehet közvetlenül alkalmazni a könyvtár információs tevékenységeihez, a tervezésükben és megépítésükben használt módszerek és részletek biztosan felhasználhatók. Mint láttuk, működésük azon alapszik, hogy külső ingereket jellé alakítanak át, más jeleket „történetileg" kialakult szabályok szerint behelyettesítenek, és ezeket új jeleket lefordítják valamilyen tevékenységre. Végül is jelek előállítása és információk továbbítása csak úgy mehet végbe, hogy néhány és időpontban az energia szintjét módosítják. Ennek a változásnak pedig – lehetőleg – a kommunikáció célállomására kell korlátozódnia. Mindazonáltal egész sor módszer áll rendelkezésre az energiaszinten a változások irányítására. A kommunikációs szakemberek mindig olyan megfigyelhető és irányítható jelenségek után kutatnak, amelyek a lehető legkevesebb energiafelhasználást igénylik. Mostani ismereteinkkel kétfajta távirányítású energiamódosítást különböztethetünk meg. Az egyiket megfelelően jelzett vagy formált anyagi tárgyakról hozzák létre, amelyeket egyik helyről vagy időpont-

ról a másikra viszünk. Ilyenek a lyukkártyák, a könyvek, a hanglemezek, a képek és a szobrok. Ezeket megfelelően jelölhetjük a régi „bizonylat" („egységnyi adathordozó") terminussal és fontos jegyeiket, formájukat stb. a „bejegyzés" szóval. Amikor a bejegyzéseket a bizonylatokra rögzítjük, akkor válnak mozgathatóvá és tartóssá. A másik végletre példa, hogy az információt múlandó és anyag nélküli jelenségek is hordozzák – nyomás, áramerősség, feszültség stb. – amelyek technikailag „jel-események", de itt a „jel" szó közönséges értelemben is alkalmazható rájuk.

Az információt gyakorlatilag és gazdaságilag a legcélszerűbben bejegyzések formájában rögzíthetjük. Így az olyan információfeldolgozó rendszerek, amelyek nagy mennyiségű információt tartalmaznak, tárolnak és rendeznek, bizonylatokkal működnek. Ahogy ezek egyre gyorsabbak lesznek, a szükséges energia a sebesség köbével növekszik és eljön az idő, amikor az anyagi tárgyak mechanikus kezelése lehetetlenné válik és puszta jelek váltják fel a bizonylatokat. Vannak olyan könyvtári tevékenységek, ahol ugyanezen okok miatt, bár különböző mértékben, a jelek kívánatosabbak, mint a bizonylatok. A kölcsönzőket, különösen a műszaki dokumentumok esetében, általában nem a könyv vagy a dokumentum anyagi, tárgyi megjelenése érdekli, hanem a (feltehetően) benne foglalt információ. Ezt az információt teljes mértékben közvetíteni lehet akkor is, ha valaki nem a dokumentumot magát, csupán annak képét közvetíti. Éppen ezért a facsimile kiküszöbölheti az egyre növekvő könyvtári előjegyzési, csomagolási és postázási munkát, valamint a kicsomagolást és az adminisztrációt a kölcsönzésnél, az egyéni kölcsönzők értesítését a kölcsönvevő könyvtár részéről, illetve e műveletek megismétlését, amikor a kölcsönzött művet visszaadják. Ráadásul az idő nagy részében a dokumentum nem hozzáférhető még a kölcsönző számára sem. Nagy létesítményben, vagy egy ritka népsűrűségű országban a kölcsönzési idő régen lejár, mielőtt az egyéni kölcsönző a könyvet megkapja. A könyvtárak közötti televíziós vagy telefacsimile kapcsolatok hamarosan megtérülnének, hiszen csökken az adminisztrációs és kezelési munka és megszűnik a rongálódás miatti veszteség.

A londoni bankok évek óta a televíziót használják kisebb információcsomagok közvetítésére. Újabban ezt vezették be a rajzok és diagramok továbbítására a repülőtéri irányításban is. A többlet, amelyre ennek a módszernek a könyvtári alkalmazásához szükség van, egy olyan praktikus hivatali berendezés, amelyben a fényérzékeny papírt kapcsolatba hozzuk egy lépcsővel és azonnal, helyileg előhívjuk, hogy a közvetített kép jól olvasható változatát megkapjuk. A papírt – a jogi bonyodalmak elkerülésére – később meg kell semmisíteni. Röviden, a modern technológia kiterjesztette az olvasás határait és lehetővé tette a könyvtárak számára, hogy bizonyos tevékenységekben a szállításról a képernyőre váltsanak át. Amit a modern technológia kínál, azt ma még a költségekre tekintettel talán vissza kell utasítani; valami mégis elkezdődött, amit már nem lehet megállítani.

Bizonyos ideig a könyvtárakban az automaták használatának legfontosabb előnye jóval inkább felépítési elveiknek, mint maguknak a tényleges berendezéseknek a felhasználásából fog származni. Ahhoz ugyanis, hogy egyet megépítsünk, először is tudnunk kell, hogy mit várunk tőle. Ez meglehetősen nehéz feladat, mivel a meghatározásnak pontosnak kell lennie. Miután már tudjuk, hogy mit kell majd csinálnia, meg kell fogalmaznunk a műveletnek vagy megtanulásának a szabályait. Általában előnyös a feladat végrehajtására kész automatákat beépíteni, de gyakorlati munkára készítettek néhány olyat is, amelyek tapasztalat alapján tanulnak. A kísérleti telefonkapcsolatok például a használói szokások feltárásával megkímélnek bennünket a nagyobb berendezések költségeitől. A hagyományos telefonközpontok arra a feltételezésre építenek, hogy bármely előfizető egyforma valószínűséggel hívja fel bármelyik másikat. A kísérleti kapcsolásokkor regisztrálják az előfizetők által leggyakrabban keresett számokat és a központ munkáját ennek megfelelően irányítják.

A használók szokásaihoz igazodó módszerek szükségessége a könyvtárakban nem új felfedezés. Hagyományosan ezt emberek oldják meg, olyan szabályok szerint, amelyeket ők maguk nem tudnak megfogalmazni. Ezt talán még akkor sem lenne érdemes mechanizálni, ha már nem így lenne. Bizonyára más, alapvetően emberi tevékenységek körébe sorolt könyvtári munkafolyamatokra is így igaz. Még ha nincs is szó egy valódi gép megépítéséről, a működési szabályok megfogalmazására irányuló kísérletek meglepően hasznosak lehetnek. Mint láttuk az eszközök csak növelik a tevékenységek végrehajtásához szükséges képességeinket. Így az ezeket irányító szabályokat kipróbálhatják olyan emberek, akik különböző jelekkel, tollal és ceruzával végeznek a szabályoknak pontosan megfelelő műveleteket. Azok az általában nagyon bonyolultnak tűnő szabályok, amelyek az adatok gépi formára fordítását irányítják, kihagyhatók. Így csak azok maradnak meg, amelyek logikailag megegyeznek az alaptevékenységgel.

Miután ezeket kézzel kipróbáltuk, három kérdést tehetünk fel: Ha az automata nem is készül el, felhasználhatók-e működésének szabályai az emberi tevékenység alapjául? Kifizetődőbb-e az alapvető feladatokra gépet szerkeszteni, ha emberi, illetve ha gépi fordítók alakítják befogadható formára az adatokat? Ha rendelkeznénk a géppel, amelybe a feladatokra vonatkozó összes releváns tapasztalatunk és információnk be lenne építve, mi a legtöbb, amit kihozhatnánk belőle?

Az első két kérdésre látszólag intellektuális tevékenység (matematikai szövegek németről angolra fordítása) gépi szabályainak kidolgozásakor pozitív választ kaptunk. Az ilyen szövegek jelentése pontosan meghatározott és a mondatok jelentése nagymértékben független a szomszédos mondatokétól. A szabályokat egy bizonyos automatára, a SWAC-ra (U.S. Bureau of Standards West Coast Computer) alkottuk meg. Így az egy időben kezelhető szótár nagysága és a szövegkörnyezet hossza pontosan meghatározott. Ez a specifikus feladat meg-

mutatta, hogy a fordítási nehézségek a várt helyett máshol jelentkeznek. A legnagyobb feladat nem a szintaxissal és a kifejezésekkel való megbirkózás, hanem a szótár összeállítása. A szintaxist a „beszédrészek" fogalmának általánosításával ragadták meg. Az eredeti szöveg minden egyes szavához meghatározott szabály szerint egy számot csatoltak, amely ugyanúgy függött a szó közvetlen szomszédaitól, mint magától a szótól. Amikor az automatikusan meghatározott mondatot vagy az önálló mellékmondatot a számok sorrendjében újrarendezték, a szavak az angolnak megfelelő sorrendbe kerültek. A valódi nehézséget a szavak egyenkénti lefordítása jelenti. Gépi használatra a leghatékonyabb szótár az, amelyben minden szót átlagosan egyformán gyakran használnak. Másképpen némelyik „potyautas" lesz és értékes helyeket foglal el a mágneses táron. A szokványos szótárak nem ilyenek, mivel azok vagy az összes használt szó felsorolására tesznek kísérletet, vagy olyan embereknek készülnek, akik már bizonyos szellemi szótárral rendelkeznek.

A gépi szótárt minden egyes szónál használni kell. Ez és az egyformán gyakori használat elve többek között azt is jelenti, hogy a gépi szótárnak a beszéd különböző részeit helyes arányban kell tartalmaznia. Kézenfekvő, hogy ha ezt a szótárt összeállították, ezzel és a szintaktikai számozási szabályok segítségével németül nem tudó emberek is képesek rutinfordításokat készíteni.

Más tevékenységek automatikus kivitelezés céljából elvégzett hasonló elemzése éppen ilyen hasznos lehet, még akkor is, ha a gépet sohasem készítjük el.

Az egyenletes megoszlású átlagos használat „kényelmes" elve alapvető fontosságú azon a területen, amelyet meglehetősen félrevezető módon információelméletnek – azaz a jelek elméletének – nevezünk, de ennek alkalmazása a könyvtári munkában jelen tanulmányunk határain kívül esik. A fordítási kísérlet azt is bemutatja, hogy a szövegkörnyezet problémái rutintevékenységek keretei között is megoldhatók. A könyvtári osztályozásban mindig is különös aggodalomra adott okot az, hogy a szavak jelentése előállításuk gyakorlati kontextusától függ. *Pierce* terminusával élve „indexek", nem pedig „szimbólumok" vagy „ikonok". *Richens* ettől függetlenül támadást indított a kontextus egy másik aspektusa ellen. Ugyanez az alapelv húzódik meg *Shannon* munkájában is, amely az angol szövegek felismerhetőségével foglalkozik.

Nem kell ezt a kérdéskört teljesen feldolgoznunk, mielőtt megkísérelnénk egy sürgető probléma gyakorlati megoldását: a kísérleti technikák bevezetését az információkeresésbe. Olyan téma ez, amely a magukat az ismeretek és jelenségek osztályozásához kötő, vagy éppen azzal összekeverő könyvtári osztályozási rendszerek számára voltaképpen sohasem lesz megközelíthető.

Shannon idézett munkája rávilágít az összetett kísérletezés és elmélet szükségességére az információs területen. Sőt az olyan kísérlet szükségességére is, amelyhez nem kell „zománcozott és krómcsillogású" berendezés. A kommunikációt kódoló eszközök teljesítményének ideális szintjét keresve arra volt kíváncsi, hogy egy géppel (amely a nyomtatott angol nyelv helyesírásának és szerkezetének

felfogásában egy felnőtt ember szintjén áll) maximálisan mennyire tömöríthető egy angol szöveg. Az eredmény egyértelműen 50% alatti. Hiszen ha egy nyomtatott oldalon akár a szavaknak, akár pedig a betűknek a felét véletlenszerűen elhagyjuk, a szöveg még mindig érthető lesz. Önként jelentkezőkkel végzett eredeti kísérleteivel azt próbálta meghatározni, hogy milyen valószínűséggel találnak ki helyesen egy betűt, ha előtte vagy mögötte (a valószínűség a két esetben körülbelül azonos) száz betű adott. Olyan közvetítőket és vevőket tételezett fel, amelyek a nyelvészeti tapasztalatok szerint az átlagos emberi kísérletező ikertestvérei. Ezek alapján részletes listát készített ama ingerekről és válaszokról, amelyek egy logikailag ekvivalens automata meghatározásához szükségesek, és képes volt az automata teljesítményét számszerűen megbecsülni. A listának tartalmaznia kell az összes lehetséges szóközökből, írásjelekből és betűkből álló száz karakteres angol nyelvű sorozatot. Ezeket csak töredékesen közölték. A kísérletek és elemzésük azt mutatja, hogy az angol szöveget negyedénél rövidebbre lehet tömöríteni információveszteség nélkül. Ez nem közömbös a helyesírási reform szempontjából, sem olyan gyorsírás kialakításakor, amely nemcsak a feljegyző emlékezetének meghosszabbítása, hanem mások által is olvasható. A tömörített szövegek szükségszerűen tartalmazhatnak átírási hibákat. Ahogy egy számsorban, itt is minden egyes jel független a többitől. Így nemcsak a hibák javítása, hanem még a szövegkörnyezet alapján történő kiderítésük is lehetetlenné válik.

Ezúttal megint csak az automaták mögött rejlő munka, és nem maga az automata lesz az, amely a könyvtárak számára értékes. Mindent összevetve, a kibernetika az információs szolgáltatásoknak azzal teszi a legnagyobb szolgálatot, hogy az elmélet és a könyvtárakra kiterjedő kísérletek segítségével feltárja azokat az elveket és jelenségeket, amelyek mind a kettőben közösek.

Nem várható olyan szenzációs előrehaladás, amilyennel a szabályozó berendezések és a többfunkciójú intézmények közötti felszínes analógiák kecsegtetnek. Először az alma mozgását kell tanulmányozni, csak azután a Holdét, de elértük azt a szintet, amelyen az elmélet a kis és világosan körülhatárolt gyakorlati problémák megoldására a siker némi reményével felhasználható.

Információelmélet és ügyviteli rendszerek[19]

Szemantikai információ

Képzeljünk el egy jelet, amelyet megfordítható szabályok szerint úgy alakítanak át (például nyomtatásból Morse-jelekre, vagy kifejezésekről Bentley-jelekre), hogy visszakódoláskor az információ tökéletesen visszanyerhető legyen. Mivel a fordítás visszakódolható, bárhogy is határozzuk meg, a jel infor-

19 Information theory and clerical systems. – In: Towards information retrieval, p. 22–41.

mációtartalmának mindkét formában azonosnak kell lennie. A visszakódolható fordítások között biztosan létezik egy, amely a legkevésbé komplikált. Ennek a legegyszerűbb szinonimának az összetettségi mértéke „bitekben" egyben minden ekvivalens fordítás információtartalmának a mértéke.

Természetesen szemantikai ismeretekkel kell rendelkeznünk ahhoz, hogy eldöntsük, mely jelek minősülnek szinonimáknak és melyek „kivonatoknak" abban az értelemben, hogy jelentősen rövidítettek vagy „elvesztették jelentésüket". Ezekre a szemantikai ismeretekre csak a kommunikáció legelején van szükség. Ha a fordítási szabályokat már megállapítottuk, feleslegessé válnak. Ez a szemantikai tudás szorosan kötődik az adott szövegkörnyezet szóhasználati módjához. „Jelentésük" bármely más összefüggésben elvész, mert semmi esetre sem tudja megváltoztatni a kommunikációs folyamatot és az ezáltal előidézett cselekvéseket. A könyvtárosnak vagy más tájékoztatási szakembernek tudnia kell, hogy a dokumentációs gyakorlatban a különböző „Thompsonok" és „Thomsonok" közül melyik azonos „Lord Kelvinnel", holott a személyes ismeretség az emberekkel nem elengedhetetlen.

Ilyenfajta meggondolások indították *Bar–Hillelt* és *Carnapot* a szemantikai információ olyan elméletének megfogalmazására, amelyben mind a „forma", mind a „tartalom" számszerű mértékkel rendelkezik. E könyv szerzője ezt a fejlődés ígéretes útjának tartja, de mivel ez a carnapi logikai valószínűségelemzés igen specializált eljárásait használja fel, részletes kritikája a szerző képességeit meghaladja. Lényege, hogy a valószínűséggel mint alapvető tényezővel számolva, logikailag pontosítható és számszerűsíthető az a mindeddig ingatag talajon álló feltételezés, hogy a konnotáció (intenzió) és a denotáció (extenzió) változásai ellentétesek. A valószínűség – mint határozatlanság – a kommunikáció egyik legalapvetőbb eleme. Ha előre ismernénk az elküldendő üzeneteket, nem kellene elküldenünk őket. A könyvtári kommunikáció főleg abban különbözik a táviratitól, hogy már minden üzenetet elküldtek, és ezek közül kell azt kiválasztani, amelyik megfelel egy eddig ismeretlen kérdésnek. A távírász arra törekszik, hogy a leggyakoribb üzeneteknek a legrövidebb jeleket feleltesse meg; a könyvtáros pedig azt próbálja megoldani, hogy a leggyakrabban kért dokumentumok lehessenek a legkönnyebben elérhetők. Sajnos a legtöbb módszer és elmélet egyszerűen nem számol a valószínűséggel és az összes eshetőség bekövetkezését hallgatólagosan egyformán valószínűnek tételezi fel.

Könyvtári szemantika

A tárgyi osztályozás szemantikája csak a bibliográfiai tájékoztatásban előforduló szinonimákkal foglalkozik. Nincs kapcsolata a természeti jelenségekkel vagy a tudományos képzésben előforduló szinonímiával; ezek a tudományos osztályozásra és a tudományos terminológiára tartoznak. Sőt, a szinoni-

mák a dokumentumokra mint fizikai tárgyakra vonatkoznak, amelyeket kezelni és rendezni kell, el kell küldeni és lehetőleg vissza kell kapni. Tudományról, mint olyanról, igazságról és hamisságról, erkölcsről stb. beszélni ebben az összefüggésben értelmetlen és haszontalan. A könyvtári osztályozás egyik célja például, hogy a szinonimákat egyetlen, egyszerű megnevezéssel helyettesítse, amely az összes releváns dokumentumra vagy az azokról készült leírásokra alkalmazható. Ez esetben a szinonimák olyan tájékoztatási feladatokhoz vezetnek, amelyek egy meghatározható könyvtári művelettel oldhatók meg.

Ez a helyzet az összes többi könyvtári tevékenységgel is, amelynek szemantikája érvényes az adott könyvtári felhasználásra, de nem vonatkozik az általánosabb vagy határozatlanabb értelemben vett „jelentés"-re. A könyvtári eljárások célja az információ hozzáférhetőségének és az információról adható információnak a biztosítása. Ha azt követelnénk, hogy a specifikus információval is foglalkozzék, gyakran a lehetetlent és sohasem az alapvetőt kérnénk.

A könyvtári tevékenység szemantikáját a használók tájékozódási helyzetekben tanúsított viselkedésének gyakorlati tanulmányozásával, körültekintő előrejelzésével alakíthatjuk ki. Ebből következően a kódolás szintaktikai problémái a kódolás és a tevékenységek (mint például a rendezés, a kiválasztás, a dokumentumok és a tételek elhelyezése) összehangolásának pragmatikus kérdései egyértelműen a konkrét feladatok költségének, idő- és munkaráfordításának csökkentésével kapcsolatos nehézségekként foghatók fel.

Az információ hálóelmélete

A fentiekben bemutattuk, hogy adott, legegyszerűbb jelsorozatnak és az aktuális szövegkörnyezeten belüli megfelelőinek minden visszakódolható fordítása a minimum jel (vagyis a legegyszerűbb szinonima) összetettségéhez viszonyított bitekben mérhető információtartalommal rendelkezik. Egy peremlyukkártyán ez az információtartalom a jelsorozat lyukasztásához szükséges pozíciók számának felel meg. Hollerith-, IBM- vagy Powers-kártyán pedig a pozíciók minimális száma szorozva log 10-zel (nagyjából tíz harmad), vagy log 12-vel (nagyjából negyvenhárom tizenketted), mivel ezek a gépek 10 vagy 12 lyukasztási pozíciót képesek egy helyen (kártyaoszlopon) belül megkülönböztetni.

Így a különböző jelek „jelentésük" kifejezett ismerete nélkül sorba rendezhetők. Nem olyan teljes sorba, mint például az A, 2, 3, . . . , 10 , J, Q, K, (A, …, V, D, R, A, … , B, D, K) láncok, hanem a rendezésnek abban az értelmében, hogy bármely két jelsorozatnak van egy legnagyobb közös osztója, a mindkét sorozatban előforduló közös részlet információértéke; és van egy legkisebb közös többszöröse, az a legrövidebb jelsorozat, amely mindkettőt tartalmazza. Ez a jelek információtartalmának logikai szorzata, metszete, egybeesése vagy legna-

gyobb alsó határa; illetve logikai összege, uniója, vagy legkisebb felső határa. Az ezek segítségével kialakított rendszer hálóhoz vagy rácshoz hasonló, és ennek a rendnek és információs mértéknek az algebrája (szimbolikus logika) a hálóalgebra.[20] Ezt mutatta ki néhány évvel ezelőtt e tanulmány szerzője, és Shannon elméletileg ezt fejlesztette tovább, majd Mooers közvetlenül könyvtári módszerekre alkalmazta.

A hálóalgebrának az itt tárgyalt összefüggések megvilágításához szükséges része nem különösen bonyolult. Sajnálatos módon a legtöbb magyarázat matematikailag jól képzett olvasókat tételez fel. Az átlagolvasó a fontosabb részek összefoglalását megtalálja Skrásek tanulmányában, amely az elméletet a tudományos (nem könyvtári) osztályozásra alkalmazta. Számára a legnagyobb problémát az jelenti, hogy egy téma osztályozó kifejezések több különböző láncával is kifejezhető. Ha egy láncból néhány szem hiányzik, hogyan kapcsolható az össze egy másik lánc elemeivel úgy, hogy ne keletkezzék zavar, amikor a hiányzó leíró adatot megállapítják és hozzáfűzik. Ez a két egymáson lévő labirintus problémája, amelyeket úgy helyeztek el, hogy ha az egyik valahol elzáródik, csapóajtón kijuthatunk a másikba, és később olyan helyen bukkanhatunk fel ismét, amely akkor is útba esett volna, ha nem teszünk kitérőt.

A lyukkártyarendszerek hálóalgebrája szigorú Boole-algebra, logikai összegük és szorzatuk eleve meghatározott. Az összeg egy olyan kártya, amelyet az összes többi kártyán kilyukasztott valamennyi pozícióban és csak azokban lyukasztanak ki. A szorzatot az a pozícióegyüttes adja, amely akkor látható vagy állapítható meg, ha a kártyákat egymásra helyezve a fény felé tartjuk. A segédeszközökben és a módszerekben akkor és csak akkor következik be jelentős egyszerűsödés, ha felismerjük, hogy az ilyen modellek is egyenrangú tagjai az információhálónak.

A logikai összeadás egy másik változata lehetséges a lyukkártyás számológépeken. Az ugyanarra a helyre (oszlopba) lyukasztott két lyuk közül csak a nagyobb számot jelölőt vesszük figyelembe, például ugyanannak az oszlopnak a 3-as és 6-os pozíciójába lyukasztott lyukak 6-ot jelentenek. Így az „12345" és az „54321" logikai összege „54345" lesz. E sajátosságnak gyakorlati haszna van. Ugyanez nem mondható el a számológépek bármely aritmetikai képességének dokumentációs felhasználásáról.

Az információ hálóelmélete közvetlenebb haszonnal jár a könyvtáros, mint a kommunikációs szakember számára, mivel a könyvtáros a jelet teljes egységként kezeli. A kommunikációs szakembernek minden vonatkozásban bitről-bitre kell foglalkoznia a jellel. Kevés lehetősége van a nem teljes jelek

20 Az információhálók általában nem modulusosak és gyakran nem disztributívak. Az a hiedelem, hogy ezek kivétel nélkül a Boole-logikának felelnek meg, csupán az elektrotechnikai szakemberek naiv természetéből táplálkozik.

tárolására, és a késlekedést sem fogadja szívesen. Helyzete olyan mint azé az emberé, akinek a német külpolitikáról vallott téves nézetei abból fakadnak, hogy sohasem jutott el az újságok utolsó oldalára, ahol pedig az összes NICHT-et megtalálta volna. Hasonló probléma merül fel a nyelvészetben is. Mandelbrot elemzése szerint a szokványos beszélgetésekben a természetes nyelvek a pontatlan és zavaros mindennapi nyelv használatában optimális kompromisszumot alakítottak ki az emlékezetben megőrzendő anyag és a kommunikáció között.

Bár a könyvtáros az információval mint egésszel foglalkozik, a jelek rendjét időnként meg kell változtatnia. A kommunikációs szakemberek teljesen elhanyagolták mind az olyan helyzetek elemzését, amelyben a jelek sorrendjét meg kell változtatni, mind maguknak a jeleknek a vizsgálatát. A szerző kiadatlan műve azt sugallja, hogy egész kevés matematikai elemzéssel hasznos mennyiségi kapcsolatok mutathatók ki a különböző ügyviteli folyamatokra fordított munka és az információk kódolásának módja között. Ezek közül néhányat később ismertetünk.

Az osztályozás delegálása (az osztályozás megosztása és áthelyezése más – automatizált – szintekre)[21]

A könyvtári osztályozások a szövegekhez csatolható jelzetek jegyzékét képviselik, de ritkán adnak utasítást ahhoz, hogyan rendelendő a megfelelő jelzet a megfelelő szöveghez. A könyvtári osztályozás elmélete, amelynek az a célja, hogy megmondja, hogyan fogalmazzuk meg ezeket a jelzeteket, még ennél is szemérmesebb. Az osztályozás művészet, amelyről szégyen beszélni, és maguk a szövegek sem jobbak, mint amilyennek lenniük kell. Beszélgetésekben óvatosan a „tudás világát" emlegetjük, amin legfeljebb a „nyelv világát" értjük, de gyakran csupán a „kiadott szövegek verbális tartalmát"; vagy a „fogalmakról" fecsegünk, és közben csak „az osztályozási rendszer fordítása után változatlanul maradó részekre" gondolunk, vagy nemegyszer a „szöveges tartalom leírásának módjára" stb.

Súlyos ára van ennek a szégyenlősségnek: az a mélyen gyökerező hit – amely annál is veszélyesebb, mert igyekszünk nem észre venni –, hogy a fogalmak világa abszolút örök és változatlan, amit lépésről-lépésre tárunk fel.

Egy speciális osztályozás a „tudás világának" bizonyos részeit térképezi fel, az általános osztályozás pedig mindannak a térképe, ami a mai napig elénk tárult. Az osztályozó, ha egyáltalán szóba kerül, felfedező és nem feltaláló. Bármely éles szemű és helyes észjárású megfigyelő egy „gondolatot", állítást

21 Delegation of classification. In : Towards information retrieval, p. 124–134. [Eredetileg megjelent 1957-ben.]

vagy dokumentumot (sohasem egészen egyértelmű, hogy milyen entitásokra irányul a megfigyelés) bárhol, bármikor, korábbi olvasmányaitól függetlenül ugyanúgy fog osztályozni.

Más szavakkal, úgy képzeljük, hogy a dolgok nyakában névtábla lóg és csak annyit kell tennünk, hogy elkapjuk őket és elolvassuk a táblákat. Ezért a könyvtári osztályozás központi problémája az, hogy ezeket a feliratokat – jelzeteket – hogyan rendezzük el olyan módon, hogy azok különböző célokra, a leginkább megfeleljenek. Ez nem egészen felel meg a gyakorlati tapasztalatnak, amelyben a fő nehézség először a helyes osztályozás, másodszor a helyesen osztályozott dolgok megtalálása. Az első feladat átadása, megosztása – delegálása – nagyon nehéz, a második – az információkeresés – „átruházhatósága" pedig nagymértékben az elsőtől függ.

Egy másik, talán kevésbé támadható álláspont szerint az elnevezendő vagy leírandó dolgok nevét vagy leírását nem fedezhetjük fel úgy, hogy megvizsgáljuk őket, hiszen ezek nem a dolgokban rejlenek, hanem mi ruházzuk rájuk őket. Sok megfigyelést végeztek az asztronómia területén, de ezek eddig semmiféle információhoz nem vezettek a csillagképek neveit illetően. Ezek kiderítésére azt kell megfigyelnünk, hogy az emberek hogyan és miért beszélnek a csillagokról, nem pedig magukat a csillagokat. A dolgok nem saját maguk leírásai; ez az alapja Zénon paradoxonjának Achillesről és a teknősbékáról. A verseny leírása az, és nem maga a verseny, ami végtelen. Ismét, ha a dolgok önmaguk lennének, minden információkeresési problémánk tökéletesen megoldható lenne, korlátlan számú facsimile másolattal.

Mindez nyilvánvaló. Kevés osztályozáselmélettel foglalkozó szakember hiszi – világosabb pillanataiban –, hogy a szövegek már azelőtt is osztályozottak lennének, hogy valaki azokat osztályozta volna, vagy még inkább, mielőtt azokat valaki megírta volna. Mégis sokan viselkednek úgy, mintha valóban azt hinnék, nem mintha ostobák vagy perverzek lennének, hanem mert bizonyos fontos helyzetekben az osztályozó jelenléte nem mindig nyilvánvaló. Akkor sincs láthatóan jelen, ha a feladata eléggé szűk ahhoz, hogy valaki másnak a nevében dolgozzon. Az osztályozó a könyvtárelméletben ugyanazt a szerepet tölti be, mint amit a viszkozitás az aerodinamika elméletében. Bármennyire is elhanyagolható mennyiségileg, kihagyása a tapasztalat minőségileg abszurd modelljéhez vezet.

Vegyük először az információkeresés pusztán ügyviteli részét, ahol a jelölt tételeken végzett műveleteket teljes egészükben maguk a jelek határozzák meg. Ha a rendszerbe semmilyen új tétel nem kerülhet, és ha minden tételt megfelelően jelöltek és mindegyik hozzáférhető, elfelejthetjük, hogy valakinek az osztályozást is el kellett végeznie. A tulajdonságok, kategóriák és hasonlók terjedelme teljesen meghatározható az ezekkel jelzett tételek halmazaival. Az is állítható, hogy egy tétel vagy beletartozik adott kategóriába, vagy nem, mivel elvileg a tétel elővehető és a jelekből kideríthető, hova tartozik.

Ezen a területen a Boole-algebra érvényesül, különösen a kettős komplementum: azon tételek halmaza, amelyek nem „nem A”-k azonos azzal a halmazzal, amelybe az összes A-k tartoznak.

Szigorúan Boole-algebrai rendszer ritka, még a számítógépek esetében is. Általában ez a nem Boole-algebrai rendszer pillanatfelvétele premier plánból, és emiatt csak helyileg és időlegesen érvényes. Az osztályozót azonban a fejlődő rendszerben is el lehet egy bizonyos pontig hanyagolni, de erre a használók fizetnek rá. A fejlődő rendszerben mindig lesznek olyan tételek, amelyeket még nem osztályoztak, és még nem hozzáférhetők. Nem mondhatjuk el róluk, hogy adott kategóriában osztályozták azokat vagy sem, de a könyvtár használójának legalább a létezésükről tudnia kell. Tehát hozzáadjuk ezeknek a tételeknek a listáját minden egyes listához, amely a már valamelyik kategóriába egyértelműen besorolt és osztályozott tételeket tartalmazza. A rendszer nem tudja ezeket plusz csak a kért egységeket keresni. Ehelyett vagy a teljes pillanatnyi állományt számba veszi és keresi, és nemcsak azt, amit kértek, vagy elhanyagolja a nem osztályozott vagy nem hozzáférhető állományrészt, és csak ezeket hívja vissza, és nem mindent amit kértek.

Ez az elkerülhetetlen tudatlanság elkerülhetetlen következménye, és éppen ezért egyetlen használható elmélet sem feledkezhet meg erről. Nem tudunk eleget ahhoz, hogy egyetlen pontos megkülönböztetést tegyünk az A-k és „nem-A”-k között, csak azt, hogy két lazább megkülönböztetést adhassunk. Vagyis: most egy helyett kétfajta komplementumunk van, és kettő is lesz, nem pedig egy, bármennyire kicsi is legyen a kettő közötti különbség.

Az egyik komplementum a dolgok legnagyobb halmaza, amely biztosan nem tartalmaz egy A-t sem. Ez a „pszeudo-komplementum”, amelyet itt A*-gal jelölünk. Ha tudatlanságunk csökken, ez a halmaz növekedhet, de sohasem csökkenhet. Egy kis gondolkodás után rájövünk, hogy a kétszeres pszeudo-komplementum A** a dolgok legkisebb halmaza, amely biztosan tartalmazza az összes A-t.

A másik komplementum dolgok legkisebb halmaza, amely biztosan tartalmaz minden „nem A-t”. Ez a „Brouwer-féle komplementum”, amit ‘A-val jelölünk. Ha tudatlanságunk csökken, ez a halmaz is csökkenhet, de sohasem növekedhet. A kettős Brouwer-féle komplementum ‘’A a legnagyobb halmaz, amely bizonyosan kizárólag A-kat tartalmaz.

A „nem A” valamennyi értelmezése következetesen használható, még akkor is, ha a dolgok halmaza az idő múlásával változik. Ezek mindegyikének egy különleges algebra felel meg, amely hasonló a Boole-algebrához, de lazább annál, így bizonyos kiegészítéseket tehetünk anélkül, hogy tönkretennénk a korábban vagy máshol feldolgozottakat. Mindkét algebra alkalmas a keresésre, és a részlegesen rendezett halmaz hálói disztributívak. Mint már korábbi tanulmányokban bemutattuk, ez a megközelítés kényelmes, még ha nem is szükségszerű. A két algebra szemmel láthatóan közeli kapcsolatban áll

egymással. Az egyik a másikkal párba állítható, vagyis, ha az egységek halmazáról beszélünk, az egyik halmazainak a közös tagsága megfelel a másik halmazai teljes tagságának.

Mivel sem a halmazok, sem a műveletek nem azonosak, tudnunk kell, hogy mikor melyik módszert használjuk, hogy kellene használnunk. Nem biztos, hogy az osztályozó, az információkereső rendszer és a használó megközelítése egyforma. Általában az egyszerűbb ügyviteli rendszerek sem alkalmazhatóak mindkét módon ugyanazzal a könnyedséggel. Azért egyszerűek és időnként olcsók, mert csak egyetlen tevékenység elvégzésére tervezték őket, vagy a rendszerek közös részének, „szorzatának", „metszetének" vagy „találkozásának" a kialakítására, vagy a rendszerek szuperpozíciójának, „összegének", „egyesítésének" vagy „uniójának" megalkotására. Minden résztvevőnek, az osztályozótól a felhasználóig, tudnia kell, hogy teljesítenie kell-e és hogyan a „minden, de nemcsak" és a „csak, de nem minden" feltételt.

Ezzel a módszerrel az osztályozási feladatok csak időlegesen adhatók át. Ha az osztályozó lemarad, nem kell visszatartania minden a nem osztályozott dokumentumra vonatkozó hivatkozást. A rendszer ezeket automatikusan „minden, de nemcsak"-nak osztályozza. Ez a kettős pszeudo-komplementum minden kategóriában. Később az osztályozó megmondja, hogy ezek közül melyiket kell a kettős Brouwer-féle komplementum státusára emelni; tehát a „minden de nemcsak A"-ról a „csak, de nem minden A"-ra, amelynek minden tagja biztosan A.

Eddig úgy kezeltük az osztályozót mint egy készséges, csalhatatlan, bár késlekedő kívülállót. Mielőtt tevékenységét közelről megvizsgáljuk, be kell vezetnünk egy fontos fogalmat, amelyet jobban be lehet mutatni elemek egy halmazára alkalmazva, mint egy elvontabb síkon. A Brouwer-féle és a pszeudo-komplementum algebrákat halmazokra alkalmazva mutattuk be az információkereső rendszerekben, nemcsak azért, mert ezek az alkalmazások alapvetőek, hanem azért is, mert a józan ész is erre késztet bennünket, és jobban megfelelnek a valóságnak, mint a Boole-algebrai modell. Általánosabb szinten is realisztikusabbak, de ha nem kezeljük őket óvatosan, metafizikai irrelevanciákat kelthetnek fel. Igazában nagyon földhözragadt ok miatt kell használnunk őket. Az osztályozási tevékenységekben, amelyekben a megfigyelő vagy a közvetítő szerepe lényeges, minden áron el kell kerülnünk az olyan elméleteket, amelyek feltételezik, hogy olyan információt is használhatunk, amely nem áll rendelkezésünkre, és olyan dolgokat is megtehetünk, amelyekhez nincsenek meg az eszközeink. Ez az elv vonatkozik a nyelvészetre és ezért a könyvtári osztályozásra is legalább annyira, mint a fizikára, amely a relativitás és a kvantummechanika létrehozásával mozdult ki a holtpontról. Most a „távolság" fogalmát vesszük szemügyre, a szövegek tartalmára vonatkoztatva. Ennek természetesen ebben az összefüggésben is van értelme, vagy több értelme, mivel beszélhetünk olyan témakörökről, amelyek „kö-

zel" vagy „távol" állnak, és ezzel valami hasznosat mondhatunk az embereknek. Világossá tehetjük ezt azáltal, hogy valamilyen szabály segítségével megállapítjuk, hogy egy témapár közelebb vagy távolabb van-e, mint egy másik? Ha igen, felállíthatunk-e a témák között objektív távolsági skálát? Ez az „irrelevancia" skálája lenne. Ez esetben az osztályozás feladatai átruházhatók, amennyiben valaki másnak meg tudjuk mondani, hogy milyen mértékű irrelevancia engedhető meg egy szövegnek egy témához történő hozzárendelésekor, és mekkora irrelevanciát tűrünk meg a keresésben. Ez utóbbi határozza meg az egységek elvesztésének (vissza nem keresésének) kockázatát, ami így egy adott rendszeren belül kiszámítható.

A távolság minden összefüggésben a „különbség" fogalmán alapszik. Általánosságban két dolog közötti különbözőség az egyiknek az a része, amely nem része a másiknak. Pozitívabb meghatározással, ez az, amit hozzá kell tennünk az egyikhez, hogy a másikat is magába foglalja. A jelen összefüggésben a távolság az elemek egyik halmaza és a másik között azon elemek halmaza, amelyek csak az egyik halmazhoz tartoznak és a másikhoz nem. Hasonlóképpen, a különbség a másik halmaz és az egyik között az elemek azon halmaza, amely csak a másikhoz tartozik és az egyikhez nem. Például a különbözőség a „minden, de nemcsak" halmaz és a „csak, de nem minden" halmaz között az elemek ama halmaza, amelyik ki van téve annak a kiegészítésnek, amely a halmazok közötti határt képezi. A „csak, de nem minden" és a „minden, de nemcsak" halmazok közötti különbség nulla, mivel nem létezik olyan elem, amely úgy tartozhatna az elsőhöz, hogy a másodikhoz nem tartozik.

Két halmaz közötti távolság ezen két különbség halmaza. Vagyis, ama elemek halmaza, amelyek az egyikhez vagy a másikhoz tartoznak, de sohasem a kettőhöz együtt. Ezt gyakran a két halmaz „szimmetrikus különbségének", „kizáró diszjunkciójának" vagy néha „exjunkciójának" nevezzük. A szimmetrikus különbség biztosan mutatja mindazokat a formális tulajdonságokat, amelyekkel a távolság egy megbízható mértékének rendelkeznie kell. Először: a halmaz és önmaga közötti távolság nulla, mivel nincsen olyan elem, amely egyszerre bele is tartozna, meg nem is. Másodszor: az egyik és a másik halmaz közötti különbség azonos a másik és az egyik halmaz közötti különbséggel. Harmadszor: az első és a második, valamint a második és harmadik halmaz közötti távolság összege nem lehet kevesebb, mint az első és a harmadik halmaz közötti távolság. Ez a „háromszög egyenlőtlenség".

Furcsának tűnhet két tárgy közötti távolságot ugyanolyan fajta tárgynak vennünk, még akkor is, ha az a formális követelményeknek megfelel. Kénytelenek vagyunk ezt megtenni, mivel más tárgyak nem állnak rendelkezésünkre. Abszurd lenne a relevancia kereteként ama területek geometriai tulajdonságait használni, amelyekben a fizikai tárgyaknak vagy bejegyzéseknek tekintett elemeket tároljuk vagy leírjuk. A tevékenység magasabb szintjén törekednünk kell arra, hogy a szövegek és bejegyzések fizikai és tipográfiai szem-

pontból – amennyire lehetséges – az osztályozás szerkezetét tükrözzék vissza, a kérdéses feladatnak leginkább megfelelő módon. Ez határozottan a jelzetelés problémája: a könyvespolcok, kártyák és számítógépek ugyanúgy jelzetelési elemek lehetnek, mint a papírlapok vagy az ábécé. Ezek szükségszerűen három fizikai és egy időbeli dimenzióra korlátozódnak. Az általuk képviselt osztályozási rendszer vagy rendszerek matematikai értelemben majdnem biztosan sokkal többdimenziósak. Még ha tudja is valaki, hogy mit akar és mit tud tenni, megmarad a többdimenziós perspektivikus modellezés megfogható gyakorlati és elméleti problémája.

Jelenleg azonban csak azzal törődünk, hogy a lehető legmesszebb eljussunk azokat az elemeket felhasználva, amelyekkel rendelkezünk, és ahogy azokat csoportosítottuk. Feltételezzük, hogy saját szerkezetüket teremtik meg, ahogy a rendszer növekszik, és nincsenek beágyazva semmilyen már létező rendszerbe.

Ebből a szempontból a halmazok szimmetrikus különbsége kielégíti a távolság mind intuitív, mint formális követelményeit. Két téma olyan messze áll egymástól, amennyire csak lehet, ha nincs közös elemük, és a távolság nő, ha olyan új egységek érkeznek, amelyek csak az egyikről vagy a másikról szólnak, de nem mindkettőről. A két szöveg közötti távolságot nem változtatják meg azok a szövegek, amelyek mindkettőről szólnak. A távolság legkisebb egysége egy témából az a szöveg, amelynek a témája különböző. Ha kifinomultabb mértéket akarunk, a szöveg kisebb egységét kell vennünk, és a megkülönböztetést élesítenünk. Nincs ugyanis értelme annak, hogy egy olyan szöveg tartalmában tegyünk megkülönböztetéseket, amelyet nem tudunk kisebb egységekre bontani, vagy, hogy egy szöveget felosszunk, amelynek részeit nem tudjuk megkülönböztetni.

Mostanáig az egyszerűség kedvéért feltételeztük, hogy a távolságok a Boole-algebrai összegzésnek felelnek meg, ahol minden dolog osztályozott és elérhető. Valójában két távolság létezik, amely a Brouwer-féle, illetve a pszeudo-komplementumnak felel meg. A Brouwer-féle távolság a nagyobb, mivel ez azoknak az egységeknek a legkisebb halmaza, amely így biztosan tartalmazza az összes egységet, amely az egyik halmazhoz tartozik, de nem mindkettőhöz. A pszeudo-komplementumban jelentkező távolság a kisebb, mivel ama egységek legnagyobb halmaza, amelyek biztosan a halmazok egyikéhez tartoznak, de nem mindkettőhöz. A rendszerek matematikája vagy geometriája, amely ezekkel a távolságokkal dolgozik, már kidolgozott és megtalálható a matematikai folyóiratokban.

E két távolság különbségével mérhető egy adott rendszerben a két téma relevanciájára vonatkozó bizonytalanság. Bár eddig csak a késedelemből vagy a hozzáférhetetlenségből eredő bizonytalanságokról beszéltünk, most már hozzávehetjük az osztályozás néhány bizonytalanságát is. Az „A” vagy „nem A” döntéseket még mindig visszavonhatatlannak tekintjük, de az osztályozó-

nak joga van átmeneti döntéseket hozni, amelyek később módosulnak. Az ilyen módon jelölt szövegek a kettős pszeudo- és kettős Brouwer-féle komplementumba tartozó jelzett egységek – A** és ‘’A – különbségébe esnek, a késleltetett és hozzáférhetetlen hivatkozásokkal együtt.

Mivel most már elég információval rendelkezünk a rendszer állapotáról, kiszámíthatjuk azt a hibát vagy kiesést, amelyet a használó a témák bármely kombinációjában várhat. Ezek, és a témák közötti különbségek könyvtárról könyvtárra és napról napra változnak. Mindazonáltal amit az ember a homályos „szerkezet” névvel illet, a mennyiségi mértékek változtatásai dacára, állandó maradhat. Ekkor mondjuk, hogy a rendszer, az osztályozók és használók, ugyanazt a „speciális osztályozást” alkalmazzák.

A legtöbb osztályozási tevékenység esetében hallgatólagosan feltételeznek valamilyen speciális osztályozási rendszert használó osztályozót, nem is alaptalanul. Ha nem tudjuk a problémát egy könyvtárra, egy nézőpont alkalmazásával megoldani, nem tudjuk majd megoldani több könyvtár számára, több nézőpont alkalmazásával. Ez természetesen nem azt jelenti, hogy ha elképzelünk különlegesen nagy és bonyolult szuperosztályozást használó szuperosztályozót, akkor megtaláljuk a megoldást. Amint azt éppen láttuk, a különböző helyi rendszerek a különböző osztályozók szemében torznak látszanak, mivel a témák közötti távolságok különbözőek. Csupán a nulla távolsághoz vezető döntésekben érthetnek egyet, amikor bizonyos egységről megállapítják, hogy egy adott témában releváns. Az irrelevancia megítélése, még ha az megtámadhatatlan is, gyengébb, mivel az irrelevancia mértéke, amit „távolságnak” neveztünk, időről időre és rendszerről rendszerre változik, míg a nulla távolság mindig nulla marad. Egy speciális osztályozásban a téma egybeesésének e változatlansága és egy speciális relativitásban a fizikai egybeesés változatlansága közötti hasonlóság nem teljesen esik egybe.

Fel kell tételeznünk, hogy az osztályozó így el tudja dönteni, hogy egy szöveg adott témakörben releváns-e, hogy – a melléfogásoktól eltekintve – sem a jövőbeni fejlődés, sem a máshol hozott döntések nem vonnak maguk után módosításokat. A későbbi fejlődés semmiképpen sem mentesítheti a relevanciáról hozott döntéseket; ha egy egység bizonyos témakörben releváns, mindig is releváns lesz, bár a relevancia elveszthetí jelentőségét és új relevanciák adódhatnak hozzá. Más osztályozók azonban, még azok is, akik ugyanazt a speciális osztályozást alkalmazzák, máshol, máskor, ellentmondó döntéseket hozhatnak. Ez az egy ember általi helyi osztályozást kivéve minden mást lehetetlenné tenne. Ez mind térben, mind pedig időben korlátozott, hiszen egyetlen könyvtáros sem működik tovább osztályozóként, mint amíg hivatali megbízása vagy élete tart, bármelyik legyen is a rövidebb. Ha mindent magának kell csinálnia, beleértve az osztályozást és az információkeresést, nem kell tudnia, hogyan csinálja, csak tennie kell. Ha munkájának egy részét másra akarja bízni, meg kell alkotnia a relevancia ellentmondásmentes eldöntésének szabályait.

A szabályok különböző szintűek lehetnek. Ha semmilyen szabályt nem ismerünk, de munkánkat el tudjuk végezni, csak az utánzás szabályait adhatjuk meg: „a nehéz út megtanulása”. Ez a módszer egyben hosszadalmas és költséges is és a pontos utánzásra képes, jó memóriájú és a negatív visszacsatolásokra azonnal reagáló embert kíván meg. Ha van valami elképzelésünk arról, hogyan végezzük a munkát, az utánzást helyettesítheti a tapasztalat. Ha eleget tudunk, csak készséget kívánó világos utasításokat adhatunk, nem kell a tapasztalatra vagy kezdeményezőkészségre várnunk. Ezt a lehetőséget erősíti az a feltételezés, hogy bizonyos döntések megismétlődnek, bárki is hozza őket, vagyis az a feltételezés, hogy a több ember által használható osztályozások egyáltalán lehetségesek.

Az ilyen szabályok megfelelnek a szövegek osztályozására szolgáló gyakorlati eljárásoknak. Az egy terület szövegeit jól ismerő osztályozó ránézésre meg tudja határozni egy szöveg relevanciáját. A kevesebb tapasztalattal rendelkező osztályozók valamilyen listában vagy táblázatban megadott terminusok alapján ismerik fel a relevanciákat. E rendszerek bizonyos értelemben az osztályozók tapasztalatának átadását szolgálják. Végül a relevanciát összehasonlítások által is megállapíthatjuk. A szöveget általában a szókincs szintjén összehasonlítjuk már osztályozott szövegekkel. Ezekben és a róluk született döntésekben szükségszerűen testet öltenek az osztályozás szabályai. Ez alkalmazható osztályozás, mivel az osztályozás alapvetően a szöveg szavaiból áll, és azok alapján történik. Az osztályozás olyan rendkívüli kezelési, olvasási és emlékezeti képességeket igényel, hogy az embernek azonosítók és felismerők formájában intelligenciájával és műveltségével kell e képességeket helyettesítenie. Ennek ellenére, amíg felismerés és azonosítás nélkül végre nem hajtunk összehasonlítást, nem valószínű, hogy felfedezzük a szövegek osztályozásának szabályait, a szövegekét, amelyeket szavak gyűjteményeinek tekintünk. Ezek az alapvető szabályok, még akkor is, ha az elvontabb entitásokra való fordítás mindig kényelmesebb lehet.

Az összehasonlítás logikai és matematikai szempontjait általánosságban nagyon élénken vizsgálják mostanában. Az osztályozáselmélettel foglalkozók is tanulhatnak ebből valamit, különösen ami azt a szerkezetfajtát illeti, amelyet ez és az egybeesések állandósága a speciális osztályozásokban eredményez. Ez a különböző rendszereket konzisztensebbekké tehetné, és hozzásegítenek ama egyértelmű szabályok megfogalmazásához, amelyek lehetővé teszik az osztályozói munka megosztását és áthelyezését alacsonyabb szintekre.

Egészen mostanáig olyan speciális osztályozásokkal foglalkoztunk, amelyekben a szövegeket releváns vagy nem releváns voltuk alapján különböző „tárgyszavakhoz” rendeltük, amelyek – kicsit pongyolán fogalmazva – az emberi tevékenység és érdeklődés területeit képviselik. Bármik legyenek is a relevancia meghatározásának szabályai, minden alkalommal újra kell dönte-

nünk egy szövegről, ha egy másfajta tárgyi osztályozásban vagy akár információkereső rendszerben használjuk. Ez a felfogás bírálható ama józan megfontolás alapján, hogy a szövegek azok, amik, és olyanok is maradnak, így ez puszta gondolat- és időpocsékolás.

A problémát akkor sem oldjuk meg, ha feltételezzük, hogy a tárgyi osztályozás egy azonos logikai típusú, kiterjedtebb és részletezőbb osztályozásba ágyazódik. Bármilyen szerkezetben is térképezzük fel a speciális osztályozásokat, annak feltétlenül lazább a rendje, mint a részlegesen rendezett speciális osztályozásoknak. Mivel egy tárgyszó, amely – az információkeresés elméletében most már ismerő értelemben – „tartalmaz" egy másik tárgyszót az egyik speciális osztályozásban, egy másik osztályozásban a második tárgyszó része, és egy harmadikban lehet, hogy nincs közöttük kapcsolat. Egy általános osztályozásban nem létezhet a „belefoglalás" vagy ennek megfelelő kapcsolat, mint például az „előtt" és az „után". A lehető legszorosabb rendezést az „ekvivalencia", a „hasonlóság", a „távolság" és hasonlók hozhatnak létre. A speciális osztályozások analógok egy térkép alapján kialakított perspektivikus látásmóddal, és a térképeknek nincs tág perspektivikus látásmódjuk. A speciális osztályozás matematikája az általános osztályozáséhoz képest valószínűleg egy megfelelő algebra kongruens relációinak felel meg.

Az ilyen térkép elemei az egyszer és mindenkorra elhelyezett leírások. Mivel állandóak, egy szöveg vagy tárgy leírása bármely ésszerű típusának lefordíthatónak kell lennie egy hasonló finomságú másik típusra. A leírások alkalmazhatóak szövegekre és a speciális osztályozások tárgyszavaira, hiszen a „tárgyszó" nem egy önmagában álló egység, hanem azoknak a szövegcsoportoknak a címkéje, amelyek „ezen a speciális tevékenységen belül relevánsnak" ítéltettek. Sok leírás fog egyetlen tárgyszónak megfelelni és egy leírás sok tárgyszó alá bekerülhet, különböző speciális tevékenységek alapján. Mind a leírásokat, mind a tárgyszavakat, ha annak vesszük azokat, amik – azaz szöveghalmazoknak –, elméletileg kibővíthetjük a szövegek összehasonlításával. Azok a csoportok, amelyeket a leírások összehasonlításával alkottunk, állandóbbak, de szükségszerűen kevésbé rendezettek, mint a fentiekben hangsúlyoztuk. Csak a kis gyűjtemények tudják, vagy a többcélú gyűjteményeknek kell lerövidíteni a keresési kört tárgyszavak elhagyásával és a leírások keresésével, mivel ehhez minden egyes egységet meg kell vizsgálni. A tárgyi osztályozás és a rendezés teszi lehetővé, hogy az egységek egyes csoportjait a rendszer felosztásának különböző szintjein ki lehessen zárni, a legfinomabb felosztást a végére hagyva.

A leírások „rövidre zárhatók" úgy is, hogy felsoroljuk az összes speciális osztályozás valamennyi tárgyszavát, amelyre nézve egy szöveg most és a jövőben releváns. Ez az eljárás amellett, hogy a használata kényelmetlen, állandó bővítést és mindentudó osztályozót kíván. A háromszöget például olyan listával írhatnánk le, amely az eddig kitalált és a jövőben lelkes, geometriával

foglalkozó emberek által kitalálandó háromszögek minden tulajdonságát tartalmazza, és amelyet ki kell egészíteni a háromszögek összes lehetséges felhasználásával. Ez a leírás valószínűleg tágabb lesz, mint bármelyik leírt tárgy. Egy háromszög úgy írható le a legjobban, hogy megmutatjuk, hogyan kell lerajzolni.

A szövegek konstruktív leírása lehetséges és sok területen használják is ezt. Ezek a szabványosabb leírási módok megfelelői. A leírt dolog részeinek pontos specifikációinak az összege helyett ezek az egész pontatlan specifikációit kapcsolják össze. A leírás két fajtáját „végrehajtónak" és „irányítónak" nevezhetjük, de itt „hagyományosnak" és „tezaurusznak" hívjuk őket. A két típus elméletileg lefordítható, de a tezaurusz-leírás elemei nem közönséges nyelvi szavak vagy kifejezések, hiszen a mindennapi nyelv nem az ellenőrzött pontatlanság laza kifejezésére szolgál. Az elemek olyan tezaurusz szóklaszterei, amely a leírás egy adott elemének megfelelő szövegklaszterekeknek felel meg. A szavak nyelvről nyelvre változnak, de a szövegek klaszterei azonosak ama határokon belül, amelyeket a fentiekben bevezetett „halmaztávolsággal" mérhetünk. Mivel a leírás eme elemei pontatlanok és a pontosságot közös előfordulásuk adja, kétértelműségek és különösebb képességek nélkül is kijelölhetők. Ez az alapja a „tulajdonság felismerésen" alapuló működő rendszereknek, bár ez természetesen egy szűkebb osztályozási probléma.

Az utóbbi években a tezaurusz módszereket többen is alkalmazták egymástól függetlenül. Az információkereső rendszerekben hol világosnak, hol alkalmazhatatlannak bizonyultak. Kémiai indexeket egyszerűsítettek olyan tezaurusz csoportosításokkal, amelyek retorikai formákon alapulnak. Különösen a gépi fordítás bizonyos módszereiben használták fel azt közvetlenül és fordítottan. A Cambridge Language Research Unit munkatársai szótársorozatokat alkottak tezaurusz deszkriptorokból. A leírások és osztályozások azért ellenőrizhetők, mivel azoknak a mondatoknak, amelyeknek fő szavait a speciális tezaurusz-csoportokból vonták ki, csak bizonyos meghatározott értelemben kell felismerhetően megfelelniük a leírt szövegeknek. A kérdéseket, hogy a szövegek milyen „messze" lehetnek a klaszter centroidjaitól, és egyáltalán mik ezek a centroidok, a halmazok távolságával és a szerkezetek rokonságával pontosíthatjuk, oldhatjuk meg és közölhetjük másokkal. A válasz természetesen változik, attól függően, hogy a tezaurusz szavakból vagy szövegekből áll-e, valamint a nyelv, a leírás típusa vagy a speciális tárgy szerint is. Mivel bizonyos szövegcsoportosítások változatlanok lehetnek, a variációktól eltekintve valamilyen „alapfogalommal" foglalkozunk. A csoportosítás és a mérték változásait a távolsággal mérhetjük, és így objektív információt adhatunk az interpretáció és a leírás toleranciájáról speciális célok érdekében. Az egyszerű, de lehetetlen „A" vagy „nem A" döntéseket felcseréljük az „A" vagy „több mint A ennyi egysége" döntésekkel, és a döntések helyességét objektíven ellenőrizhetjük.

Így az osztályozás terheit részben levehetjük a vállunkról úgy, hogy a szövegekkel, szótárakkal és speciális témákkal kapcsolatos rabszolgamunkát a különböző tezauruszok, indexek összeállítóira hárítjuk át. Hogy ez hatékony ügyviteli segédletek hiányában mennyire kivitelezhető, nem lehet megmondani. A konkordanciák szöveg-előkészítésében megkövetelt hatalmas munka jelzi a betűről betűre vagy szóról szóra emberi döntéseket kívánó átírások reménytelenségét. Még a lehetséges gépi segédletekkel is akadozni fog a munka, mint ahogy a lexikográfia mindig is akadozik. Mindazonáltal ennek egy részét egyszer s mindenkorra el fogják végezni. A maradékot mint rabszolgamunkát söprik majd egyik helyről a másikra. Ez a tanulmány azt sugallja, hogy jobb a törmeléket egyetlen látható halomba söpörni, mint úgy tenni, mintha ott se lenne.

Információkeresési modellek[22]

Témám nem is annyira maga az információelmélet, mint inkább bizonyos fehér foltok, melyekre az információelmélet hatékonyon alkalmazható. Nem szándékozom azonban az alkalmazás hogyanját tárgyalni, részben, mert az ezzel kapcsolatos matematikai apparátus bonyolult, de még inkább azért, mert magam sem tudom a válaszokat. Olvasóim között vannak avatottabbak, akik gyors léptekkel haladnak ezen a területen, ahová én csak belopakodtam. Mégis veszem a bátorságot, hiszen örömteli küzdelemmel köteleztem el magam a „könyvtári osztályozás" ügye mellett. Bízom a jövőben, abban, hogy a könyvtári osztályozás nemcsak szavakban alakul át dokumentumkereséssé.

A könyvtári osztályozásnak és az információelméletnek első látásra kevés közös vonása van, némi gyanús metafizikai vonzódáson túl. A könyvtári osztályozás állítólag az ismeretek rendszerezésére irányul, amit én kétlek. Az információelmélet, durva meghatározás szerint bizonyos jelölést hordozó események – jelek – kezelésével és reprodukálásával, illetőleg az ezzel összefüggő fizikai problémákkal foglalkozik. A jelek továbbítását zavarhatják, vagy éppenséggel jelnek látszhatnak bizonyos ellenőrizhetetlen események, különböző hibák és zajok. Néhány tipikus probléma: milyen mennyiségű energia szükséges ismert, átlagos bonyolultságú jelek továbbításához, ismert, átlagos erősségű zajokon keresztül? A jelek milyen kódolása véd meg leginkább valamilyen adott típusú hibától?

A jeleket kizárólag fizikai események kombinációjaként tudjuk kezelni. Egyetlen „jelentésük" az információelmélet számára az, hogy milyen tevékenységek fűződnek hozzájuk a rendszeren belül: gombok és billentyűzetek

22 The patterns of retrieval. In: Toward information retrieval, p. 83–93. [Eredetileg megjelent 1956-ban.]

nyomogatása, villamos áram modulációja, fizikai tárgyak rakosgatása egyik helyről a másikra, vagy egyéb energiafogyasztó fáradozások. A jelekhez az adók vagy a vevők által hozzárendelt magyarázatok nemcsak irrelevánsak, hanem a kommunikációs rendszer szempontjából hozzáférhetetlenek is. Egy olyan számítógép, amely képes magukkal a számokkal foglalkozni, nem csupán fizikai jelekkel vagy eseményekkel, amelyekkel mi csak reprezentáljuk a számokat, valóban csodálatraméltó eszköze lehet nem is az automatikának, hanem a „platomatikának". De ha a jelek jelentése bizonyos fizikai tevékenységekre korlátozódik, akkor a kommunikációs rendszer eleve nem képes különbséget tenni az értelem és a badarság között. Előfordul az ilyesmi másutt is. A kommunikációs rendszerek esetében viszont az a jellemző, hogy tőlük nem várja el senki, hogy különbséget tegyenek. Akár értelmes az, hogy „Hamlet", akár értelmetlen, valamennyi létező fizikai jelentése közvetíthető bizonyos jelek meghatározott sorba rendezésével. A jelek és sorrendjük az, amivel a telefonnak, a nyomdásznak, a gépelőnek vagy a kamerának foglalkoznia kell, és ennél többre nincs is szükségük. A jelek értelmezése már nem az ő dolguk, ez a vevő magánügye.

Nos, helyben is vagyunk, hiszen ugyanez áll a könyvtári osztályozásra is. Vitatható, hogy a könyvtári osztályozás foglalkozik-e valóban az emberi tudással, igazsággal és hamissággal, a hegeli abszolútumokkal, és hasonlókkal; az viszont tagadhatatlan, hogy nyomtatott anyagokkal igenis foglalkozik. Kézzelfogható nagy és nehéz tárgyakkal, amelyek jeleket viselnek magukon.

A környvtári osztályozás „delegálható" területein ezek a jelek szabályozott fizikai tevékenységeket vonnak maguk után, amelyek másokkal is elvégeztethetők. Ha szerencsénk van, a szabályok teljesek, odaillőek és nincs köztük ellentmondás.

Például: „Az 'I–X' (9) jellel jelölt köteteket a 'V–I–I–I' (8) jelűek jobb oldalára kell tenni."

Vagy „Vizsgáljon meg minden kártyát: emelje ki a '62'-es számmal jelölteket, aztán emelje ki azokat, amelyek '8'-cal folytatódnak."

Vagy „Gyűjtse össze mindazokat a dokumentumokat, amelyek sorozati száma e listák mindegyikén előfordul."

Vagy „Hasonlítsa össze a címlapokat ezzel a mintakollekcióval. Ha valamelyik hasonlít, tegye a folyóiratot az elérhető legmagasabb polcra."

A dokumentumkeresésnek, és általában a könyvtári munka nagy részének vajmi kevés köze van a szemantikai tartalomhoz, csak megfigyelésből, azonosításból, megjelölt tárgyak meghatározott szabályok szerinti kezeléséből áll, vagy meghatározott szabályok szerint meghatározott szabályok készítéséből.

Magától a számítástechnikától nem idegen mindez, de a mindennapi számításoknál jóval bonyolultabb és rugalmasabb tevékenységet kíván. Itt a gyö-

kere a könyvtárgépesítés számos nehézségének és viszonylagos sikertelenségének.

A leírt tevékenységfajtákat én „ügyviteli" tevékenységeknek nevezem; nem kell hozzájuk különösebb ötletesség, sem felfedezés, sem invenció, csak gyakorlat és megbízhatóság. A könyvtárosoknak nem kell emiatt szégyenkezniük. Az egyedi intellektuális munka igényesebb tevékenységei nem ruházhatók át, de elvégzésükhöz, mások bármilyen intellektuális munkájához a könyvtárak teremtik meg a szükséges feltételeket. Arra valók. A gyakorlat és megbízhatóság egyáltalán nem lebecsülendő tulajdonságok.

Nézzük meg, melyek e dokumentumkeresés egyértelműen ügyviteli mozzanatai. Először is, minden információkereső rendszerben bizonyos fajta jelöléseket használnak: bejegyzéseket a hivatkozási kártyákon vagy jegyzéken, vagy magukon a dokumentumokon. E jelöléseket valakinek elő kell állítania; ez olyan kemény tény, amelyről hajlamosak vagyunk megfeledkezni. A jelöléshez jelhordozó tárgyak is kellenek: írógépek, gumibélyegzők, másolóberendezések, a másolatok eredetije és mindezekhez számos ügyviteli tevékenység is. A tárgyakat úgy lehet megjelölni, hogy valamilyen érzékelhető módon megváltoztatjuk őket, megfestjük, kilyukasztjuk, hozzádörzsöljük egy görényhez stb. Ezt nevezem „beírásnak" vagy „bejegyzésnek".

De megváltoztathatjuk a tárgynak a környezetéhez viszonyított helyzetét is: fejetetejére állíthatjuk, oldalt fordíthatjuk, beletehetjük egy kijelölt rekeszbe stb. Tágabb értelemben nevezhetjük ezt „rendezésnek", de kevésbé hivatalos célra kifejezőbb terminusok a „megjelölés" és a „helyretétel".

Akár megjelöljük, akár „helyére tesszük" tételeinket, az mindenképpen időt, energiát, munkát, felszerelést és raktári helyet igényel. Az információelmélet segítségével kiszámíthatjuk, hogy adott nagyságú gyűjtemény és adott megoldás esetén mennyi lesz a minimális költség, vagyis átlagosan mennyibe kerül adott nagyságú szövegblokkok megkülönböztető jelölése. Világos, hogy minél nagyobb a gyűjtemény, és minél kisebbek a megkülönböztetendő szövegegységek, annál gazdagabb választékot kell nyújtani a jelkészletnek, annál bonyolultabb kombinációk szükségesek ahhoz, hogy eltérő jelzeteket hozhassunk létre.

Ez olyan feladat, amivel az információelmélet viszonylag könnyen megbirkózik, ha – de vigyázzunk, ez elég nagy „ha" – rendelkezésre áll a műveletek és költségek teljes leírása. Ennek tartalmaznia kell a bizonylatok útját egyik helyről a másikra (fölösleges állomásokkal együtt), a jelzések időközi másolását részben, vagy egészben, helyesen vagy helytelenül, tehát olyan tevékenységeket, amelyeket sikerült mások számlájára elvégeztetnünk stb.

Röviden, a láncnak minden egyes szemét figyelembe kell vennünk, nemcsak azokat, amelyek érdekesek vagy presztízs okokból fontosak. E körben talán szükségtelen ezt a pontot tovább feszegetni, de engem rászoktatott a környezetem, ahol hidegfejű számítógépes mérnökök és hasonlók dolgoznak.

Az elmélet az ügyviteli tevékenység költségeinek csak a minimumáról ad számot. Tudjuk, hogy az információkereső rendszer jelzetrendszere, mint minden nyelv, önkényes. A legmagasabb költség, ami még elfogadható, attól függ, hogy a rendszer megalkotója mivel képes kivívni a használók elégedettségét. Gyakorlatilag ez a határ nem a csillagos ég. Ott már valami baj van, amikor az információkereséshez több információ kell, mint amit nyerünk belőle.

Nos, minden tűrhető rendszerben arra kell törekedni, hogy ha költségekként merülnek föl a jelölések, akkor több haszon származzon belőlük. Lehetnek bonyolultak a jelzetek, de akkor pontos leírásokat kapcsoljanak szűk keresési területhez. Élesebben fogalmazva: tegyenek hozzáférhetővé olyan koncentrált területeket, amelyekbe szinte biztosan bele kell, hogy tartozzon a kívánt dokumentum és kicsi legyen a kockázata annak, hogy kicsúszik belőle.

A keresési költségek és a vissza nem hívott releváns tételek miatt fellépő veszteség megengedhető aránya függ attól is, hogy mi a keresés tárgya. Tankönyvszerű információk esetében ez a veszteség nem olyan tragikus, hiszen egy többé-kevésbe azonos tartalmú másik szöveg ugyanolyan jól kielégítheti az igényeket. Jogszabályok esetében vagy a tudományos kutatásban azonban az információveszteséggel járó költségek nagyon magasak lehetnek – különösen, ha nagyszabású kutatásról van szó. Ebben az esetben az információkereső rendszernek olyan „letisztult" információval kell ellátnia a használót, amely majdnem biztosan tartalmazza a szükséges szöveget, ha az egyáltalán létezik, még olyan áron is, hogy át kell rágni meglehetősen sok dokumentum szemantikai tartalmát.

Az információelmélet mindezt számszerűen értékelhetővé teszi, így egyensúly alakítható ki a teljes, az információkereső rendszerben a gyűjtemény mérete, a leírás finomsága, a jelzetelés és a keresési műveletek ára és munkaigénye, valamint a talált, illetve elvesztett információk árának relatív mértéke között.

Térjünk vissza az információkeresési jelölésekre. Ezek kötik össze a leírásokat, egyrészt a nekik megfelelő dokumentumokkal, másrészt a dokumentumok kereséséhez szükséges fizikai műveletekkel. Legalábbis így kellene lennie. A valóságban örülhetünk, ha az osztályozási jelzetek jól-rosszul visszatükrözik a leírás belső szemantikai szerkezetét, de hol van még ez attól, hogy útbaigazítsanak a leírt anyag megszerzésében? Vannak persze sikeresebb rendszerek, amelyekben a keresési leírások arról is tudósítanak, hogyan keressünk vissza. Az ETO-ban például a „629.13" nem csupán „aeronautikai műszaki ismeretek„ leírását jelenti, hanem ez a jelzet áll a leírásnak megfelelő dokumentumokon is, és egyben meghatározza a könyvek megtalálásához szükséges műveleteket is. Azaz odamegyünk a „6"-os jellel jelzett nagyobb zónához, annak egy kisebb „62"-vel jelzett részéhez stb.

A gyűjteményt tehát mintegy a „helyrendjével" jelzeteljük, amely ugyanakkor a leírások rendjével is összhangban van. Ugyanilyen leírások vezérlik a

válogatási folyamatokat, s ezek éppen azt hívják elő, amire szükségünk van, hiszen a jelzetelés kapcsolatot teremtett a leírás és a dokumentum között. A dokumentumokon általában rajta van a bejegyzés is, de ezzel csak a helyretevés könnyen felbomló rendjét védjük.

A véletlenszerű kavargást a fizikusok „Brown-féle mozgásnak" hívják, a könyvtárosok viszont jóval földhözragadtabb nevekkel illetik.

A helyretétel az ilyen típusú rendszer lényege, e nélkül esetenként dokumentumról dokumentumra át kellene vizsgálnunk az egész gyűjteményt, hogy megtaláljuk valamilyen jelölés alapján az ezzel megjelölt valamennyi szöveget.

Hiába határozza meg a jelzetelés az információkeresés menetét, illetve a leírásokat és az ezekhez tartozó dokumentumokat, addig nem vehetjük ennek hasznát, míg a jelzeteket fel nem ismertük, „le nem olvastuk". Maga az olvasás fizikai művelet és nagyon megkönnyíti a helyzetet az, ha ez egyszerre zajlik az információkeresés műveletével, vagy legalábbis szorosan kapcsolódik hozzá.

A peremlyukkártyák esetében például az alapvető olvasási tevékenység megegyezik a kiválasztási tevékenységgel. Nem keressük a hornyolást először, hogy utána kiválasszuk a kártyát, hanem tudjuk, hogy ha kiesik, akkor hornyolva van. A felismerés és a kiválasztás megegyezik. Minden hatékony rendszer, amely peremlyukkártyát használ leírásként, ezt a tulajdonságot aknázza ki. Némi leleményességgel másféle leírások is kezelhetők így és lehetőség és szükség is van ilyen jellegű új rendszerek kidolgozására.

Ha a jelzetelést és az információkeresést ilyen módon összekötjük, kevesebb műveletre van szükség, rövidebbek lesznek az ezeknek megfelelő jelölések. Másrészt az is nyilvánvaló, hogy minél durvább a leírás, annál kevésbé szűkíthető egyáltalán a keresés, ezzel is csökkenthető az információkereső műveletek száma. Az sem közömbös, hogy a részletes leírások szövege hosszabb, mint az elnagyoltaké. A legkevésbé körülhatárolt dokumentumokról csak annyit mondhatunk, hogy „a gyűjteményhez tartoznak". Ilyenek a még nem osztályozott új szerzemények és az olyan egységek, amelyekről nem tudjuk eldönteni, hogyan kell osztályozni őket. Ilyenkor a leghelyesebb mindenféle jelölést elhagyni. Az ilyen dokumentumok értelemszerűen minden információkeresés eredményéhez hozzá kell, hogy tartozzanak, hiszen potenciálisan bármely kérdésre választ tartalmazhatnak. Ha ezt nem biztosítjuk, a könyvtár bizonyos része – általában a legfrissebb részek – az információkereső rendszer számára „láthatatlanok" maradnak.

Az ETO, a Dewey-féle tizedes osztályozás és egyéb a „fa-struktúrán" alapuló rendszerek meghatározott bizonytalansági szint felett mégsem így járnak el. Minél kevésbé tudunk választani a lehetséges alternatívák között, annál több jelölést kellene készítenünk, s a megfelelő információkeresési művelet is hiányzik hozzá. Ha szigorúan alkalmaznánk a szabályokat, beérkezéskor minden új művet plusz jellel összekapcsolva az összes főosztály jelzetével el kellene látnunk.

Nem gondoskodnak az ilyen dokumentációkról a fénylyukkártyás és az Uniterm rendszerek sem, ahol egy leírás jelzete a leírás által képviselt dokumentumok teljes listája. Ha egy dokumentumot nem sorolunk be valamilyen ismérv alá, a dokumentum „láthatatlan" marad, hiába keresünk rá arra az ismérvre.

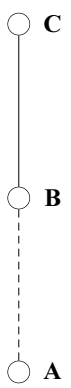
A megoldás kézenfekvő, bár ritkán alkalmazzák. Kezdetben minden ismérvkártyán minden dokumentumot fel kell sorolni. A fénylyukkártyán például kezdetben minden pozíciónak lyukasnak kellene lennie. Osztályozáskor e dokumentum azonosítóját kihúzzuk, leragasztjuk vagy átlátszatlanná tesszük az összes olyan ismérvkártyán, amely nem érvényes rá. Vagyis a dokumentumokat nem bevisszük e rendszerbe, hanem kitöröljük a felesleges helyekről. Az ügyviteli munka mennyisége a leírások kívánt mennyiségével egyenes arányban nő és a rendszer mindig felöleli az összes dokumentumot. Igaz, az osztályozó lustasága esetén a kereső zajosabb válogatást kap, viszont egyetlen dokumentum sem vész el számára.

Annak a régi jó elvnek az egyszerű alkalmazása ez, hogy egy konjunkció komplementuma megegyezik a komplementumok alternációjával.[23]

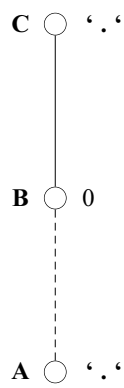
Másik végletként rendelkezünk olyan dokumentumokkal, amelyek elemzésekor kiderül, hogy nem illeszthető bele leírási (osztályozási) rendszerünkbe. Ezeknek potenciálisan bonyolultabb jelölést kell adnunk, mint azoknak az objektumoknak, amelyek a rendszernek megfelelően osztályozhatók. Ezzel biztosítjuk, hogy e rendszer későbbi kibővítésekor azokat szervesen beilleszthessük rendszerünkbe, anélkül, hogy esetleg egybeesnének már meglévő jelöléseinkkel.

A beírások módosításának egyetlen, gyakorlatilag járható útja a kiegészítés – a törlés és a helyettesítés túl költséges volna.

Vegyünk egy egyszerű, egyetlen leíró kategóriából álló könyvtári osztályozást. Az új szerzemények három eset valamelyikével írhatók le. Jelöljük ezeket (a leírás bizonytalanságának csökkenő sorrendjében) **C, B, A** eseteknek.



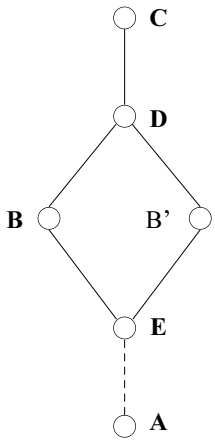
1. ábra



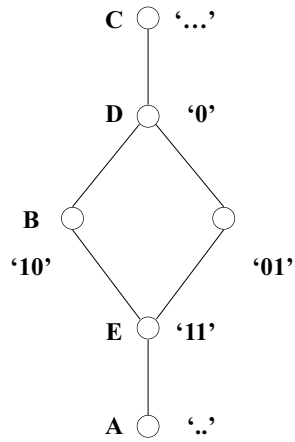
2. ábra

23 Az ún. De Morgan törvényről van szó: ab = ab (a szerk.).

C a következőt jelenti: ‘Lehet, hogy **B**, lehet, hogy nem az, még nem foglalkoztunk vele’. A **B** eset jelentése a következő: ‘Beleillik a **B** osztályba, mert ráillik az, hogy „bolondság”’. **A** pedig a következőt jelenti: ‘Megnéztük, de úgy találtuk, hogy nem **B**’.



3. ábra



4. ábra

Vizsgáljuk meg az ezeknek megfelelő jelöléseket. Most csak a formális struktúrával foglalkozom, ezért csak helytöltő karaktereket használok: 1 (egy) és 0 (zéró). Peremlyukkártyán ezek a hornyolás meglétét vagy hiányát fejezik ki. A zéró jel-hornyolással „eggyé” alakítható, igazából csak „üres hely”. A pontok **(.)** betöltendő pozíciókat jelölnek. Meggondolásaink érvényesek lesznek bonyolultabb jelzésekre is, hiszen bármilyen ábécé leegyszerűsíthető ilyen formára.

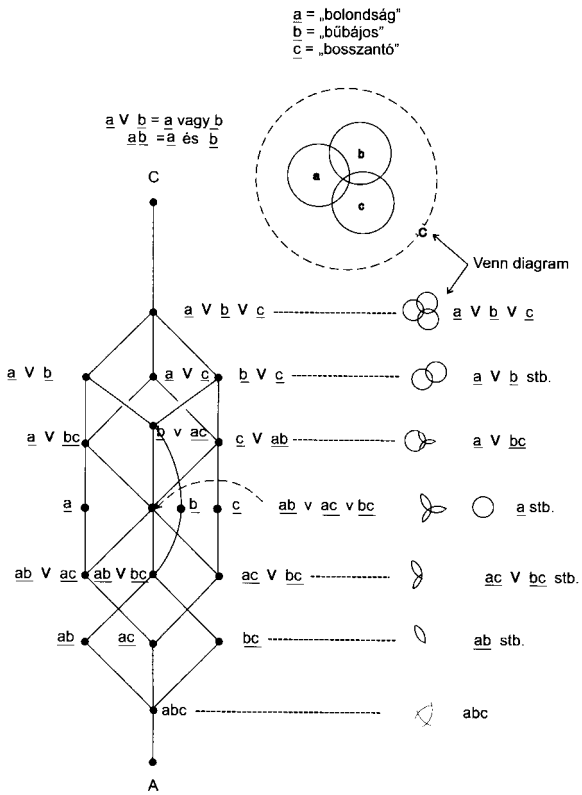
A leírásoknak megfelelő jelölések a 2. ábrán láthatók.

Minél specifikusabb a leírás, annál bonyolultabb a jelölése. A B-től A-ig vezető vonal mindkét diagramon szaggatott, ami azt mutatja, hogy az A-val jelzett dokumentumoknak mind a státusza, mind a jelölése ideiglenes.

Tegyük fel most, hogy megnövekszik az érdeklődés az eddig egyetlen kategóriával osztályozott dokumentumok iránt, és elhatározzuk, hogy bevezetünk egy új kategóriát, mondjuk **B’**-t („bűbájos”). Pusztán a kívánt új kategória felvételével módosítható a rendszer a 3. ábrán látható formára.

A **C**, a **B** és az **A** még mindig ugyanazt jelenti, mint eddig: **C** – ‘még nem tudjuk’, **B** – ‘bolondság – de esetleg más is’ és **A** – ‘az osztályozási rendszeren kívüli’. Az új kategória jele, a **B’** azt jelenti: ‘ráillik az, hogy „bűbájos” ‘. Még két összetett esetet kell figyelembe vennünk: ‘vagy „bolondság”, vagy „bűbájos”, vagy mindkettő’ (**D** eset); továbbá: ‘„bolondság” és „bűbájos” egyszerre’ (**E** eset).

Elemzésünk azt mutatja, hogy az egyszerűbb rendszer minden jelzete könnyen leírható az új kategóriákkal kibővített rendszer követelményei szerint, csupán ki kell egészíteni a régi jelöléseket néhány újjal. Csak a kategóriák megváltozásakor kellene a jelzeteket is átírni, a dokumentumok fokozatosan mélyülő osztályozása bármikor megoldható e nélkül. Itt is érvényes tehát az a tapasztalati igazság, hogy helyesebb „aláosztályozni", mint „túlosztályozni". Hasonlóképpen bővíthetjük a rendszert egy harmadik kategória bevezetésével, a megoldás kézenfekvő (5. ábra).



5. ábra

A teljes rendszer itt már húsz leírást tartalmaz. A húsz leírás csökkenő bizonytalanság szerint rendezhető, attól kezdve, hogy nincs osztályozva, azon át, hogy vagy „bolondság"', vagy „bűbájos", vagy „bosszantó", de lehet ez is, az is, egészen addig, hogy „bolondság" is, „bűbájos" is, „'bosszantó" is, illetve még ennél is tovább addig, hogy egyik tulajdonság sem jellemző a felsoroltak közül.

A középső szint nemcsak az eredeti generáló leírásokat tartalmazza („bolondság", „bűbájos", „bosszantó"), hanem az ezekből összetett, „kevert"' leírásokat is: „bolondság és bűbájos", „bűbájos és bosszantó", „'bosszantó és bolondság". Így a kiinduló kategóriáknak nem kell kölcsönösen kizáróaknak

lenniük, hétköznapi fogalmak esetén ez teljesíthetetlen követelmény is lenne. Persze, ha sikerül kölcsönösen kizáró fogalmakat választanunk, annál jobb. Sok fáradtságtól és kellemetlenségtől kímélhetjük meg magunkat.

Hat kiinduló kategóriából már körülbelül nyolc millió különböző leírás generálható. Ez a valóságban nem annyira ijesztő, mert bár külön-külön valamennyi kategóriára szükségünk lehet, logikai kapcsolataik között számtalan olyan van, amely felesleges. Ráadásul a leírások meghatározott bizonytalansági szinten túl gyakorlatilag megkülönböztethetetlenek az osztályozatlan tételektől, s a túlságosan finom leírások annyira specifikusak, hogy példát sem igen találunk rájuk.

Ahogy finomodnak az információkeresés módszerei, a túlrészletezés veszélye egyre nő. A könyvtárosok – ha az információkeresésről van szó – általában hajlanak arra, hogy túlzásba vigyék a finom megkülönböztetéseket.

Vizsgáljuk meg most magukat a dokumentumokat. Diagramunk szemléletesen mutatja a leírásoknak megfelelő dokumentumhalmazok között fennálló tartalmazási kapcsolatokat. A lefelé tartó vonal, amely két leírást összeköt, azt is kifejezi, hogy a feljebb álló leírásnak megfelelő dokumentumhalmaz tartalmazza az alacsonyabban álló leírásnak megfelelőt.

Átvihetjük ezt a szóhasználatot a leírásokra is. De tovább is mehetünk. Az információkeresés folyamatát ugyanolyan jól kifejezi a diagram, mint a leírások egymáshoz való viszonyát, így azt mondhatjuk, hogy az egyik információkeresési művelet akkor „tartalmazza" a másikat, ha megtalálható vele az összes dokumentum, amely a másikkal is megtalálható. Így minden információkeresési műveletnek afféle „desztillációs", „szűrő" műveletnek kell lennie, melynek segítségével – a jelölésükön alapuló szabályok szerint – a dokumentumok bizonyos része kiválasztható. Ez a rész azután tovább szelektálható a részszabályok alapján. Ahogy az ember a diagramon lefelé halad, szemantikailag is, fizikailag is finomodik a felbontás, élesebbek lesznek a megkülönböztetések. A finomítások lehetőségének az ára: a jelölések bonyolultsága egyre növekszik.

A tartalmazási diagram egyaránt megfelel a leírásoknak, a dokumentumhalmazoknak (valamint a dokumentumhalmazokkal végzett műveleteknek) hiszen ugyanazok a jelzetek kapcsolják össze őket. A megfelelés voltaképpen fordított, gyakorlatilag a jelölésekben az eredeti rendszer „duálisát" használjuk. Ezáltal teljesül az a követelmény, hogy a diagramon lefelé haladva növekedjék a jelölések bonyolultsága, holott a dokumentumhalmazok terjedelme lefelé haladva éppen ellenkezőleg, csökken. A jelölések bonyolultsága tehát ellentétes a tartalmazási viszonyokkal: egyetlen jelölés sem lehet kevésbé bonyolult azoknál a jelöléseknél, amelyek az átfogóbb, szélesebb dokumentumhalmazoknak felelnek meg. Az egyre specifikusabb leírásokhoz így egyre bonyolultabb jelölések kapcsolódnak.

A diagramunkból nem tűnik ki, hogy a jelölések a tartalmazási viszonyokkal ellentétesen alakulnak, mivel szimbolikus jeleket használtunk és va-

lamennyi lehetséges tartalmazási viszonyt feltüntettük. Élő rendszerek esetében a teljes diagram egyszerűsödik, mert elhagyhatók belőle mindazok az élek, amelyek a finomítás ki nem használt útjait jelölnék. (Világosan látható ez olyasféle rendszerek példáján, amilyen az ETO.)

A tartalmazási viszonyoknak megfelelően két leírás jelöléséből előállíthatjuk annak a leírásnak a jelölését, amely mindkettőnek a másik leírás szerinti finomítása. Egyszerűen a leírásoknak megfelelő jelölésekből ki kell választani a legnagyobb közös részt, vagyis a leírások átfedését vagy „szorzatát". Két fénylyukkártya kódjainak szorzata például azoknak a lyukaknak a kombinációja, amelyeken átlátunk, ha a kártyákat egymásra helyezzük és a fény felé tartjuk. Fordítva, két leírás összevonása, közös durvítása annak a leírásnak a legkisebb jelölésével képviselhető, amely mindkét leírást tartalmazza – ezt nevezhetjük a leírások „összegének".

Két fénylyukkártya kódjainak összegét úgy képezhetjük, ha mindazokon a pozíciókon lyukat veszünk fel, ahol valamelyik kártya át van lyukasztva.

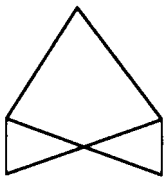
Két objektum szorzatát és összegét úgy jelölhetjük, hogy neveiket „és"-sel, illetve „vagy"-gyal kötjük össze. Az „és" meg a „vagy" itt puszta szimbólumok, nem okvetlenül a szokásos nyelvészeti jelentést hordozzák.

Minden kiterjesztésre és módosításra képes információkereső rendszernek jól meghatározott struktúrával kell rendelkeznie, s a struktúra meghatározásának kiinduló fogalma éppen a „tartalmazás", az „összeg" és a „szorzat". Csínjával egy további művelet is megengedhető, a „komplementumképzés", amelyet néha (eléggé sajnálatosan) a „nem"-mel szoktak jelölni. Ha a komplementumképzést az egész rendszerre kiterjesztjük, az befagyasztja, bővíthetetlenné teszi a rendszert. Gyakorlatilag egy bővíthető információkereső rendszer struktúrája nem lehet erősebb, mint egy szabad, disztributív háló. A komplementumos disztributív háló (Boole-háló vagy Boole-algebra) a mi szempontunkból merev.

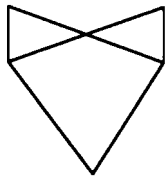
Sok rejtett feltételezés húzódik meg e fejezet állításai mögött. Feltételezem például, hogy ha a lovak állatok, akkor a lovak farka állatfarok. Ez meglehetősen egyszerű állítás, mégsem bizonyítható az arisztotelészi logika eszköztárával. Kiegészítő feltevésekkel kiküszöbölhetjük a problémákat és ezek eléggé elfogadható feltevéseknek tűnnek a dokumentalisztikában.

A miénkhez hasonló rendszerekben, ahol a leírásokat tulajdonképpen az extenziókra, nem az intenziókra vonatkoztattuk, ez a következtetési forma különösen fontos, minden formális pusztulátummal egyetemben, amely a tartalmazással kapcsolatos. Részletes vizsgálatuk megtalálható az algebrakönyvekben. Itt elég, ha arra hívjuk fel a figyelmet, hogy diagramunkban nem fordulhatnak elő vízszintes összekötő vonalak, más szóval a rendszer nem tartalmazhat szinonimákat. Ezeket előre ki kell szűrni, hogy a rendszerbe került leírások már szinonima nélküliek legyenek.

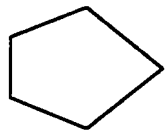
A többi rejtett feltételezés arra vonatkozik, hogy a diagramok pontjai között bizonyos együttes kapcsolatok nem engedhetők meg. Nevezetesen nem lehet a diagramban öt olyan pont, amelyre fennállnak a következő kapcsolatok:



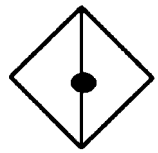
6. ábra



7. ábra



8. ábra



9. ábra

Ezzel kizárjuk az önelnyelő kategóriákat – amelyeket az amatőr osztályozók annyira szeretnek – és biztosítjuk az osztályozás formális konzisztenciáját. Az utolsó diagram (9. ábra) megtiltása talán nem is olyan lényeges. Szemléletes példával megvilágítva: megkövetelhetjük a rendszertől, hogy ha a férfiakat, a nőket és a lovakat egyaránt állatoknak tekintik, akkor különbséget tudjunk tenni a különböző nemű kentaurok között. Ez a kikötés inkább csak kényelmi szempontokat szolgál.

Hol kapcsolódik mindehhez az információelmélet? Először is: a diagram különböző pontjaihoz számértékeket, súlyozást rendelhetünk, s ennek segítségével nagyobb valószínűséggel eljutunk a hálóban a kívánt szöveghez. Másodszor: szabályokat ad annak meghatározásához, hogy adott típusú keresésben egy leírás mikor tartalmazza a másikat. Minden dokumentum egyértelmű leírás kell, hogy legyen, de hogy ez a leírás átfogóbb-e, mint egy másik, tartalmazza-e azt, az a kereső szempontjaitól is függ, például attól, hogy az „elektrotechnika" vagy mondjuk az „egyháztörténet" felől közelít-e a témához. Meg vagyok róla győződve, hogy az információelmélet segítségével súlyozni lehet majd a leírások szemantikai összetevőit és kapcsolatait; a súlyozás függővé tehető a megközelítéstől és a teljes súly alapján számíthatók ki a tartalmazási viszonyok.

Végezetül még egy megjegyzés. A könyvtári munkát úgy tekintettem, mint megjelölt dolgok szabályokon alapuló – vagyis automatikus – kezelését. Ez az induláshoz elég. Nem szabad beleragadni még akkor sem, ha a tudományosságot időnként a publikációk fizikai, és nem intellektuális súlyával mérik. Azok, akik a kreatív munkát féltik az automatizálástól, megnyugodhatnak. A térbeli távirányítás – akár drótos, akár drót nélküli, az összes gombjaival és billentyűivel együtt – nem érinti az alkotó munka lényegét. Az automatizáció pedig nem egyéb, mint időbeli távirányítás.

HANS PETER LUHN (1896–1964)

A Németországból Amerikába kivándorolt Hans Peter Luhn 1941-ig textilgépészmérnökként tevékenykedett, ettől kezdve az IBM mérnöke és feltalálója, több mint 20 szabadalom fűződik a nevéhez. 1946-ban az American Airlines automatizált helyfoglaló rendszerét készítette el, 1948-ban a vegyész *William Perryvel* a vegyületek adatainak gépi feldolgozásával fog-

lalkozott. 1956-tól az IBM yorktown-i kutatóközpontjában az információkeresés menedzsere. Érdeklődése ebben az időben fordult a teljes szövegek gépi feldolgozása felé. A nevéhez fűzött KWIC indexet 1958-ban mutatta be és őt tekintik a Szelektív Információszolgáltatás (Selective Dissemination of Information; SDI) kezdeményezőjének is.

Először 1953-ban a CIA (az amerikai felderítő ügynökség) dokumentátorai kezdték gépi segítséggel permutálni a címszavakat. Ezzel szinte egy időben a Milánói Egyetemen Aquinói Szent Tamás Teológiáját preparálták ki analitikus mutató formájába egy IBM számítógép segítségével. 1958-ra már több szervezet is előállított gépi indexeket, Luhn volt az, aki megvizsgálta matematikai hátterüket és nevet (KWIC, KWOC index) adott az eljárásnak.[24]

Luhn úttörő szerepet játszott az automatikus információkeresés kialakulásában. Munkáiban a dokumentum szövegében (elsősorban a címben és a referátumban) szereplő szavak gyakoriságát vizsgálta, hogy meghatározza, mely szavak eléggé szignifikánsak ahhoz, hogy a dokumentumot a számítógépben reprezentálják. Az ilyen kulcsszavakból álló lista minden dokumentumhoz elkészíthető. A szavak előfordulási gyakorisága szövegben felhasználható a szignifikancia fokának jelzésére is, ez pedig igen egyszerű eszköz a jegyzéken belül a kulcsszavak hangsúlyozására, és lehetővé teszi, hogy a dokumentumot „súlyozott kulcsszavas leírással" reprezentálják. Azaz a gyakorisági adatok fölhasználhatók a dokumentumot reprezentáló szavak és mondatok kiválasztásához. Egy korai írásában erről így fogalmaz:

„Az a véleményünk, hogy valamely cikkben szereplő szavak előfordulási gyakorisága a szavak fontosságának jól használható mérőeszközét adja. Úgy véljük továbbá, hogy a mondaton belül a fontossági értékkel ellátott szavak relatív helyzete a mondatok fontosságának mérésére ad lehetőséget. A mondatok fontossági tényezője ily módon e két mérték kombinációján alapul."[25]

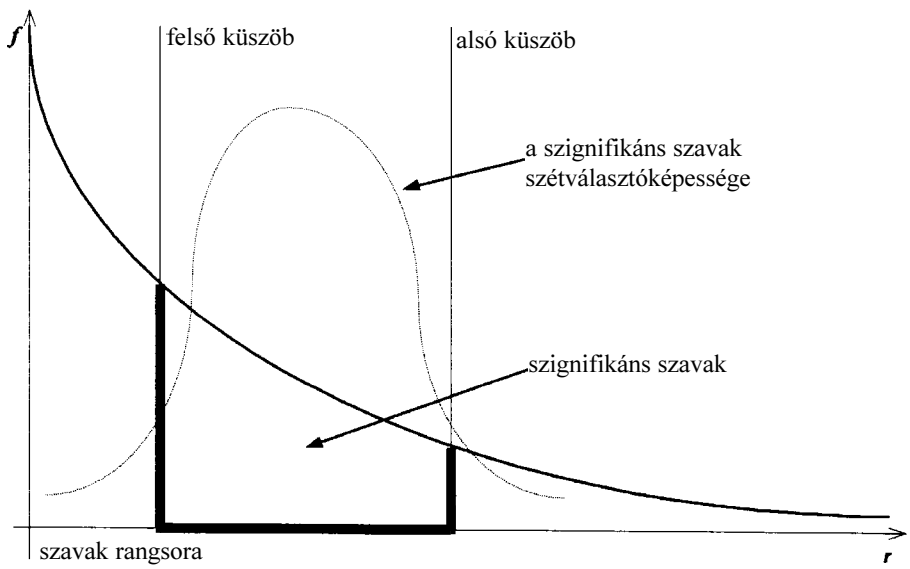
Luhn alapeszméjét *Cornelis van Rijsbergen* a következőképpen foglalja össze:

„Jelölje ***f*** a szöveg adott pozíciójában különféle szótípusok előfordulási gyakoriságát és ***r*** ezek rangsorban elfoglalt helyét, azaz előfordulási gyakoriságuk rangját; ekkor az ***f*** és az ***r*** kapcsolatát grafikusan ábrázolva olyan görbét kapunk,

24 Rajan, T. N.: Indexing Systems. – Calcutta: Indian Association of special Libraries and Information Centres (IASLIC), 1979. p. 25.

25 Luhn, H. P.: The automatic creation of literature abstracts. In: IBM Journal of Research and Development, 1958, No. 2, p. 159–165.

amely hasonlít a következő ábrán látható hiperbolikus görbéhez. Ez a görbe voltaképp *Zipf* törvényét demonstrálja, amely szerint a szó használatának gyakoriságát és rangsorban elfoglalt helyét összeszorozva közelítőleg állandó értéket kapunk. *Zipf* az amerikai–angol »újságnyelv« alapján igazolta törvényét. Luhn ugyanezt nullhipotézisként használta, hogy meghatározhasson két küszöbértéket, egy alsót és egy felsőt, kizárva így az inszignifikáns szavakat. A felső küszöböt meghaladó gyakoriságú szavakat az általánosaknak, az alsó küszöböt el nem érő szavakat ritkáknak tekintjük, s így egyike sem járul hozzá szignifikáns módon a cikk tartalmához. A számlálásos eljárás így vezetett a szignifikáns szavak megtalálásához. Ezzel teljes összhangban feltételezte, hogy a szignifikáns szavak »szétválasztó képessége«, amelyen azt értette, hogy a szavak mennyire képesek meghatározni a tartalmat, a két küszöbérték közti félúton éri el a csúcsát, és ettől mindkét irányban csökken, a küszöbértéknél csaknem elérve a nullát. A küszöbértékek meghatározásánál bizonyos önkényesség is fellép. Nincs az a bölcs, aki megadja nekünk ezeket az értékeket, próba–szerencse alapon kell meghatároznunk őket.



Érdekes megemlíteni, hogy ezek az elvek az információkeresés későbbi munkáiban jórészében is alapvetőek maradtak. Luhn maga automatizált referálási módszer kialakítását alkalmazta, majd a mondatok szignifikanciájának numerikus mérési eljárását fejlesztette ki a mondat egyes részeiben lévő szignifikáns és nem szignifikáns szavak számát véve alapul. A mondatokat numerikus értékük szerint rangsorolta, s a legmagasabb rangúak kerültek be a referátumba."[26]

26 Van Rijsbergen, C.: Információ visszakeresés. – Budapest: Múzsák Közművelődési Kiadó, 1987. p. 20–21.

Az alapprobléma az indexekben már kezdetben a kifejezések sorrendje volt, áttételesen tehát a szintaxis. Luhn után a kutatások szétváltak. A kutatók egy része a sokváltozós matematika – főleg a klaszteranalízis – segítségével tökéletesítette az automatikus indexelést. (Ezekkel a fejleményekkel a kötet későbbi részében, Gerard Salton kapcsán foglalkozunk.) Másik részüket a felvetődő összekapcsolási és relációs problémák nem hagyták nyugodni, és nyelvészeti eszközöket alkalmazva keresték a megoldást a megbízhatóbb kulcsszóláncok kialakítására. Ezekből – és a gépi fordítási – kísérletekből született meg idővel a számítógépes nyelvészet. (Lásd kötetünkben a gépesítés szemantikai problémáival foglalkozó részt). Az erőfeszítések egyik legismertebb eredménye dokumentációban az 1971-ben üzembe állított PRECIS (PREserved Context Index System) szövegkörnyezetet megőrző félautomatikus indexelési rendszere volt (a PRECIS-zel az első kötetben foglalkozunk).

Automatizált tájékoztató rendszerek

In: Varga Dénes: A dokumentáció nyelvészeti kérdései I. Szemelvénygyűjtemény. – Budapest : Országos Műszaki Könyvtár és Dokumentációs Központ, 1966. p. 5–17.
Eredeti: Automated intelligence systems. In: Information retrieval management. Ed. by Lowell H. Hattery and Edward M. McCormick, 1962. p. 92–100.

Ebben a korai cikkben az intézményen belüli automatizált rendszer struktúrájáról van szó. A rendszer feladata a szelektív információterjesztés, az információkeresés, a kérdések (profilok) egyeztetése, találatkiadás és a kettősségek kiszűrése. Ismerteti az egy mondaton belül egymástól két-három szónál nem távolibb szavak gyakoriságán alapuló automatikus indexelést. Leszögezi, hogy az intellektuális indexeléssel szemben, amelyben a feltárás igényel intellektuális erőfeszítést, az automatikus eljárás során a kereséskor van szükség intellektuális közreműködésre. Megfogalmazza, hogy szükség van a felhasználó és a gép között álló, megfelelően képzett „közvetítő” szakemberre. Az indexeléshez hasonlóan automatizálható referátumkészítés is.

AZ INTELLEKTUÁLIS ÉS GÉPI INDEXELÉS KÖZÖTT, AVAGY AZ ÖSSZEGEZŐK

Miközben 1960 után az automatikus indexelés és osztályozás területén is elkezdődtek a kutatások, és egyre több gépi információkereső rendszer jelent meg, az információkeresés kérdéseinek összegezésére, a kereső rendszerek átfogó, általános tárgyalására is egyre nagyobb szükség lett. Az 1960–85 közötti időszakban megjelent legismertebb kézikönyveket és monográfiákat az alábbi táblázatban mutatjuk be.

Az információkereső rendszerekkel (velük összefüggésben az indexeléssel) foglalkozó művek

1961 Vickery, B. C.: On retrieval system theory (2. kiad.: 1965; német kiad.: 1970; francia kiad. 1971)
1963 Bourne, M. F.: Methods of information handling
1965 Bernštejn, E. S.–Lahuti, D. G.–Černăvskij, V. S.: Informacionno-poiskovye sistemy
1965 Sharp, J. R.: Some fundamentals of information retrieval (2. kiad.: 1968)
1965 Taylor, R. S.–Hieber, C.: Manual for the analysis of library systems
1966 Hieber, C. E.: An analysis of questions and answers in libraries
1966 Jesse H. Shera: Documantation and the organization of knowledge144.
1967 Cigánik, M.: Informačné fondy vo vede, technika a ekonimike (2. kiad. 1969; 2. német kiad. 1969)
1967 Kochen, M.: The growth of knowledge; readings on organization and retrieval of information
1968 Goldhor, H.: Research methods in librarianship
1969 Ungurian, O.: Information retrieval systems. Methodological guidelines, exercises and tests (eredeti lengyelül)
1970 Bundy, M. L.–Wasermann, P.: Reader in research methods for librarianship
1970 Meetham, R.: Information retrieval
1970 O'Brien, J. J.: Management information systems

A táblázat folytatódik!

A táblázat folytatása:

1970 Sharp, J. R.: Information retrieval
1970 Vickery, B. C.: Technique of information retrieval
1971 Thomas, P. A.: Task analysis of library operations
1973 Herrmann, P.: Informationsrecherchesysteme
1973 Vickery, B. C.: Information systems
1975 UNISIST indexing principles
1978 Dabrowsky, M.–Laus-Maczynska, K.: Information retrieval and classification: a survey of methods (eredeti lengyel)
1978 Borko, H., Bernier, C. L.: Indexing concepts and methods
1978 Harrod, L. M.: Indexers on indexing: a selection of articles published in the indexer
1978 Engelbert, H.: Informationsrecherchesysteme in der Wissenschaft
1978 Supper, R.: Neuere Methoden der intellektuellen indexierung. Britische Systeme unter bersonderer Berücksichtigung von PRECIS
1978 Rowley, J. E.: Abstracting and indexing
1978 Heaps, H. S.: Information retrieval
1979 Kuhlen, R.: Datenbasen, Datenbanken, Netzwerke – Praxis des Information Retrieval. Band 1–3.
1980 Reusch, P. J. A. Informationssysteme, Dokumentationssprachen, Data Dictionaries
1981 Gebhardt, F:: Dokumentationssysteme

Olyan művek szerepelnek az összeállításban, melyek elsősorban nem a gépesítés, hanem az információkeresés elmélete és gyakorlata szempontjából tárgyalják a szakterülete kérdéseit, ugyanakkor nem hagyják figyelmen kívül a gépesítés következményeit sem. A felsorolásban nem szerepelnek a kizárólag a gépi információkereséssel és automatikus indexeléssel/osztályozással foglalkozó szerzők (például *Joseph Becker, Robert M. Hayes, Frederick W. Lancaster, Gerard Salton, Karen Spark Jones, van Rijsbergen*) fontosabb műveiket a kötet későbbi, a teljes automatizálással foglalkozó részében soroljuk föl.

A hatvanas és hetvenes évek egyik legtermékenyebb szerzője és az elért eredmények összefoglalója – legalábbis ami az információkeresés és az információkereső rendszerek kézi- és tankönyveit illeti – az angol *Brian C. Vickery*, akinek két könyvéből mutatunk be szemelvényeket kötetünkben (*Gernot Wersig* mellett *Vickery* az egyetlen nyugati szerző, akinek műveiből a korábbi években magyar fordítások megjelentek). Az amerikai *Jesse H. Shera* az információtudomány egyik úttörője, az intellektuális szakkönyvírás mestere, aki a könyvtártudomány szinte minden kérdéséhez hozzászólt.

A hatvanas években indul útjára *Allen Kent* szerkesztésében a máig legnagyobb méretű könyvtári kézikönyv.

JESSE HAUK SHERA (1903–1982)

Többek között a kötetünkben is szereplő *Harold Borkoval, Douglas John Foskettel, Brian C. Vickeryvel* együtt a „második tanárnemzedék" egyik legismertebb tagja. *Vickeryhez* hasonlóan könyvtáros végzettséggel a háta mögött a gépesítés eredményeinek egyik legjelentősebb szaktudományos feldolgozójává vált. 1947–51 között a chicagói egyetem, majd a Western Reserve University könyvtáros iskoláján tanított. Itteni dékáni időszakának elejére esnek *William J. Perry* és *Allen Kent* kísérletei a szemantikai kódokkal; ezek a kísérletek többek között az ő ösztönzésére kezdődtek el. *Samuel C. Bradford* és *Mortimer Taube* mellett a dokumentáció egyik úttörőjeként tartják számon. Hosszú ideig a Library Quaterly és a Journal of Cataloguing and Classification szerkesztője volt, 1972 után az amerikai könyvtárosok társaságának információtudományi és automatizálási osztályát vezette.

Műveire az intellektualitás és a szellemi felkészültség jellemző; a könyvtártudomány legkülönbözőbb elméleti és filozófiai kérdéseivel és ezek szociológiai szerepével foglalkozott. Témái a közkönyvtáraktól az egyetemi könyvtárakig, a hagyományos könyvtári technikáktól az információs rendszerekig, a bibliográfiai leírástól az osztályozásig szinte minden kérdést átfogtak. Angol nyelvterületen a könyvtári publicisztika egyik legismertebb művelője; tanulmányai stiláris szempontból szépirodalmi igénnyel íródtak és nem nélkülözik a szellemességet. Sherának mind az első, mind a második kötetünkben helye volna; azért itt szerepeltetjük, mert gazdag munkásságából az információkereső nyelvek és a könyvtári munka viszonyáról írt egyik részletet választottuk ki.

Felfogása szerint a modern osztályozásnak specializáltnak és nem egyetemesnek kell lennie, mivel nincs univerzális felhasználó. Itt bemutatott könyvében a könyvtári osztályozás szempontjából, a ranganathani eredményekre (is) támaszkodva újraértelmezi a nyelvet, és vele összefüggésben a rendszerező, mesterséges nyelven alapuló osztályozási rendszert. Arra a megállapításra jut, hogy az osztályozási rendszerek/információkereső nyelvek közvetítő, átkapcsoló, a katalizátor szerepét játszó rendszerek a különböző kultúrájú, nyelvű emberek között.

Az alábbiakban bemutatott szemelvényen kívül 1970-ben magyarul is megjelent a tájékoztatástudomány történeti fejlődéséről és hatvanas évekbeli helyzetéről írott tanulmánya, melynek adatait a szemelvény után adjuk meg.

A dokumentáció és az ismeretek szervezése[1]

[...]

II. Nyelv és könyvtártudomány[2]

Általában az vélemény, hogy a nyelv a humán tudományok körébe tartozik, amely a filológusok és a nyelvészek vadászterülete, akik magával a nyelvvel csupán mint a szavak rendszerével és a szavak jelentésével foglalkoznak. A XIX. század végének német filológusai úgy vélték, hogy a nyelvészet feladata a szintaktikai szerkezetek, a szókapcsolatok, a szóeredetek és a mássalhangzó váltakozásokban megfigyelhető minták törvényeinek, elveinek és szabályainak a megfogalmazása. Mindezek segítségével feltérképezték a nyelvcsaládokat és igyekeztek megállapítani a nyelvek eredetét. Csak az olyan területeken, mint az irodalomkritika és a szövegmagyarázat vizsgálták a nyelvnek kommunikáció szerepét, és jóformán teljesen elhanyagolták a nyelv szemiotikai szerepének és a kultúra gyarapodására kifejtett hatásának a vizsgálatát.

Idővel a nyelvészet felkeltette a logikával foglalkozó kutatók figyelmét is, akik – különösen a hagyományos logika területén – górcső alá vették a szintaktikai szerkezetek és a kijelentő mondatok különféle formáit és összefüggéseit. A logikával foglalkozó szakemberek azonban elfelejtik, hogy a nyelv megállapodás, amely a használat évszázadai közben fejlődött és használói nem sokat törődtek a logika meglehetősen stilizált szerkezeteire, a feltevések és a következtetés között világos kapcsolat követelményére és azokra a leegyszerűsített gondolati sémákra, amelyek a logikai szakembert foglalkoztatják. Az ember a nyelvet úgy használja, ahogy az közvetlen szükségleteinek megfelel, ez a használat pedig mind az igazság eltitkolásához, mind annak felfedezéséhez elvezethet. A nyelvészet problémáinak logikai kezelésében a szakemberek legnagyobb hibája az agyonegyszerűsítés. Így aztán arra kényszerülnek, hogy értekezéseik szűk világából kiirtsák azt a nyilvánvaló tényt, hogy az emberi elmében csaknem minden szóra a legkülönbözőbb válaszok lehetségesek. Nem veszik figyelembe, hogy a nyelv a valóság bizonyos vonatkozásait sokkal szabatosabban fejezi ki, mint ahogyan az a formalizált logikai minták alapján lehetséges.

A közelmúltban a mérnökök is kezdték felfedezni a nyelvet, bár ők a nyelv kommunikációs vonatkozásai iránt érdeklődnek elsősorban, a jelelmé-

---

1 Documentation and the organization of knowledge / Jesse H. Shera. – London : Crosby Lockwood, 1966. 185 p.

2 Language and librarianship In: Documantation and the organization of knowledge, . p. 139–144.

letekkel és azokkal a hatásokkal, amelyek adott jel vagy jelhalmaz információközvetítő képességét növelik vagy csökkentik. Kétségtelen, hogy ezek a szakemberek a nyelv tanulmányozása előtt új távlatokat nyitottak, maguk is jelentős szerepet vállaltak a vizsgálódásban azzal, hogy matematikai apparátust alkalmaztak a nyelvnek, mint a mondanivaló közvetítőjének a jobb megértéséhez. Nem tették meg azonban a jeltől a szimbólumig vezető utat és jobban érdekelte őket a mondanivaló közvetítésének a módja, mint az, hogy mi a közlés célja, vagy melyek a közlés szociológiai, pszichológiai hatásai. A nyelv végül is sokkal több mint csupán a kommunikáció eszköze, bármennyire alaposan elemezzük is a mondanivaló közvetítésének technikáját.

Ilyen módon elérkezünk a pszichológusokhoz és a kultúrszociológusokhoz, akik – a szerző véleménye szerint – a nyelv teljesebb megértéséhez jobban hozzá tudnak járulni, fontosabb dolgokat tudnak róla mondani, mint bármely más csoport, minthogy a nyelv – végső fokon – az emberi elme társadalmi közegben létrejött terméke. Bizonyos, hogy a pszichológusok sok segítséget kapnak a mérnököktől és a matematikusoktól, és ez a segítség a jövőben valószínűleg csak növekedni fog. A kultúrszociológusoknak is vannak további feladataik a nyelv tanulmányozásában. Alighanem *Melvil Dewey-nak* volt igaza, aki a nyelvet osztályozási rendszerében a természettudományok és a társadalomtudományok közé helyezte. Mindamellett csak a pszichológiának és a kultúrszociológiának az együttes alkalmazásával juthatunk a nyelv természetének és annak a kapcsolatnak a megértéséhez, amely a nyelv és az emberi tudás gyarapodása között fennáll.

A nyelv problémáinak interdiszciplináris megközelítéséhez és ahhoz, hogy a kommunikáció átléphesse a földrajzi és politikai határokat, a könyvtárnak életbevágó érdeke fűződik. A könyvtáros a grafikus leírások világában élve hagyományosan intellektuális és esztétikai értékeket közvetít és céljai eléréséhez nemcsak a kommunikáció folyamatait kell ismernie, hanem magának a tudásnak a jellegét, a tudás keletkezésének a módját és azokat az eszközöket is, amelyek révén az ismeretek a társadalomban elterjednek. Feladata a szintézis megteremtése, és ebben a szintézisben minden diszciplína szerepet kap. Ez a szintézis maga is alkotó tevékenység, olyan szimbólumrendszer megalkotása, amely a tudomány eredményeit azokon az akadályokon keresztül is közvetíteni tudja, amelyeket az ember a maga ügyefogyott módján saját előrehaladásának az ösvényén emelt.

Öntudattalanul talán, de a könyvtárosok fektetik le ennek az újabb szimbolizmusnak az alapjait. Azáltal, hogy a jelentősebb – főképp egyetemes – osztályozási rendszereket afféle szabályként fogják föl, néhány nagyon speciális, de azért eredményes lépést tettek afelé, hogy egységes szimbólumrendszer szülessék az ismeretek logikai elrendezéséhez. Az ilyen egyetemes szimbólumrendszer nagymértékben előmozdítaná a kommunikációt és a megértést a *tudomány egész világában:* talán elkezdődhetne általa a gondolkodás különböző területei-

hez kapcsolódó tények, törvények és elméletek összehangolására alkalmas fogalmak és módszerek cseréje. A hagyományosan elfogadott bibliográfiai osztályozásból kialakuló rendszerezés lehetővé teszi, hogy a tudomány különböző területei között kibontakozzék a nyelvi szinten megfogalmazott analógiák, illetve az azonos tartalmú képzetek és fogalmak cseréje. A különböző nyelven beszélő emberek közötti kommunikációhoz arra van szükség hogy olyan átkapcsoló nyelvi formákat találjanak, amelyek révén a különböző nemzeti eredetű és kultúrájú tudósok elméjében az emberiséget egységesen jellemző gondolkodási minták felismerhetővé válhatnak. Az ilyen szabályozott minták kialakításában – legalábbis a biológiai szakterületeken – a rendszertanban érték el eddig a legnagyobb sikereket. A könyvtárosok viszont ezeknek a mintáknak az elfogadását jóval szélesebb körben mozdítják elő.

A könyvtárosok az osztályozás területén érték el a legnagyobb sikereket az említett nemzetközi szimbólumrendszer kialakításában, különösen az olyan jelzetrendszerek, mint amilyenek például az Egyetemes Tizedes Osztályozás vagy a Ranganathan-féle kettős pontos osztályozás. Ez nem jelenti azt, hogy ezek a rendszerek minden szakaterületen tökéletesek, vagy hogy optimális a hatékonyságuk, bár a kettőspontos osztályozás már olyan átgondolt, elméletileg következetes rendszer, amely figyelmet érdemel, szemben például az ETO-val. De mindkettőt nemzetközileg elfogadták, és mindkettő a maga egyszerű formájában olyan jelzetrendszer-elemeket tartalmaz, amelyeket az egyetemes kommunikáció, illetve párbeszéd minden eszközének tartalmaznia kell.

III. Szempontok a megfelelő szimbólumrendszer kialakításához

Ha a nyelvet nem egyezményes jelek állományának, hanem egymással szorosan összefüggő szimbólumok szövedékének tekintjük, akkor ebből az következik, hogy minden olyan mechanizmusnak, amely a szabályozott kommunikációt szolgálja, tekintettel kell lennie az ismeretekkel összefüggő három egyetemes szempontra, nevezetesen (a) a dinamikus, evolúciós és funkcionális folyamatokra, (b) a jelenségek statikus és strukturális vonásaira és (c) a teleologikus vagy célirányos magatartásmódokra. Figyelembe kell venni még (d) az időbeliséget, amely a változáshoz, a növekedéshez, és a funkcióhoz kapcsolódik, (e) a térbeliséget, amely az alakzatokban, a szerkezetekben, az izomorfiában és a taxonómiában megnyilvánuló jellemző, és (f) a szintetikus vagy szimbolikus (a jelzetekkel összefüggő) jelleget. Ugyanakkor a térbeliség és az időbeliség bizonyos értelemben a statikus és a dinamikus közötti különbséget is kifejezi, a szintetikus pedig a teleologikusnak a jellemzője.

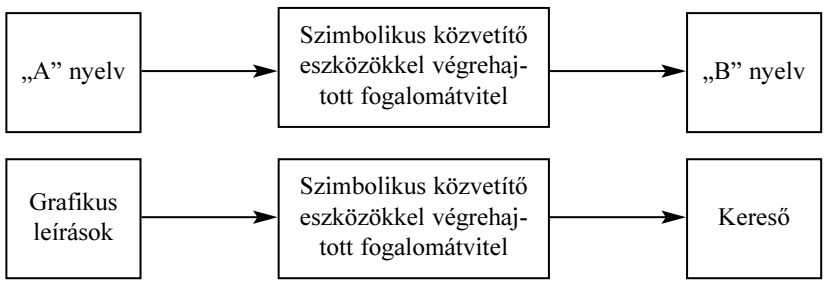
Az ismeretek eme szempontjai felismerhetők az Egyetemes Tizedes Osztályozás jelzeteinek és szimbólumainak a rendszerében is, bár sokkal világosabban és egyértelműbben öltenek testet a ranganathani idő, tér, anyag, ener-

gia és személyiség fazettákban. Tetten érhetők például *Perry* és *Kent* szemantikai kódjaiban is, különös tekintettel a rendszerük szerepjelölőire. Hasonló célokat tűzött ki *Henry E. Bliss* is, amikor rendszerében a természet rendjét kísérelte meg megszerkeszteni, amely számos alrendszerből – például a történelmi, a fejlődési és a pedagógiai alrendszerekből – tevődik össze. Ezek után érthető, ha azt mondjuk, hogy a könyvtárosok már hosszú ideje úgy dolgoznak a jelzetrendszerekkel, ahogy Moliere Jourdain ura beszélt prózában – anélkül, hogy tudott volna róla.

A kívülállók számára a jelzetek és általában a kódolás gyakran olyan szimbólumalkotásnak tűnik, amelynek csupán két célja lehet: vagy az, hogy az üzenet tartalmát elrejtse mindenki elől, kivéve azokat a kiváltságosakat, akiknek kulcsuk van a megértéséhez; vagy az, hogy a kommunikációs folyamatot hatásosabbá tegye, egyfajta gyorsírássá redukálva a jeleket. Most azonban a kódokkal, jelzetekkel nem mint a természetes nyelv rövid helyettesítőivel foglalkozunk, hanem olyan – átkapcsoló – szimbólumrendszerként vizsgáljuk őket, amely összehangolja a dokumentum vagy leírásának fogalmi rendszerét annak a személynek a fogalmi rendszerével, aki a dokumentumot előveszi. Következésképpen a kódolás, a jelzetalkotás célja nem az, hogy gazdaságosabbá tegye a kommunikációt, vagy hogy az üzenet tartalmát elrejtse, hanem éppen az, hogy biztosítsa a szélesebb körű hozzáférést. Figyelmünk középpontjában ezért a szimbólumrendszer (a jelzetrendszer) katalizátor-szerepe kerül. Az információkereső nyelvre, a jelzetekre való fordítás (az indexelés, osztályozás) és az információkeresés szimbólumrendszerének jóval többet kell jelentenie, mint a nyelvi egységek kodifikálását. A szimbólumrendszer alapját képező kódnak lehetővé kell tennie az üzenet gondolati tartalmának olyan összekapcsolását és összehasonlítását, amely megfelel az emberi agy működésének. Mind a generikus, mind a specifikus fogalmakat kezelnie kell és képesnek kell lennie ezek összekapcsolására. Több szemantikai szinten kell működnie, hiszen az emberi közlésben – akár közvetlen, akár közvetett – a relációk éppen olyan fontosak a jelentés szempontjából, mint az összetevők tartalma. Ezért a szerkezetet, a rendszert és a szavak elhelyezkedésének egyéb megnyilvánulásait ugyancsak ki kell fejezni a kódokban.

Így valójában a kód, a jelzetrendszer maga is nyelvnek, pontosabban metanyelvnek tekinthető, *Morris* szerint olyan halmaznak, amely szóban vagy grafikusan megjeleníthető; ennek a halmaznak a „pluriszituatív jelei" lehetővé teszik, hogy a jelentéseket az adott kultúrában vagy szubkultúrában egyértelműen interpretálják. A kódolás (a jelzetekre „fordítás") pedig, akárcsak a szokásos, nyelvek közötti fordítás, olyan tevékenység, amelyben az egyik jelhalmazt egy másikkal helyettesítenek, miközben az üzenet belső jelrendszere (szimbolizmusa), gondolati tartalma változatlan marad. Végül az információkeresés folyamata is olyan tevékenységek sorozata, amelyek egyrészt a kereső elméjében lévő, másrészt a dokumentumokban rögzített fogalmak közelí-

tését, megfeleltetését célozzák az egyezményes és kodifikált szimbólumok segítségével, amelyeknek a kezelését az emberi agy végzi külső, gépi támogatással vagy anélkül. Így mind a fordítás, mind az információkeresés szimbolikus közvetítő eszközökkel végrehajtott fogalomátvitelt jelent, amint az az alábbi ábrán is látható:



Az alapelvek szempontjából lényegtelen, hogy az átvitel gépesített-e vagy sem: az egyetlen követelmény az, hogy valóban végbemenjen. Ha a fogalomátvitelt gépesítik fordítás vagy információkeresés céljából, akkor az ismereteket kifejező szimbólumrendszernek bizonyos alapvető sajátosságokkal kell rendelkeznie. Mind a fordítás, mind az információkeresés közvetítő apparátust igényel, kommunikációs hidat a rögzített ismeretek halmaza és az ezekben az ismeretekben a maga számára információt kereső ember elméje között. Ezt a közvetítő eszközt vagy katalizátort szimbolikus formában kell kifejezni. Mindebből az következik, hogy a gépi rendszernek, amellyel ezt az átvitelt végrehajtják, ezt a kódot vagy szimbólumrendszert úgy kell tudnia kezelni, hogy e kezelésmód a lehető legnagyobb mértékben összhangba kerüljön az emberi elme működési rendszerével. Minél tökéletesebb ez az összhang, annál hatékonyabb az átvitel. Ahhoz, hogy ilyen kódot akár a gépi fordítás, akár az információkeresés céljaira kialakítsanak, meg kell oldani az interpretáció relációs és szintaktikai problémáit: meg kell állapítani a fogalmak közötti, a dokumentumok tartalmától független (szövegfüggetlen) szemantikai vagy analitikus összefüggéseket, továbbá azokat az összefüggéseket, melyek a dokumentumok taralmának reprezentálására kiválasztott fogalmak között állnak fenn e tartalom összefüggései szerint (szövegfüggő, szintaktikai összefüggések). Valószínű, hogy mindezt különböző szinteken, különböző részletességgel kell megoldani. Mind a fordítás, mind pedig az információkeresés szükségessé teszi a szókincs színvonalas ellenőrzését, vagyis a terminológiai ellenőrzést, mivel e nélkül a szóban forgó szimbólumrendszer sokat veszítene kifejezőképességéből, s ez a hiány egyszersmind redundanciát, félreérthetőséget, általános zavart és információveszteséget okozna. Ebből következik, hogy a közös terminológiai problémák megoldásában konszenzusnak kell létrejönnie, különösen a jelentések és a fogalmak közötti szemantikai összefüggések szabványosítása terén, beleértve az ezeket

szimbolikusan reprezentáló kódokat (jelzeteket, illetve jelöléseket). Végül mind a gépi fordításban, mind a gépesített információkeresésben a rendszernek vagy rendszereknek olyan kicserélhető, illetve konvertálható programokkal kell rendelkezniük, amelyek általános vagy speciális célú, nagy, illetve kis, személyes használatú számítógépeken egyaránt használhatóak. Ugyanakkor az emberi tudás specializálódásával párhuzamosan előreláthatólag egyre több lesz a speciális célú számítógép és egyre szűkebb körben használják majd az általános célú számítógépeket.

- Sheratól magyar nyelven az alábbi mű jelent meg:

Könyvtárosság, dokumentáció, tájékoztatástudomány

In: Könyvtári Figyelő, 1970, évf., 16. évf., 3. sz., p. 222–229.
Eredeti: On librarianship, documentation and information science.
In: Unesco Bulletin for Libraries, 1968, No. 2, p. 58–65

A tanulmány történeti áttekintés a tájékoztatástudomány amerikai fejlődéséről, kezdve a századfordulón a Szakkönyvtárak Társaságának (Special Library Association; SLA) önállósulásától harmincas és ötvenes évek könyvtári és dokumentációs csatározásain át a hatvanas évekbeli helyzetig.
Jól felismerhető a fejlődés két vonala: kezdetben a könyvtárosság tudós foglalkozás volt. A közművelődési könyvtárak megjelenésével párhuzamosan fölerősödött a közösségi szolgálat ideája, lejátszódott az eltolódás a szolgálat irányába. A dokumentáció jelentőségének növekedésével törés következett be a könyvtárosok és a dokumentátorok között. „A területet elözönlő nem-könyvtárosok nyíltan megvetették magát a könyvtárosságot... Azt hitték, ha megváltoztatják a terminológiát, megváltoztatták a gyakorlat jellegét is... a »deszkriptor« kifejezéssel például tudományos méltósággal ruházták föl a »tárgyszavakat«." A dokumentáció mindent átfogni akaró radikalizálódását mi sem jellemzi jobban, mint *Suzanne Briet* kijelentése 1951-ben: „Az állatkertben az állatok is dokumentumok."
„Egyre több egyesület alakult, de tagjainak jelentős hányada nem könyvtáros, hanem tudós, aki azért fordult a dokumentációhoz, mert foglalkoztatta a szakirodalom problémája; közülük sokan nyíltan megvetették a könyvtárosokat."
A „dokumentáció" egyre kevésbé látszott alkalmasnak az új szakterület megnevezésére, mivel többértelmű volt. Lényegében ezért

kezdett – eredetileg elsősorban angol nyelvterületen – terjedni helyette az „információtudomány” („tájékoztatástudomány”) kifejezés. A könyvtárosság és az információtudomány merev szétválása szerencsétlen dolog, és csak átmeneti jelenségnek tekinthető. Valójában az elmélet és a gyakorlat ellentéte húzódik meg mögötte. A két szakterület egymás természetes szövetségese.

BRIAN CAMPBELL VICKERY (1918)

Többek között a kötetünkben is szereplő *Harold Borkoval, Jesse Hauk Sheraval, Douglas John Foskettel* együtt a „második tanárnemzedék” egyik legismertebb tagja, a brit osztályozáskutató társaság (Classification Research Group, CRG) egyik alapítója. Osztályozási rendszerszerkesztő pályafutását tudományos intézetek könyvtáraiban kezdte, talajtani, asztronómiai, élelmiszer-ipari rendszerek készítésével. Egyike volt azoknak, akik a legtöbbet tették a fazettás osztályozás európai elterjedése és továbbfejlesztése érdekében azáltal, hogy a gyakorlati követelményekhez igazította *Ranganathan* rendszerét. Rendkívül termékeny szakíró. Azok közé a könyvtáros, tehát eredendően társadalomtudományi képzettségű szaktudósok közé tartozik, akik a kezdődő gépesítéssel párhuzamosan az elsők között tartott lépést az új technológiákkal és közvetítette a tágabb, európai és amerikai nem mérnök–matematikus végzettségű könyvtárosok számára az új felismeréseket. Az angliai szakkönyvtári szervezet, az Association of Special Libraries and Information Bureaux (Aslib) kutató és fejlesztő osztályának vezetőjeként – jórészt saját oktatói tevékenységére támaszkodva – sorra jelentek meg a részben egymásra épülő, egyre kiérleltebb, több kiadást is megért könyvei az indexelésről (1958), a fazettás osztályozásról (1960), információkeresési technikákról (1970) és információs rendszerekről (1961) (1973). Közülük többet más nyelvre is lefordítottak. 1973-tól egy londoni könyvtári és információtudományi főiskola (University College School of Library Archive and Information Studies) igazgatója.

Vickerytől magyar nyelven szöveggyűjteményben az alábbi művek jelentek meg:

A tárgyszavak és tárgyjelek kérdéseinek újabb eredményei

In: Babiczky Béla: Szöveggyűjtemény az osztályozás és indexelés kérdéseinek tanulmányozására. – Budapest : Tankönyvkiadó, 1970. p. 117–128. [Átvéve a Könyvtári Tájékoztató 1956. 4. számából.]
Eredeti: Development in subject indexing. In: Journal of Documentation, 1955, Vol. 2, No. 1, p. 1–10.

A tárgyszavak közötti szemantikai összefüggések[3] szerepének jelentősége az ötvenes évek elején került napirendre az osztályozási szakirodalomban. Vickery korabeli vizsgálatok nyomán ismerteti az összefüggések feltüntetésének grafikus, kategoriális és fazettás elemzéseken alapuló módszereit, *John W. Perry* elemzéseit azokról a fogalmi összetevőkről, melyekből néhány év múlva a *Perry* és *Allen Kent* nevéhez fűződő szemantikai kódok nőttek ki, végül *Ranganathan* fazettás osztályozási rendszerét, gyakorlati példákkal illusztrálva. (A címben szereplő „tárgyjelek" helyesen: „tárgyi jelzetek" vagy egyszerűen csak „osztályozási jelzetek", a cím helyes magyar fordítása pedig: A tartalom indexelésének újabb eredményei.)

Visszakereső rendszerek elemzése[4]

In: Varga Dénes: A dokumentáció nyelvészeti kérdései I. Szemelvénygyűjtemény. – Budapest : Országos Műszaki Könyvtár és Dokumentációs Központ, 1966. p. 18–23.
Eredeti: 2. The analysis of retrieval systems. In: On Retrieval System Theory. – London: 1960. p. 8–13.

Az ötvenes években kezdték először megfogalmazni az információkereső rendszerek összetevőit és az összefüggéseket a tároló, az

3 Az angol szakirodalomban inkább az „analitikus relációk" kifejezés fordul elő, olykor „báziskapcsolatok". A nyelvészetben „paradigmatikus összefüggésekről" beszélnek. A 60-as évek orosz szakirodalmában az „értelmi összefüggések" kifejezést is használták. Minden esetben szövegfüggetlen, a fogalmak, ill. a szavak között eleve meglévő összefüggésekről van szó.

4 A kereséssel összefüggő teljes folyamatot jelentő „retrieval" kifejezésnek – melybe beleértjük a keresőkérdés elemzését, a keresőprofil és a keresési stratégia és taktika kialakítását, a tárolóban végzett keresőműveleteket és a találatok képzését és kiadását – magyarul az „információkeresés" kifejezés felel meg. Korábban elterjedt a „visszakeresés" és az „információ-visszakeresés" kifejezés is (lásd még az „Információkereső gondolkodás kezdetei" című fejezet bevezetőjében a „retrieval" kifejezés használatának történetét, a szerk.).

invertált fájl (Vickery korabeli könyvében: deszkriptorlista) és a keresőkérdés között, az invertált fájl deszkriptorainak és a kérdés deszkriptorainak összehasonlítási problémáit. Mindezt elemi fokon, a korai megfogalmazások viszonylag egyszerű nyelvén írja le a szerző.

Deszkriptornyelvek

In: Varga Dénes: A dokumentáció nyelvészeti kérdései I. Szemelvénygyűjtemény. – Budapest : Országos Műszaki Könyvtár és Dokumentációs Központ, 1966. p. 24–60.
Eredeti: 4. Descriptor languages. In: On Retrieval System Theory. – London: 1960. p. 23–57.

Olyan „korai" időszakban, amikor még „a deszkriptornyelv felépítési elvei tekintetében nincs általánosan elfogadható megegyezés", Vickery részletesen elemzi kora minden, a deszkripotornyelvekkel és a koordinált indexeléssel összefüggő eredményeit. (A deszkriptornyelvet a legáltalánosabb értelemben használja, beleértve az osztályozási rendszereket is.) Ismerteti a deszkriptornyelv ellenőrzésének, a deszkriptorok közötti szinonim, faj–nem és egyéb analitikus relációk elemzésének, a fogalmak elvont kategóriákba sorolásának kérdéseit. Kitér *Perry* és *Kent* szemantikai kódjaira, a keresési relevancia és zaj mérésére, a hierarchikusan szervezett katalógusra, illetve a koordinált (általa korrelatívnak nevezett) indexelésre. A szemantikai kódokkal összefüggésben foglalkozik a szerepjelölőkkel is. Végül konkrét tartalmi leírás öt különböző deszkriptornyelvvel végzett feltárását mutatja be összehasonlító elemzéssel.

A könyvtár és a tájékoztatástudomány oktatása és a tudományos kutatás

In: Könyvtári Figyelő, 1976, 22. évf., 4–5. sz., p. 399–402.
Eredeti: Academic research in library and information studies. In: Journal of Librarianship, 1975, Vol. 7, No. 5, p. 153–160.

Elsősorban gyakorlati jellegű tárgyak oktatására van szükség, mint a dokumentumtipológia, a közvetítő tevékenységek (elemzés, tárolás, keresés), a dokumentumok előállítása és másolása, könyvke-

reskedelem, kommunikációs szokások, az információközvetítő folyamat rendszerei. Végül ki kell térni az oktatásban a kutatási célokra és módszerekre.

Információs rendszerek[5]

8. fejezet. Információkereső nyelvi modellek[6]

Az ötödik fejezetben az információkereső nyelvek két fontos tulajdonságát tárgyaltuk. Az első, hogy az osztályozáshoz/indexeléshez[7] használt kifejezések között relációk[8] vannak, akár a szövegből emelik ki, akár ellenőrzött tezauruszból választják ki őket. A második, hogy a kifejezéseket a dokumentumképen vagy a keresőképen belül koordinálják, mégpedig vagy szabadon egymás mellé rendelve, csoportokba összekapcsolva, rendszerezetten összekapcsolva vagy súlyozva. Fejezetünkben ezeknek a tulajdonságoknak a modelljeit tekintjük át általánosabban. Mint már a hetedik fejezetben említettem, a modellek leírásához nagymértékben használunk logikai szimbólumokat.

Meadow három diagrammal modellezi az indexkifejezések (osztályozó fogalmak) alkotta struktúrát. Felfogásában minden kifejezés a „tárgyi (tematikai) univerzum" meghatározott területét foglalja el. A közvetlenül a természetes nyelvű szövegből vett kifejezések (1. ábra) mindaddig „nem osztják fel az univerzumot, amíg a szót nem használják fel… Az osztályozók akkor foglalják le a tárgyi univerzum egyes területeit, amikor a kifejezéseket felhasználják" (a sötét területek).

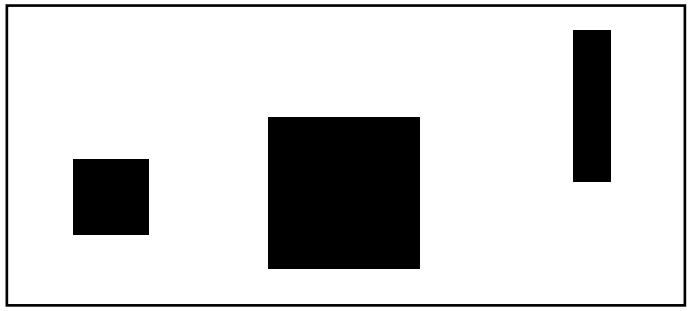
---

5 Vickery, B. C.: Information Systems. London : Butterworth, l973. 350 p.

6 Retrieval language models. In: Information Systems, p 203–222.

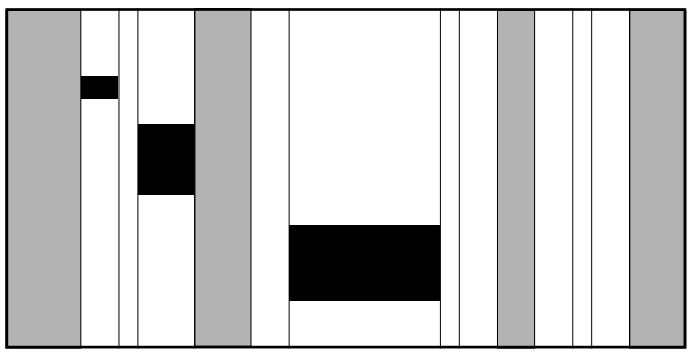
7 A szerző mindig az „indexing" (indexelés) kifejezést használja, amely szűkebb értelemben a dokumentum természetes nyelvű kifejezéseinek közvetlen felhasználását jelenti a tartalmi feltárásban (vö. Dahlberg, I.: Osztályozás és/vagy indexelés című tanulmányával első kötetünkben). Következtetései e könyvben a hierarchikus, prekoordinált, mesterséges nyelven alapuló információkereső nyelvekre is vonatkoznak, ezért az „indexing" szó egyenértékű az általánosabb értelmű „osztályozás" kifejezéssel. Továbbá a szerző mindig az információkereső nyelv kifejezést használja a dokumentációs nyelv kifejezés helyett. A szóhasználatra vonatkozóan lásd az első kötet „A dokumentációs célú osztályozás" című részének bevezetőjét és *Gernot Wersig* szemelvényét (a szerk.).

8 Az összefüggések relációk abban az esetben, ha egzakt, matematikailag meghatározott értelmük van; kapcsolatok viszont, ha további, már csak körülményesen meghatározható jelentésük (is) van. Az összefüggések valójában csak a legritkább esetben állnak fenn a kifejezések (terminusok) és kifejezések között (például szinonímia esetén). Általában vagy a kifejezésekkel képviselt fogalmak (például generikus, nem–faj reláció), vagy a fogalmakkal képviselt dolgok, jelenségek (például az egész–rész reláció, hasonlósági összefüggések) között állnak fenn. Vickery leegyszerűsítve mindig csak a kifejezések (terminusok) közötti összefüggésekről, kapcsolatokról beszél (a szerk.).



1. ábra. Természetes nyelvű kifejezések

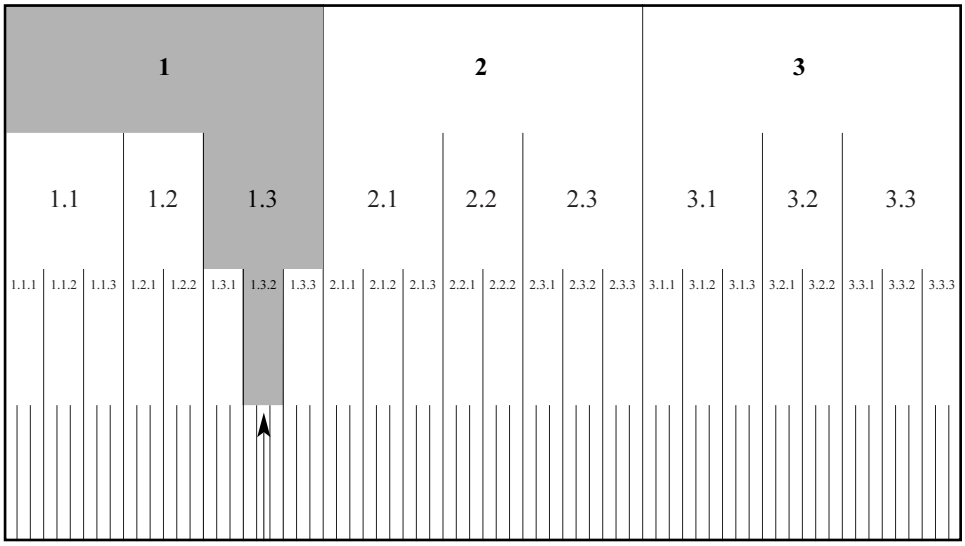
A betűrendes tezaurusz[9] természetes nyelvű kifejezései (2. ábra) „eleve meghatározott módon osztják fel a tárgyi univerzum területét... [de felhasználáskor] nem kell a teljes univerzumot lefedniük definiált deszkriptorokkal" (a fekete területeket használták fel osztályozáskor, a vonalkázott területekre a tezauruszban nincsenek kifejezések).



2. ábra. A tezaurusz kifejezései

A strukturált tezauruszban (3. ábra) „bármely kifejezés kiválasztása a tárgyi univerzum adott területének alsó sávját határozza meg és automatikusan elkülöníti az univerzumnak a sáv fölé eső részleteit, amelyeket a vonalkázás mutat".

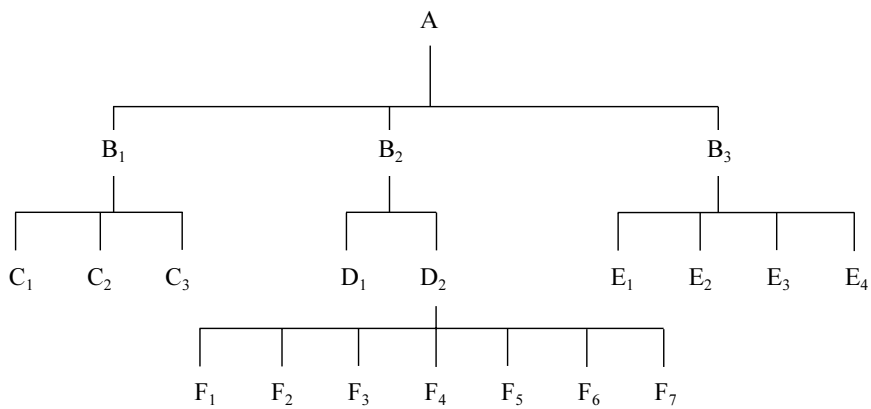
9 Vickery a tezaurusz és deszkriptor fogalmakat általánosabb értelemben használja, ahogy ez a deszkriptornyelvek gyakorlati alkalmazásakor szokásos (és ami összhangban van a nemzetközi szabványokon alapuló magyar MSZ 3418 jelű szabvány meghatározásával). Vickery lényegében minden információkereső nyelv (= osztályozási rendszer) szótárát tezaurusznak és minden ebben szereplő kifejezést deszkriptornak nevez. Ez a szóhasználat nemcsak az elvont megközelítés esetén gyakori (például a kötetünkben szereplő *Jurij A. Šrejdernél*), hanem az automatizált információkereső rendszerek szótárait használók körében is (a szerk.).



3. ábra. Hierarchikus osztályozás

Hierarchikus osztályozás

A tezaurusz eme utoljára modellezett formája az információkereső nyelv legrégebbi fajtája: a hagyományos enumeratív osztályozás, az „ismeretek fája”, amelyet (többek között) *Ranganathan* ábrázolt a 4. ábrán látható módon.



4. ábra. Az „ismeretek (hierarchikus) fája”

Ebben a modellben a törzsből a B_1, B_2, B_3 fő ágak indulnak ki, amelyek egyre több ágban folytatódnak. A „fát” a könnyebbség kedvéért gyakran megfordítják, így az A törzs (az ismeretek „univerzuma”) kerül fölülre és az ágak lógnak lefele. Ez a modell csak két logikai kapcsolatot fejez ki: (1) az alacsonyabban elhelyezkedő elemek (osztályok) generikus kapcsolatát[10] a fölöttük lévő osztályokkal (például B_1–A; C_1–B_1 és –A; F_4–D_2, –B_2 és –A között), és (2) az alárendelt osztályok egymás közötti kollaterális kapcsolatát (pl. E_1, E_2, E_3 és E_4 egymással kollaterális). Ebben a modellben egyetlen „univerzum” (summum genus, A) van és minden osztály közvetlenül csak egyetlen másik osztály alárendeltje (a D_2 egyetlen fölérendeltje a B_2), továbbá minden dokumentumegységhez csak egy osztályt rendelnek hozzá. Az olyan könyvosztályozás például, mint amilyen *Dewey-é*, majdnem teljesen megfelel ennek a modellnek.

Ennek a modellnek a leírási terminológiája a következő: Az *univerzum* az *entitások* (dolgok vagy fogalmak, „ideák”) vizsgált együttese. Ezt a *tulajdonságok* sorozata vagy *lánca osztja fel*, melyek mindegyikéből az osztályok egy-egy sora indul ki (a B_1 felosztása egy C jellemző szerint a C_1, C_2, C_3 sorozatot eredményezi). Az osztály *rendje* az univerzum felosztására az osztály kialakítása érdekében felhasznált jellemzők száma (C_1 a második rendbe tartozik). Az osztály *rangját* az osztálynak az adott elrendezésben elfoglalt helye adja (az indexek jelzik a rangot).

Megállapítottuk, hogy *Dewey* könyvosztályozása *majdnem* megfelel ennek a modellnek. A „majdnem” szót azért használtuk, mert csak viszonylag szűk ismeretterületet átfogó osztályozási rendszer esetén remélhetjük, hogy teljesül a feltétel: minden osztály csak egyetlen másik osztály közvetlen alárendeltje legyen. Általában az tapasztalható, hogy egy-egy kifejezés több osztályba is besorolható. Az „ismeretek fája” csupán a kifejezések között fennálló kapcsolatok leegyszerűsítése. *Perry* és *Kent* olyan modellt hozott létre, amelyekből meghatározott fák származtathatók.[11]

Modelljük az entitások összekapcsolására alkalmas tulajdonságok halmazából állt. Példaként álljon itt a következő tíz tulajdonsághalmaz *(1. táblázat),* amelyek felhasználhatók az építési blokkok és hasonló entitások összekapcsolására:

---

10 A generikus kapcsolat a beletartozást képviseli (a logikában inklúziónak nevezik), szigorú értelemben csak az általános, fölérendelt, magasabb szintű, generikus nemfogalom és a speciális, alárendelt, alacsonyabb szintű, specifikus fajfogalom között, például: bútor–asztal. A legtöbb hagyományos rendszerező rendszerben a hierarchikus összefüggés nem generikus, hanem egész–rész, leszármazása stb. típusú, például: jog–bíróságok (a szerk.).

11 *Perry* és *Kent* szemantikai kódjait első kötetünkben részletesen tárgyalja *Eric de Grolier.*

Halmaz	Tulajdonság	Példa
1	alak	kocka, gömb, gúla
2	anyag	fém, műanyag, üveg
3	szín	vörös, sárga, zöld
4	felület	érdes, sima, csorbult
5	méret	numerikus érték
6	első jelölés	arab szám
7	második jelölés	római szám
8	harmadik jelölés	kisbetű
9	negyedik jelölés	nagybetű
10	ötödik jelölés	írásjel

1. táblázat

Ha minden entitást minden halmaz egy-egy tulajdonsága jellemez és minden halmazban tíz tulajdonság van, akkor 1010 lehetséges entitás van. Ezeket az entitásokat különféleképpen csoportosíthatjuk, például feloszthatjuk az entitások univerzumát először az „alak" tulajdonság szerint, majd az „anyag, szín" stb. szerint; ez egyfajta hierarchiához vezetne. Egy másik hierarchia alakítható ki a „szín, méret, alak" stb. tulajdonságok ilyen sorrendű felhasználásával. A lehetséges hierarchiák száma (az osztályok sorozaton belüli *rangját figyelmen kívül hagyva*) $10\times9\times8\times7\times6\times5\times4\times3\times2\times1 = 10! = 3\,628\,800$. Ha az alternatív rangsorolást is figyelembe vesszük ez a szám 100!

Az inkább a tulajdonságokon mint az entitásokon alapuló ilyen modell olyan forrás, amelyből megszámlálhatatlan speciális hierarchia származtatható, tehát jóval általánosabb, mint az egyszerű hierarchikus modell. *Perry* és *Kent* a $D_G = N'/N$ hányadost javasolta a tulajdonságok általánosságának mértékét képviselő „generikussági szint"[12] kifejezésére; az N' = azon entitások száma, amelyekre a tulajdonság érvényes, és N = a rendszerhez tartozó entitások összege.

Koordinált rendszerek[13]

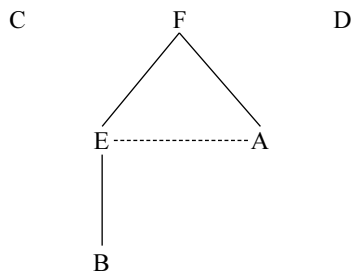
A *Perry–Kent* modell másik szemszögből is vizsgálható. Ha a dokumentumokat tekintjük „entitásoknak" és a „tulajdonságokat" generikus kifejezéseknek, akkor olyan dokumentumgyűjteményt kapunk, amelyben minden dokumentumot néhány koordinált generikus kifejezés ír le. A kifejezések közötti kapcsolatok megszámlálhatatlan hierarchiába illeszkedhetnek aszerint, hogy

12 Angolul: „degree of generic character" – D_G (a ford.).

13 A szerző ezen a prekoordinált (hierarchikus) osztályozási rendszer ellentétét, a posztkoordinált (mellérendelő) rendszert érti, melyben a kifejezéseket a dokumentumban vagy a kérdésben megfogalmazott tárgykörök szerint, tehát utólag rendelik egymás mellé (a szerk.).

milyen sorrendben használtuk őket a hozzárendeléskor vagy a keresés során. A rendszer nem egyetlen, hanem sok lehetséges hierarchiából épül fel. Ilyen módon a dokumentum leírására[14] és keresésére elvben különböző „tulajdonságláncok" használhatók fel és e láncok relatív hatékonysága is értékesíthető. A kifejezéseknek ama szekvenciáját választják, amelyik az adott dokumentumot a legkevesebb lépésben individualizálja.

A *Perry–Kent* modell adott példájában minden tulajdonság vagy generikus kifejezés halmazát kollaterális (mellérendelő) kapcsolatban álló kifejezésekből állítottam össze. Valójában mindegyik halmaz tartalmazhat hierarchiát, és az indexelés során a hierarchiák osztályait koordinálják. E típus kis modelljéhez hat kifejezés tartozik, amelyeket A-tól F-ig nagybetűkkel jelöltünk és három csoportba soroltunk. Az egyik csoport csak C-t, a másik csak D-t tartalmazza, a harmadik pedig E és A alárendeltje F-nek, B meg E-nek. A relációk halmazát az 5. ábra mutatja. A generikus és specifikus kifejezéseket folytonos, a kollaterálisokat szaggatott vonal köti össze.



5. ábra

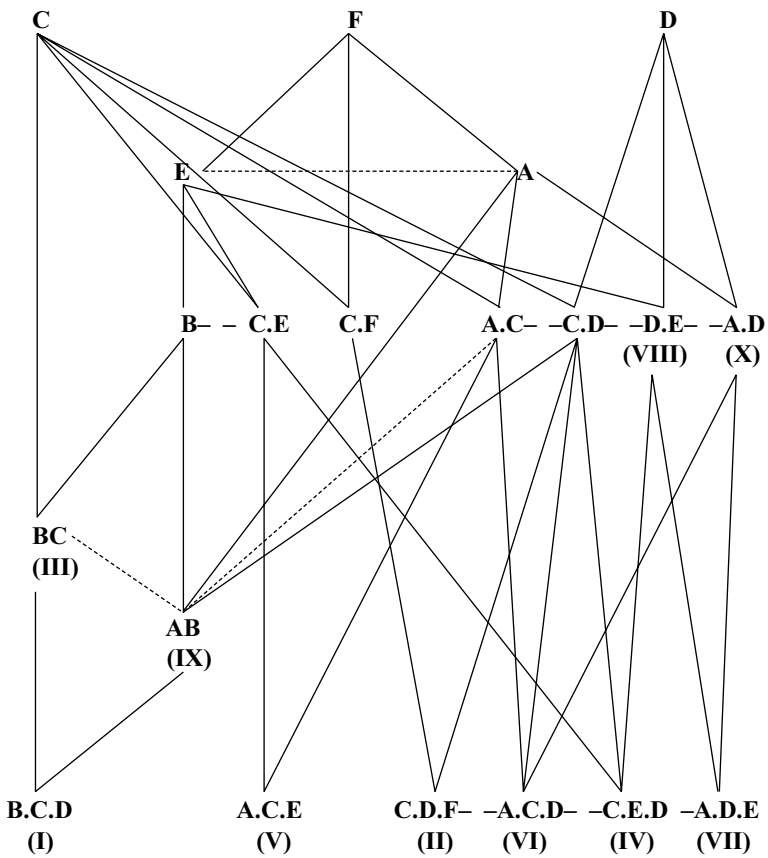
Ezekből az elemekből tíz specifikus témát származtatunk. Feltételezem, hogy minden téma egybeesik egy dokumentációs egységgel, bár ez nem szükségszerű. Az egységeket római számok képviselik *(2. táblázat).*

14 Leíráson (description) Vickery a formai (például a bibliográfiai) főleg pedig a tartalmi elemzés (feltárás) eredményének – a tárgyköri leírásnak (subject description) – legáltalánosabban értelmezett *rögzítését* érti. A leíráshoz a tulajdonságokat jelölő kifejezéseket használják fel, amelyek ettől kezdve a dokumentumtétel ismertetőjegyeit, az ún. ismérveket képviselik. Ilyen leírás például az ETO jelzetlánca, tárgyszavak vagy deszkriptorok csoportja, de természetes nyelven megfogalmazott tartalmi kivonat, referátum vagy akár cím is. Lásd még „Laisiepen, K., Lutterbeck, E., Meyer-Uhlenried, K-h.: A szakirodalmi tájékoztatás alapjai. Bevezető kézikönyv a gyakorlati dokumentalisztikába" c. szemelvényünk M 8.1.2.2 fejezetét. A német terminológia szerint a leírás (Beschreibung) csak indexelés vagy tartalmi kivonatkészítés lehet, osztályozás – azaz hierarchikus osztályozási rendszerrel végzett jelzetelés – nem (a szerk.).

Egység	Téma	Egység	Téma
I	B, C ,D	VI	A, C, D
II	C, D, F	VII	A, D, E
III	B, C	VIII	D, E
IV	C, E, D	IX	A, B
V	A, C, E	X	A, D

2. táblázat

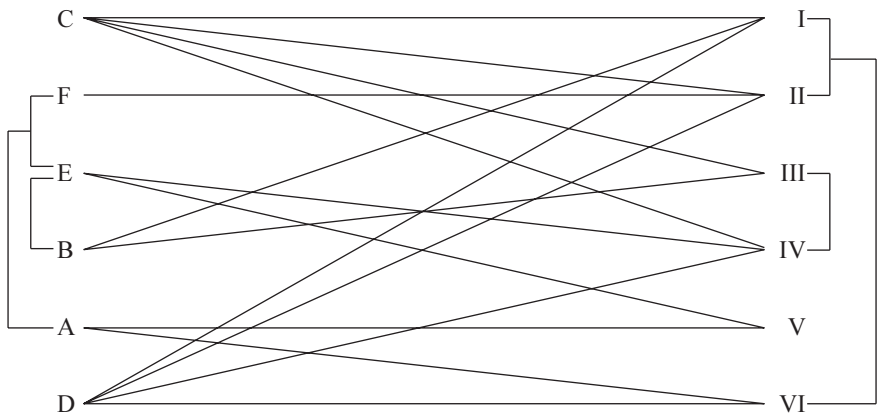
A 6. ábra mutatja, hogy fejezhetők ki diagramon a fájlon belül egymáshoz kapcsolódó elemek. A relációk bonyolultsága (bár nem jelöltük az összes kollaterális relációt) még e kis modell alapján is szembeötlő.



6. ábra. Hierarchikus kapcsolatok gráfstruktúrában

Salton mutatott rá, hogy ha több fát összefűzünk az elemeik között jelölt további kapcsolatokkal, akkor a matematikából ismert *gráfstuktúra* keletkezik. A fazettás osztályozásra épülő katalógusok és más, hierarchikusan kapcsolódó kifejezéseket koordináló rendszerek e modellnek felelnek meg. Ez a modell ennél fogva közelebb áll a modern keresési gyakorlathoz, mint az „ismeretek fája".

Az A. D. Little, Inc. kutatócsoportja szerint a fájlon belüli összefüggések még ennél is komplexebbek lehetnek. Először is kapcsolatok vannak az – ismérvként használt – kifejezések között, mint az 5. ábrán. Másodszor: minden kifejezés egynél több dokumentumhoz rendelhető hozzá, így például a C a 2. táblázaton mind az I., mind a VI. egységhez kapcsolódik. Harmadszor: a dokumentumok tartalmi leírásuktól teljesen függetlenül is kapcsolódhatnak egymáshoz; két dokumentumot például összekapcsolhat a közös hivatkozás egy harmadikra, és ez felhasználható az információkeresésben. A teljes kapcsolathálót a 7. ábra mutatja.



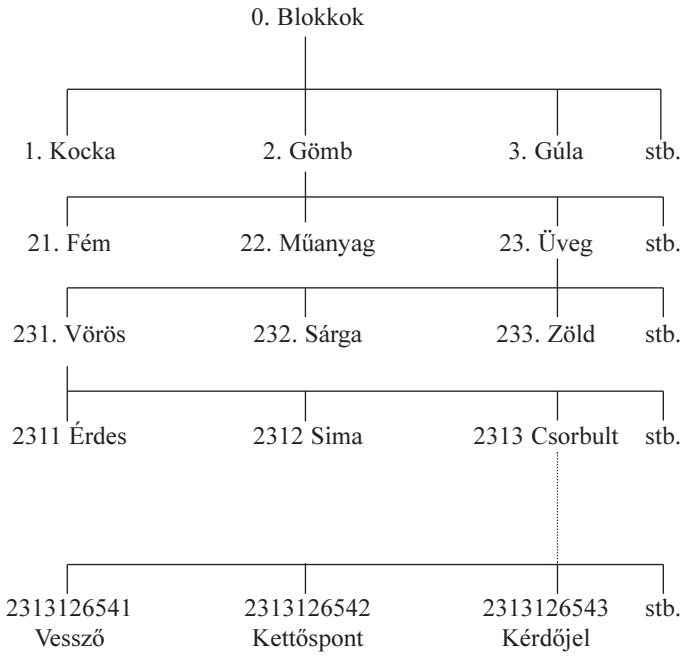
7. ábra. Kapcsolatok az ismérvként használt kifejezések és a dokumentumok között

A deskriptív kontinuum

A *Perry–Kent* modell egyszerű hierarchiájából származó részletet a 8. ábrán látható módon reprezentálhatjuk.

Szakkatalógusban a „Gömb alakú, Üveg, Vörös, Érdes… Vesszővel jelölt" blokkot a 2313126541 kifejezés képviseli, amely egyesíti magában a blokk összes tulajdonságát. Uniterm mutatóban a blokkot tíz független kifejezés egymás mellé rendelésével (korrelációjával) állítanák elő: Gömb, Üveg, Vörös … Vessző. E két végpont között sok átmeneti rendszer képzelhető el, amelyek a kifejezést olyan kombinációkból hoznák létre, mint „Üveggöm-

bök", „Sárga gúla", „Vörös műanyag kúpok", „Sima fémgömbök, nagy A-val és vesszővel jelölve" stb.



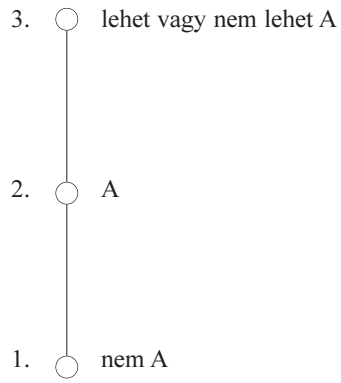
8. ábra. Hierarchiarészlet

F. Jonker ezt a jelenséget tekinti „deskriptív kontinuumnak". A paraméter, amely a kontinuumban elfoglalt helyet meghatározza – Jonker szerint – „az ismérvként felhasznált kifejezések átlagos hossza, ami egyszerűen a kifejezésekre eső betűk számával mérhető". Célszerűbb a paramétert úgy tekinteni, mint a kifejezések *Perry–Kent*-féle „generikussági szintjét" jellemző mutatót. A „Gömb" generikusabb mint az „Üveggömb", és ez generikusabb mint a „Sima fémgömbök „A-val jelölve", és ez ismét generikusabb, mint a 2313126541 által képviselt gömbfajta.

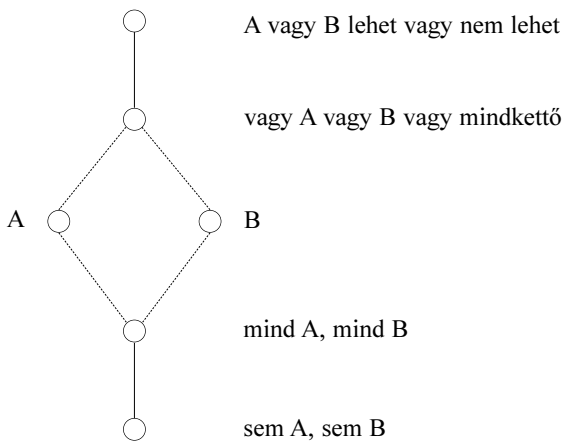
Jonker azt állította, hogy a rendszerjellemzők másik része a kifejezések generikus szintjének mértékével kapcsolatos. Felfogása szerint mennél generikusabbak lesznek a kifejezések, annál inkább növekszenek bizonyos paraméterek (az osztályozás lehetséges mélysége, a kifejezések permutáltsága, a teljes rendszerben végzett kereséshez szükséges egyedi keresések száma, a hamis kombinációk valószínűsége, a szemantikai bizonytalanságok kiküszöbölésének képessége), mások pedig (a hierarchikus meghatározottság szintje, az alapszótár nagysága) csökkennek.

A kifejezések hálója

Az eddig diagramokon bemutatott szerkezeti modellek alig többek mint az ismérvként használt kifejezések közötti kapcsolatokat illusztráló vizuális segédletek. Mások az osztályok algebrája felöl vizsgálták ezeket a kifejezéseket. *Fairthorne* például a lehető legegyszerűbb, egyetlen A témára vonatkozó szemantikai struktúrát elemezte.



Egyetlen téma esetén csak három állítás létezik a szöveggel kapcsolatban: az A leírás releváns (2. pont), irreleváns (1. pont), relevanciája eldöntetlen vagy nem határozható meg (3. pont). Ha egy második témát is bevonunk az elemzésbe a háló a következőképpen alakul:



Két vagy több elem kombinációja „és” kapcsolatban az elemek metszete vagy *szorzata* (így AB az A és a B szorzata); két vagy több elem kombinációja „vagy-vagy” kapcsolatban az elemek *uniója* vagy *összege* (így A\/B az A és

a B összege). Mennél több témát vonnak be, annál bonyolultabbá válik a háló; hat téma esetén a diagram 7 828 354 pontot tartalmaz.

Minden információkereső rendszer ilyen hálón alapszik. Ahogy *Fairthorne* rámutat „… a hálót az állomány nagyságának megfelelően bővítik vagy szűkítik a leírások pontosságának rovására; azaz egyes leírásokat ekvivalensnek tekintenek és egyforma megnevezéssel jelölik őket … Bizonyos dokumentumkereső rendszerekben nem annyira a háló felbontóképességét, mint inkább a relációszerkezetét áldozzák fel. Inkább a specifikusabb elemek kiválasztásának lehetőségét, mint ezek számát változtatják meg és eltörölnek összekötő vonalakat… Így az ETO-ban és más szigorú rendezőrendszerekben eltörlik a vonalakat, hogy meghagyják a lefelé terjeszkedő, egymást át nem fedő fákat". (Mint e fejezetben korábban már szemléltettem.) „… A tulajdonságok mellérendelő összekapcsolásán alapuló rendszerekben, mint amilyen a koordinált indexelés, az összes kapcsolóvonalat eltörlik, kivéve azokat, amelyek a specifikusabb leírásokból ágaznak szét, hogy általánosabb tartalmú, tehát kevesebb számú leíráshoz vezessenek."

Fairthorne a háló két tulajdonságát határozza meg: a dimenzionalitást és a konnektivitást. Értelmezésében a *konnektivitás* a háló elemeinek összefüggéseit vagy előfordulási gyakoriságát jelenti. A hagyományos osztályozási fa elemei vagy nem, vagy csak egyetlen szálon kapcsolódnak bármely más elemhez. A konjuktív vagy tulajdonság összekapcsoló (mellérendelő) rendszer leíró elemei viszont többszörösen kapcsolódnak összetevő tulajdonságaihoz. Egy elem *kiterjedését* (dimenzióját) a belőle kiinduló leghosszabb lefelé irányuló lánc hosszúsága (a kapcsolatok száma) képviseli. Más szóval a kiterjedés az elemnek a diagram legalacsonyabb pontjához mért „magassága". Ezt a paramétert *Ranganathan rendjével* lehet összevetni, ami az elem „mélysége" az osztályozási fa legmagasabb pontjához képest.[15]

Deszkriptorhálóban az egyik elem tartalmazza a másikat, ha az elemmel jellemezhető szövegek között az összes, a másik elemmel jellemezhető szöveg szerepel. Így, az „A vagy B vagy C" együttes tartalmazza az „A vagy B", ez az „A vagy (B és C)", ez az A, az A pedig az „A és B" együttest, és így tovább. Az ilyen belefoglalási reláció könnyen követhető a 6. ábráéhoz hasonló diagramokon, amelyeken a területek dokumentumok halmazait jelentik. A hálódiagramban minden halmaz tartalmazza az összes *lefelé* irányuló vonallal hozzákapcsolt együttest. A hálóban a belefoglalási relációnak három fajtája van:

(i) két tetszőleges elem *összege* tartalmazza mind a két elemet (A v. B mind A-t, mind B-t tartalmazza),

(ii) két tetszőleges elem *szorzatát* mind a két elem tartalmazza (A és B egyaránt tartalmazza AB-t),

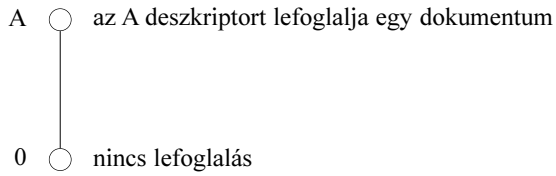
15 A kiterjedés a fogalom terjedelmét képviseli. A konnektivitás – áttételesen, közelítőleg – a fogalom tartalmára utal (a szerk.).

(iii) ha az A kifejezés *generikus* B-hez képest, akkor A tartalmazza B-t (de az A-val jelölt dokumentumok halmaza nem tartalmazza feltétlenül a B-vel jelzettekét, hacsak A-t és B-t nem kapcsolják össze a rendszerben).

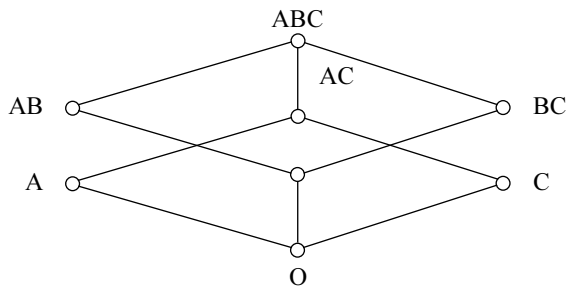
A modell megnevezési egységei („karakterei")

A deszkriptorhálót vizsgálta *Mooers* is, aki elsősorban a különböző információkereső rendszerek közötti szerkezeti eltérések feltárására törekedett. Olyan rendszerből indult ki, amelyben az osztályozó fogalmakat képviselő kifejezések legegyszerűbb formáját használják. A keresőnyelv elemeit *karaktereknek*[16] nevezi, olyan szimbólumoknak, melyek a keresés során egymástól függetlenül kezelhetők, nem bonthatók fel két vagy több független szimbólumra és véges jegyzékből származnak. *Mooers* a „deszkriptor" szót csak a hierarchia és logika nélküli karakterek megnevezésére használta.

Mooers a deszkriptoros rendszer bázisszerkezetét a következőképpen adja meg:



Ebben a rendszerben nem használják ki a „nem A" lefoglalást. *Mooers* „nincs lefoglalás"-a ekvivalens *Fairthorne* „lehet vagy nem lehet A"-jával; megjegyezzük továbbá, hogy *Mooers Fairthorne*-höz képest fejjel lefelé rajzolja diagramjait. *Mooers* háromdeszkriptoros hálóját a 9. ábra mutatja:

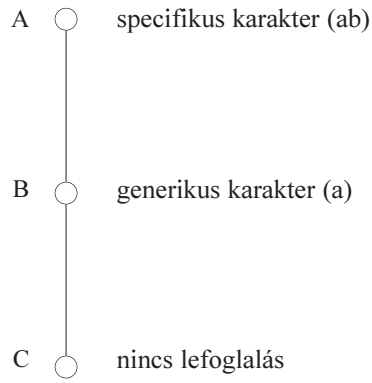


9. ábra. A deszkriptorok hálója

16 *Calvin Mooers* egyéni szóhasználatában a karakter az információkereső nyelv szavának felel meg (a szerk.).

A deszkriptorokkal csak szorzatok állíthatók elő, összegek nem: egy dokumentumot csak „A és B”-vel indexelhetünk, „A vagy B”-vel nem. *Mooers* diagramjából ezért *Fairthorne*-nak minden összeget tartalmazó eleme kimarad.

Mooers ezután továbbment és megvizsgálta az olyan „hierarchikus karaktereket”, mint amilyenek az osztályozási jelzetek. Itt a bázisszerkezet így írható le:



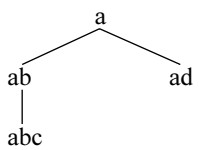
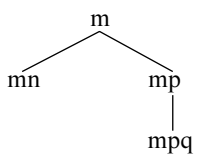
Itt az A generikus fölérendeltje B-nek, például Állat és Madár. Ha a rendszerben az Állatot és a Madarat generikusan összekapcsolják, a Madárral osztályozott dokumentumot is megkapjuk, amikor az Állatra keresünk, és a keresés a Madárról az Állatra kiterjeszthető. *Mooers* ezt úgy fejezte ki, hogy a Madarat az **ab** karakterrel reprezentálta, amelyben az **a** az Állat tulajdonság, és **b** a Szárnyas tulajdonság. Az **a**, **ab** szimbólumok jól tükrözik a hierarchikus osztályozás hagyományos jelzetelési eljárását, ahol az Állat = a = 59, és a Madár = ab = 5982. Minden további betű vagy szám a hierarchikus felosztás egy új tulajdonságát képviseli.

Mooers a következő lépésben a hierarchikus rendszerek két fajtáját különböztette meg. Mindkettő **a**, **b**, **c** stb. tulajdonságok rendezett sorozatával operál és minden egyes karaktert (hierarchikus kifejezést) egy ebből a sorozatból rendezetten kiemelt elemlánccal fejeznek ki, például (bed). Az erős hierarchiájú[17] rendszerekben azonban a sorozat minden elemét egyetlen elem előzi meg, a gyenge hierarchiájú rendszerek elemeit azonban egynél több elem előzheti meg. Az erős hierarchia a *Ranganathan* által is vázolt osztályozási fának felel meg, amelyben minden osztály csak egyetlen generikus fölérendelt alárendeltje. A gyenge hierarchia a *Perry–Kent*-modell megfelelője, amelyben minden kifejezés (pl. Vörös műanyag kocka) több, generikusabb kifejezés (Vörös kocka, Műanyag kocka, Vörös műanyag) közvetlen alárendeltje lehet.

17 Elterjedtebb és logikailag is megalapozottabb terminológia szerint monohierarchia (= erős hierarchia) és polihierarchia (= gyenge hierarchia) (a szerk.).

Mooers rámutatott, hogy „a gyengén hierarchikus karakterrendszer annál hajlékonyabb lesz, minél több közös fölérendeltje lehet a különböző elemeknek… Ha ezt a folyamatot végigvisszük, olyan állapot alakul ki, amelyben az elemeknek nincs kötött fölérendeltje és bármely más elem közös fölérendeltjévé válhat[18] …Amikor a feltételek ennyire lazák, olyan karakterrendszert kapunk, amely majdnem azonos a deszkriptorrendszerrel".

A korábbi gráfstruktúrával ábrázolt fazettás osztályozás *Mooers* szerint olyan rendszer, amelyben a fazettákon belül a karakterek között erős a hierarchia, a fazetták között azonban gyenge. Ezt a *3. táblázat* szemlélteti.

Fazetta	Elemek	Fölérendelt („előrendelt")	Fa
I.	a	0	
	b	a	
	c	b	
	d	a	
II.	m	0	
	n	m	
	p	m	
	q	p	
III.	x	0	
	v	x	

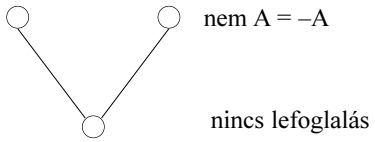
3. táblázat

Az I. fazetta bármely karakterét követheti a II. vagy III. fazetta bármely karaktere; és II. bármely karakteréhez járulhat III. bármely karaktere. Így az (abc) (mpq) (xy) forma bármely kombinációját megkaphatjuk.

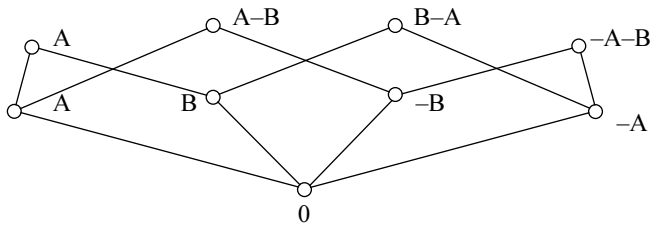
Végül *Mooers* foglalkozott a „logikailag kapcsolt karakterekkel", azaz olyan karakterekkel, amelyek nemcsak logikai szorzatokat (A és B), hanem összegeket (A vagy B) és különbségeket (A és nem B) is alkothatnak. Megkérdőjelezte az összegek használatát dokumentumok leírására, mert „ha dokumentu-

18 *Calvin Mooers* szóhasználatával: „előrendelt elem" (predecessor) (a szerk.).

mot kell leírni, a leírásnak pontosnak kell lennie, és ez alternatívákkal nem érhető el". (Ez nem zárja ki összegek használatát a keresőkérdésekben.) Következésképpen *Mooers* diagramjaiban szorzatok és különbségek csak együtt szerepelnek. A bázisszerkezet a következő:



Két ilyen stratégia kombinációja a 10. ábrán látható hálóhoz vezet.



10. ábra. Deszkriptorháló logikai szorzatok és különbségek felhasználásával

Mooers azzal a meggyőződéssel zárta munkáját, hogy „a gyakorlatban ma használatos információkereső rendszerek szinte kivétel nélkül e családok és variációk valamelyikéhez tartoznak… Bár ez a modell… a legtöbb jelenlegi keresőrendszer leírására alkalmas, az elméleti feladatok még megoldatlanok… A modellből hiányoznak bizonyos statisztikai adatok… például a különböző karakterek használatának gyakorisága… E modell másik fontos érdeme, hogy az új nyelvi rendszerek részletes struktúrájának kidolgozását módszertanilag segíti az új, kombinálható bázisszerkezetek megadásával".

Nem titkolja azt sem, hogy a modell egyelőre csak *kommutatív* karakterekre érvényes. A kommutatív karaktereknél közömbös a sorrend, amely szerint az állításokat felépítjük: AB ugyanazt jelenti mint BA.[19] Vannak azonban rendszerek, például a fazettás osztályozás, amelyekben a deszkriptorok nem kommutatív kombinációit használják, és az így kialakított háló az előbbitől eltérő lesz. A kémiai szerkezeti képletek készítésekor használt karakterek ugyancsak a hálóelv alapján kapcsolhatók össze. Ennél fogva más modellszerkezetek is megvizsgálandók.

19 Az ilyen információkereső nyelveket ma szintaxis nélküli – azaz mondattani eszközöket nem használó, szövegösszefüggéseket nem jelölő – nyelvnek nevezik. (a szerk.)

Transzformációk[20]

Az eddig megtárgyalt struktúrák az információkereső nyelvet az információs rendszerekben használatos módon modellezték. Nem a nyelv összeállítási folyamatának, csupán a folyamat végtermékének a leírására törekedtek. Pedig, ha az információkereső nyelvek készítését akarjuk segíteni, éppen ezt a folyamatot kell elemeznünk. Információkereső nyelvek szerkesztésének és a keresésben való alkalmazásának a műveletei tulajdonképpen transzformációsorozatok. A dokumentumok szövegéből indulunk ki; ezt alakítjuk át a tartalom ellenőrzött nyelvű leírásává; az egyes tartalmi leírásokat összevetjük a kérdéssel, és a megegyezőket ismét szöveggé transzformáljuk. E lépéseket kell részletesen megvizsgálnunk!

Kezdjük rövid mintaszöveggel, egy referátummal: „Nagy hőmérsékleten használható metallográfiai vizsgálóberendezések: olyan hőálló mikroszkóptárgyasztalt készítettek, amelynek segítségével 1800 °F-ig terjedő hőmérsékletig fémek szerkezeti változásai figyelhetők meg; a berendezés fő részei a vákuum kemence, az optikai egység, a hűtőbélés és a vákuumszivattyú egység; a szerkezetet nagy hőmérsékletű urán, cirkónium és SAE 1008 acél szerkezetének vizsgálata közben próbálták ki".

Minden szöveg felfogható kijelentések sorozatának; egy-egy kijelentésben az alanyt és az állítmányt a „van" (vagy „vannak") kopula köti össze.[21]

Például a fenti referátumot az alábbi kijelentésekkel helyettesíthetjük:

(1) A hőálló mikroszkóp tárgyasztal [„van"] nagy hőmérsékletű metallográfiai vizsgálóberendezés.
(2) A hőálló mikroszkóp tárgyasztal [„van"] fémek 1800 °F alatti hőmérsékleten lejátszódó szerkezeti változásainak vizsgálóberendezése.
(3) A hőálló mikroszkóp tárgyasztal részei [„vannak"] a vákuum kemence, az optikai egység, a hűtőbélés, a vákuumszivattyú egység.
(4) A hőálló mikroszkóp tárgyasztal [„van"] az urán, cirkónium, SAE 1008 acél szerkezetének vizsgálóberendezése.

Ha az *alanyokat* elnevezzük S_1, S_2 stb.-nek, az *állítmányokat* pedig P, Q stb.-nek, akkor a szöveget egy sor S_1 [van] P, S_2 [van] Q, S_3 [van] R stb.[22] alakú kijelentéssé alakíthatjuk át.

---

20 A transzformációk kérdésével – más nézőpontból – *Gernot Wersig* is foglalkozik első kötetünkben (a szerk.).

21 A magyarban a 3. személyben hiányzik a kopula, így a „van, vannak" kifejezést zárójelben közöljük. Ezek nélkül is könnyű belátni, hogy a kijelentések logikailag ekvivalensek (a szerk.).

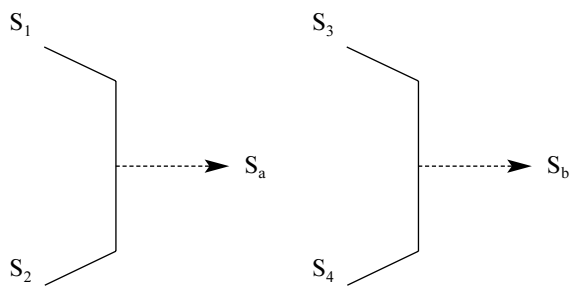
22 A kopula („van") itt kétféle viszonyt fejez ki. Az 1., 2., 4. állításokban azt, hogy a „hőálló mikroszkóp tárgyasztal" különböző fogalmak terjedelmébe tartozik, a 3. kijelentés pedig ekvivalencia relációra utal. Logikai kifejezésekkel az első típus $S_1 = P$ (ha S_1 akkor P), a második pedig $S_3 = R$-rel írható le (a szerk.).

A transzformációnak ekkor két alternatív módja van. A legtöbb információkereső rendszerben az állítmány a keresett információ, az alany pedig a kérdés, amelyre információt kérnek. Az *egyik transzformációs lehetőség* ebből következően az, hogy az állítmányokat elhagyva csak az alanyokat indexelik, például csak a „nagyhőmérsékletű metallográfiai vizsgálóberendezés" alanyt (vagy – másik megoldásként – csak az állítmányt indexeljük, elhagyva az alanyt; a döntés a mutató céljától függ, de alapvetően önkényes).

Szimbólumokkal: S_1 [van] P $\rightarrow$ S_1, S_2 [van] Q, $\rightarrow$ S_2 stb. Néhány rendszerben a teljes állítást felhasználják a tartalmi leíráshoz, de csak bizonyos kijelentéseket tárolnak. Ez a helyzet *Luhn* szerzői referátumaival. A magam részéről az első megoldást tartom tipikusnak, melyben a transzformáció eredménye S_1, S_2, S_3 stb.

A *másik transzformáció* egyes alanyok kiválasztása lehet indexelés céljaira, így mondjuk $S_1 \rightarrow S_1$, $S_2 \rightarrow 0$, $S_3 \rightarrow S_3$, $S_4 \rightarrow 0$. Ez történik minden olyan esetben, amikor a szöveg túl hosszú és az alanyok közül csak néhányat választanak ki az ismérvek szerint rendezett fájl[23] számára. Ha azonban a szöveg túl rövid, mint a mi példánkban, valamennyi alanyát meg lehet tartani. Bizonyos alanyok elvetése és mások megtartása helyett választhatjuk azt az alternatív megoldást is, hogy két alanyt újjal helyettesítünk, így a tartalmi leírásban kisebb a veszteség.

Formalizálva:



S_a és S_b az általuk helyettesített alanyokhoz a következő belefoglalási reláció (inkluzió) révén kapcsolódik: $S_a > S_1$, $S_a > S_{12}$ stb.

E helyettesítés mechanizmusa világos lesz, ha elemezzük az egyszerű alany S szerkezetét. Tipikus esetben ez az A, B, C stb. deszkriptorok soroza-

23 Az ismérvek szerint rendezett fájlon (retrieval file) a szerző hivatkozási információk (például katalógustételek, bibliográfiai rekordok stb.) minden olyan állományát érti, melyet a tartalmi ismérvek (deszkriptorok, tárgyszavak, kulcsszavak, jelzetek stb.) szerint rendeztek, mint például a szakkatalógust, az „invertált" mágnesszalagos deszkriptorfájl stb. (a szerk.).

tából áll, amelyeket az a, b, c stb. relációs kifejezések kapcsolnak össze AaBbCc… formában.[24]

Például:

(a) Az urán(nak a) szerkezetét vizsgáló berendezéssel végzett megfigyelés
A_1 α B β c C χ d D

(b) A vas(nak a) szerkezetét vizsgáló berendezéssel végzett megfigyelés
A_1 α B β c C χ d D

Az ilyen alany az AaBbCcD stb. elemek „szorzatának" tekinthető. Bármely két vagy több alany egyesíthető. Az új a következő átalakításokkal mindkettőt tartalmazni fogja:

(1) Két vagy több alany párhuzamos kifejezései összegükkel helyettesíthetők, például urán (A_1) és cirkónium (A_2) helyettesíthető „urán vagy vas" kifejezéssel (A_1+A_2). Így keletkezik az „urán(nak a) vagy a vas(nak a) szerkezetét vizsgáló berendezéssel végzett megfigyelés" (A_1+A_2) aBbCcD alany.
(2) A párhuzamos kifejezések közös generikus fölérendeltjükkel helyettesíthetők, például A_1 és A_2 helyettesíthető A-val (Fémek), így AaBbCcD lesz az alany.
(3) Egy alany bármely eleme elhagyható, ha így egyszerűbb, de az eredeti alanyt tartalmazó szorzat marad, például AaBbCcD lerövidíthető AaCcD-re (Fémek vizsgáló berendezésével végzett megfigyelés).

Az első transzformáció összevonja a deszkriptorokat, de nem jár az elemek megváltoztatásával. A második generikus kifejezésekkel helyettesíti a specifikusakat. A harmadik bizonyos kifejezéseket elvet. Az utóbbi két megoldás következtében tehát a kifejezések száma csökken.

Hasonló módon csökkenthetők a relációs kifejezések is több transzformációs lépésben. A harmadik átalakítás a B kifejezéssel együtt a b relációt is megszünteti. Ahogyan a kifejezések összevonhatók egy generikusabb kifejezésbe[25], a megengedett relációk száma is csökkenthető szabványos relációegyüttesek al-

24 Az elszigetelő, izoláló nyelvekben (mint az angol) a relációkat is önálló szavak fejezik ki. A ragozó, agglutináló nyelvekben (mint a magyar) toldalékok, képzők stb. fejezik ki a relációkat, így a jelölés bonyolultabb. A nagybetűk a deszkriptorokat, az azonos kisbetűk pedig a másik deszkriptort hozzájuk kapcsoló toldalékokat képviselik. A „relációs kifejezések" gyakorlatilag ezeknek felelnek meg (a szerk.).

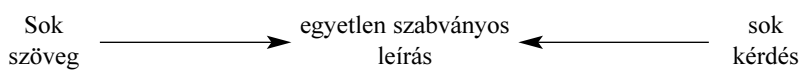
25 Pontosabban: a kifejezések által képviselt fogalmak összevonhatók a generikusabb deszkriptorral képviselt általánosabb fogalomba (a szerk.).

kalmazásával (mint amilyenek a WRU indikátorok[26]). E folyamat végigjárásával valamennyi jelölt relációs kifejezés eltörölhető: a → 0, b → 0 stb. Erre a transzformációra sok információkereső rendszerben sor kerül. A jelölt relációk eltörlése azonban nem jelenti feltétlenül azt, hogy a kifejezések között semmiféle kapcsolatot nem tüntetnek fel. Az összekapcsolás[27] különböző formái léteznek. Ezt a CdD → (DD) átalakítással szimbolizálhatjuk. Végül még az így keletkezett összekapcsolások is elhagyhatók, és maradnak a szabad, puszta deszkriptorok; például az összekapcsolt (ABC) → A, B, C lesz.

A leírt transzformációk a következőkben foglalhatók össze:

(1) Szöveg → Kijelentések S_1 [van] P, S_2 [van] Q stb.
(2) Az állítmányok elhagyása: S_1 van P → S_1 stb.
(3) Néhány alany elhagyása: $S_n \rightarrow 0$
(4) Néhány alany összevonása: $S_1 + S_2 \rightarrow S_a$ stb.
az alábbi transzformációk által ($S_1 = A_1aBbCcD$, $S_2 = A_2$ aBbCcD)
(4a) A kifejezések összegzése: $S_1+S_2 \rightarrow (A_1+A_2)$ aBbCcD
(4b) Generikus kifejezések használata: $S_1+S_2 \rightarrow$ AaBbCcD
(4c) Kifejezések elhagyása: $S_1+S_2 \rightarrow$ AaCcD
(4d) Jelzett relációk elhagyása: AaCcD → (ACD)
(4e) Összekapcsolások elhagyása: (ACD) → A, C, D

Ezek a transzformációk egytől egyig egyszerűsítések, „több" (részletes) tartalmi leírás helyettesítései „eggyel". El kell ismernünk, hogy a szöveg legpontosabb és legteljesebb leírása a szöveg maga, ezért az említett „több → egy" transzformációk sorával járó egyszerűsítések a tartalmi leírások halmazában a pontosság és a „változatosság" csökkenéséhez vezetnek. Minél messzebb megyünk ebben az irányban, annál kevésbé pontosan tükrözi leírásunk a szöveget. Hasonló módon fordíthatjuk a kérdéseket is pontatlan tartalmi leírásokká. Megállapíthatjuk, hogy minél messzebbre megyünk a transzformációk sorozatában, annál kisebb a valószínűsége, hogy a tartalmi leírás alapján kiválasztott szöveg pontosan megfeleljen a kérdezőnek, akinek kérdését – ugyancsak – leírássá alakítottuk át.



26 A WRU indikátorokat – *Perry* és *Kent* szemantikai kódjait – részletesen ismerteti első kötetünkben *de Grolier*, továbbá Vickery: Deszkriptor-nyelvek című tanulmánya (a szerk.).

27 Az összekapcsolás vagy összecsatolás (interlocking) eredetileg *Calvin Mooers* által használt kifejezése valójában a szintaktikus osztályozás legegyszerűbb formáját jelenti: az adott dokumentum egyes résztárgyköreit képviselő deszkriptorokat egyszerű kapcsolatjelölővel (angolul: link) összekötik, külön-külön csoportokat képezve ezáltal belőlük. Az összekapcsolás tehát még csupán a szintaktikai kapcsolat létezésének puszta tényét jelenti, de nem mond semmit e kapcsolat fajtájáról, tartalmáról (a szerk.).

Más szóval, az egymást követő transzformációk egyre növelik a zajt, vagyis a kérdező számára a keresett és talált irreleváns dokumentumok arányát. Másrészt viszont az ilyen transzformációk segíthetik a kérdezőt abban, hogy növelje a visszahívott releváns dokumentumok részarányát, azaz a teljességet. A kérdező néha túl aprólékosan határozza meg a keresendő témát, ilyenkor az alanyok összevonásával (pl. szinonimák egyeztetése, 4b transzformáció) elkerülhető releváns információk elvesztése. Máskor meg a kérdező nem tudja pontosan megnevezni a témáját; ebben az esetben a kevésbé pontos tartalmi leírás vezethet inkább el a releváns anyaghoz.

A szövegek „több → egy" transzformációi nemcsak a zaj forrásai, hanem a keresésre pozitív hatásuk is van. Ezt bizonyítja egy további, az információkereső rendszerekben alkalmazható transzformáció, amelyet eddig még nem említettünk, nevezetesen a „fölérendelt karakterek" bevonása az indexelésbe és a keresésbe, azaz specifikus és generikus kifejezések összekapcsolása oly módon, hogy a kereső szükség szerint összevonhatja őket. *Mooers* nyomán ezt az A → (abc…) transzformációs formulával jelölhetjük, ahol A a deszkriptor és a, b, c a deszkriptort együttesen reprezentáló generikus karakterek. Ez az alanyok összevonásának utolsó szakasza.

Az ismérv–dokumentum mátrix és felosztása

Vickery ebben a fejezetben a hivatkozási adatok állományának – így például a katalógusnak, a tárgymutatónak – olyan általánosított leírását kísérli meg, mely egyformán érvényes a kézi és automatizált rendszerekben. *Egyedi jellemzőkön* (item specifications) bibliográfiai adatelemeket, valamint olyan – nem tartalmi – jellemzőket ért mint a dokumentumtípus, nyelv, titkossági fokozat stb. *Címeken* (address) lelőhelyadatokat ért, például raktári jelzetet, oldalszámot (tárgy- és névmutatóban). *Dokumentumtételen (dokumentációs egységen)* (item) azt a teljes egységet érti, amely hivatkozási adatokból áll és a dokumentumot mind formai, mind tartalmi szempontból képviseli, így tehát a katalógustételt vagy az adatbázisok akár referátumot is tartalmazó tételeit. *Tételen,* felvételen (entry) a hivatkozási adatok valamilyen rendezett állományában – mint amilyen a katalógus, a bibliográfia vagy a mutató – adott dokumentumot jellemző adatok csoportját (például a katalógus-, a bibliográfiai vagy mutatótétel) érti. („Entry" tehát lehet „item" is, azaz a *dokumentációs egység* is *tétel.*) *Fizikai tárolási egységen/*egységnyi adathordozón (tally) azt a fizikai egységet érti, amely a tételt hordozza (például a katalóguscédulát), vagy bibliográfia, illetve mutató esetén ama lapok összességét, amelyeken a bibliográfiai, illetve mutatótételek szerepelnek. *Megjelölésen* (mark)

pedig a legáltalánosabb értelemben vett megnevezést, jelölést, szimbólumot érti, amely lehet természetes nyelven kifejezett szó vagy szöveg, vagy mesterséges nyelven kifejezett kód, jelzet.

Az ismérvek szerint rendezett fájl összeköti a tartalmi leírásokat a dokumentációs egységek (a továbbiakban: dokumentumok a ford.) egyedi jellemzőivel és címeivel. Minden dokumentumhoz ismérv kapcsolódik. A leírás lehet egyelten egy kifejezés, kifejezések önálló együttese, vagy kapcsolódó kifejezések sorozata. Az ismérveket képviselő kifejezések és a dokumentumok közötti kapcsolatok halmaza a *4. táblázaton* látható kétdimenziós mátrixszal ábrázolható[28].

		Ismérvek								
		A	B	C	D	E	F	G	H	stb.
	I	X		X	X		X			
	II		X			X	X			
	III	X	X		X			X		
Dokumentumok	IV			X			X		X	
	V	X			X	X		X		
	VI		X	X	X				X	
	stb.									

4. táblázat

Amikor egy kifejezést ismérvként rendelnek hozzá a dokumentumhoz (például G-t a III. és V. tételekhez) a mátrix megfelelő cellájába egy töltőelem (itt egy X) kerül.

A Vickery által tárgyalt rekord (dokumentum)–ismérv mátrixszal kötetünk további részében a „Könyvtári feldolgozás kézikönyvéből" vett részletben Becker és Hayes is foglalkozik.

Elvileg az ilyen mátrix ismérvek szerint rendezett fájlként használható, de a rengeteg sor és oszlop között nagyon nehézkes volna a vizuális keresés. A gépi keresésre a mátrix már alkalmasabb, de egy hiányossága így sem küszöbölhető ki, mivel a mátrix celláinak jó része betöltetlen maradna, nagy üres terület vesznék kárba, ami a mátrix használatát igencsak gazdaságtalanná tenné. A legtöbb valóságos információkereső rendszerben ezért a mátrixot átalakítják.

28 „Ismérven nemcsak a dokumentumhoz rendelt osztályozó fogalmakat képviselő kifejezéseket, hanem a közöttük fennálló szintaktikai kapcsolatokat jelölő kifejezéseket, jeleket is értem (kapcsolatmegnevezések)." (Vickerynek az előző kiadásból származó kiegészítése.)
Az ismérv–dokumentum mátrixszal foglalkozik a kötet későbbi részében *Robert M. Hayes* és *Joseph Becker* is (a szerk.).

A mátrixot a gyakorlatban vagy függőlegesen vagy vízszintesen egységekre vagy *tételekre* bontják és az egyes egységeket önálló rekordok formájában adathordozóra rögzítik. A vízszintes felosztás alapján olyan tételek keletkeznek, melyek egy-egy dokumentumra vonatkoznak és az ezzel a dokumentummal összekapcsolt ismérveket tartalmazzák. A függőleges bontás nyomán viszont olyan tételek keletkeznek, amelyek egy-egy ismérvet tartalmaznak, valamint azokat a dokumentációs egységeket, amelyekhez az adott ismérvet hozzárendelték. A mátrix felbontható a sornál vagy oszlopnál nagyobb egységekre is; bizonyos esetekben ez kifejezetten szükséges is. Végül a mátrix felbontható elemi egységekre is, amelyek egy-egy töltőelemnek felelnek meg. Ilyen módon a mátrixból ötféle tétel alakítható ki.

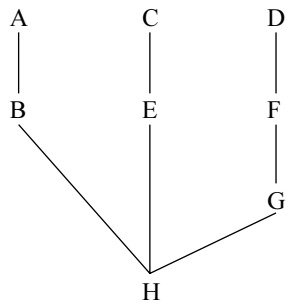
(1) ismérv-hozzáférésű tétel például A: I, III, V stb.
(2) dokumentum-hozzáférésű tétel például III: A, B, D, G stb.
(3) csoportos ismérv-hozzáférésű tétel például
A, C: I, III, IV, V, VI stb.
(4) csoportos dokumentum-hozzáférésű tétel például
III+IV: A, B, C, D, F, G, H stb.
(5) elemi tétel: A: III.

Az első típus az ismérv-hozzáférésű tétel, az „invertált” forma, melyet például az uniterm rendszerekben és más koordinált mutatókban, valamint invertált fájlokban használnak. A dokumentum-hozzáférésű tétel a cédulakatalógusok és peremlyukkártyák hagyományos formája, a hierarchikus szervezésű adatbázisokban a „bibliográfiai törzsfájlok”, a relációs adatbázisokban a bibliográfia alapleírások (rekordazonosítót, főcímet, egyéb nem ismétlődő bibliográfiai adatelemet tartalmazó rekordok) állománya.

Mindegyik tétel egységnyi adathordozóra (fizikai tárolási egységre) kerül, melyen az ismérveket és a dokumentumokat jelölések képviselik és ezek rekordokat alkotnak. Kereséskor kiválasztják azokat a fizikai tárolási egységeket (illetve a rajtuk rögzített rekordokat), melyek az adathordozón az előre meghatározott ismérveknek megfelelő jelöléseket tartalmazzák, majd leolvassák a dokumentumok egyedi jellemzőit és/vagy a lelőhelyét képviselő „címet”. A válogatás tehát az ismérveket képviselő jelölések között történik.

A 4. táblázat mátrixa a dokumentumokhoz rendelt ismérveket egymástól függetlenül tartalmazza: az A, C és F ismérvek egyaránt az I. dokumentumhoz tartoznak. A közöttük lévő esetleges összefüggések nincsenek feltüntetve, A mátrix ennél fogva a „puszta deszkriptorokkal” való indexelés modellje. A gyakorlatban természetesen két relációval is számolnunk kell: az ismérvek által képviselt fogalmak közötti generikus kapcsolatokkal általában, valamint az egy dokumentumhoz vagy dokumentumrészletekhez tartozást kifejező ismérv-összekapcsolással (interlocking).

Tételezzük fel, hogy a mátrix nyolc kifejezése a 11. ábrán látható belefoglalási – nem–faj relación, inkluzión alapuló – háló szerint kapcsolódik egymáshoz.



11. ábra. Belefoglalási háló

Az A kifejezés által képviselt fogalom tartalmazza a B kifejezés által képviselt fogalmat. Ezt A > B-vel jelöljük és ugyanígy C > E, D > F, F > G. A H kifejezés által képviselt fogalmat a B, E és G tartalmazza. Magában a mátrixban ezeket a relációkat a „keresztutalások" jelzik, így A(>B), B(A, H), C(E), D(F), E(C, H), F(D, G), G(F, H), H(B, E, G).

Az összekapcsolódást a mátrixban a kapcsolatrögzítőnek (interfixing) nevezett mutató jelöli. Például: ha a II. dokumentum teljes tartalmi leírása „Baktériumok(nak az) elpusztítása festékanyagokkal", és B = baktériumok, E = festékanyagok, F = elpusztítás, R = -nak az … a birtokos személyrag, birtokviszony és S = által, -val, -vel, eszközhatározó rag. Annak a kifejezésére, hogy a baktériumokat és nem a festékanyagokat pusztítják el, az ismérveket képviselő kifejezéseket kapcsolatrögzítőkkel kell ellátnunk: F(1), R(1, 2), B(2, 3), S(3, 4), E(4). A mátrix bejelölése helyett kapcsolatrögzítő számokat tehetünk ki. A mátrix sora ezután így fest:

	Ismérvek									
	A	B	C	D	E	F	G	H	R	S
Dokumentum	(>B)	(<A,>H)	(>E)	(>F)	(<C,>H)	(<D,>G)	(<F,>H)	(<B,<E,<G)	–	–
		2,3			4	1		1,2	3,4	

A modellek haszna

Mi a jelentőségük e modelleknek az információkereső rendszerek készítésekor? Láttuk, hogy egyik sem teljes és – főként – egyik sem tartalmaz az ismérvként használt kifejezések előfordulási gyakoriságával kapcsolatos statisztikai adatokat. Következésképpen egyiket sem tarthatjuk az ismérvek szerint rendezett fájl tökéletes modelljének.

Ugyanakkor a gyakorlati munkának mindig előnyére válik, ha pontosan értjük, hogy mit is csinálunk, márpedig ezt a leegyszerűsített, elvont modell jól megvilágítja. A *Perry–Kent* modell például rádöbbent bennünket, hogy milyen önkényes és korlátozott lehet egy-egy osztályozási fa, valamint arra is, hogy mennyivel rugalmasabb a generikus tulajdonságok összekapcsolásának felhasználásával kialakított rendszer, mint a hagyományos hierarchikus „entitásoké". A gyenge hierarchián alapuló saját fazettás osztályozási modellem a fa- és a *Perry–Kent* modellek közötti kompromisszum. *Jonker* olyan tényezőket emel ki, amelyek segítségével eldönthető, hol érdemes kompromisszumot kötni, és kiválasztható a mindenkori „generikus szint".

Fairthorne és *Mooers* általánosabb modelljei további segítséget jelentenek. Az osztályok algebrájával leírható lehetséges struktúrák elvont elemzésével olyan lehetséges jellemzők (például az összegek és különbségek) használatát javasolják, amelyek különben elsikkadtak volna. Egyben le is írják használatuk következményeit a rendszerben. E modellek a gyakran teljesen elütő alapelvűnek tartott rendszerek (például koordinált indexelés szemben a hagyományos osztályozással) közös vonásait is feltárják és megmutatják, hogy strukturálisan valóban létezik „deskriptív kontinuum". *Mooers* szerint további elemzések valószínűleg elvezethetnek a keresésben még ki nem próbált új struktúrákhoz.

Az a felfogás, mely szerint a tartalmi leírás készítése az eredeti szövegen végrehajtott transzformációk sorozata, segíthet a transzformáció különböző feltételek közötti optimális szintjének meghatározásában. A „több → egy" transzformációk az ismérv-fájlon belül a változatosságot csökkentik. Ez a változatosság elvileg mérhető *(Ashby).* Adott rendszermutatók mellett meg tudjuk mondani, hogy milyen mértékű a változatosság és ennek megfelelően hány transzformáció szükséges?

Fairthorne a másik véglet felöl közelítette meg a kérdést. A „végső" transzformációval kezdte, melyben az összes szöveget ugyanazzal a kifejezéssel írják le. „Ez az egyetlen eljárás, amellyel a kívánt dokumentum teljes biztonsággal megkereshető. Bár ez a módszer egyszerű és nem akadályozza a dokumentumok keresését (a teljesség 100%-os), mégis használhatatlan, hiszen a zaj is 100% körüli lesz. A még hasznos finomítás határának keresésekor se feledkezzünk meg arról, hogy minél pontosabban különböztetjük meg a leírásokat (azaz minél kevésbé alakítjuk át a szöveget), annál kevesebb szöveg felelhet meg egy-egy leírásnak…"

„A kereső, aki »valamiről« rögzített, vagy további információt kér, részletesen meghatározhatja igényét (keresési igény). A keresett információhoz azonban csak az irreleváns szövegek kásahegyét átrágva juthat el (hiszen – amikor újra meg újra transzformáltuk a szöveget – összevontunk alanyokat, durvábban fogalmaztuk meg a leírásokat). A legtöbb, amit információkereső rendszerrel elérhetünk, hogy ne legyen több zsákutca (zaj), mint

amennyi a kérés többértelműségének ellensúlyozásához feltétlenül szükséges. Ez pedig kiszámítható, ha a szöveg és a leíró (információkereső) nyelv sajátosságai ismertek. Ugyancsak kiszámítható ebből a releváns egységek elvesztésével és az irrelevánsak felhasználásával járó költség, az átlagos és a kivételes kérdések megválaszolásának egymáshoz viszonyított költsége, és – ami a legfontosabb – mind a báziselemek, mind a dokumentációs gyűjtemény formátuma".

Fairthorne rámutatott, hogy néhány működő információs rendszer sokkal finomabb szerkezetű és körültekintőbben rendezett információkereső nyelveket használ, mint amilyent velük kezelt gyűjtemény formátuma valaha is megkívánna. Ezekben az esetekben a transzformáció túl kis méretű, túl sok összetett témakör írható le, túl sok a specifikus jelentésű kifejezés, túl sok az explicit formában feltüntetett reláció. Másrészt vannak elnagyolt rendszerek. Ha a finomítás folyamatát transzformációk sorozataként fogjuk fel, akkor egyben azt is jobban fel tudjuk mérni, hogy mi a pontos szerepe a transzformációknak a leírás finomításában és milyen mennyiségi kihatása van a finomításnak.

Végül megvizsgáltuk, az ismérv–dokumentum mátrixot. Ez az információkereső termékeny modelljének bizonyult, amely megvilágítja a tételek különböző elrendezésén alapuló fájlszervezési eljárásokat és jobb kihasználásukra is ösztönzően hat.

Az információkeresés technikái[29]

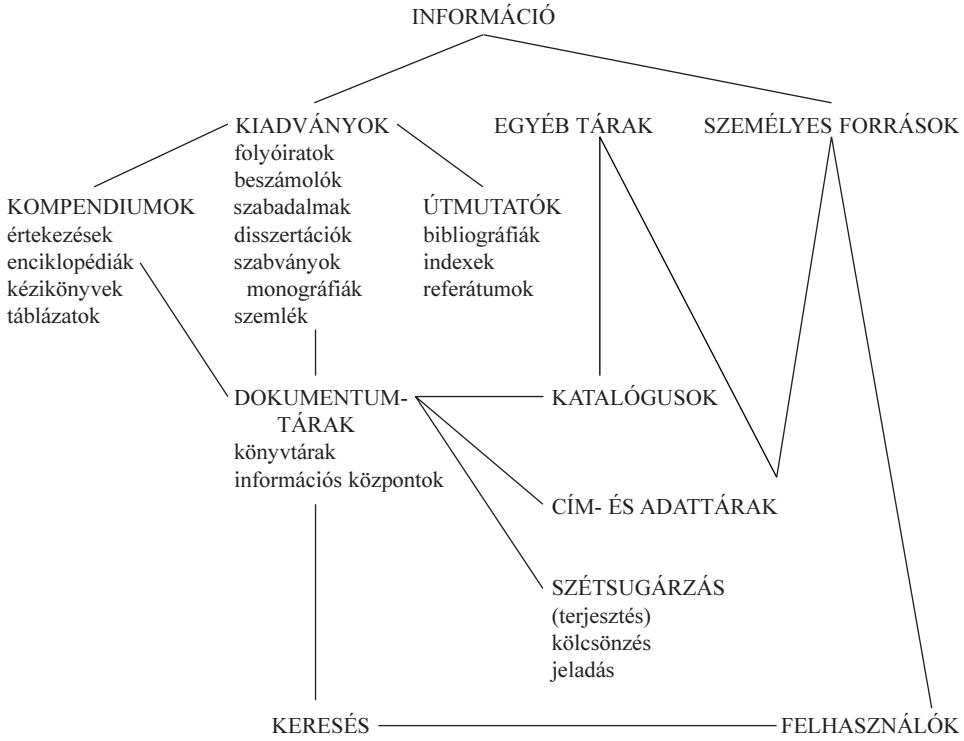
4. Az információkeresés modelljei[30]

A természettudományi és műszaki szakemberek felé áramló információ útja az 1. ábrán látható. Az információt rögzítheti kiadvány (kiadványnak tekinthető minden a nyilvánosságnak szánt vagy bizalmas dokumentum) vagy az emberi emlékezet, de akár jegyzetfüzet is. Az elsődleges kiadványokat tárakban, általában könyvtárakban gyűjtik össze. A könyvtárakban ezen kívül megtalálhatók az ún. kompendiumok és a keresést segítő különféle hagyományos útmutatók is. Egyetlen ilyen tár sem elégíthet ki minden igényt, ezért szükség van egyéb információforrások feltárására katalógusok, valamint címés adattárak segítségével. A dokumentumtárak és felhasználók között két ellentétes irányú folyamat játszódik le: a tártól a felhasználó felé irányul a terjesztés, a felhasználótól a tár felé megy a keresés.

---

29 Techniques of information retrieval / Vickery B. C. – London : Butterworth, 1970. 264 p.

30 Patterns of retrieval. In: Techniques of information retrieval, p. 33–47.



1. ábra. A tudományos információ áramlása

Általános értelemben a dokumentumtár az információk dokumentumokban rögzített gyűjteménye. Példaként a könyvtárat említettük, de ebben az értelemben dokumentumtárnak minősülnek a referáló folyóiratok, sőt a kompendiumok is. Ha keresésen azt a műveletet értjük, amelynek során valamely tárból kiválasztjuk a megfelelő tárgyakat, akkor nemcsak a könyvtári katalógus, hanem bármilyen mutató vagy bibliográfia is keresőeszköznek tekinthető. A keresés művelete minden információkeresés alkalmával megismétlődik.[31]

Vizsgáljuk meg a következő keresési modelleket:

1. Tisztáznunk kell a keresés tárgyát, meg kell határoznunk megnevezéseit, ki kell jelölnünk kiterjedését, körét. Ehhez lexikonokat, szótára-

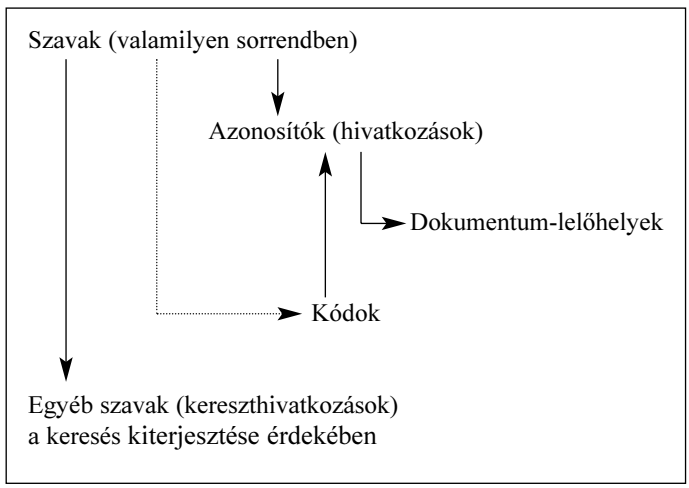
31 A nemzetközi irodalomban megkülönböztetik az egyszerű keresést, lekérdezést (search, Suche, recherche) a rátalálás és elkülönítés, a találatképezés összetett műveletétől (retrieval, Retrieval, recherche automatique), amelyben a kódolt – információkereső nyelvre fordított – kérdést az állomány ugyancsak kódolt tételeivel hasonlítják össze. Ez utóbbit magyarul – megkülönböztetésül az egyszerű kereséstől – információkeresésnek vagy visszakeresésnek nevezzük (lásd még a kötet első részének kommentárját) (a szerk.).

kat és egyéb referenszkönyveket kell átnéznünk. E könyvek szövegét vagy mutatóját valamilyen szabály szerint – betűrendben vagy szisztematikusan – rendezték el, és ezért könnyen áttekinthető.

2. Meg akarunk ismerkedni valamilyen témával és ehhez meg kell találnunk a megfelelő tankönyveket és monográfiákat. Ennek érdekében betűrendes vagy szisztematikus elrendezésű könyvjegyzékeket vagy katalógusokat nézhetünk át.

3. Hivatkozásokra (másodlagos információkra, dokumentumleírásokra) van szükségünk valamilyen tárgyról szóló folyóiratcikkekről. Ehhez referáló folyóiratokat, kurrens gyarapodási jegyzékeket és friss folyóiratszámokat kell átnéznünk. Minden esetben betűrendben, szakrendben vagy véletlenszerűen elrendezett tételeket kell átvizsgálnunk.

A felsorolt esetekben mindig bizonyos tételeket választunk ki a tárból. Ha bármelyik hagyományos információkereső rendszert vizsgáljuk, mindig a következő közös szerkezetre bukkanunk:



Az információkereső rendszerbe szavak listáján – mutatótételeken, tartalomjegyzéken stb. – keresztül léphetünk be. Ezek vagy közvetlenül vagy kódokon keresztül kapcsolódhatnak a dokumentumazonosítókhoz. A kód lehet egyszerűen oldal vagy tételszám, esetleg szakjelzet. Az azonosító tartalmazhatja a dokumentum címét, illetve lelőhelyének megjelölését vagy összekapcsolódhat ezzel (mint például a könyvtári katalógusban). A keresőszavakat valamilyen módon rendszerint összekapcsolják más szavakkal (keresztutalásokkal), így a keresés kiterjeszthető.

Az információkeresés műveletének négy fázisa különböztethető meg:

(a) ***Szókeresés:*** meghatározzuk azokat a szavakat, amelyek megfelelően írják le a keresett információt.
(b) ***Hivatkozáskeresés:*** meghatározzuk azokat a hivatkozásokat (dokumentumleírásokat), amelyek valószínűleg a keresett témára vonatkoznak.
(c) ***Dokumentumkeresés:*** megállapítjuk a szóban forgó dokumentumok lelőhelyét.
(d) ***Adatkeresés:*** a dokumentumokból kiválasztjuk a keresett adatokat.

Sok olyan információkereső rendszer létezik, amely négy fázis közül csak az egyikkel foglalkozik. A szótár például a szókeresés eszköze; a legtöbb mutató csupán a hivatkozások keresését teszi lehetővé; a raktári (helyrajzi) katalógus a dokumentumkeresést segíti; a műszaki kézikönyv adatkeresésre használható. Vannak rendszerek, amelyek két vagy több fázist egyesítenek. Legtöbbjük azonban megtorpan a dokumentumkeresésnél, és az adatkeresést, ami sokkal több szellemi és műszaki problémát vet fel, mint az összefüggések kikeresése, teljesen a felhasználóra bízza. Könyvünkben elsősorban a szavak és a hivatkozások keresésével foglalkozunk.

Az információkereső rendszereknek két funkciója van: az egyik a tájékoztatás, amely ráirányítja a felhasználók figyelmét a frissen beszerzett és feltehetőleg érdeklődési körükbe tartozó kiadványokra. A másik a retrospektív keresés, amely gondoskodik róla, hogy a felhasználó az őt érdeklő dokumentumokat a teljes tárból kikeresse. (A retrospektív keresés szintén kétféle lehet: meghatározott tények keresése, vagy valamely témakör, szakterület mindenre kiterjedő vizsgálata.) Sok rendszer úgy működik, hogy a két fő feladat közül csak az egyiket látja el, vagy pedig két egymástól független részállományokból áll, amelyek egy-egy funkciót látnak el. A könyvtári katalógus például csak retrospektív keresésre alkalmas, a heti gyarapodási jegyzék pedig csak az új szerzeményekről tájékoztat. Célszerű volna a gyakorlatban olyan rendszereket tervezni, amelyek mind a két funkció ellátására alkalmasak, hiszen mindkét esetben ugyanolyan típusú műveletek elvégzéséről van szó. E fejezetben vázlatosan áttekintjük a teljes folyamatot, a részleteket majd a további fejezetekben fejtjük ki. Áttekintésünk sok általánosítást tartalmaz, és az olvasó számára hasznos lehet, ha a részletek áttanulmányozása után visszatér ehhez az elvontabb részhez.

Az információkeresés folyamata

A dokumentumokban rögzített információk felhasználása előtt összetett műveletsorozatot kell elvégezni: dokumentumokban rögzíteni az információt; a dokumentumokat könnyen hozzáférhető módon, lelőhelyüket nyilvántartva

tárolni; megállapítani minden egyes dokumentum ismertetőjegyeit a dokumentumkép vagy -profil (a dokumentumleírások) megalkotása céljából, és a dokumentumképeket valamilyen rendezett állományban – fájlban – rögzíteni; a mindenkori felhasználónak azokkal a kifejezésekkel kell megfogalmaznia kérdését vagy érdeklődési körét, amelyeket a dokumentumokról ismertetőjegyekként rögzítettek: ezt a keresőképet össze kell hasonlítani a dokumentumképekkel, és megállapítani az összeillő dokumentumok lelőhelyét; a lelőhely alapján meg kell keresni és a felhasználó rendelkezésére kell bocsátani a dokumentumokat.

A dokumentumkép szükséges volta nyilvánvaló. Minden dokumentum többé-kevésbé komplex feljegyzés, terjedelme olykor 100 szónál kevesebb, néha viszont tízezer szó. A felhasználó a dokumentumokat információtartalmuk miatt keresi, és ezért a dokumentumokat olyan „címkével" – azaz kulcsszóval, ismérvvel – kell ellátni, amely kifejezi ezt az információt. Az egyes dokumentumokat azonban többnyire sok olyan ismertetőjegy jellemzi, amely keresőkulcsként – ismérvként[32] – szolgálhat, és ezért a dokumentum meghatározásához egész sor ismérv szükséges. Az ismérvek eme együttesét nevezzük dokumentumképnek.

Vegyünk egy példát! Az amerikai űrkutatási hivatal, a NASA bibliográfiai leírást készít a gyűjtőkörébe tartozó összes kutatási jelentésről és folyóiratcikkről. A dokumentumkép olyan ismérvekből áll, amelyek külön-külön vagy együttesen szerepelhetnek a felhasználói keresőképben.

Az ismérvek egy része a dokumentum keletkezésére vonatkozik (egyéni vagy testületi szerző készítette, a megjelenési ideje, nyelv, titkossági fokozat stb.), másik része a tartalmat írja le (például a 2. ábra TERM mezőnévvel jelölt oszlopában az ELECTRON = elektron, FLUX = áramlás, IONOSPHERE = ionoszféra, MAGNETIC = mágneses, RADIO = rádió és SATELLITE = műbolygó kifejezések).

A dokumentumok keletkezésére vonatkozó ismérvek viszonylag egyértelműek. Rendszerint könnyen megállapíthatók a dokumentum címoldala alapján, és maguktól szabványosodnak: a szerzők mindig egyféleképpen írják a nevüket, akármelyik kiadványukról is van szó, és a felhasználók többnyire tisztában vannak a helyes írásmóddal. A tartalmi ismérvek sokkal kevésbé egyértelműek. Még egy rövid dokumentumban is sok olyan téma fordulhat elő, amely a dokumentum jellemzésére szóba jöhet, és ez problémákat okoz a kiválasztásban. Ráadásul ugyanazt a témát a szerzők – akárcsak a felhasználók – különféleképpen fogalmazhatják meg, sőt ugyanaz a szerző sem mindig egyféleképpen fogalmaz; ezért a tárolási és keresési folyamatban néhány dolgot szabványosítani kell.

---

32 Az automatizált információkereséssel összefüggő gépi fájlkezelés szempontjából az ilyen fajta azonosítókat egyszerűen csak kulcsnak nevezik, a dokumentációs gyakorlatban – osztályozási és katalogizálási szempontból – pedig inkább ismérvnek.

Az egész folyamat hét lépésben elemezhető: (1) annak eldöntése, hogy a dokumentum azonosítására alkalmas ismérvfajták közül melyiket használják, (2) megfelelő ismérvek kiválasztása a dokumentumból vagy a dokumentum jelölése előre meghatározott ismérvekkel, (3) a kiválasztott vagy megállapított ismérvek szabványosítása vagy/és dekódolása, (4) rekord előállítása a dokumentumkép (dokumentumleírás) és a dokumentum azonosítójának összekapcsolása révén, (5) az így készült tételek nyilvántartásba vétele (fájl-szervezés), (6) a használó kérdésének vagy érdeklődési körének megfelelő, szabványosított és szükség esetén kódolt keresőkép elkészítése, (7) a keresőkép és a dokumentumképek összehasonlítása és az egyezés meghatározott mértéke esetén a megfelelő hivatkozásai vagy lelőhelyadatok megszerzése. A következőkben rövid észrevételeket fűzök az egyes lépésekhez, megjegyzéseimet a dokumentumok tartalmára vonatkozó bonyolultabb ismérvtípusokra korlátozom.

(1) ***A felhasználandó ismérvek fajtái*** attól függnek, hogy az információkereső rendszert milyen célok kiszolgálására tervezték. Ha a felhasználók érdeklődése nagyon speciális – mondjuk különleges vegyi anyagokra kíváncsiak – akkor a kiválasztott ismérvek is kizárólag ilyen típusúak lehetnek. Másrészt a felhasználók potenciálisan érdeklődhetnek bármely, a dokumentumokban előforduló téma iránt, ezért célszerűtlen a kiválasztható tartalmi ismérvek típusait korlátozni. Ha a használók csak olyan dokumentumokat igényelnek, amelyek átfogóan foglalkoznak valamilyen témával, ismérvként csak a dokumentum központi, fő témáját szabad kiválasztani. Ugyanakkor, ha a tárgykörrel kapcsolatos valamennyi részinformációra szükség van, akkor a kisebb terjedelmű tárgyköröket képviselő ismérveket is kiválasztják. Sok ilyenféle megfontolás befolyásolja az ismérvfajták kiválasztását.

(2) ***A tartalmi ismérveknek a dokumentumból való kiválasztását*** tartalmi feltárásnak, tartalmi elemzésnek, illetve indexelésnek/osztályozásnak nevezik. A keresőrendszerek egy részében az indexelés még mindig emberi tevékenység. Másik részükben számítógépes módszereket alkalmaznak (a géppel olvasható szövegeket összevetik az előzetesen kiválasztott kulcsszavak jegyzékével, vagy megszámolják a szövegben előforduló szavakat, és kiválasztják közülük a leggyakrabban előfordulókat). A gyakorlati működés során indexeléskor mindig célszerű intellektuálisan is átnézni a dokumentumokat ahhoz, hogy megállapítsák, miről szól és milyen témákkal foglalkozik. Néhány rendszerben ezt a feladatot a szerző végzi el, aki a tárgyat címszerűen megfogalmazza, és később ezt használják fel dokumentumképként. A legtöbb rendszerben azonban gyakorlott indexelőket al-

kalmaznak erre a feladatra. Munkájuk során az (1) lépésben meghozott döntések szolgálnak irányelvként arra, hogy milyen jellegű ismérveket válasszanak. Segítségükre lehet a dokumentumok meghatározására alkalmazható szabványosított kifejezések jegyzéke is. Ebben az esetben az indexelés megköveteli, hogy a dokumentumból kiválasztott témákat értelmileg összevessék a témáknak a szabványosított kifejezések szótárában található elnevezéseivel.

20–N64–2788		*(Kiadvány beszerzési szám)*
TEMP	00194885	*(Átmeneti azonosító)*
BASE	DC–1, TC–1, DG– , SC–12, NS–1, NF–1, CF–2, CS–1, PA–2, FO–2 RT–1, AN– , RE–300764, RD–100764 AB–2, AL–1, DL–01, RE–1, CO–2, MI–1, DT–00, MP–1, MS– , DC–1, HA–1, ET– , LA– , SO–02, FO– , NP–004 day mo yr	
CATL	CD–110984, RD–	*(Katalogizálási dátum)*
CORP	201 19400 Pennsylvania State U., University Park.	(*Testületi név/forrás)*
CPSP	Ionosphere Research Lab.	*(Testületi név/forrás kiegészítő)*
TITL	A study of the ionosphere at mid latitudes. Based on total elekctron content scientific report, July 1961 – June 1962	
AUTH	Hibberd, F. H.	*(Szerző)*
PAAF		*(Szerző testülete/hovatartozása)*
IMPR	10. July 1964 44 P refs	*(Megjelenési adatok)*
CNTR	NSG–114–610	*(Szerződésszám)*
REPT	NASA–CR–56935 SR–213 N–	*(Tanulmámy/„report" szám)*
HIST		*(Keletkezésre vonatkozó megjegyzés)*
NOC	Total electron content of ionosphere derived from satellite doppler measurements	
TERM	1 Content	*(1 = nem publikált kifejezés)*
	1 Disturbance	*(3 = publikált kifejezés)*
	3 Doppler effect	
	3 Electron	
	1 Flux	
	3 Ionosphere	
	1 Magnetic	
	1 Measurement	
	1 Radio	
	1 Satellite	
	3 Satellite measurement	
	1 Solar	

Alapkódok:

DC	dokumentumosztályozás	RT	beszerzés módja	MI	mikrofilm
TC	címosztályozás	AN	leltári szám	DT	dokumentumtípus
DG	titkossági osztály	RE	beérkezés napja	MP	mikrofilmkód, előtag
SC	témakategória	RD	leírás dátuma	MS	mikrofilmkód, utótag
NS	NASA támogatás	AB	referátum	DC	dokumentum osztály
CF	konferencia	AL	referátum nyelve	HA	tárgyszó
CS	testület résztestülete	DL	dokumentum nyelve	ET	et al.
PA	szerző testülete/hovatartozása	RE	sokszorosíthatóság	LA	utolsó analitikus feldolgozás
PO	külföldi	CO	copyright	SO	forrás

2. ábra. A NASA információs rendszerének bibliográfiai rekordja (dokumentumleírása)

(3) Az ***indexeléskor használt szabványosított szókincs*** pozitív szerepe a keresésben örök vitatéma. A mellettük felhozott általános érv meggyőzőnek tűnik. Mivel mind a szerzők, mind a felhasználók következetlenül használják a tárgyat megjelölő szavakat, valószínű, hogy a kereső- és a dokumentumképeket eredményesebben lehetne összehasonlítani, ha azok mindkét esetben szabványosak lennének. Sok rendszer alkalmaz szabványosított szókincset – tezauruszokat, tárgyszójegyzékeket, és/vagy szisztematikus osztályozási rendszereket –, ugyanakkor sok számítógépes rendszer a szövegben előforduló szavak bármelyikét kiválaszthatja ismérvként, és az eredmények (találatok) az ilyen keresésekkel kiegészülnek. A szabványos szókincs szerkezetének kialakításához igen jelentékeny szellemi erőfeszítésre van szükség, hogy a szövegekben előforduló sokféle és változatos szavakat szabványosított kifejezésekből álló, a lehető legrövidebb jegyzékbe tömörítsék.
A következő lépés az ismérvek átalakítása kódokká, jelzetekké. Sok rendszer nem tartalmazza ezt a lépést, hanem a természetes nyelv szavait használja, mivel az alkalmazott dokumentációs vagy információkereső nyelv nem mesterséges alapú (mint például az ETO). Kódok használatával csökkenthető az ismérvek rögzítésének helyigénye, fenntartható a kívánt nyilvántartási sorrend és kifejezhető a közöttük lévő viszony, kapcsolat.

(4) Az indexelést követően ***elkészíthető a bibliográfiai rekord***[33]. Ez lehet viszonylag rövid: a dokumentum lelőhelyének egyszerű jelzete vagy azonosítója, amelyhez kisszámú kódolt ismérv kapcsolódik. Más rendszerekben a rekordok igen részletesek, ha később komplex módon kell kezelni őket. A 2. ábrán az a rekord látható amelyet a NASA rendszereiben készítenek a dokumentumokról: sok jellemzőt rögzítenek, egy részük keresései ismérv (kulcs) (a formai leíráskor keletkező bibliográfiai elemeket egységesített besorolási adatoknak, a tartalmi leírás során keletkezők egy részét tárgyszónak, deszkriptornak, osztályozási jelzetnek, más részűket – mint pl. a dokumentumtípus, nyelv, ország – dokumentációs/információs adatoknak nevezik), más részüknek egyéb kezelési célja van, harmadik részük a felhasználóhoz jut el tájékoztatásként (leíró adatelemek).

33 A rekord kifejezés akkor használatos, ha géppel olvasható adathordozón szereplő tételről van szó. Egyéb, általánosabb esetben az angolban szereplő „entry” (= fel-/lejegyzés) kifejezésnek inkább megfelelő leírás, dokumentumleírás kifejezés használatos magyar nyelven (a szerk.).

(5) Akár egyszerű, akár összetett a bibliográfiai rekord, a keresés szempontjából két jellegzetessége van: az egyik a ***lelőhely jelzete,*** illetve a ***dokumentum azonosítója,*** a másik a keresési ismérvek. Ha megvizsgáljuk a rekordok gyűjteményét a következő – mátrixban ábrázolható – szerkezet tárul elénk:

Dokumentumok vagy tételek	Ismérvek vagy kifejezések						
	A	B	C	D	E	F	stb
1	x			x			
2		x	xy				
3		x			x		
4			x			x	
5	x				x		
6		x				x	

A nyilvántartásban (a fájlban) ez a mátrix vagy horizontálisan vagy vertikálisan helyezkedhet el.[34] Az előbbi esetben minden egyes rekord vagy dokumentumtétel jelzetéből és a hozzátartozó összes ismérv vagy megnevezés szimbólumából áll. Az utóbbi esetben minden egyes rekord egy ismérv vagy kifejezés kódjeléből és mindazon dokumentumok vagy tételek jelzetéből áll, amelyekre ez az ismérv vonatkozik. Ez a kétféle szervezési mód kézi és gépi működésű rendszereknél egyaránt használatos.

(6) A ***keresőképek*** kétfélék lehetnek: az egyik a rendszernek meghatározott időben feltett egyéni kérést, a másik tágabb és tartósabb használói érdeklődést fejez ki, amelyet rendszeresen össze kell vetni a rendszerbe beérkező új dokumentumképekkel.
Ha a dokumentumképeket szabványosított terminológiával készítik, akkor a keresőképeket is így kell megszerkeszteni. Ha a dokumentumképek ismérveit a szövegből szabadon emelik ki, akkor a felhasználónak gondolnia kell a szöveg változékonyságára: hogy biztos legyen a keresés (az összehasonlítás) sikerében, a keresőképbe be kell építenie az összes olyan szót, szóvariánst, amelyet a szerzők feltételezése szerint használhattak a keresett tárgykör kifejezésére.

---

34 A kétfajta – ismérv-hozzáférésű és dokumentum-hozzáférésű – mátrixszal részletesen foglalkozik Vickery a kötetünkben szereplő „információkereső nyelvek szerkezetének modelljei" című szemelvény „Az ismérv–dokumentum mátrix és felosztása" című fejezetében (a szerk.).

Még akkor is, ha valamilyen szabványosított terminológiát használnak, javíthatja a keresés eredményét, ha a használó keresőképét kibővíti olyan ismérvekkel, kifejezésekkel, amelyek határozott kapcsolatban állnak a keresőképben használt kifejezésekkel. Ezek kiválasztásának megkönnyítésére a szabványosított kifejezések jegyzékei, szótárai rendszerint széles körű – „lásd még" – keresztutalásokat, deszkriptorok közötti összefüggést (fölé- és alárendelt, egész és rész, egyéb rokonsági relációt) tüntetnek föl, amelyek megmutatják a tartalmi ismérvek asszociációs körét (a szemantikai összefüggéseket).

(7) A keresőkép ismérvekből áll, amelyeket ***össze kell hasonlítani*** a dokumentumképekkel. Az összehasonlítás legegyszerűbb formája, ha azt igényeljük, hogy csak azoknak a dokumentumoknak a lelőhelyét közöljék a felhasználóval, amelynek a dokumentumképében a keresőkép valamennyi ismérve jelen van. Az összevetés elfogadhatóságának feltétele különbözőképpen finomítható:
(a) Igényeljük azt, hogy a keresőkép ismérveinek csak bizonyos százaléka legyen jelen. Ennek egyik változata az, amikor az összeillő ismérvek és a dokumentumképben lévő ismérvek számának aránya el kell, hogy érjen egy bizonyos minimális értéket. (b) A keresőkép ismérveinek különböző fontosságot tulajdoníthatunk (értéküket súlyozzuk), és azt igényeljük, hogy az összeillő ismérvek értékének súlyozott összege érjen el egy bizonyos minimális értéket. (c) Logikai összeadásokat végezhetünk a keresőkép ismérveivel: igényelhetjük, hogy a dokumentumkép tartalmazza az A és (B vagy C vagy D) és az (E vagy F) ismérveket, de a G ismérvet ne (Boole-operátorokkal végzett keresés).
A két profil összevetése vizuálisan vagy gépi úton egyaránt elvégezhető. A felhasználóval közlik az összeillő dokumentumképeket: megkaphatja a dokumentumok jelzetét vagy azok teljes rekordját (leírását) (2. ábra), vagy magát a dokumentumot. Ezután kezdődhet a használó valódi munkája: a dokumentum információtartalmának a „kitermelése". Ez a tevékenység idő- és energiaigényes, és ebből adódik a tulajdonképpeni – faktografikus, vagy tény- – adatokat kereső rendszerek iránti igény a speciális szakterületeken.

A keresés folyamatának hét lépését körvonalaztam és megemlítettem néhány ezekkel kapcsolatos problémát: így a potenciális felhasználói igény megállapításának fontosságát annak eldöntésében, hogy milyen típusú ismérveket alkalmazzanak a rendszerben; a szövegek intellektuális elemzésének szükségességét annak meghatározására, hogy milyen ismérvekkel jellemezzenek egy-egy dokumentumot; azt, hogy miként segíti az indexelőt és a keresőt

a szabványosított szókincs használata az ismérvek és összefüggéseik megállapításában; a szabványosított szókincs kidolgozásához szükséges szellemi erőfeszítés nagyságát és bonyolultságát; a kódolási módszereket; az olyan részletes, sok ismérvet tartalmazó bibliográfiai rekord szükségességét, amelynek révén a keresés elvégezhető; a fájlszervezés különböző módjait, amelyek nagymértékben befolyásolhatják a rendszer gazdaságosságát; a keresőprofilok összeállításához szükséges képzelőerőt, a keresési stratégia különböző formáit.

Mindezek a problémák és lehetőségek a fájl fizikai formájától függetlenül felmerülnek, akár sokat forgatott jegyzetfüzetről, akár on-line számítógépes adatbázisról van szó. A folyamat valamennyi szempontját – beleértve azokat is, amelyeket nem érintettünk – át kell gondolni a keresőrendszer tervezésekor.

A keresés nyersanyaga

A kérdések megválaszolása céljából végzett keresés során át kell vizsgálni az információt tartalmazó dokumentumokból készített rekordok tárát. Nézzük meg most közelebbről magát az információtárat (az állományt) – a keresés nyersanyagát. Ennek az anyagnak a természetéből következnek azok a speciális problémák, amelyeket az információkeresés felvet, és amelyek különösen vonzó tudományterületté teszik.

Az információ és a kérdések főként szövegek formájában jelennek meg, és a keresés művelete a kérdés szövegének az információt tartalmazó szövegekkel való összevetéséből áll. Vizsgáljuk meg ezek természetét. Az információk szöveges hordozói, amelyekkel leggyakrabban találkozunk, elsődleges dokumentumok.

Hogyan kell a kérdést és a dokumentumokat összevetni az információkeresés során? A legkevesebb szellemi erőfeszítést és a szövegek legkisebb módosítását igénylő módszer az, ha mind a kérdéseket, mind a dokumentumokat a szerző, illetve a használó által megfogalmazott formában használják fel, és teljes terjedelmükben vetik őket össze egymással. Ez a módszer komoly technikai problémákat vet fel, az átvizsgálandó anyag hatalmas mennyisége miatt.

Sőt, még ha a végignézés megoldható is lenne – mégis gazdaságosan –, akkor sincs biztosíték rá, hogy a teljes szöveg átvizsgálása sikeres keresést eredményez. A gyakorlatban a legtöbb működő és tervbe vett információkereső rendszer nem ezen a módszeren alapszik. A dokumentumok ugyanis olyan szavak alapján is relevánsak lehetnek tartalmilag, melyek maguk a dokumentumban nem szerepelnek. A rekordok terjedelmének csökkentése érdekében a szövegeket dokumentumképekkel helyettesítik, lényegesen rövidebb szövegek formájában, amelyek a dokumentumok helyét foglalják el, azokat képviselik, reprezentálják, leírják.

A keresés eredményének javítása érdekében a dokumentumok és a kérdések reprezentánsai, helyettesítői nem lehetnek a szövegből egyszerűen kiemelt részletek, hanem csak módosított kivonatok. Vizsgáljuk meg, milyen módszereket alkalmaznak az ilyen reprezentánsok, leírások megalkotásához és eközben milyen problémák vetődnek fel.

A dokumentumképek kialakítása

A dokumentumképek vagy reprezentánsok háromféleképpen hozhatók létre: az előbbiekben felsorolt jellemző sajátságok (például a cím, vagy más adatelem) közül egyet vagy többet közvetlenül kiválasztunk a dokumentumból; szelektív kiemeléssel (például a szöveg egyes szavainak kiválasztásával); bizonyos ismérvek (például szabványos deszkriptorok) hozzárendelésével.

A ***közvetlen kiválasztást*** sokáig szerzői mutatók, címmutatók valamint források és szerzők adatait tartalmazó mutatók készítésére használták. Újabb keletű a hivatkozási indexek összeállítása.

A közvetlen kiválasztás kevés intellektuális problémát támaszt. Ha a dokumentumban megjelölték azt a szöveget, amelyet ki kell belőle emelni, a további eljárás már pusztán adminisztratív feladat, ezért, ha a szöveg géppel olvasható, a munka adatfeldolgozó berendezéssel elvégezhető. Az eredmény: a dokumentum reprezentánsa, s sokkal rövidebb, mint az eredeti dokumentum.

Bár a közvetlen szövegkiválasztás csökkenti az átvizsgálás problémáját, nem vezet szükségszerűen eredményes kereséshez. Könyvtárosok hosszú tapasztalata bizonyítja, hogy a dokumentumokról a szerző neve alatt besorolt rekordokat módosítani kell, szabványosítani a katalogizálási szabályok segítségével annak érdekében, hogy növeljük a valószínűségét annak, hogy a dokumentumon és a kérdésben azonos módon szerepeljen a szerző. Az újabb tapasztalatok azt mutatják, hogy hasonló szabványosításra a hivatkozásoknál is szükség van.

Most nézzük meg azt az eljárást, amelyet ***szelektív kiemelésnek*** neveztem. Ez többnyire a címből, képaláírásból, fejezetcímekből vagy a főszövegből vett tárgyszavakat, szószerkezeteket (összefoglalóan kifejezéseket) jelent. Ezeknek a kifejezéseknek csak egy részét emelik ki abból a célból, hogy megalkossák a dokumentumképeket. A hagyományos intellektuális (emberi) indexeléskor a válogatás azon a szubjektív mérlegelésen alapul, hogy melyek a szignifikáns kifejezések. Az automatikus indexeléskor a szavak előfordulásának gyakoriságával kapcsolatos statisztikai kritériumokat alkalmaznak. Azonnal felvetődik a kérdés: milyen kifejezéseket válasszunk ki és hányat?

Az intellektuális indexeléskor a választás gyakran szoros kapcsolatban van az előre látható használói igényekkel, amelyek meghatározzák, hogy milyen fajta kifejezéseket ítéljünk szignifikánsnak, és milyen részletező legyen

az indexelés. Rengeteg bizonyíték szól amellett, hogy az indexelést végző embereknél hiányzik a következetesség, de nincs bizonyíték arra, hogy az algoritmikusan „következetes" gépi indexelés jobb.

Nem kételkedhetünk abban, hogy az indexelő kifejezések szelektív automatizált kiemelése gazdaságos. De még mindig nem tudjuk, hogy a statisztikai szelekció elég hatékony-e és elég gazdaságos-e ahhoz, hogy feleslegessé tegye az emberi indexelést. Hogyan értékelhető az indexelés bármely formája – ez még vitatott. Ha nem tudjuk biztosan megmondani, hogy az intellektuális indexeléskor meghatározott szituációban mi az indexelésre szolgáló kifejezések intellektuális kiválasztásának optimuma, akkor nem tudjuk értékelni a gépi kiválasztást sem. Eddig még igen kevés kísérletet végeztek ennek értékelésre.

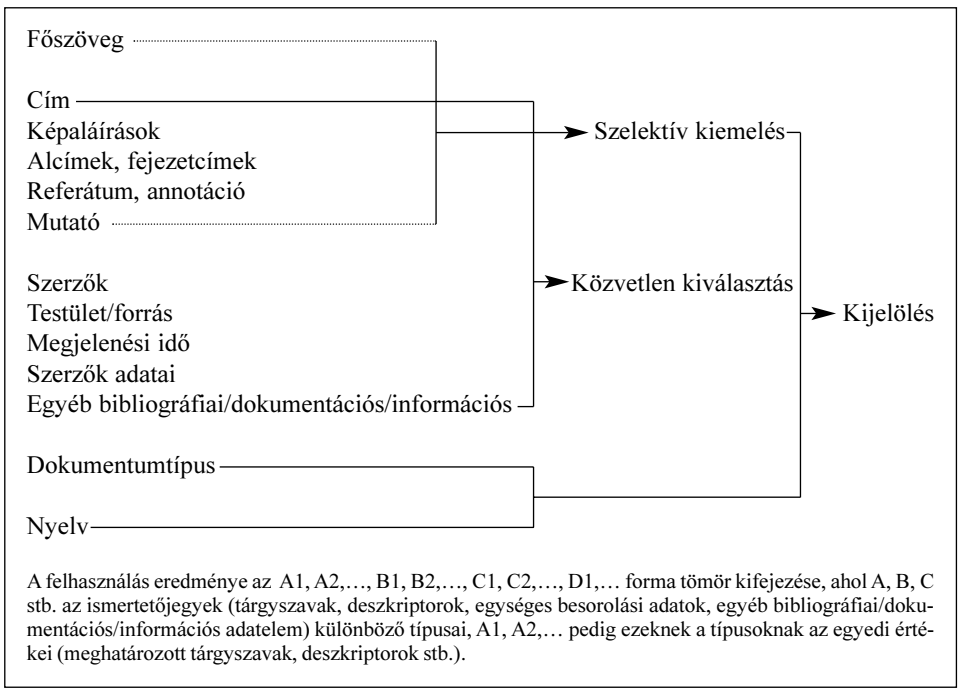
Most rátérünk a dokumentumokat reprezentáló harmadik eljárásra: a már meglévő, előzetesen összegyűjtött és ***szótárba foglalt ismérvek alkalmazására.*** Ezek az ismérvek lehetnek a dokumentum formai sajátosságai (például a dokumentum típusa: folyóiratcikk, kutatási jelentés; a dokumentum nyelve stb.). A forma rendszerint a dokumentumnak olyan implicit jellegzetessége, amelyet elvileg csak értelmes elemző képes megállapítani, de kifejleszthetők mechanikus felismerésének módszerei is (például a latin betűs írást használó nyelvek azonosíthatók bizonyos betűk jelentése alapján, de egyes rövid – kevés betűből álló – szavak előfordulási gyakorisága alapján is).

Sokkal gyakrabban használt módszer a tartalomra vonatkozó ismérvek – szabványos szavak, kifejezések vagy kódok (együttesen deszkriptorok) – hozzárendelése, amelyek reprezentálják a dokumentumban szereplő kifejezéseket. Ehhez az eljáráshoz szintén hozzátartozik a kifejezések szelektív kiemelése, mint az előző módszernél, de azt követnie kell a kifejezések alapján a deszkriptorok megállapításának. Sok működő keresőrendszerben ezt az átalakítási eljárást az indexelő veleszületett érzékére bízzák. Valamilyen módon össze kell hasonlítaniuk az általuk kiválasztott kifejezéseket az engedélyezett deszkriptorok jegyzékével, a szabványos szókinccsel.

Az ilyen terminológia összeállításához nélkülözhetetlen a szerkezeti összefüggések megállapítása abból a célból, hogy – mondjuk – több szinonimát kapcsolatba hozzanak azzal a deszkriptorral, amelyek reprezentálják őket. Ha ezeket az összefüggéseket rögzítették, az átalakítás már mechanikusan, tehát automatikusan elvégezhető. Igen sok bizonyíték van rá, hogy a kifejezések „átalakítása" deszkriptorokká – különösen a szinonimák összegyűjtésének segítségével – megjavítja a rendszer teljesítményét. Ezért nagy tere van annak a tartalmi feltárásnak, amely tezauruszt használ (abban az eredeti értelemben, amelyben *Roget* és *Luhn* használta a tezaurusz fogalmát).

A 3. ábrán összefoglaltuk a dokumentumkép kialakításának azokat a módszereit, amelyeket körvonalaztam. Az egyszerű kiválasztás alkalmazható a dokumentum bármely ismertetőjegyére a címtől a többi hivatkozási (bibliográfiai)

adatokig. A szelektív kiemelés főként a főszövegekre alkalmazható, de egyéb jellemzőkre is kiterjeszthető. Az előre meghatározott ismérvek (deszkriptorok) hozzárendelése alkalmazható a dokumentumok formális ismérvei esetében és a szelektív kiemelés produktumainál. *Mind a három eljárás alkalmazható párhuzamosan ugyanannál a dokumentumnál.* Az eredmény a dokumentum reprezentációja, leírása jellemző sajátságainak tömörítése segítségével. Ez a sűrítés az információ egy részének elkerülhetetlen elvesztéséhez vezet, és csupán a dokumentum tökéletlen, részleges helyettesítője, reprezentánsa marad, de ez is bőven elég a megkívánt feladat elvégzésére, arra, hogy az érdeklődőt releváns dokumentumhoz juttassa.



3. ábra. A dokumentumok jellemzőinek felhasználása ismérvként

A kérdések összehasonlítása a dokumentumokkal

Annak érdekében, hogy a kérdést össze lehessen hasonlítani a dokumentumképek meghatározott sorával, a kérdést tartalmilag ugyanolyan módon fel kell tárni és ugyanazzal a módszerrel kell megszerkeszteni. Sőt, ha ez megtörtént, még akkor sem lehetünk biztosak abban, hogy az egyszerű összehasonlítás optimális keresést eredményez. Mivel a dokumentumok reprezentánsai

nem tökéletesek, gyakran ki kell terjesztenünk a keresést a kérdésben megfogalmazott hiányos közlésen túl is. Ugyanakkor igen gazdaságtalan lenne a kérdést a fájl minden dokumentumának reprezentánsával összehasonlítani, előnyösebb leszűkíteni a keresést és a dokumentumképek sorát megfelelő részállományokra osztani.

Legegyszerűbb módja ennek, ha a fájlt egész sor részállományra bontják, annyira, hogy a fájlban előforduló minden egyes kulcsszónak saját részállománya legyen. A fájl A1 részállománya azokat a dokumentumokat sorolja fel, amelyeket az adott ismérv/kulcsszó reprezentál és így tovább minden kulcsszó esetében. Az eredmény invertált vagy ismérvek szerint rendezett fájl. Ha a kérdés az A1, B3, C9 és D27 kulcsszavak együttes előfordulását keresi a dokumentumképben, akkor elegendő a fájlnak ezt a négy részállományát átvizsgálni, összehasonlítani. A kérdés kibővíthető, ha egy vagy több kulcsszót kihagynak a keresés során. Nagyobb részállományok hozhatók létre úgy, hogy azokból a dokumentumképekből, amelyekben egy vagy több kulcsszó közös, részegységeket alakítanak ki. Ez a munka intellektuálisan elvégezhető, például a dokumentumkép egyik kulcsszava lehet deszkriptornak kiválasztott tág jelentésű kifejezés, és az olyan dokumentumkép, amely bizonyos tág jelentésű deszkriptort tartalmaz, a fájl részállományát fogja alkotni. A dokumentumok mindegyik reprezentánsa tartalmazhat számos válogatás alapján kiemelt kifejezést. A keresés, amely a T1, T9 és T47 kifejezések együttes előfordulására irányul, leszűkíthető arra a meghatározott részállományra, amelyet a tág jelentésű D33 deszkriptor alkot.

Ilyen részállományok létrehozhatók a kiemelt kulcsszavak statisztikai vizsgálatával és asszociációs (a kapcsolatok szorosságának mérésére szolgáló) módszerek felhasználásával. A kulcsszavak két típusa használható erre: a hivatkozási (bibliográfiai) adatok és a tárgyszavak vagy deszkriptorok. A hivatkozási adatoknak, tárgyszavaknak vagy deszkriptoroknak a dokumentumképeken belüli együttes előfordulására alapozva, rokon dokumentumképek csoportjai (klaszterek, clumpok)[35] hozhatók létre különféle matematikai eljárások segítségével. E csoportok mindegyike külön részállományt alkothat. A beérkező kérdést azokhoz a részállományokhoz irányítják, amelyekkel a kérdés ismérvei a leginkább összeillenek.

Vannak más módszerek is a kérdés kiterjesztésére. Ha a kérdésben előfordul az A1 kulcsszó, a keresés kiterjeszthető az A2, A3, A4 stb. (együttesen A1) rokon kulcsszavak csoportjára. Hogy ezt megtehessük, meg kell állapítani a tezauruszból a kulcsszavak közötti relációkat. Ezek a relációk beépíthetők az ismérveket helyettesítő szimbólumokba. Például, ha a kérdés a BENZOFENON

35 Vö. *Gerard Salton* és *Karen Spark Jones* szemelvényeivel kötetünk további részében, valamint az első kötetben a „A szakirodalmi tájékoztatás alapjai. Bevezető kézikönyv a gyakorlati dokumentalisztikába" c. kézikönyv automatikus indexeléssel foglalkozó fejezetével (a szerk.).

szakkifejezés, különböző módon csonkolhatjuk és kereshetünk minden olyan kifejezés szerint, amelyben szerepel a BENZO- vagy a FENO- vagy az -ON. Ugyanúgy, ha például a 632.954 szakjelzetet (kódot) használjuk, csonkolhatjuk ezt, és átfogóbb osztályban végezhetjük a keresést. Hasonlóképpen járhatunk el a szerzők nevével (kezdőbetűk elhagyása), a hivatkozási (bibliográfiai) adatokkal (kötet- és oldalszámok elhagyása, címek rövidítése) és a dátumokkal is. Ezeken a szimbolikus relációkon kívül kapcsolatot teremthetünk szemantikailag összetartozó kulcsszavak között is. Az ilyen összefüggések bemutatására szolgálhat az információkereső tezaurusz.

A keresőkérdés kiterjesztésének sokkal bonyolultabb formája az asszociáció. Először megkeresik a dokumentumképeket, amelyek megfelelnek a kérdésben lévő kifejezéseknek (mondjuk T24, T91, T214). Aztán megvizsgálnak más, ezekkel kapcsolatban álló kifejezéseket, amelyek az egyes összeillő dokumentumképekben találhatók, és azokból, amelyek gyakran előfordulnak, második kérdés jön létre (mondjuk T37, T52, T104), amelyet ismét összehasonlítanak a dokumentumképekkel. Az eljárás megismételhető. Nem szükséges, hogy a vizsgált ismérvek mind egyfélék legyenek – figyelembe vehetők a szerzők azonosító adatai, a hivatkozások stb., valamint a szakkifejezések vagy deszkriptorok.

Ez az eljárás azonos az intellektuális kereséskor és a géppel olvasható fájlok párbeszédes üzemmódú lekérdezésekor alkalmazott eljárással.

A 4. ábrán összefoglaltam azokat a módszereket, amelyeket említettem a dokumentumképek és a keresőképek összehasonlításáról szólva. Valamennyit használják mind a működő, mind a kísérleti rendszerek.

Valamennyi eljárás elvégezhető automatikusan, de megismételjük: nincs bizonyíték az ilyen keresés tökéletességére és arra, hogy helyettesíteni tudná az intellektuális feltárást. Azt tapasztalták, hogy a keresés nem korlátozható kizárólag az automatizált eljárásokra, és az átvizsgálást intellektuálisan is el kell végezni az optimális eredmény érdekében.

A keresés nyersanyagának elemzését a teljes szöveg átvizsgálásának nehézségeivel kezdtem. Nézzük, mivel lehet ezt pótolni: a dokumentum reprezentánsa egyetlen szóra rövidíthető le, például: „Biblia”, de ennek kevés a haszna az információkeresésben. A tendencia az, hogy az ilyen reprezentánsok hosszabb és szilárdabb szerkezetűek legyenek. A 2. ábrán bemutattunk egy dokumentumképet, amelyet a NASA (az Egyesült Államok űrhajózási hivatalának) információs rendszerében használnak. Ez a dokumentumkép olyan ismérveket, kulcsszavakat tartalmaz, amelyeket közvetlen választással (például cím), szelektív kiemeléssel (például kifejezések) és az ábra alján felsorol, hozzárendelhető alapkódokkal. Ez a dokumentumkép összehasonlítható sok alapkóddal, testületi és egyéni szerzővel, számokkal és kifejezésekkel. Az ilyen jellegű dokumentumképekből részállományok alakíthatók ki, és kiterjeszthető a keresés. Akár hatékonyabb és gazdaságosabb ez a rendszer a teljes szö-

veg-összehasonlításnál, akár nem, nagyon messze van az egyszerűségtől, és sok problémát okoz a dokumentumok rekordjainak elkészítésekor és a fájlszervezésben, valamint a keresési stratégiákban. Igazolja-e ezt a bonyolultságot az a szolgáltatás, amelyet nyújtani képes? A dokumentumok komplex entitások, de talán a döntő kérdés, ami velük kapcsolatban felvetődik, az, hogy melyek azok a legegyszerűbb eszközök, amelyekkel reprezentálhatók, helyettesíthetők, úgy, hogy a használói igényeket is kielégítsék. A keresés tárgyalása során nem szabad szem elől téveszteni annak célját. A felhasználói igényeknek kell meghatározniuk, hogy milyen rendszert dolgozzanak ki.

A dokumentum reprezentánsai:
A1, A2,…, B1, B2,…, C1, C2,…, D1, D2,…, T1, T2
A kérdések reprezentánsai:
A10, D27, T24, T91
A dokumentumfájl részállománya (szubfájl):
(1) Az egyes ismérvek részállománya
A1, A2, B1, B2
(2) Az egyes tágértelmű ismérvek részállománya, például
D1, D2, D3, D4
(3) A statisztikai módszerekkel kapcsolatba hozott helyettesítők részállománya
A kérdés kiterjesztése:
(1) Ismérvek törlése, például csak A10, T24 meghagyása
(2) Az ismérvek kibővítése, például A10-zel

$$A10 + A12 + A91 = \sum A10$$

a szimbólumok csonkolásával
tezaurusszal
(3) Az ismérvek kapcsolatainak feltárása

4. ábra. A kérdések összehasonlítása a dokumentumokkal.

6. Dokumentumképek szerkesztése[36]

Vickery a dokumentumkép (dokumentumprofil) fogalmát rendkívül általános értelemben használja. Valójában a fizikai dokumentációs egységről készült dokumentumleírást érti rajta. A gyakorlatban a dokumentumképet ennél szűkebben értelmezik: a dokumentumleírásnak csak azok az elemei alkotják, melyek a kereséshez felhasználhatók abból a célból, hogy a keresőképpel (keresőprofillal) összehasonlítsák. A fordításban azonban megőriztük a szerző szóhasználatát.

36 Construction of document profiles. In: Techniques of information retrieval, p. 58–65.

A szerző/cím szerinti leíró katalogizálásról szóló fejezetet azért vettük föl a kötetünkbe, mert ha a dokumentum személyről, testületről, vagy földrajzi helyről szól, akkor a tartalmi feltáráskor, illetve tartalomra vonatkozó kereséskor a személy-, testületi vagy földrajzi neveket a leíró katalogizálásnak megfelelő formában kell kezelni.

A dokumentumkép szerkesztését – azaz a dokumentumot tömören kifejező szóhalmaz kiválasztását a dokumentumból – általánosítva *elemzésnek* nevezhetjük. A dokumentumkép vagy dokumentumprofil felhasználható a dokumentum azonosítására, sőt helyettesítésére is. Formáját tekintve lehet kivonat, összefoglalás, referátum, katalógustétel, mutatótétel stb.

Az „elemzés" szó olyan általános kifejezés, amely számos hagyományos könyvtári tevékenységet – például katalogizálást, mutatókészítést (indexelést), osztályozást és kivonatkészítést – továbbá olyan kísérleti eljárásokat is jelent, mint az indexelés, osztályozás és kivonatkészítés. Mindezek az elemzési eljárások arra valók, hogy a dokumentumokat kiválogathassuk a tárból. Mihelyt a gyűjtemény túllépi a közvetlenül áttekinthető terjedelmet, szükség van rájuk. Még a néhány száz kötetes gyűjteményből sem könnyű kiválasztani meghatározott könyvet, ha a gyűjtemény rendezetlen. Minden könyvből ki kell emelni valamit, ami jellemző rá (például a szerzőt és a címet), rögzíteni kell ezt valamilyen hordozón (mondjuk a könyv gerincén) és a könyvgerinceket úgy kell sorba rendeznünk, hogy könnyen átnézhessük őket. A dokumentumképek a dokumentumok tömörített képviselői, amelyeket az információkeresés egyszerűsítésére és meggyorsítására használunk.

Az elemzés hagyományos eljárásainak hosszú, de eddig még megfelelően fel nem dolgozott története van. Az alexandriai könyvtár katalógusából (*Kallimakhosz:* „Pinákeszéből", i. e. 250) töredékek állnak rendelkezésre. Megvan jó néhány középkori könyvtári katalógus – kéziratok esetleges és rendezetlen leltárai. A nyomtatás feltalálása a könyvtermelés és az írott szó iránti igény gyors növekedéséhez vezetett, és ezáltal a különböző célú könyvjegyzékekre is sokkal nagyobb szükség lett. A bibliográfiai elemzés a reneszánsszal kezdődik.

Az információkeresés kulcsául (kulcsszavául)[37] szolgáló dokumentumonkénti *rendszó* ideája 1545-ben, *Konrad Gesner* Bibliotheca Universalisában merült fel. Ez a könyveket a szerzők keresztnevének betűrendjében sorolta fel, a vezetéknevek mutatójával kiegészítve. 1548-ban *Gesner* újabb kötetett adott

37 A „kulcs", kulcsszó a dokumentumtétel elérésére szolgáló eleme a leírásnak. Lehet a szerzőnév, a cím, az egységes besorolási adat, de akár a megjelenési hely és év, vagy a kiadó, továbbá a cím, referátum, sőt a teljes szöveg valamelyik szava is. Jelentése azonos az „ismérv" szóéval. Ameddig az ilyen szót mint a leíráson belüli elemet tárgyaljuk, addig beszélünk kulcsról, ismérvről. Amikor a szó a leíráson kívül, például mutatóban jelenik meg, és ott a besorolás alapja, hagyományosan rendszónak nevezik. Ha a leírás, vagy bármely szócikk élén áll, vezérszó a neve. A vezérszó általában a rendszó szerepét is betölti, de ez nem feltétlenül szükséges.

ki, amely ugyanezeket a feldolgozott műveket jelzettel ellátott tematikus osztályozási rendben tartalmazta, kiegészítve egy betűrendes tárgymutatóval a szakjelzetek feloldásához.

A katalogizálás további történetét nem tárgyalhatjuk. Az elemzés fő problémái már ebben a korai szakaszban is nyilvánvalóak. A dokumentumokból két szóhalmaz származhat. Az egyik halmaz a dokumentum *eredetével* (szerző, kiadó testület, kiadó, megjelenés helye stb.), a másik az *információtartalmával* (tárgyával) kapcsolatos. A dokumentum *címe* mindig bizonytalan átmeneti helyet foglalt el a kettő között, hiszen joggal tekinthető a tartalom jellemzőjének, gyakran mégis csak a dokumentum azonosítására alkalmas címkének használják. (Cím nélküli versek és kéziratok esetében a szöveg első sora szolgálja ezt a célt.)

E felosztásnak megfelelően a bibliográfiai elemzés szétválik két hagyományos eljárásra: (1) szerző/cím szerinti katalogizálás, ennek eredménye, hogy a dokumentumok eredetével kapcsolatos szavak lesznek a tétel rendszavai és ehhez kapcsolódva (1A) a leíró (deskriptív) katalogizálás, amelynek célja a dokumentumok egyedi azonosítását lehetővé tevő szavak – beleértve a címet is – kiválasztása; (2) tárgyi/tartalmi elemzés, amely a dokumentumok információtartalmára irányul.

Szerző/cím szerinti és leíró katalogizálás

Egy dokumentum keletkezéséhez sok ember járul hozzá. Az egy vagy több természetes szerző mellett szerepelhet kiadó testület (mondjuk a szerzőt foglalkoztató intézmény), kiadó, nyomdász, szerkesztő, fordító, illusztrátor stb. A megjelenés helye, ideje, a kiadvány nyelve ugyancsak az eredetet jellemző adatok. Esetenként ezeknek az elemeknek bármelyikét rendszóként használhatjuk a katalógusban annak érdekében, hogy bizonyos szerző által írt, meghatározott nyomdász által készített, adott évben kiadott stb. művet könnyűszerrel megtalálhassák. Dönteni kell arról, hogy az eredetre utaló lehetséges jellemzők közül melyik váljék a megtalálást biztosító rendszóvá. E döntést meghatározza (1) a katalogizálási politika: milyen felhasználói igények kielégítésére tervezik a katalógust, és (2) a rendelkezésre álló források: mennyi munkaerőt lehet erre a feladatra fordítani.

Ezeket a bibliográfiai elemeket általában a dokumentumok első lapjának („az előzékeknek") az áttanulmányozásával ki lehet deríteni, bár időnként az információkat ki kell egészíteni külső forrásból (pl. ha a kiadó nem tünteti föl a fordító nevét). Ha egy katalógus, bibliográfia, mutató vagy más információkereső rendszer adott író összes műveit, vagy testület összes kiadványát össze akarja hozni, akkor a neveket egységesíteni kell. E probléma már *Gesner* idejében is létezett: vajon az írókat keresztnevükön (ezt a középkori gyakorlatot követte *Gesner* jegyzékének főrészében), vagy vezetéknevükön (ami a 17. századig nem vált egységes gyakorlattá) vegyék föl?

A 19. századra a nevek egységesítésének szükségessége a katalógusban elvezetett az első terjedelmes szabályzathoz, amit *Anthony Panizzi* készített 1841-ben a British Museum számra. Sok szabályzat látott napvilágot azóta. Az angolszász országokban a legfontosabbak: az angol–amerikai vagy Közös Szabályzat (Joint Code, 1908), 1967-ben megjelent átdolgozott kiadásban az Amerikai Könyvtáros Egyesület Szabályzata (American Library Association Code, 1949) és a Kongresszusi Könyvtár Leíró-katalogizálási Szabályai (Library of Congress Rules for Descriptive Cataloging, 1949).

Napjainkban a legjelentősebb katalogizálási szabályzat angol nyelvterületen az „Anglo–american cataloguing rules" (AACR), német nyelvterületen pedig a „Regeln für den allgemeinen Katalog" (RAK) és a „Regeln fürd den Schlagwortkatalog" (RSWK). Magyarországon nincs átfogó, országos érvényű nemzeti katalogizálási szabályzat; mérvadó szerepet a nagyobb könyvtárak adatbázis-rendszereihez (Országos Széchényi Könyvtár NEKTÁR rendszere, Országos Műszaki Információs Központ és Könyvtár OSZKÁR rendszere) készült kis példányszámú belső szabályzatai játszanak.

Az angol személynév szabályos értelmezése az esetek többségében nem jelent problémát: a családnév kerül az élre és ezt követik az egyéni nevek (pl. Whitehill, Walter Allen). A kötőjeles név azonban már döntést követel: Baring–Gould B vagy G alatt veendő fel? Vagy a névelőzékes családnevek esetén: De Quincey D vagy Q alatt sorolandó be? Ugyanígy járjunk el de Gaulle-lal is? Minden előírás tartalmaz szabályokat az ilyen döntésekre. Ha azonban túllépünk Európa és Amerika határain, a nevek szerkezete eltér a mi kultúránkban megszokott alaktól és ezekre a nyugati előírások „olyan alaktalanok, akár a folyadékok" – mondja az indiai könyvtáros, *Ranganathan* (1955).

A nyugati világ katalogizálói számára pillanatnyilag azonban nagyobb gondot jelent a „testületi szerzők" értelmezése, az olyan intézményeké, amelyek a sokszerzős vagy anonim dokumentumok kiadásáért felelősek. Az ilyen típusú dokumentumok – mindenféle hivatalos kiadványok, intézmények jelentései, konferenciaanyagok, időszaki kiadványok stb. – mennyisége állandóan nő.

Könyvtárnyi irodalom gyűlt össze erről a problémáról, amit az alábbi összetett példával szemléltethetünk. A *nem hagyományos információkereső rendszerek*ről adott ki jelentést az alábbi testület:

Centre for Documentation Research,
School of Library Science,
Western Reserve University
AFOSRTN 58–575 Jelentés,
Directorate of Research Communication,
Air Force Office of Scientific Research,
Air Research and Development Command,
U.S. Air Force.

Melyik nevet vagy neveket válasszuk ki a jelentés testületi szerzőjének? Hogyan kell a testület nevét leírni? Minden egyes katalogizálási szabályzat másként rendelkezik. A jelenlegi gyakorlat lényege, hogy csak megkülönböztető és önmagukban is megálló neveket szabad rendszóként használni. Például több „centre for documentation research" (a dokumentációs kutatások központja) létezhet. A fenti példában szereplőt mint a Western Reserve University, Centre for Documentation Research-öt kell azonosítani.

Az időszaki kiadványokat, folyóiratokat, különféle sorozatokat lehet úgy katalogizálni, mint kiadó testületek publikációit, de a címet gyakrabban használják rendszónak. A cím lehet megkülönböztető (pl. American Documentation), vagy maga is tartalmazhat testületi nevet (pl. Bulletin of the Medical Library Association). Az időszaki kiadványok címeinek szabványosítása, különösen a műszaki és természettudományokban, rövidítések segítségével történik.

A rövidített formából általában kiesnek a névelők, kötőszavak, prepozíciók és hasonlók. Hosszú címek rövidíthetők az utolsó szavak elhagyásával, bizonyos szavakat a záróbetűk levágásával kurtítanak meg, bár néha a belső betűk is kiesnek. Például: Journal → Jo; Annals, Annalen, Annales, Annali, Annaes → Ann.; Reports, Report → Rp.; Biologia, Biologie, Biology → Biol.; Engineering → Engg.; Manufacturing → Mfg.; Journal of the American Chemical Society → J. Am. Chem. Soc.; Archives des maladies professionelles de médicine du travail et de securite sociale → Arch. mal. prof.

A szabványos rövidítéseknek még rövidebb alakját is kifejlesztették, a „codent". Ez minden címet egyedi és részben mnemonikus betűkódra redukál. J. Am. Chem. Soc. → JACS; Arch. mal. prof. → AMPM; American Documentation → AMDO. A „codenek" széles körben elterjedt listáját *Kuentzel* adta ki.

A nevek szabványosításának e rövid áttekintése után helyes lesz összefoglalni, hogy a dokumentumok eredetével kapcsolatos mely jellemzőket használják általában rendszóként. Könyvek esetében általában biztosítják a hozzáférést a természetes szerző(k)ön és/vagy testületi szerző(k)ön, időnként a címen, és azon a sorozaton keresztül, amelyhez a könyv tartozik. Intézmények által kiadott jelentések esetében a jelentés száma is elérési pontot jelenthet.

Itt következik a leíró katalogizálás másik problémája – milyen további, az azonosítást segítő és a mű természetére utaló bibliográfiai elemek épüljenek be a dokumentumképbe? Könyvek esetében hagyományosan a szerző a rendszó és a tétel tartalmazza a sorozati címet, a kiadót, a megjelenés helyét és idejét, a terjedelmet (kötetszám, oldalak száma, illusztráció stb.), a könyvben található bibliográfiára és néha a szöveg nyelvére vonatkozó megjegyzéseket. Jelentések dokumentumképe kiegészülhet még a szerződés számával, a beszerzés forrásával (amennyiben ez a testületi szerzőtől és a kiadótól különbözik) és a biztonsági osztályozással.

Időszaki kiadványokban megjelenő cikkek dokumentumképébe néhány újabb analitikus kalauz további ismérvet épít be. A szerző által felhasznált és

megjelölt korábbi dokumentumok is az eredet jellemzőinek tekinthetők. Időnként hasznos lehet kinyomozni, hogy Newton *Principiá*ja mely írásra hatott oly módon, hogy megnézzük, kik említik, hogy felhasználták. Az ilyen típusú kérdésekre a *hivatkozási indexek* adnak választ. Ezek felsorolják a hivatkozott könyveket és tanulmányokat és megmutatják, mely dokumentumok hivatkoztak ezekre.

A szerző/cím szerinti és leíró katalogizálást jelen pillanatban kizárólag emberek végzik. Nincs az a gép, amely a dokumentum címlapjáról – vagy akár ennek egy géppel olvasható átiratáról – ki tudná választani a szerzőt, a címet, a kiadót stb., ha ezeket az elemeket előzetesen nem jelölte ki elemző személy. A leíró katalogizálás automatizálása csak akkor válik lehetővé, ha a dokumentumokat eleve az eredet jellemzőinek ilyen egyedi jelölésével adják közre. Természetesen, ha emberi erővel előállított katalógustételt géppel olvasható formában rögzítenek, az adatfeldolgozó berendezések már képesek különböző célú műveleteket végezni ezekkel. A „számítógépes katalogizálás" eme aspektusa már közismert.

A tartalmi elemzés

A dokumentum eredetére utaló bibliográfiai elemekből viszonylag kevés van és ezek viszonylag könnyen azonosíthatóak. Ezzel szemben az információtartalomnak csak egyetlen indikátora – a cím – állapítható meg közvetlenül, viszont a dokumentum teljes fennmaradó szövege a tárgyi információ forrása. A tartalmi elemzés feladata tehát elvileg sokkal kevésbé egyértelmű, mint a leíró katalogizálásé, annak dacára, hogy a gyakorlatban nagyon könnyűnek látszhat, ha a tartalmi feltárást egyszerűen a dokumentum címének kiemelésével oldják meg.

A cím szerinti indexelés a dokumentum témáját megnevező szerzőre hagyatkozik. A kicsit alaposabb indexeléskor már a dokumentum részeinek, fejezeteinek címét is indextételnek választják. Ám a tartalmi elemzés céljában eltér a leíró katalogizálástól. Az eredet jellemzői a dokumentumtól elválaszthatatlanok – szerzője, kiadója stb. egyedi. Az azonban, hogy a dokumentum tartalmának melyik szempontja érdekes, használónként változik. Az adott dokumentum leíró katalogizálása többé-kevésbé azonos, ezzel szemben a tartalmi elemzése eltérő lehet, s ez különböző szavakból álló mutatót eredményez.

Az egyik olvasó életrajzi adatai miatt tartja George Bernard Shaw „Előszavait" értékesnek, a másik a színművekkel kapcsolatos utalások miatt, a harmadik társadalomkritikája miatt, a következő más írókról szóló kommentárjai miatt stb. Minden egyes érdeklődési kör olyan szempontot képvisel, amely a tartalmi elemzés alapjául szolgálhat. A valóságban a tartalmi elemzést a feltételezett felhasználói csoporthoz szabjuk. Nincs semmi garancia arra, hogy a szerzőtől származó címek éppen megfelelnek valamelyik használói kör igényeinek. Ebből

következik, hogy a tartalmi elemzés elvileg a teljes szöveg áttanulmányozását megköveteli annak érdekében, hogy a kulcsszavak olyan halmazát tudjuk kiválasztani, amely a szöveg információtartalmának feltehetően fontos, érdekes szempontjait tükrözi a készülő tárgymutató felhasználói számára.

A tartalmi elemzés terméke lehet a dokumentum kivonata (pl. címe); egy vagy több természetes nyelven kifejezett (pl. tárgyszónak, deszkriptornak nevezett) vagy mesterséges nyelven kifejezett, kódolt (pl. könyvtári osztályozási jelzetnek nevezett) ismérv; vagy terjedelmesebb referátum, tömörítvény stb. A következőkben ezekre mutatunk egy példát:

Cím:

Kollektív elektron ferromágnesesség III. Nikkel és nikkel–réz ötvözetek

Tárgyszavak:

Nikkel ötvözetek: Ferromágnesesség
Nikkel–réz ötvözetek: Ferromágnesesség
Szuszceptibilitás: hőmérséklet-változás
Osztályozási jelzetek: 538.114:669.245

Referátum:

„A *d* és *a* elektronsávok átfedésének elméleti következményeit vizsgálják. Tárgyalják az elektroneloszlás függését a hőmérséklettől. Számításokat végeztek az átviteli hatással kapcsolatban és megvizsgálták a hőtágulás hatását. Az elméletet a szuszceptibilitásnak a Curie-pont feletti hőmérséklet függésére alkalmazzák. Megállapításaik a nikkeldús ötvözetekre kielégítően megegyeznek a kísérleti eredményekkel. A rézdús ötvözetek magas hőmérsékletű változatát is magyarázzák, de az alacsony hőmérsékletűt is érintik. Az összetételtől függő változás kielégítően magyarázható a nikkeldús ötvözetekben, de a rézdús ötvözetek alacsony hőmérsékletű területein ismét ellentmondások mutathatók ki."

A referátum összefoglalás, a dokumentum információtartalmának rövid kivonata. A referátumok egy része „informatív" – elegendő információt tartalmaz a cselekvéshez, bizonyos fokig helyettesíti a dokumentumot. Az előző példa többékevésbé informatív, bár nem idézi a mennyiségi eredményeket, mint az alábbi:

A szerves vegyületek gyorsan jelölhetők tríciummal, a trícium gázzal, csendes elektromos kisülés közben történő inkubációval. Pl. 500 ml benzol 0,67 mc tríciumot vett fel egy 40 mc tríciumot tartalmazó hidrogén–trícium keverékből, 20 kV/mA kisülés mellett egy óráig tartó érintkezés során. Ez az érték 104-szer nagyobb a csendes kisülés nélküli felvételnél. A Cobalt–60 gammasugarak jóval kevésbé voltak hatékonyak, amellett roncsoló a hatásuk.
R.M. Lemmon–B. M. Tolbert–W. Strohmeier–M. Whittemore: Az ionizáló energia mint a tríciumos jelölés cseréjének elősegítője. = Science 129 (1959) pp. 1740–41.

Az indikatív referátumok rövidebbek és csak a dokumentum tartalmának jelzése a céljuk.

> Strohmeier–M Whittemore; Science 129 (1959) pp. 1740–41. – A csendes elektromos kisüléseknek vagy a Cobalt–60 gammasugaraknak hatását ismerteti a szerves vegyületek, pl. benzol trícium-felvételére.

Még rövidebb az annotáció, ami a cím egyszerű kibővítése.

> Az ionizáló energia mint a tríciumos jelölés cseréjének elősegítője. R. M. Lemmon–B. M. Tolbert–W. Strohmeier–M. Whittemore; Science 129 (1959) pp. 1740–41. – A csendes elektromos kisülés és a Cobalt–60 hatása.

Borko és *Chatman* 130 referáló szolgálat gyakorlatát áttekintve megállapítja, hogy 18% informatív, 37% indikatív, 25% mindkét féle referátumot szolgáltat. (A fennmaradó 20% nem adott egyértelmű választ erre a kérdésre.) A felmérés arra a közös megállapodásra épült, hogy a referátumnak tartalmaznia kell:

(a) a referált dokumentációs egység célját, esetleg az érintett probléma határait, nagyságrendjét,
(b) az alkalmazott módszereket, beleértve a felszereléseket, anyagokat, teszteket,
(c) az elért eredményeket, időnként számadatokat,
(d) a levont következtetéseket.

A referátumokból tanácsos kihagyni bizonyos dolgokat, például részletes leírásokat, megfontolásokat, jól ismert tényeket.

Az elemzés problémái

A katalogizáláshoz, az osztályozáshoz a gyakorlatban számos segédeszközt használnak. A tartalmi elemzés számára a legfontosabbak a szabványos terminológiák, a szabályzatok és a strukturált bizonylatok. Ötven műszaki információs központból álló mintán *Korotkin* kimutatta, hogy 90%-ukban használtak valamilyen szabványos szójegyzéket. 75% dolgozott szabályzatokkal és kézikönyvekkel, amelyekhez indexelési és osztályozási szabályok, házi szabványok és az osztályozás mélységére vonatkozó elvek stb. tartoznak. Strukturált osztályozási bizonylatokat 38% alkalmazott a mintában. Ezek a segédletek a közfelfogás szerint növelik az osztályozás pontosságát és következetességét.

Ennek ellenére az elemzés gyakorlatának vizsgálata a könyvtárakban és információs szolgálatokban nagy különbségeket fog felszínre hozni a dokumentumok leírására kiválasztott elemek és ezek szabályozottsága között. Távol vagyunk még a gyakorlatot irányító világos elvektől és az elvek, amelyek mellett némelyek kiállnak, mások szerint nagyon vitathatók. A tartalmi elemzés problémái közül az egyik legégetőbb a mélységé. Mennyire legyen részletező az osztályozás? Hány kifejezés reprezentálja a dokumentumot? A tendencia az egyre alaposabb osztályozás felé mutat. Vagy harminc modern információs rendszer példája a következő megoszlást mutatja:

Kifejezés/dokumentum	1–4	5	6–9	14–19	20–30	30 fölött
A rendszerek száma	1	4	9	3	7	6

A mélységet mérhetjük az eredeti szöveg tömörítettségének mértékével. A *McGraw–Hill Encyclopédia of Science and Technology* közel 10 millió szót tartalmaz. Könyvtári katalógus számára ez a dokumentum két szóval – természettudomány, műszaki tudomány – indexelhető és a tartalmi leírás aránya ekkor 500 000:1 lenne. Részletezőbb információs rendszer mind a 7000 cikket osztályozhatja egyetlen szóval; a szöveg–leírás arány ekkor kb. 1400:1 lesz. Az enciklopédia mutatója szövegoldalanként 8 kétszavas mutatónevet tartalmaz. Az oldalak átlagosan 1000 szót tartalmaznak, így a szövegszavak és a mutatóneveket képviselő kifejezések aránya 60:1. A két szélső érték – 500 000 és 60 – között melyik arány a leghatékonyabb meghatározott körülmények között? Mindeddig alig rendelkezünk adatokkal, amelyek megmutatnák, hogy az elemzés milyen mélysége kívánatos.

A tartalmi elemzés másik gondja a terminológia szabályozottsága. Egyesek amellett kardoskodnak, hogy a különböző vizsgált szövegekben folyamatosan felmerülő szavakat eredeti alakban kell kiválasztani és az osztályozási–indexelési gyakorlatot is ehhez kell igazítani.

Más elemzők szigorúan kötött szójegyzékeket – tárgyszójegyzékeket, tezauruszokat – dolgoztak ki és következetesen ragaszkodnak ezekhez. Ismét lehetetlen az eljárások viszonylagos értékét egyértelműen meghatározni. Ez a probléma tetéződik azzal, hogy ha a terminológia kötött, el kell dönteni, hogy az információs rendszer mely pontján kell azt alkalmazni. Az alkalmazás történhet az információelemzés során, de néhány rendszer szabad szövegszavakat használ az elemzés során és a kereséskor vezeti be a terminológiai változatokat. A kötött szabványosított szójegyzék ilyenkor a keresőnek megmutatja, milyen alternatív szövegszavakat használhattak még az elemzés során.

Nézzünk most túl az elemzés szubjektív tevékenységén és tekintsük helyét az egész információs rendszerben. E rendszeren belül különféle szervezetek – kiadók, könyvtárak, információs intézetek, bibliográfusok stb. – újra és újra elem-

zik ugyanazt a dokumentumot. Nagyarányú a többszörös vagy legalábbis átfedő feldolgozás. A szabványos előírások és jegyzékek dacára ez a párhuzamosság a rendszerek közötti végtelen számú variációhoz vezet. Ami a leíró katalogizálást illeti, azt mondják, hogy ha megnézzük tíz jelentős amerikai könyvtár katalóguscéduláit ugyanarról a könyvről, bár azonos katalogizálási szabályzatokat használtak, a cédulák mégis más-más rendszóval fognak kezdődni. 938 tárgyszó közül, amelyeket kilenc műszaki könyvtárban használtak, 57% csak egy könyvtárban fordult elő és 23% jelentősen eltérő formában jelent meg. Hasznos és szükséges az ilyen többszörösség és „sokszínűség"?

Hosszú ideje elfogadott az a nézet, hogy a leíró katalogizálásban az egységesség értékesebb, mint azok az apró eltérések, amelyek a speciális helyi szükségleteket elégítik ki. Az Egyesült Államokban a Kongresszusi Könyvtár kezdeményezte és támogatta a központi és egységes katalogizálás bevezetését. Abban már jóval kisebb az egyetértés, hogy kell-e és elérhető-e az egységes tartalmi feltárás? Összehasonlítva különféle tezauruszokat és kódolt – mesterséges nyelven alapuló – szógyűjteményeket, nyilvánvaló, hogy az egyes intézményekben készült szabványosított szótárak alig egyeznek meg egymással. Sokan érvelnek úgy, hogy ez természetes, mondván: a szabványosított terminológiának meg kell felelnie a mindenkori felhasználói csoport érdekeinek, hiszen minden csoport a saját szóhasználatát akarja szabványosítani. Nem tudható, vajon az ilyen egyedi érdekeken alapuló megközelítés előnyei megérik-e a beléjük fektetett munkát?

Az elemzés – *Fairthorne* szerint – az információkeresés alapkérdése és egyben legszűkebb, leginkább költséges keresztmetszete. Szükség van tehát a gyakorlati elemző munkát vezérlő elvekre és olyan szabványokra, amelyek elősegítik az információs rendszerek közötti együttműködést, adatforgalmat és kompatibilitást.

ALLEN KENT (1921) ÉS A KÖNYVTÁRTAN ENCIKLOPÉDIÁJA

Az eredeti foglalkozására nézve vegyész Kent 1947-től kezdett a dokumentáció problémáival foglalkozni. 1955–56-ban *James W. Perryvel* együtt dolgozta ki a Western Reserve University dokumentációs rendszerének szemantikai kódjait (lásd az első kötetben *Eric de Grolier* szemelvényét). 32 éves korában az Egyesült Államok Tájékoztató és Kommunikációs Kutatóintézetének igazgatója, később elnöki szaktanácsadó, számtalan nemzetközi konferencia elnöke és kormányzati információs program kidolgozója, professzor a Pittsburghi Egyetem Könyvtár- és Tájékoztatástudományi Főiskoláján.

A szó fizikai értelmében is nagy műve a jelenleg 59 kötetes könyvtári és információtudományi enciklopédia.

1. Encyclopedia of library and information science

Ed. by Jack Belzet, Albert G. Holzman and Allen Kent. – New York : M. Dekker Inc., 1968–1972. – Vol. 1–8.
Ed. by Allen Kent and Harold Lancour. – New York : M. Dekker Inc., 1973–. Vol. 9–59
Az 59. kötet (supplement) 1997-ben jelent meg.

Első kötete 1968-ban jelent meg, s 1997-ben az 59. kötetnél tartanak, a kiadás nincs lezárva. A 33. kötetig bezárólag a szócikkek egyetlen betűrendezett állományt alkotnak, ezt követően kötetenként kezdődik újra a betűrend, időnként pedig önálló kötetben (3–35., illetve 46–47.) a mutatók jelennek meg. Az egyes szócikkek különféle hosszúságú, önálló, tömör tanulmányok. Az első és az utolsó kötet között eltelt hosszú idő következtében a szócikkek szerkesztésének szemlélete is fokozatosan változott. Az osztályozással és információkereséssel összefüggő fontosabb szócikkek a következők:

Absztrakt osztályozás [Classification, abstract]	1. köt	p. 12–16
Automatikus kulcsszavas osztályozás [Clumps, theory]	5. köt.	p. 209–229
Cutter-féle tárgyszavas osztályozás [Expansive Classification]	6. köt.	p. 297–312
Szemantikai összefüggések [Association trails]	2. köt.	p. 55–87
Hierarchikus permutált mutató, lépcsőzetes mutató [Chain index]	4. köt	p. 423–439
Kettőspontos osztályozás [Colon classification]	5. köt.	p. 316–340
Kongresszusi könyvtár osztályozási rendszere [LC classification]	15. köt.	p. 93–181
Mutatók, indexelő, indexelés [Index, indexer, indexing]	11. köt.	p. 305–311
Osztályozás és kategorizálás [Classification and categorization]	5. köt.	p. 43–141
Osztályozáselmélet [Classification theory]	22. köt.	p. 197–198
Osztályozáselmélet [Classification theory]	5. köt.	p. 147–174
Osztályozási jelzetek [Notation, classification]	19. köt.	p. 194–197
Osztályozási táblázatok mutatói [Index to classification schemes]	11. köt.	p. 305–311
Sears közművelődési könyvtári tárgyszójegyzék [Searst List of Subject Headings]	27. köt.	p. 160–177
Természetes nyelven alapuló osztályozási rendszer [Natural language classification]	5. köt.	p. 44–50
Természetes osztályozás [Natural classification]	19. köt.	p. 186–205
Tezaurusz [Thesaurus]	30. köt.	p. 416–461
Tizedes osztályozás [Dewey Decimal Classifikation]	7. köt.	p. 128–141

Magyarul az alábbi szócikkek jelentek meg:

Permutált indexek

In: Bibliográfiai olvasókönyv : Szakirodalmi szemle / [szerk.] Varga Ildikó ; [közr. az] Országos Széchényi Könyvtár Könyvtártudományi és Módszertani Központ. – Budapest : NPI, 1979. p. 157–187.
Eredeti: Charles R. Bernier: Permuted indexes. In: Encyclopedia of library and information science, Vol. 22. p. 36–65.

A permutált KWIC, KWOC stb. mutatók meghatározásai, története, példái, előállítása és használata.

Hivatkozási indexek

In: Bibliográfiai olvasókönyv : Szakirodalmi szemle / [szerk.] Varga Ildikó ; [közr. az] Országos Széchényi Könyvtár Könyvtártudományi és Módszertani Központ. – Budapest : NPI, 1979. p. 157–187.
Eredeti: Melwin Weinstock: Citation indexes. In: Encyclopedia of library and information science, Vol. 22. p. 16–40.

A hivatkozási indexek története, szerkezete, használata, a keresési módszer, az eredmények értékelése, szemantikai problémái.

2. Szaktájékoztató központok

In: Bibliográfiai olvasókönyv : Szakirodalmi szemle / [szerk.] Varga Ildikó ; [közr. az] Országos Széchényi Könyvtár Könyvtártudományi és Módszertani Központ. – Budapest : NPI, 1979. p. 157–187.
Tömörítve
Eredeti: Specialized information centres. Washington, D. C. ; London: 1965. 297 p.

A szervezeten belüli munkamegosztás, műveleti egységek leírása után a beszerzés, az elemzés, adatrögzítés és a források tárolása következik. Részletesen foglalkozik a kérdések elemzésével, a kereséssel és a felhasználó kiszolgálásával.

HAROLD BORKO (1922)

Harold Borko a pszichológia doktora, a kaliforniai egyetem könyvtár és információtudományi iskolájának (Graduate School of Library and Information Science) professzora. Szakterülete az automatizált információs rendszerek tervezése és értékelése, az automatikus nyelvfeldolgozás és az indexelés. Korábban rendszerfejlesztőként dolgozott, számos kutatási programot irányított. Szerkesztője az Information Processing and Management című folyóiratnak, elnöke az amerikai információtudományi társaságnak (American Society of Information Science, ASIS). Munkásságára az átfogó szemlélet, a nyelvészeti érdeklődés jellemző, elemzéseinek középpontjában még számítástechnikai könyveiben is a tartalmi feltárás kérdései állnak.

Műveiben Borko – az angol nyelvű szerzők jelentős részéhez hasonlóan – indexelő nyelvről beszél. Ez a dokumentációs nyelv, információkereső nyelv megnevezése a feldolgozás, indexelés szempontjából. E terminológiai kérdéssel részletesen az első kötetben *Gernot Wersig* szemelvényének 7.2 c) fejezetéhez fűzött kommentárban foglalkozunk.

Útban az átfogó indexeléselmélet felé[38]

1. Az indexeléselmélet szükségessége

A tudomány feladata, hogy a tárgykörébe tartozó jelenségeket értelmezze, ellenőrizze és előre jelezze. Az információtudomány az információ viselkedését és tulajdonságait, az áramlását meghatározó erőket és az információfeldolgozás technikáját vizsgálja, az optimális tárolás, keresés és terjesztés érdekében. Az információtudományi vizsgálódások során sok specifikus adat gyűlt már össze az információátvitelről. Ezekből azonban mindmáig nem alkottak átfogó elméletet.

Az indexelés, azaz a természetes nyelven alapuló mellérendelő osztályozás az információtárolási és -keresési folyamat lényeges része. A mutatók (indexek) biztosítják, hogy a tárolt bibliográfiai tételeket később tartalmi szempontok alapján és a természetes nyelv segítségével kereshessék. Sok mindent tudunk az indexelésről és osztályozásról, tudjuk, hogyan kell indexelni, de hiányoznak az ismereteink, hogy pontosan e lőre jelezzük a különböző indexek használhatóságát. Szükségünk van egy elméletre, amely segít megmagyarázni, ellenőrizni és előre jelezni az információáramlást. Ennek az

38 Toward a theory of indexing. In: Information Processing and Management. 1977, Vol. 13,. No. 6, p. 355–365.

elméletnek ki kell terjednie az indexelésre és osztályozásra, valamint a mutatókra. *Landry* szerint az indexelés és osztályozás az információtárolás és -keresés „lelke" és hozzáteszi: „Elképzelhető, hogy egy átfogó indexeléselmélet segítségével jobban megmagyarázhatjuk annak a tevékenységnek a természetét is, amelyet »információkeresésnek« nevezhetünk." *Landry* felsorol jó néhány megválaszolandó kérdést:

– Milyen természetű döntéseket hoznak az indexelők és tartalomelemzők tevékenységük során?
– Mi a szerepe az indexelő nyelveknek (dokumentációs nyelveknek) és eszközöknek?
– Hogyan választhatók ki a tételek tartalmát jellemző indexkifejezések?
– Milyen kritériumok szerint minősíthetők a mutatók?

Landry listáját néhány specifikusabb kérdéssel kiegészíthetjük:

– Mennyiben javítja az ellenőrzött szótár használata az információkereső rendszer hatékonyságát, teljességét és pontosságát?
– Az indexelő nyelv nagyságának 10%-os növelése milyen mértékben módosítja az információkeresés teljességét és pontosságát?
– A nagyon specifikus kifejezések számának 10%-os növekedése az indexelő nyelvben hogyan befolyásolná a teljességet és a pontosságot?
– Ha az egy dokumentumhoz rendelt ismérvek (indexkifejezések) átlagos számát 3-ról 8-ra növeljük, hogyan befolyásolja ez a teljességet és a pontosságot?

A felsoroltak korántsem alkotják a kérdések teljes körét, de példaként szolgálhatnak arra, hogy milyen fajta kérdéseket kell egy információtároló és -kereső rendszer tervezése, illetve az indexelő nyelvként felhasználható szókincs kiválogatása során megválaszolni. Ezekre a kérdésekre pontos és számszerű válaszok szükségesek, hogy kiszámítható legyen a költség, a teljesítmény és a haszon. Jelenleg e kérdésekre – ha egyáltalán felmerülnek – csupán a minőségre vonatkozó, tapasztalati válaszok adhatók.

Néhány kutató megkísérelte megfogalmazni az információkereséshez használt indexeléselmélet szempontjait. Ebben a dolgozatban négy szerző elméleti munkáját tekintjük át vázlatosan, arra törekedve, hogy az előzőekben feltett kérdésekre adott válaszaikat megismerhessük.

2. Jonker: az ismérvként használt indexkifejezések rendszerének elmélete

Az indexeléselmélettel az elsők között *Frederick Jonker* foglalkozott, akit joggal nevezhetünk az információtudomány és az információkeresés és -tárolás úttörőjének. *Jonker* találta fel az információtároláshoz és -kereséshez a Trematrex rendszerű fénylyukkártyákat és egyéb keresőeszközöket. Az indexeléselméletről az első publikációja 1957-ben jelent meg, majd 1964-ben könyvet írt „Indexing theory, indexing methods and search devices" (Indexeléselmélet, indexelési módszerek és keresőeszközök) címmel. Ebben a munkájában az indexelés általános elméletét fogalmazta meg, bevezetve a terminológiai és a kapcsolási kontinuum fogalmát. Elmélete feltárja a szabványosított indexelő nyelvek természetét, funkcióit és az élő nyelvekkel fennálló kapcsolatukat.

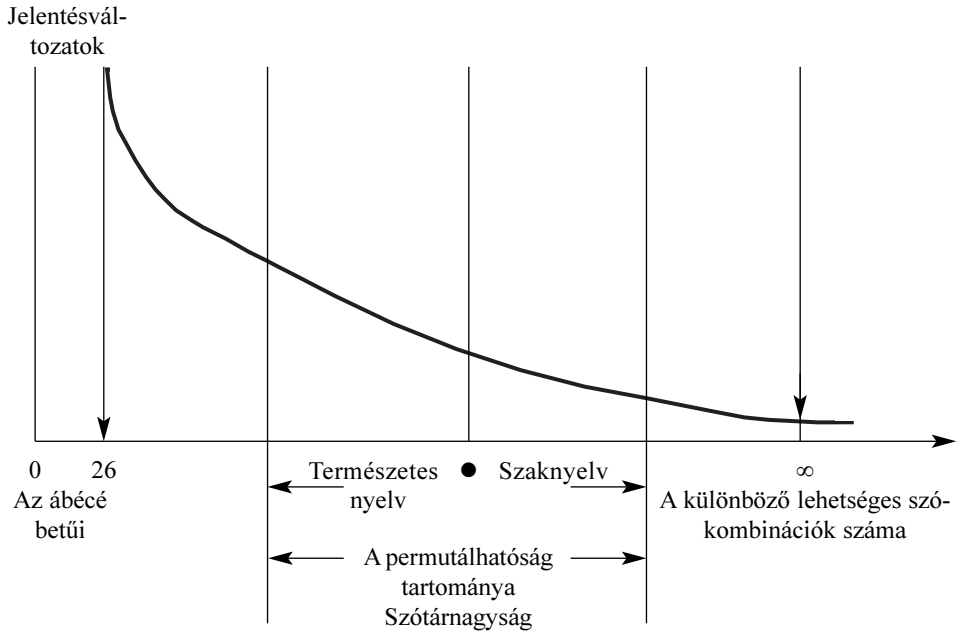
2.1. Terminológiai kontinuum

Jonker diagramját a terminológiai kontinuumról kis módosításokkal az 1. ábra mutatja. *Jonker* elképzelése szerint a terminológiai kontinuum az ábécé betűiből, szavakból és teljes információegységekből épül fel. A kontinuum bal oldalán a latin ábécé 26 betűje áll, melyekből a szavak épülnek fel. Jobbra haladva a kontinuum következő állomását a köznapi beszédben használt „alapszavak" alkotják. Ezeket az alapszavakat ahhoz a szókincshez hasonlíthatjuk, amelyre egy utazónak külföldön szüksége lehet az egyszerű kommunikációhoz. Jobbra tovább haladva elérjük a „közvetítő nyelvet", amelyet az újságokban, a szépirodalomban, a dokumentumprózában és a népszerű tudományos művekben használnak. A kontinuum jobb szélén szerepelnek az egyes foglalkozások, tudományok és mesterségek szaknyelvei – szaknyelvek vagy szakzsargonok –, amelyeket a szakemberek használnak a pontos és egyértelmű kommunikációra. Szakmai szempontból egy repülőgép pontosabban meghatározott valami, mint például delta-szárnyú, szuperszonikus, jet-rendszerű légi szállítóeszköz. Míg a közvetítő nyelv szavai több jelentéssel is rendelkezhetnek, a szaknyelvben arra törekszenek, hogy minden szónak egy és csakis egy jelentése legyen. Ennek érdekében új szavakat alkotnak (radar), régi szavakat újradefiniálnak (program – számítógépes program), a közhasználatú szavakból összetételeket alkotnak (adatbázis) és növelik a szókincs nagyságát.

A diagram felső részén láthatók a könyvtári osztályozásban és tárgyszavazásban, illetve az indexelésben használatos szókincs specifikussága és nagysága közötti arányok. Általában az indexeléshez nagyobb és speciálisabb szókincseket használnak, mint a szakozáshoz.

1. ábra. A terminológiai kontinuum (*Jonker* nyomán)

A szótárnagyság és a szavak specifikussága közötti összefüggést jól szemlélteti a 2. ábra görbéje. Ismét balról indulva az ábécé 26 betűjéből szinte végtelen számú szó alkotható. A szótárnagyság növekedésével nő a szavak specifikusága, és a szavak egyre ritkábban rendelkeznek több jelentéssel, egyre kevésbé fordul elő a poliszémia, a többértelműség. A görbe jobb oldala azt az elméleti esetet ábrázolja, amikor a szótár olyan nagy, hogy minden fogalmat egy és csak egy szó jelez. Mivel az új fogalmak köre folyamatosan bővül, új kifejezéseket kell alkotni, és így elméletileg végtelen számú kifejezés és fogalom létezik. Az ábrán látható, *Jonker* terminológiai kontinuumából származtatható görbékből kiderül, hogy a szaknyelvek specifikusabbak a természetes nyelveknél. A terminológiai kontinuum *Jonker* indexelméletének vázát alkotó alaptörvénye: a szótárnagyság és a fogalom leírásakor szükséges specifikusság közötti összefüggést írja le. Ha ezt az elméletet alkalmazzuk az információtároló és -kereső rendszer, az indexelő rendszer és az ellenőrzött szótár tervezésekor, akkor először azt kell meghatározni, hogy a felhasználói csoport nyelvét hol helyezzük el a terminológiai kontinuumban; azután kerül sor az ellenőrzött szótár megtervezésére, amelynek szókincse ne legyen se általánosabb se specifikusabb, mint a felhasználóé. A gyakorlatban rendkívül nehéz meghatározni a nyelv e szintjét és azt a pontot a terminológiai kontinuumban, amely az átlagos felhasználót jellemzi. További kísérletekre és adatgyűjtésre van szükség. Mindazonáltal az elmélet elősegíti az indexelési problémák megértését és a szótárnagyság, valamint a jelentés specifikussága közötti összefüggés meghatározása segíti az indexelőt a szókincs növekedésének szabályozásában.

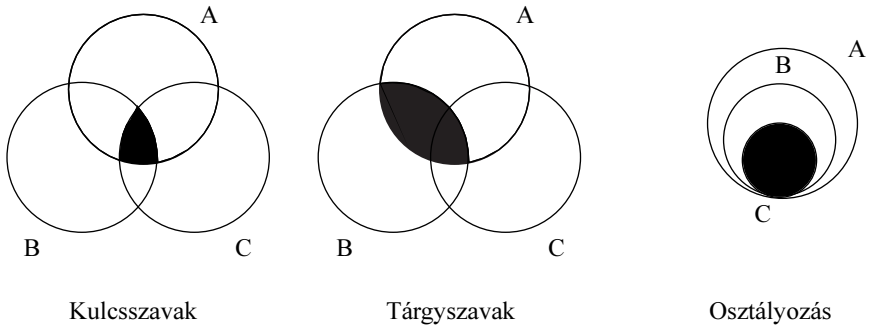


2. ábra. Az egyes kifejezések jelentésbeli variációinak és a szótárban található kifejezések számának az összefüggését ábrázoló grafikon

2.2 Kapcsolási kontinuum

A terminológiai kontinuum a szavakra és nem azok kapcsolatára vonatkozik, a kapcsolási kontinuum pedig a szavak közötti kapcsolatokra. A változót az indexkifejezésre eső szavak átlagos száma képviseli. A szavak közötti kapcsolatok mellérendelésen, ún. posztkoordináción alapulhatnak, ahogy az *Jonker* fénylyukkártyás rendszerében megtestesül, de alapulhatnak hierarchikus osztályozáson vagy a kettő kombinációján is.

A 3. ábrán a kontinuumot a Venn-diagram jelöli. A baloldali besötétített rész az indexelés specifikusságát jelzi, amelyet akkor érhetünk el, ha a három ismérvet „és" kapcsolatba állítjuk, ahogy ezt a következő Boole-algebrai kifejezés mutatja: $A \cap B \cap C$ (olvasva: A és B és C). Ez egyszavas kifejezések – ún. unitermek – teljes posztkoordinációját – utólagos összerendelését – képviseli. Az ábra középső részén A kifejezést B-vel és C-vel rendelik össze előzetesen (prekoordináció) Boole-algebrai kifejezéssel: $A \cap (B \cup C)$ (olvasva: A és vagy B, vagy C). A jobboldali ábra hierarchikusan rendezett szókincset mutat, amelyben A a legáltalánosabb fogalmat képviseli, B A-nak alosztálya, C B-nek az alosztálya, A-nak pedig az al-alosztálya. A keresés $A \supset B \supset C$ kifejezés szerint lehetséges, attól függően, hogy milyen szintű specifikusság kívánatos.



3. ábra A kapcsolási kontinuum diagramja (*Jonker* nyomán)

A keresés kapcsolódási mértéke a kontinuum baloldalán éri el maximumát, ahol egyszavas kifejezéseket használnak, és minden szó kapcsolódhat minden másik szóhoz. A kontinuum másik végén a kapcsolhatóság mértéke a nullához közelít, mivel minden kifejezés része egy általánosabbnak, ahogy ez egyszerű hierarchikus osztályozási rendszer esetében szokásos. A hierarchikus alárendelés a kontinuum jobb oldali végén éri el maximumát, bal szélen pedig közelít a nullához. A kontinuum közepe nagyjából megfelel annak a területnek, amelyet a legtöbb összetett tárgyszóval kifejezett ismérv meghatároz.

A kapcsolási kontinuum kifejezi a leírandó fogalom specifikussága és az ismérvekben használt szavak száma közötti funkcionális kapcsolatot, továbbá a rendező elvet, amely segítségével a szavakat kifejezésekké kombináljuk. *Jonker* írja: „Bár matematikailag pontosan még nem mutatható ki, világosnak látszik, hogy nagyobb mértékű hierarchikus meghatározottság csak a keresési hatékonyság rovására érhető el és viszont. Amennyiben X a keresési hatékonyságot és Y a hierarchikus meghatározottságot jelöli, e kapcsolat nagyjából így fejezhető ki: XY^2 konstans".

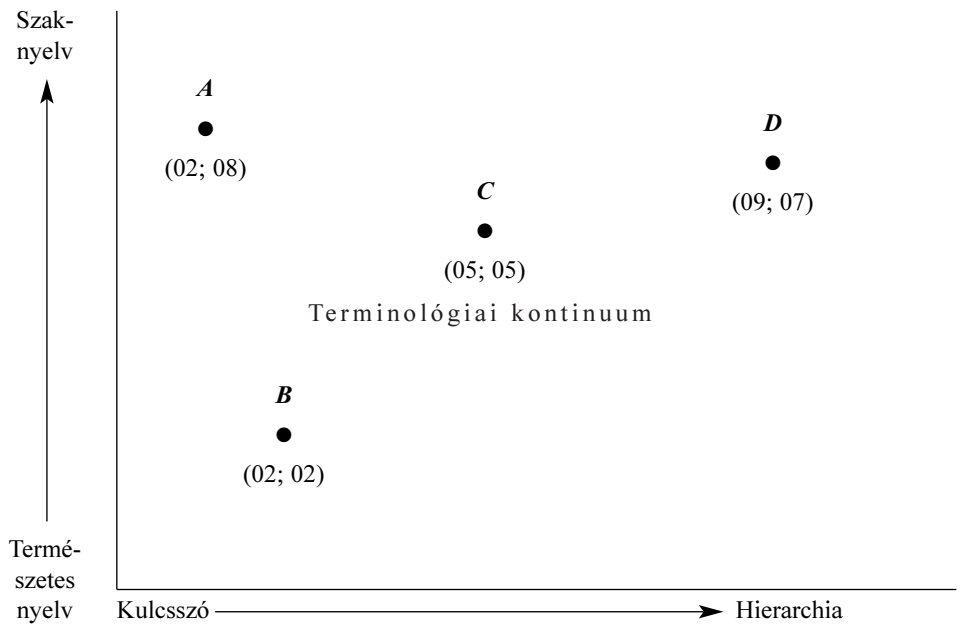
2.3. A terminológiai és kapcsolási kontinuum összekapcsolása

Az általános indexeléselmélet mindkét megfogalmazott „kategóriát": a terminológiai és kapcsolási kontinuumot egyaránt magába foglalja. Kapcsolatukat a 4. ábrán látható kétdimenziós diagram mutatja. Az információkereső rendszert úgy írhatjuk le, ha e rendszert mindkét kontinuumban elhelyezzük. A terminológiai kontinuumban elfoglalt helyet az alkalmazott szókincs milyensége határozza meg: természetes nyelvről, szaknyelvről vagy a kettő valamilyen kombinációjáról van-e szó? A szókincs komplexitási fokát 0 és 1 közötti számértékkel fejezik ki, mely érték az y tengelyen elfoglalt helyet jelöli ki.

Az indexelési rendszer típusa, vagy pontosabban az indexelési rendszerben található hierarchikus kapcsolatok mértéke határozza meg a rendszer helyzetét az x tengelyen. A kapcsolási kontinuumban elfoglalt helyzet értékei szintén 0 és 1 között változhatnak, attól függően, hogy a kifejezések közötti hierarchikus rendezettség milyen mértékű. Ez adja tehát az értéket.

Visszatérve a 4. ábrához: az A rendszerről leolvasható, hogy egyszavas (uniterm) kifejezéseket tartalmazó szótárral rendelkezik, melynek nyelve specifikus szaknyelv. A két kontinuumban helyzetét a (0,2; 0,8) jelöli. A rendszer szintén kulcsszavakat használ, de túlnyomórészt alapvető szókincset tartalmaz. Helyzetét a (0,2; 0,2) értékek fejezik ki. A C rendszer tárgyszavakat használ az indexelésre, amelyek bizonyos hierarchikus kapcsolatokat is tükröznek. A szótár nyelve közvetítő nyelv, (0,5; 0,5) értékek jelölik ki. A D rendszerben szaknyelven alapuló hierarchikus osztályozási rendszert használnak. Helyét a (0,9; 0,7) értékek jelölik ki.

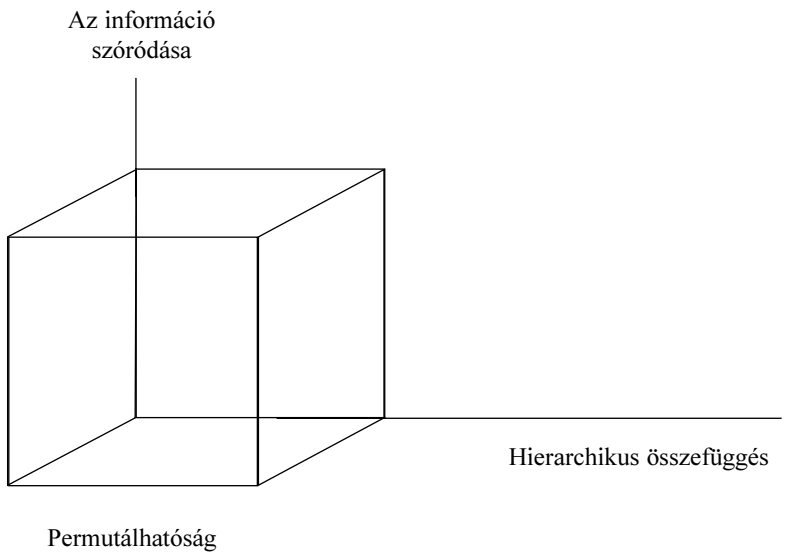
Jonker általános indexeléselmélete lehetővé teszi, hogy a terminológiai és kapcsolási kontinuum szempontjából leírjuk és elkülönítsük a különböző indexelő rendszereket. Így láthatjuk, megérthetjük és ábrázolhatjuk a szótárnagyság, a szókincs hierarchikus szervezettsége és a fogalmak leírásához szükséges specifikusság közötti kapcsolatot. Mindazonáltal nem kapunk minden kérdésre választ, amely az indexeléssel kapcsolatban felvetődik.



4. ábra A terminológiai és a kapcsolási kontinuum együttes ábrázolása (Jonker nyomán)

3. Heilprin: Jonker indexeléselméletének módosítása

Lawrence Heilprin ugyancsak sokat tett az információs rendszerek elméletének tudományos megalapozásáért. Röviddel azután, hogy *Jonker* publikálta írását, *Heilprin* megkísérelte *Jonker* modelljét formálisabb szerkezetűvé és matematikailag megalapozottabbá tenni. Bevezette a keresőút fogalmát, amely megfelel a keresőkérdésektől a dokumentumokhoz vezető lehetséges utak számával. Ő vezette be a zaj fogalmát is, amely akkor keletkezik, ha az indexelés permutálhatóságának vagy hierarchiájának elméletileg lehetséges határait bármiképp átlépjük. Ami a legfontosabb: *Jonker* kétdimenziós modellje helyett az indexelés háromdimenziós modelljét javasolta. Rámutatott, hogy amikor a rendszer modellbeli elhelyezkedését keressük, inkább görbékkel, mintsem egyedülálló ponttal ábrázolhatjuk „n" értékét, amely a független kifejezések dokumentumonkénti átlagos számát jelenti a deskriptív kontinuumban.



5. ábra Az indexelés tartománya (*Heilprin* alapján)

Heilprin modellje három változót tartalmaz:

D = az információ szórása, amelyet indexeléskor a dokumentumhoz rendelt független kifejezések átlagos számával mérnek;

F = permutálhatóság, amely az ismérvek lehetséges kombinációinak száma;

H = hierarchikus összefüggés, amely az (általa indexelő rendszernek nevezett) információkereső rendszer hierarchikus szervezettségének mértékén alapul.

Az 5. ábra *Heilprin* háromdimenziós modelljét ábrázolja sematikusan. A doboz az indexelés tartományát jelenti, amely a D, P és H értékek által alkotott lehetséges információkereső rendszert magába foglalja. Bár a rendszer még nem teljes, az elemző ábrázolás ígéretesnek látszik.

4. Landry: Az átfogó indexeléselmélet

Körülbelül 10 év telt el *Jonker* és *Heilprin* kezdeti munkája és *Landry* újabb keletű kísérlete közt, amely az általános indexeléselmélet megalkotására, illetve arra irányult, hogy szilárdabb alapot teremtsen az információtároló és -kereső rendszerek számára. *Landry* vizsgálatait az Ohioi Állami Egyetemen folytatta, részben doktori disszertációja előkészítésekor. Arra nincs lehetőség és nem is lenne célszerű, hogy e helyen a teljes elméleti modellt ismertessük. Inkább, ahogy ezt a többi modellnél is tettük, az ismertetést néhány lényeges jellemzőre korlátozzuk.

Landry azzal kezdi modelljének kifejtését, hogy rámutat az indexelési folyamat és a kommunikáció általános menetének hasonlóságaira. Egy sor elvi definíciót és posztulátumot fogalmaz meg. Munkája három fogalmi osztály meglétének előrejelzésén alapul bármely információtároló és -kereső rendszerben. Ezek: a dokumentumállomány, a dokumentum tartalmi tulajdonságainak állománya és a kettő közötti kapcsolatot kifejező relációk állományai. Az elmélet az adatelemek fogalmára épül. E fogalom három fontos jellemzőjét adja meg.

(1) Az adatelem függetlenül kezelhető. Az adatfüggetlenség azt jelenti, hogy az adatokkal összefüggő jellemzők változása nem igényli a kezelő rendszer – például a programrendszer – szerkezetének megváltoztatását (a szerk.).
(2) Az adatelem nem bontható két vagy több egységre.
(3) Az adatelemnek határozott jelentése vagy értelmezése van.

Bár az adatelem bármilyen dolog lehet, szólánc, cím stb., az osztályozó (indexelő) rendszer azoknak az adatelemeknek a meghatározására irányul, amelyek felismerhetők és feldolgozhatók a rendszerben.

A dokumentum az adatelemek célszerűen rendezett állománya. Az indexelő rendszer inputjait képviselő dokumentumok alkotják a dokumentummezőt. Az index – a mutató – pedig az indexelési folyamat outputja. Ez elvezet *Landry* 2.2 tételéhez, amely megállapítja: a mutatónak – definíciójából fakadóan – meg kell őriznie az alapdokumentumok és a dokumentummezők célszerű rendezettségét. A mutató tehát az adatelemek célszerűen szervezett (közönségesen: ábécé szerint rendezett) készlete.

Az indexelési folyamatnak és az indexelő rendszernek egyaránt az a szerepe, hogy a dokumentum szövegén belüli adatelemeket és relációkat a rendezettséget

megőrző transzformáció segítségével teljes mértékben specifikálja. E ponton a legcélszerűbb *Landry* saját elméleti összefoglalójához fordulni. Mivel Landry beszámolójában speciális tételszámokra is utal, az alábbi listát alkalmaztuk:

(1) $I = f(\mathcal{D},\mathcal{I})$ ahol

I – index,
f – indexelési folyamat,
$\mathcal{D}$ – dokumentummező: a dokumentumok rendezett állománya,
$\mathcal{I}$ – mutatómező: az adatelemek és az indexelő rendszer relációinak ábrázolása.

E tétel állítása szerint a mutató a megengedett adatelemek és relációk megjelenítése, és egyben az adatelemek (és ezáltal a dokumentumok) jól szervezett állománya.

(2) A pontos keresés az indexelés pontosságától függ. (Ez *Landry* 2.2 tételének összefoglalása; lásd előbb.)
(3) A kommunikációban elemenként átvitt egységek az adatelemek és a hozzájuk kapcsolódó relációk.
(4) A kommunikáció bármely elméletét vagy gyakorlatát, amelynek során adatelemek vesznek el, akár azok téves ábrázolása, akár a folyamat korlátozása miatt, inadekvátnak kell tekinteni.
(5) Az indexelő rendszer biztosítja a megállapított kapcsolatteremtő halmazt, és a hatékony kommunikáláshoz szükséges transzformációkat.
(6) A O transzformáció az adatelem-modelleket határozza meg.
(7) Az indexelő rendszer felismeri az egyes dokumentumokon belüli és dokumentumok közti adatelem-kapcsolatokat.
(8) A jelenlegi indexelési gyakorlat során a dokumentumokban az adatelemek közötti rendezettség összekuszálódik.
(9) Az indexelő rendszer felismeri és kifejti a dokumentumon belüli adatelem-kapcsolatokat.
(10) Az adatelemek adott állományából nyerhető információ valószínűsége az indexelő rendszer által ráfordított tevékenység függvénye.
(11) Adott D-nél ténylegesen felhasznált indexkifejezések száma kisebb, mint az elméletileg lehetségeseké, mivel a dokumentummező indexelése nem teljes.
(12) A keresés (lekérdezés) maximális és minimális útjainak előfordulási valószínűsége kicsi.
(13) Az adatelemek hasznosságcsökkenésének mértéke Poisson-eloszlást mutat.
(14) Az adatelemek értékmegoszlásuk alapján nem különböztethetők meg.

(15) Az újonnan keresett és talált adatelemek hasznosságcsökkenési hányada csökken, ha a H–D–G szerkezetben a keresési út hossza nő.

Ezek a posztulátumok jelzik, hogy *Landry* indexeléselmélete mennyire átfogó. Az elmélet egzakt jellegét matematikai képleteivel mutathatnánk be. Az elmélet az indexelés és a mellérendelt osztályozás általánosítását nyújtja. Ám sok kérdés maradt megoldatlan. A reprezentáció és transzformáció speciális operációs rendszereket igényel. Ki kell próbálni az elméletet, hogy alkalmazható-e operációs rendszer esetében. Több adatot kell gyűjteni különösen a célból, hogy meghatározzuk az alternatív indexelési módszerek hatását a rendezettség mértékére és az adatelemek típusára, amelyeket különböző indexelési folyamatok megőriznek vagy elvetnek. Talán egy szimulációs modellt kellene felépíteni és tesztelni különböző feltételek között.

5. Salton: indexeléselmélet

Az indexeléselmélet legújabb művelője *Gerard Salton,* aki informatikusként az információtárolás és -keresés automatizálását kutatja. 1975-ben fejezte be monográfiáját „A Theory of Indexing" (Indexeléselmélet) címmel, amelyben összefoglalja a számítógépes dokumentumtároló és -kereső rendszerek értékeléséről és fejlesztéséről szóló saját és más szerzők írásait, különös hangsúlyt helyezve a nyelvek hatékony szótárainak szükségességére. Ennek során (1) meghatározza az ismérvek vagy kulcsszavak szerepét a dokumentumállomány strukturálásában, (2) módszereket javasol a jó és rossz mutatókialakítás jellemzőinek mérésére, (3) javaslatot ad indexelő nyelvek szótárának módosítására, a keresés hatékonyságának javítása érdekében. Salton munkája egyszerre elméleti és gyakorlati. A monográfiában az elméletet matematikai kifejezésekkel írja le, hogy bizonyítsa általános érvényét és belső összhangját. Beszámol az elmélet kísérleti teszteléséről három különböző adatbázison. Az itt következő fejezet *Salton* indexeléselméletének újabb oldalait világítja meg a matematikai formulák és a kísérleti bizonyítékok minimális alkalmazásával.

5.1 Az indexelés formális (képletszerű) definíciója

Indexeléskor dokumentumokat és kifejezéseket rendelnek egymáshoz. Ezek a kifejezések tulajdonságtípusokat[39] képviselnek, és a dokumentum információtartalmának kifejezésére választják ki őket. A tulajdonságtípusokhoz

39 Az eredetiben „attribute", tulajdonság; a tulajdonságtípus azonban pontosabb, ezért ezt alkalmaztuk. (a szerk.)

értékek is tartozhatnak, így például a személyzeti nyilvántartásban a tulajdonságtípus lehet az alkalmazotti kategória elnevezése, munkaköri besorolása, fizetése stb., a tulajdonságok értékei pedig az alkalmazott személy neve, konkrét besorolása, tényleges fizetése stb. *Ranganathan* kettőspontos osztályozási rendszerében a tulajdonságtípusokat a következő rövidítés tükrözi: PMSET – melynek feloldása: Egyediség (**P**ersonality) – Anyag (**M**atter) – Tér (**S**pace) – Mozgás (**E**nergy) – Idő (**T**ime). A tulajdonságértékek pedig azok a specifikus kifejezések, amelyekkel ezeket a tulajdonságtípusokat egy-egy dokumentumban leírjuk. Még általánosabban: a tulajdonságértékek lehetnek kulcsszavak, ismérvek vagy deszkriptorok stb. Mindenesetre az indexelési folyamat minden dokumentum esetében dokumentumvektort eredményez

$$D_i = (a_{i1}, a_{i2}, a_{i3}, a_{i4})$$

ahol D_i dokumentum a_{i1} pedig a D-hez rendelt tulajdonságértéket – a konkrét ismérvet – jelentő érték.

Ha a tulajdonságértékeket is, azaz ha az ismérveket súlyozzuk a dokumentumazonosítás szempontjából való fontosságuk alapján, a dokumentumvektor így alakul:

$$D_i = (a_{i1}, w_{i1}; a_{i2}, w_{i2}; \ldots ; a_{in}, w_{in}).$$

A fenti definíció, amely az indexelési folyamatra és a dokumentumvektorra vonatkozik, képletszerű, az indexelést egyszerűen, formális matematikai kifejezésekkel írja le. Semmivel sem bővíti vagy szűkíti az eddigi ismereteket. A formális kifejezés értékét az adja, hogy felhasználható más értékek kiszámítására, amelyek azután betekintést adnak az indexelés folyamatába.

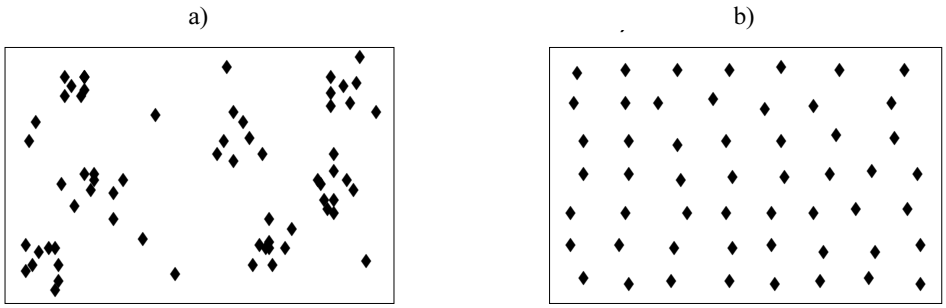
5.2 A dokumentumok közötti hasonlóság vizsgálata

Egy adott indexelt állományon belül a dokumentumpárok közötti hasonlóság mértéke kiszámítható a megfelelő tulajdonságok vektorpárjainak összevetésével. Az alábbi képlet használható erre

$$s (D_i, D_j) = \sum_{k=1} a_{ik}\, a_{jk}$$

ahol s (D_i, D_j) a D_i és D_j dokumentum közötti hasonlósági együttható. Számértéke 0-tól 1-ig változhat: 0 a teljes különbözőséget, 1 pedig a teljes hasonlóságot jelenti. Elképzelhető a –1 együttható is, amely szintén hasonlóságot jelez, de az ellenkező irányban, azaz a_{i1} = hátrányos, a_{j1}= előnyös.

Tegyük feI, hogy egy adott állományban minden dokumentumot összehasonlítottunk egymással, és kiszámítottuk a párok közötti hasonlóság mértékét, az eredményeket pedig ábrázoltuk.



6. ábra. Tipikus mezőszerkezetek: (a) klaszterált elrendeződés; (b) elkülönült tételek

Tipikus állományelrendeződés látható a 6. ábra (a) részén. Ebben a példában w jelöli a dokumentumokat, és a két dokumentum közötti távolság fordítottan arányos a dokumentumvektorok hasonlóságával.

A dokumentumok halmokat, „rakásokat" ún. klasztereket képeznek, amelyekben a hasonló egységek osztályai jól elkülöníthetők a többi egységtől. *Salton* rámutat arra, hogy ez a csoportos eloszlási forma annak eredménye, hogy az indexelő nyelv szókincsét alkalmazták jó hatásfokkal. Megismételve a lényeget: *az indexelő nyelv szótára akkor hatékony, ha a tárgyi teret olyan klaszterekre bontja, hogy egy tétel keresése ugyanazon klaszter hasonló egységeihez vezet.* Ugyanakkor a nem releváns tételek nagyobb távolságra helyezkednek el, így könnyen kizárhatók, ezáltal nagyfokú pontosság érhető el.

Bár eddig csak a dokumentumok közötti hasonlóság mértékével foglalkoztunk, könnyen kiszámítható a dokumentumok és a kérdések közötti hasonlóság mértéke is. A kérdést egyszerűen dokumentumként kezeljük, és így dokumentumvektorral le tudjuk írni. A keresési folyamatban a kérdésvektort hasonlítjuk össze a dokumentumvektorokkal, és kiszámítjuk a hasonlóság mértékét. Ha a kérdés a dokumentummezőn belül valamelyik klaszterbe esik, valamennyi dokumentum, amely ebben a klaszterben található, hasonlóságot mutat a kérdéssel és kereséskor megjelenik a válaszok között.

A klaszterelemzéssel kialakított mezőszerkezettel szemben (6/a ábra) mivel nem tudjuk előre mérlegelni, hogy mely kérdésekhez mely releváns dokumentumok kapcsolódnak, a legjobb eljárás, ha a dokumentumokat – amennyire csak lehetséges – elkülönítjük egymástól (6/b ábra). Ha a dokumentumok egy mezőn belül elkülönülnek, nagyfokú pontosság, sőt nagyfokú teljesség érhető el, mivel minden egyes dokumentum megtalálható, anélkül, hogy összes lehetséges nem releváns szomszédja is felbukkanna.

5.3 Az indexelő nyelv szókincsének jellemzői

Az indexelő nyelv szókincsének jellemző tulajdonságai a specifikusság és a kimerítő teljesség. A specifikusság a részletezésnek azt a szintjét jelöli, amelyen a fogalmak a vektorban szerepelnek. A kimerítő teljesség pedig azt a teljességet jelenti, ahogy a dokumentum releváns tartalma az indexelő nyelv szótárában és a dokumentumvektorban szerepel. Amikor specifikus indexeléssel keressük a dokumentumokat, az állományból viszonylag kevés, de nagyon releváns dokumentumhoz jutunk. A kimerítő indexelés sok találatot idéz elő, amelyben sok releváns és sok nem releváns dokumentum is szerepel. Az indexelő nyelv szókincsének tulajdonságai, a specifikusság és a kimerítő teljesség részben magyarázzák a teljesség és pontosság gyakran emlegetett fordított kapcsolatát.

Megállapítottuk, hogy az ismérvek specifikusságukkal és/vagy teljességükkel jellemezhetők: most fordítsuk figyelmünket a jó ismérvek jellemző tulajdonságaira. *Salton* szerint „a legjobb kifejezések azok, amelyek az állomány bizonyos specifikus tételeiben hangsúllyal szerepelnek, ugyanakkor előfordulási gyakoriságuk a teljes állományban általában kicsi". Ebből az következik, hogy egy ismérv szerinti keresési hatékonysága az adott állományon belüli használatának gyakoriságával mérhető. Ennek egyik mérési jellemzője a kommunikációelméletből vett jel/zaj viszony. Egy kifejezésnek akkor nagy a jel/zaj aránya, ha nagyon specifikus és az állomány nagyon kevés dokumentumában fordul elő, ugyanakkor egy kifejezés akkor nagyon „zajos", ha az állományban a dokumentumok nagy százalékának jellemzésére használták. A legszélsőségesebb esetben, amikor a kifejezést az állomány minden dokumentumánál alkalmazták, a jel nulla, a zaj pedig maximális.

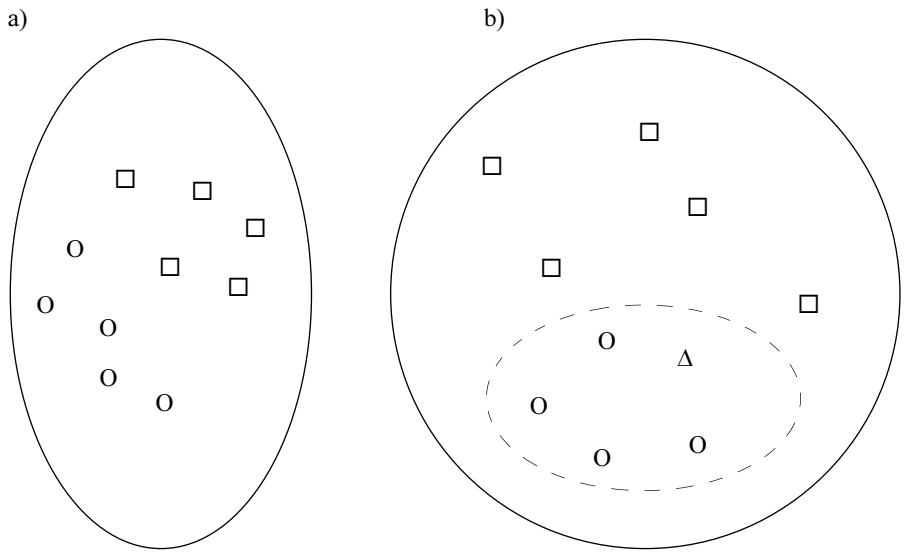
A zaj és a kifejezés specifikussága között olyan kapcsolat van, hogy az átfogó nem-specifikus kifejezések meglehetősen egyenletesen oszlanak el az állományban és ezáltal nagy a zaj, míg a specifikus kifejezések nagyon elszórtan és ritkán fordulnak elő és magas a jelértékük. Ebből *Salton* szerint az következik, hogy a „jel/zaj" számítások felhasználhatók az indexelő nyelv optimális szókincsének létrehozásakor, úgy, hogy törölni kell azokat a kifejezéseket, amelyek jel/zaj értéke rendkívül alacsony.

5.4 Az ismérvek diszkriminációs értékének javítása

Egy mutató diszkriminációs értéke megmutatja, hogy egy kifejezés milyen mértékben segíti elő a releváns dokumentumok megkülönböztetését a nem relevánsaktól. A diszkriminációs érték modelljét célszerű grafikusan ábrázolni (7. ábra), hogy az indexelési folyamatot fizikailag értelmezni tudjuk.

A 7/a ábra tipikus dokumentumkörnyezetet ábrázol. Megjegyzendő, hogy itt kevéssé különülnek el a releváns és nem releváns dokumentumok. A 7/b ábrán a nagyobb diszkriminációs értékű (tehát specifikusabb jelentésű) kifejezések beépültek a dokumentumvektorokba, és a dokumentummező kiterjedtebb. Ha a kérdésvektort alkalmazzák, a keresésnek nagyobb a megkülönböztető hatása, azaz nagyobb teljességgel és pontossággal hajtható végre.

Az ábrák jól szemléltetik az elméletet, de nem sugallnak megoldást. Megállapíthatjuk, hogy mit kell tenni, de nem feltétlenül állapítható meg, hogyan. *Salton* számos módszert javasol az ismérvek specifikussági fokának optimalizálására, a jel/zaj arány javítására egy adott állományban. Azt javasolja, hogy a túl általános, túl gyakran előforduló kifejezéseket specifikusabbá kellene tenni oly módon, hogy más kifejezésekkel kombináljuk őket összetett kifejezéseket alkotva. Azokat a kifejezéseket pedig, amelyek túl specifikusak (túl kicsi az előfordulási gyakoriságuk), általánosabbá kell tenni, szinonima vagy rokon kifejezések hozzáadásával, amelyeket a tezauruszból választhatunk ki a dokumentumvektor számára.



7. ábra Modell a kifejezések megkülönböztetésére
□ – nem releváns dokumentum; O – releváns dokumentum;
Δ – kérdés; O – keresési tartomány
(a) eredeti mező; (b) diszkriminátorok kijelölésével kiterjesztett mező

Salton formális és gyakorlati bizonyítékokat tett közzé az indexeléselmélet eme aspektusának hatékonyságára vonatkozóan és leírta a fent már említett feltételeket.

6. Előrelépés és a jelenlegi helyzet

Az átfogó indexeléselmélet kidolgozása komplex feladat, amely számos információtudományi szakember munkájából alakul ki, és a különféle koncepciókat egységes szerkezetbe integrálja. Az elmélet felépítése erősíteni fogja az információtudomány alapjait és elősegíti a dolgok megértését és a kísérletezést. Az indexeléselmélet segíteni fog továbbá az indexelés természetének és jó hatásfokú alkalmazásának megmagyarázásában. Elősegíti az egyes vagy talán az összes indexelési munkafolyamat automatizálását, és előmozdítja azokat a törekvéseket, amelyek az indexelési előírásokra és szabványosításra irányulnak. A jó elmélet legalábbis megkönnyítené az indexelés tanítását és tanulását.

E tanulmány *Jonker, Heilprin, Landry* és *Salton* elméletépítő törekvéseit vizsgálja. Azért választottuk e szerzőket, mert munkáikban főként az indexelés természetével és az indexelő nyelv szókincsének jellegzetességeivel foglalkoztak. Más szerzők az indexeléselmélet egyéb aspektusairól írtak, vagy talán más nézőpontból vizsgálták ugyanezeket, illetve a kapcsolódó szempontokat. *Maron, Kuhn, Swets, Cooper, Bookstein* és *Swanson* munkáit nem érintettük ebben a bevezető elemzésben, jóllehet fontos adalékokkal járultak hozzá az indexeléselmélethez. Mindezeket a munkákat beépíteni ebbe a tanulmányba túl nagy és nehéz feladat lett volna. Ugyanezért nem vizsgáltuk *Garfield, Kessler* és *Liptz* a hivatkozási indexek elméletével foglalkozó munkáit sem.

6.1 Definíciók

Az indexelés formális definíciójáért akár *Landryhez,* akár *Saltonhoz* folyamodhatunk, kettőjük definíciója ugyanis hasonló.

A mutató (az index) az adatelemek célszerűen szervezett állománya. Az adatelem egy szó vagy kifejezés, amelynek (a) meghatározott a jelentése, (b) nem bontható szét két vagy több értelmes egységre, és (c) más adatelemektől függetlenül kezelhető.

A jó mutató megőrzi az eredeti dokumentum adatelemeinek rendszerességét. A rossz indexelés adatelemek kihagyásának vagy nem reprezentatív adatelemek használatának a következménye.

6.2 Jellemzők

Jonker szerint az ismérvet – az indexkifejezést vagy mutatónevet – a terminológiai kontinuumban elfoglalt helye jellemzi, amely az általánostól a specifikusig terjedhet. *Salton* a kimerítő teljesség és a specifikusság kifejezéseket használja, de koncepciójuk hasonló.

Ezek a fogalmak közvetlenül kapcsolódnak a szótárnagysághoz és a kifejezések használati gyakoriságához. Ha az indexelő nyelv szókincsét elsősorban általános kifejezésekből állították össze, akkor a nyelv terjedelme kicsi lesz, és úgy növelhető, hogy a specifikusságát növeljük.

Az ismérvek kombinálhatók. *Jonker* leírja a kifejezések közötti lehetséges kapcsolatok modelljét a kapcsolási kontinuum segítségével, és kétdimenziós mezőben ábrázolja ezeknek a kifejezés-kombinációknak a hatásait: a terminológiai és a kapcsolási kontinuumokat, valamint a nyelvi jellemzőket a Boole-algebrával összekapcsolva. *Salton* kimutatja, hogy a specifikusság többszavas kifejezésekkel növelhető. *Heilprin* ezzel szemben háromdimenziós indexmező-modellel operál, amelyben a fő változók: a szórás, a permutálhatóság és a hierarchia.

6.3 Hatékonyság és minőség

Ha formálisan meghatároztuk a mutatót, és azonosítottuk az indexelő nyelv szótárának legfontosabb jellemzőit, az indexeléselmélet segítségével meghatározhatjuk a mutató minőségi tulajdonságait, és különböző kifejezések jellemzőinek változtatásával ezeket megjavíthatjuk.

Jó minőségű az a mutató, amellyel kereséskor nagy teljesség és nagy pontosság érhető el. Ez azonban csupán a cél megfogalmazása, de nem mond semmit a hogyanról. Mi a feltétele a jó mutatónak, és mi jellemzi azt? *Salton* szerint jó minőségű az a mutató, amellyel nagy teljesség és nagy pontosság érhető el, és bármely dokumentumállományban alkalmas arra, hogy a hasonló vagy egymással összefüggő tételeket klaszterelemzésre alkalmas csoportokba rendezze. Gyengébb minőségű az a mutató, amely a tárgyi mezőben egyenletes dokumentumeloszlást idéz elő, mivel ennek a teljesség alacsonyabb foka a következménye; a pontosság ebben az esetben is nagy lehet.

A mutató minőségének eme meghatározásai tanulságosak: arra következtethetünk belőlük, hogy az indexelés minősége a klaszterelemzésre alkalmas csoportokon belüli kohézió és a csoportok közötti elkülönülés növelésével javítható.

A klaszterelemzés az alkalmazott ismérvek jellemzőin – általános és specifikus jellegükön alapszik. Ha adott állomány valamennyi dokumentumát ugyanazzal a kifejezéssel – például kémia – indexelnénk, az eredmény egyetlen nagy csoport lenne. Ha viszont: az indexelő nyelv olyan specifikus lenne, hogy egy-egy kifejezést csak egyetlen dokumentumhoz rendelnénk az állományban, akkor nem képződnének csoportok. Minden ismérv csoportképző hatása és közvetve a kifejezés keresési hatékonysága a jel/zaj aránnyal mérhető, amely az állományban előforduló kifejezések használatának gyakoriságán alapszik.

Az ismérvek keresési hatékonysága növelhető a jel/zaj arány módosításával, azaz a kifejezések specifikusságának vagy általános jellegének változtatásával. A túl általános ismérvek úgy tehetők specifikusabbakká, ha kombináljuk őket más kifejezésekkel, többszavas kifejezéseket képezve. A túl specifikus kifejezések pedig úgy tehetők általánosabbakká, hogy szinonimákat vagy rokon kifejezéseket adunk hozzájuk a tezauruszból.

6.4 Helyzetkép

Ezek között az elméletek között figyelemreméltó az összhang, ami arra enged következtetni, hogy az indexelés általános elméletére nem kell már sokat várni. De minden elméletet bizonyítani is kell, és ezen a téren viszonylag kevés történt. *Jonker* és *Heilprin* alapvető koncepciókat fogalmazott meg. *Landry* bebizonyította, hogy posztulátumai ésszerűen teljesek és következetesek. *Salton* néhány kísérleti bizonyítékot tett közzé elméletének alátámasztására, de még sokkal többre lenne szükség.

A fejezet elején felvetettünk jó néhány kérdést, amelyeket szándékosan mennyiségileg fejeztünk ki. Ezek csak kísérletekből származó empirikus adatokkal válaszolhatók meg, amely kísérletek célja az indexeléselmélet kipróbálása, következtetések levonása kell, hogy legyen. Erre vonatkozó adatok még nincsenek, sok kísérlet szükséges még. Az indexeléselmélet – akármilyen kezdetleges is – segíti a kísérletek tervezését, amelyek révén jobban megérthetjük az indexelés természetét, előre jelezhetjük és szabályozhatjuk az indexelő nyelv szótárának a hatásait.

Irodalom

Landry, B. C.: A theory of indexing: indexing theory as a model for information storage and retrieval. Ph. D. Dissertation, Ohio State University, Clumbus, Ohio, 1971.

Jonker, F.: The descriptive continuum, a „generalized" theory of indexing. Jonker Business Machines, Inc., AD–132–238. 1957.

Jonker, F.: Indexing theory, indexig methods an search devices. The Scarecrow Press, New York, 1965.

Heilprin, L.: Mathematical model of indexing. Documentation Assoc., AD–136–477. 1957.

Salton, G.: A theory of indexing. Regional Conference Series in Applied Mathematic, Society for Industrial and Applied Mathematics. Philadelphia, 1975.

BERTRAM CLAUDE BROOKES

A szerző az University College (London) munkatársa. Tanulmányában egyrészt általános szinten elemzi az információrendszerek fejlődését, megkülönböztetve a gépesítés első – korukban megvalósuló – szakaszát, és a második - magára még váratő, jövőbeli - szakaszát. Ebből a – prognosztikai – szempontból írása *Vannewar Bush* 1945-ben írt programadó, „Úgy, ahogy gondolkodunk" (As we may think) című tanulmányának szerényebb megismétlésének is tekinthető. Másrészt – *Norbert Wieneri*-alapokon – behatóan foglalkozik az információ és a tudás viszonyával. Ebből a szempontból tanulmánya az információkeresés szakterületének egyik alapkérdésére világít rá.

Az informatika mint alapvető társadalomtudomány[40]

1. Az információs rendszerek gépesítésének első szakasza

Az informatika akkor került először igazán az érdeklődés homlokterébe, amikor húsz évvel ezelőtt *Cyrill Cleverdon* és munkacsoportja közzétette híressé vált cranfieldi jelentését, amelyben az akkor ismert információkereső nyelvek keresési hatékonyságának vizsgálatáról számoltak be.[41] *Cleverdon* meglepő következtetései meglehetősen különböző elméleti kérdéseket vetettek fel, amelyeket azonban csoportosítani lehetett három különböző szinten.

A legközvetlenebbül elérhető szinten azok a statisztikai és matematikai problémák álltak, amelyek az alkalmazott mérési módszerekre és az információkereső rendszerek matematikai modelljeinek a kidolgozására vonatkoztak úgy, hogy a méréseket és azokat az információkereső folyamatokat, amelyekhez a mérési eljárások kapcsolódtak, részleteikben jobban meg lehetett érteni. Akik rendelkeztek az elemzéshez szükséges matematikai ismeretekkel, azoknak vonzóak voltak ezek a problémák. Rengeteg adat állt rendelkezésre, s a matematikus anélkül láthatott neki a munkájának, hogy meg kellett volna kérdőjeleznie az információkereső folyamatra vonatkozó, a matematika és a statisztika alapján tett feltevéseit.

A második és valamivel mélyebb szinten *nyelvi* problémák álltak. Az alkalmazott nyelvtudomány problémái nehezebbek, mint az alkalmazott matematikáé, mivel az alkalmazott nyelvésznek sokkal kevesebb közvetlenül hasz-

40 Informatics as the fundamental social science. In: New trends in documentation and information: Proceeding of the 39th FID Congress, University of Edinburgh, 25–28 September 1978. Ed. by Peter J. Taylor. London: ASLIB: FID, 1980. – (FID Publicaton 566.) p. 19–29.

41 A kísérletekről magyarul lásd Horváth Tibor: A második cranfieldi jelentés. In: Könyvtári Figyelő, 1968, 14. köt. 5. sz., p. 351–369. (a szerk.).

nálható egzakt elmélet áll a rendelkezésére. Bár ebben az időben a nyelvtudomány elmélete látványosan kibontakozott, a *Chomsky* nevéhez fűződő új elmélet fejlődése még máig sem zárult le, és ezért az információkeresés elméletében még nem használható fel.

A harmadik és még mélyebb szinten a felhasználói igények és az információs rendszerekkel elérthető információk kölcsönhatását érintő kérdések állnak. Ezek a problémák összefonódnak a lelki, a társadalmi és a gondolkodási folyamatok megismerésének alapvető és még jórészt megoldatlan kérdéseivel. Hogyan tanulunk, hogyan bővíthetjük a legjobban tudásunkat? Ezeket a kognitív, a gondolkodással összefüggő problémákat 1000 év óta tanulmányozzák a filozófusok és a pedagógusok és még inkább – az utóbbi időben – a pszichológusok. Az információs rendszerek további gyakorlati fejlődése azonban hirtelen újra szembeállít bennünket velük.

Az információs munka gyakorlati világa mégsem vár választ ezekre az elméleti kérdésekre. Ahogy a technika gyakorlati világában gőzzel működtetett vasutakat és több flottányi óceánjáró gőzhajót tudtak építeni jóval azelőtt, hogy a gőzgépek termodinamikai elméletét kidolgozták volna, ugyanúgy az információs munka gyakorlatában is rájöttek, hogy kis hatékonyságuk ellenére is elboldogulnak a 20 évvel ezelőtti, primitív információs rendszerekkel, mert elég jól működnek ahhoz, hogy vonzzák a felhasználókat.

Azután arra is rájöttek a gyakorlatban, hogy az abban az időben már kereskedelmi forgalomba kerülő számítógépekkel automatizálható néhány olyan folyamat, amelytől az információkeresés függ. Így az információs rendszerek gépesítése lett a legfontosabb gyakorlati célkitűzés. A távközlésnek is volt ugyanolyan eredménye, amelyet az információs rendszerekben felhasználhattak. Ezeknek a technológiáknak az alkalmazása látványos eredményekkel járt. Ma már interkontinentális hálózatba bekapcsolt, hatalmas, működő bibliográfiai adatbázisok léteznek, amelyek szinte mindenki számára elérhetők, akik használni akarják a szolgáltatásukat. Nagy eredmény ez.

Van még olyan irodalom és nyelv (például az arab), amelyet nem ért el az automatizálás. A gépesítés további kiterjesztése azonban nem új elméleteket igényel, hanem már ismert és bevált technológiák további alkalmazását.

Az információs rendszerek látványos fejlődését a *gépesítés első szakaszának* nevezem. Ebbe a szakaszba tartozik az is, hogy húsz évvel ezelőtt a számítógépeket és távközlési technológiát kézi rendszerekbe építették be. Hangsúlyoznom kell, hogy a gépesítés első szakaszában nem volt szükség új elméleti informatikai elvek alkalmazására. *A gépesítés első szakaszának rendszerei az általuk helyettesített kézi rendszerek automatizált másolatai.* Az elméleti informatika mindössze az információkereső nyelvek és a keresési eljárások némi finomításával járult hozzá a jelenlegi rendszerekhez. Az elméleti és gyakorlati informatika ez idáig gondolkodás nélkül elfogadta azokat a józan ész diktálta elveket, amelyeken a korábbi manuális rendszerek nyugodtak.

Nagyobb történelmi távlatból nézve úgy találom, hogy a gépesítés első szakasza egyben annak a programnak a befejezése is, amelyet 1895-ben *Henri La Fontaine* és *Paul Otlet* javasolt az Institut International de Bibliographie-nek (Nemzetközi Bibliográfiai Intézet), s amelyet később – 1931-től – az Institut International de Documentation (Nemzetközi Dokumentációs Intézet) majd 1938-tól a FID – a Nemzetközi Dokumentációs Szövetség – folytatott.

Az eredeti program akkor alakult ki, amikor a *bibliográfia* és a *dokumentáció* fogalmak pontosan fedték mindazoknak az elsődleges érdekeit és törekvéseit, akik a világ bibliográfiai forrásainak megszervezésével foglalkoztak és azzal, hogy minden érdeklődő számára hozzáférhetővé tegyék őket. A legújabb gépesítési program során viszont a bibliográfusok és a dokumentációs szakemberek kapcsolatba kerültek a számítógép-tudomány és a távközlés szakértőivel, akiknek az *információk* feldolgozása és továbbítása a fő érdeklődési területe. Ennek az együttműködésnek az az eredménye, hogy a szóhasználat a *bibliográfiától* az *adatbázis,* a *dokumentációtól* pedig az *információ* felé tolódott el.[42]

> Mélyebb elvi értelemben azonban a gépesítés első szakasza nem igazolja ezt a terminológiai változást, mohó elfogadása és széles körű használata azonban, érzésem szerint, jelentős érdeklődésbeli eltolódást jelez. A dokumentumok fizikai termékek, a dokumentáció pedig ezeknek a termékeknek a szervezésével foglalkozik a használat megkönnyítése érdekében. Az *információ* viszont sokkal megfoghatatlanabb, elvontabb, átfogóbb terjedelmű fogalom, mint a dokumentáció. A számítógép-tudományban és a távközlésben az információ fizikailag meghatározó és mérhető a Shannon-elmélet fogalmai alapján. A mi szakterületünkön néha szándékosan használjuk ugyanebben a korlátozott fizikai értelemben az *információ* fogalmát, ugyanakkor azonban általánosabb értelemben, minősítés nélkül, a tudati hatására utaló értelemben is használjuk a kifejezést. Még meg nem értett kapcsolat áll fenn a számítógép által kiadott információ és eme információ felhasználói értelmezése között. A kétféle információ közötti kapcsolat az informatika (vagy az információtudomány) központi elméleti kérdése.
>
> Szigorúan az adatfeldolgozás szempontjából azt, amit a számítógép kiad adatnak, azt pedig, ami ennek hatására a felhasználóban keletkezik információnak nevezzük. Az információ ebben a leszűkített szá-

42 A szóhasználat eltolódásának van azonban tárgyi alapja is, mivel a mai gépi tárakban nemcsak egyetlen, bibliográfiai tételeket tartalmazó fájl, hanem egyéb információtételeket (például referátumot, továbbá e tételek részeit is) visszakereshetően tartalmazó fájlok összehangoló rendszere (az adatbázis) tárolható. Ezzel párhuzamosan nemcsak dokumentumokat képviselő tételek, hanem egyéb számszerű és tényadatok, valamint szöveges adatok is kereshetővé váltak (a szerk.).

mítástechnikai megközelítésben tehát inkább valamiféle hatást jelent, vagy legalábbis annak eredményét a tudatban. A számítástechnikában a két kifejezést ebben az értelemben igyekszenek szabványosítani is. Eszerint: „Adatnak nevezzük a tények és elképzelések nem értelmezett, de értelmezhető formában való közlését. Az adat tehát reprezentált, de nem értelmezett ismeret." Továbbá: „Információnak nevezzük az adatokon végrehajtott gondolati műveletek értelmezett eredményét. Az információ tehát értelmezett ismeret."[43] Eme értelmezések szerint tehát mind az Adat, mind az Információ az általánosabb Ismeret egy-egy fajtája csupán. (Lásd még a következő kommentárt!)

A mi jelenlegi információs rendszereinket – a szó szoros értelmében véve – helytelenül nevezték el. Feladatuk általában abból áll, hogy referenciajegyzéket adnak a felhasználónak, aki azután a jegyzék ismeretében megkeresi és elolvassa a dokumentumokat. A rendszerek bibliográfiai jellegű információkat nyújtanak. Elméleti szempontból tehát jelenlegi információs rendszereinket nem *információs* rendszereknek, hanem *gépesítetett dokumentációs* rendszereknek tartom. Az *információs rendszer* kifejezés széles körű használata nem a valóságos helyzet, hanem inkább a mélyen átérzett vágyak kifejeződése.

Bár a működő gépesített dokumentációs rendszerek jelenlegi világméretű hálózata az első szakasz végét jelzi, egyben a második szakasz kezdetét is. Mik legyenek a második szakasz céljai? Szerintem kétségtelen, hogy a *bibliográfia* és a *dokumentáció La Fontaine* és *Otlet*-féle világától az *információ* új világa felé tartunk, s nem vitás, hogy át kell gondolnunk ennek a lényeges eltolódásnak a következményeit.

Mielőtt néhány lényeges célkitűzést javasolnék a második szakaszhoz, tisztáznom kell az információ fogalmát és az információs rendszerekhez fűződő viszonyát.

2. Információ, információs folyamatok és tudás

Mit értünk információn? Mindenkinek van többé-kevésbé összefüggő egyéni tudása. Ez az egyéni tudás magában foglalja a világról alkotott személyes elképzeléseinket, a benne betöltött társadalmi szerepünket és minden különleges szaktudásunkat, amelyet akár szakmai tevékenységünk során, akár valamilyen művészettel – mint például a zene – foglalkozva, akár a játékból vagy sok más készség gyakorlásából szerezhetünk.

Életünket mindannyian igen csekélyke tudással kezdtük, de valamilyen velünk született kényszer élt bennünk, hogy életben maradjunk, s rájöjjünk, mik va-

43 Halassy Béla: Az információs rendszerek alapfogalmai. Budapest: Számítástechnika-alkalmazási Vállalat, 1982. – (Rendszerfejlesztési Kiskönyvtár sorozat.) p. 18. (a szerk.).

gyunk, hol vagyunk és kik vagyunk. Az első néhány évben azzal voltunk elfoglalva, hogy felfogjuk a közvetlen környezetünket és megtanuljunk emberi nyelven beszélni és érteni. Azután iskolába kerültünk, összevegyültünk a korunkbeliekkel és tanulni kezdtünk egy nagyobb világról. Elsősorban a nyelv és a számok jobb megértését és azok használatát fejlesztettük ki és elsajátítottunk egyéb olyan készségeket, amelyek hasznosnak bizonyultak a másokkal való érintkezésben. Azt alakítgattunk, amit én a „tudás struktúrájának" nevezek.

Mire elvégeztük az iskolát, tudásstruktúráink is kialakultak és formát öltöttek. Ahogy azután egyre többet tapasztaltunk a társadalmi és a fizikai világban, tudásstruktúráink is tovább nőttek, változtatták formájukat, s talán merevebbek is lettek. Az öregséggel együtt is változtak, de már egyre lassabban. Amikor meghalunk egyéni tudásstruktúránk is velünk hal.

Szerintem *információ az, ami bármely értelemben módosítja a tudásstruktúrát.* A módosító információk közül néhány a környezetünk közvetlen megfigyeléséből származik, néhány abból, amit mások mondanak el nekünk (bár ez talán kevesebb, mint amennyit a tanáraink szeretnének), néhány pedig olvasmányainkból. Lehetnek bizony olyan hatások is, amelyeknek nem vagyunk a tudatában. Minden tudásstruktúrát módosító információ információs folyamat eredménye. A megismerő lát, hall, ízlel, szagol vagy érez valamit. Néha szándékosan keressük az információt, néha ránk kényszerítik, de minden megszerzett információ olyan folyamat eredménye, amelyben az idegrendszerünket valamilyen tudatunkon kívüli forrás működésbe hozza.

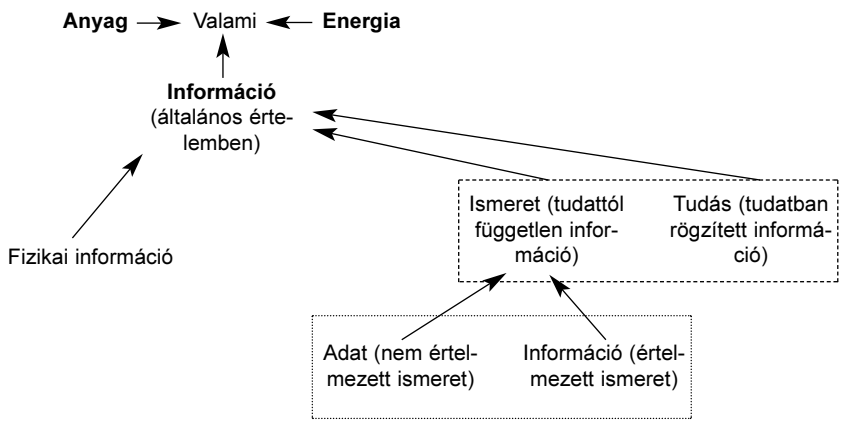
Az információk elárasztanak bennünket. Bizonyos mértékig mégis összpontosíthatjuk figyelmünket konkrét forrásokra. Válogatnunk kell abból, ami számunkra elérhető, különben elmerülnénk benne és összezavarodnánk. A válogatás részben érzékeinktől függ. Csak azokat az információkat tudatosítjuk, amelyekre érzékszerveink válaszolnak, ez pedig csupán töredéke a bennünket elözönlő információknak. Az emberi szem például, amely már önmagában is csodálatos szerkezet, az elektromágneses spektrum 60 vagy több oktávjából csak egyetlen egyet érzékel. Az emberi fül hangfrekvencia-érzékenységének is korlátai vannak. Így tehát a rajtunk kívül lévő világról kapott információk szintén meghatározottak. Elegendőek azonban ahhoz, hogy mindennapi életünkben tevékenykedni, játszani vagy dolgozni tudjunk a Földön. Hétköznapi célokra *megfelelőek.*

Brookes láthatóan a legáltalánosabb, immanens értelemben használja az információ fogalmát. Szűkebb – számítástechnikai – értelemben ugyanis nem az információk, hanem az adatok – az „értelmezhető ismeretek" – árasztanak el bennünket és közülük azok, amelyeket értelmezünk, az információk. A Brookes által használt általános értelemben az Adat, az Információ és az Ismeret fogalmak összeolvadnak, ami nem véletlen. A szerző ugyanis ugyanolyan ontologikusan adott, alapvető kategóriának tekinti az Információt, akár az Anyag és

az Energia fogalmát, amelyek ugyancsak nem vezethetők vissza semmi másra. A Tudás kategóriájának a bevezetése *Ranganathanra* emlékeztet, még ha nem is utal rá a szerző. *Ranganathan* megkülönbözteti a tudattól független Ismeretek és a tudatban rögzített ismeretek, a Tudás világát. Az utóbbi mindig csak részhalmaza az előbbinek. (Ez a kettőség az adat–információ kettősségével analóg, csak általánosabb, magasabb szinten.) Brookes értelmezéséből következik, hogy az Ismeretek az információ fogalmának egyik – tudatilag felfogható – fajtáját képviselik, a shannoni értelemben vett információk pedig a fizikai fajtáját; a Tudás pedig – ugyanúgy ahogy például *Ranganathannál* – a tudatban rögzített ismereteket képviseli. Innen nézve tehát nincs lényeges különbség értelmezett és nem értelmezett ismeret (adat és információ) között. Szigorúan számítástechnikai szempontból viszont nincs szükség az információ ontologikusan értelmezett, legáltalánosabb fogalmára és a Tudás fogalma az ismeret fogalmával esik egybe: a tudatban keletkező Tudás esetén információról, a tudatban tárolt Tudás esetén pedig adatról van szó. Az utóbbi is válhat értelmezett információvá, vagyis a tárolt tudás visszahathat a tudatra (például „beléhasított valakibe annak tudata, hogy…").

Fordításunkban a „tudás" kifejezést használjuk, ha tudati, és az „ismeret" kifejezést használjuk, ha a tudattól független információról van szó. A szakértői rendszerekben kezelt információt éppen ezért nem tudásnak, hanem ismeretnek nevezzük, és a vele összefüggő szakterületet, a „knowledge enginnering"-et pedig ismerettechnikának (lásd még kötetünk végén Miranda Lee Pao szemelvényét).

Az alábbi ábrán a tárgyalt fogalmak gráfja látható. A pontvonalas keret a *Ranganathan* által használt, a szaggatott pedig a számítástechnikában használt kifejezéseket határolja.



1. ábra

A tudomány egyik nagy vívmánya, hogy olyan eszközök elkészítéséhez segített hozzá, amelyek érzékelik a mi érzékszerveinkkel fel nem fogott információkat, mint amilyenek az infravörös és az ultraibolya hullámok, a röntgensugarak stb. Ezek az eszközök az érzékeinknek felfogható vizuális vagy auditív formába alakítják át a kapott információkat. A teleszkóphoz, a mikroszkóphoz hasonló átalakító eszközök nagyon megnövelték a fizikai világ megfigyelésének nem is annyira az egyéni, mint inkább a *társadalmi* lehetőségeit.

Minden információ fizikai csatornákon és fizikai folyamatokon keresztül éri el agyunkat. Míg az eredeti forrásból eljut hozzánk, felfogható műszerekkel átvihető egyik kontinensről a másikra, rögzíthető és tárolható, s végül, amikor hajlandók vagyunk foglalkozni vele, mégis olyan formában kapjuk meg, amely felfogható az idegrendszerünk számára. Mindaddig, amíg nem tudatosítjuk ezt az információt, minden állapotváltozását nyomon követhetjük és állapotát pusztán fizikai fogalmakkal leírhatjuk.

Amikor a beérkező fizikai információt tudatosítjuk, újabb kiválasztás történik. Az információáradat átszűrődik a tudásstruktúrákon. A tudásstruktúra olyan élő információkereső (emmorfotropikus) egységnek fogható fel, amely folyton önmaga változtatására törekszik, hogy dinamikus egyensúlyban legyen az éppen befogadott információval. Vannak információk, amelyek semmi újat nem nyújtanak ahhoz képest, amit már tudunk: a struktúra részei megerősödnek, de nem tudatosan. Más információk újnak bizonyulhatnak és segíthetnek néhány anomália megoldásában, csökkenthetik a struktúrán belüli feszültséget; az ilyen információk fontos szerepet játszanak és az is megtörténhet, hogy tudatosan éljük meg a hatásukat. Végül vannak információk, amelyek annyira újak, hogy nem kapcsolódnak a már kialakított struktúra egyetlen részéhez sem. Ezek átszűrődnek a struktúrán és beépülés nélkül tároljuk őket, hogy később esetleg visszahívhassuk őket, ha segíthetnek egy anomália feloldásában.

Nagyon nehéz lenne valamilyen közvetlen, empirikus módszerrel ellenőrizni az egyéni tudásstruktúrák fejlődéséről szóló elmélkedéseket. Még ha műszakilag elképzelhető is lenne, hogy mások *agyában* fizikai értelemben vett információs folyamatokat figyeljünk meg, akkor sem hiszem, hogy a megfigyelendő jelenség befolyásolása nélkül végezhetnénk beható megfigyeléseket mások *tudatában.* Az önvizsgálatban sem bízhatunk. Bár a gondolkodó emberek azt hiszik, hogy tudatuk kognitív része teljesen nyitva áll saját maguk és az önvizsgálat előtt, a pszichoanalízis erről mást mond. Így a szubjektív tudásstruktúrák közvetlen vizsgálata semmivel sem ígér többet, mint évszázadok óta bármikor. Szerencsére tanulmányozható a kérdés egy másik területen. Ez a terület szerintem nemcsak fontosabb az információs tevékenység számára, hanem nyitva áll a nyilvános megfigyelés és a tudományos vizsgálódás előtt is.

3. Társadalmi informatika

Néhány évvel ezelőtt *Jesse Shera* és *Margaret Egan* azt javasolta, hogy „tanulmányozzák azokat a folyamatokat, amelyek révén a társadalom *mint egész* próbál kapcsolatot teremteni a teljes – fizikai pszichológiai és szellemi – környezettel", hogy felfogja és értelmezze azt. A javasolt tudományágat „társadalmi *ismeretelméletnek"* nevezték. Nemrégiben *Patrick Wilson* elemezte általánosabb szinten a publikált ismeretek és az egyéni felhasználók közötti kölcsönhatásokat; gondolatokban gazdag tanulmánya ugyancsak a fenti kérdést vizsgálja meg egy másik szempontból. A könyvtár- és információtudományban nagy figyelmet szentelnek a „felhasználói igényeknek". Az igényvizsgálatokban egyes – néha gépesített – információs szolgáltatások és a felhasználók csoportjai közötti kölcsönhatásokkal foglalkoznak. Ezek a beszámolók „társadalmi ismeretelmélettel" kapcsolatos megfigyelési adatokat tartalmaznak. Mégis összefüggéstelenek maradnak, mivel nem kapcsolódnak egyetlen általános elmélethez sem.

Azért említem mindezt, hogy hangsúlyozzam: a mi szakterületünkön nagy hagyománya van az ilyen tanulmányoknak, s rendelkezésünkre állnak azok az adatok, amelyek a hiányzó általános elmélet megalapozásához és megfogalmazásához szükségesek. Az ilyen elmélet valóban nagyon fontos. Először is szakmai szempontból van rá szükségünk a gépesítés második szakaszának gyakorlati megvalósításában. Másodsorban azért kell, mert az egyéni „felhasználói igények" és a publikált irodalom közti összefüggések jobb, *tudományos* megértése alapvetően érinti az összes társadalomtudományt.

Abban is biztos vagyok, hogy az általános elmélet sohasem születhetne meg a bibliográfia és a dokumentáció keretei között, csak most, hogy elhagyjuk ezeket a szűkös körülményeket és átmegyünk az információ és a tudás szélesebb és mélyebb világába, kerülhet sor a kidolgozására. Most az a dolgunk, hogy az információ új szóhasználatával megfogalmazzuk problémáinkat, sajátos és ellenőrizhető hipotéziseket állítsunk fel és teljes mértékben kihasználjuk a rendelkezésünkre álló összes eljárást.

A tudás abban különbözik az információtól, hogy kognitív formában épül fel. A tudásstruktúrában minden információ tudatilag kapcsolódik a struktúrán keresztül a struktúrában lévő összes többi információhoz. Ezeknek a tudati kapcsolatoknak a szintézise olyan struktúra, amelynek – térben kifejezve – számtalan dimenziója lehet.

Amikor *A* és *B* tudós kapcsolatba lépnek egymással, *A*-nak a konkrét cél érdekében szavakkal (vagy más szimbólumokkal) kell kifejeznie tudásstruktúrájának lényeges részeit, vagyis az n-dimenziójú struktúrákat egyenes vonalú (lehetőleg egymásután következő) nyelvi megnyilvánulások folyamatába kell szorítania. Magukat a szavakat egyéb, időben lineáris folyamatok közvetítik, mint a hanghullámok a levegőben, az elektromágneses hullámok a tele-

fonban vagy a rádióban, vagy a *B* tudósnak postán küldött szöveg. A fizikai csatornákon belül az információközvetítésnek nincsen *kognitív* struktúrája. Amint azonban elérik *B* tudós agyát, a fizikai információk áradata átszűrődik *B* n-dimenziójú tudásstruktúráján és – sikeres vétel esetén – a helyüket megtalált információk megváltoztatják a struktúrát.

A tudomány fejlődése a tudásstruktúrák rendszeres összehasonlításától függ. Ha *A* gondosan megfigyelt néhány új jelenséget, próbaképpen változtatnia kell talán tudásstruktúráján, hogy helyre tehesse új felfedezését. Utána tanulmányt ír róla, amelyet *B* elolvas. Mivel a jelenség *A* új megfigyelései közé tartozik, könnyen előfordulhat, hogy *B* nem tudja könnyen beilleszteni saját tudásstruktúrájába a dolgozatban kifejtett új elképzeléseket. Így *B* tudós félreteszi *A* írását, hogy megfigyelje magát a jelenséget képviselő dolgot. Majd összehasonlítja megfigyeléseit *A* beszámolójának állításaival. Ha *B* elismeri, hogy *A* dolgozata azt írja le és azt magyarázza, amit ő maga megfigyelt, akkor *A* és *B* tudásstruktúrájának megfelelő része egyezni fog. Ha azonban *B* anomáliákat talál, most ő ír valamit, amiben bírálja *A* beszámolóját és megpróbálja kijavítani azt. Miután a második írás is megjelent, más *C, D, E* stb. tudósok is megismételhetik a megfigyeléseket és összehasonlíthatják *A* és *B* beszámolóját. Ez az ellenőrző és bíráló folyamat addig folytatódik, amíg *A* eredeti, bár a rákövetkező bíráló megjegyzések fényében szükség szerint módosított állításait illetően az érintett tudósok egyetértésre nem jutnak. Ebben a formában a tudományt a többi vitatkozási formától az különbözteti meg – s ez nagyon fontos vonás – hogy ellentmondás esetén a vitatkozók szögre akasztják előző írásaikban közzétett álláspontjukat, hogy mielőtt újabb bíráló megjegyzést tennének, *újra kritikusan szemügyre vegyék* magát a dolgot.

> A jelenségről a tudatban kialakított fogalmi struktúra a tudás. Amikor valaki ezt mások számára is érthetően leírja, ismeret keletkezik. Összehasonlítani valójában csak a nyelvi formában kifejezett tudásokat (=ismereteteket) lehet. Következésképp B tudósnak is valamiféleképpen nyelvi szinten kell kifejeznie a tudását ahhoz, hogy az A tudós által kifejezett ismerettel elvégezhesse az összehasonlítást. Az egyes tudati tudások tehát a nyelvi szinten kifejezett ismeretek összehasonlítása révén egyeztethetők.

Ily módon a tudósok nemcsak tudományos ismereteiket bővítik, hanem ellenőrzik és javítják is megfigyeléseik nyelvi leírását. A tudomány megköveteli, hogy ha hozzáértő tudósok egymástól függetlenül megvizsgálnak valamit, az általuk adott leírások megegyezzenek. Ha új fogalmakra van szükség a leíráshoz, akkor azokat olyan szavakkal kell kifejezni, amelyeknek a használatát már elfogadták az érintettek. Ezért az új jelenség leírása nemcsak azzal jár, hogy a jelenséggel foglalkozók tudásstruktúrájához új alkotóelemek

társulnak, hanem, hogy ezek az új elemek a korábbi tudásstruktúrába is beépülnek.

Úgy vélem, hogy ez a tudományos *módszer* lényege: a *látottak* és *mondottak* közötti folyamatos kritikai összehasonlítás. Abban is hiszek, hogy pontosan ezt a tudományos módszert használjuk, amikor gyermekként nyelvünket tanuljuk. De a tudományos tevékenységgel együtt jár, hogy a vizsgált jelenségekkel összefüggő írások megjelenhessenek, s így minden érv és ellenérv megismerhető.

Mint mondottam, ha a tudós tanulmányt vagy monográfiát ír, akkor tudásstruktúrájának valamelyik részét fejezi ki nyilvánosan. Ha ez tényleg így van, akkor mondjuk az 'S' tudományos témának szentelt közlemények összessége az adott témára vonatkozó tudásstruktúra nyilvános kifejezése. Az egyes tudósok tudásstruktúrája magánjellegű, nagyrészt hozzáférhetetlen, de összefüggő; a megjelent közleményekben rögzített tudásstruktúrák, nyilvánosak, elérhetők, de nem összefüggők. Az 'S' témában megjelenő minden egyes dokumentum tartalmaz némi összetevőt az 'S' témára vonatkozó tudásstruktúrából, néhány dokumentum tartalma azonban ellentmond egymásnak, másoké elavulóban van vagy már elavult; az irodalom összességében pedig nagy lesz a redundancia.

Az első szakasz rendszerei ebből a nem összefüggő és redundáns összességből dokumentumok nagy választékát ajánlják az olvasónak. A felhasználóra vár a feladat, hogy különválassza az ocsút a tiszta búzától (ha van egyáltalán tiszta búza). A második szakaszba feltétlenül bele kell tartoznia annak a lehetőségnek, hogy a felhasználó *közvetlenül* fedezze fel az 'S' tudásstruktúrájának lényeges részeit.

Bár ezeket a kérdéseket a tudományos irodalomról szólva tárgyaltam, minden más szakterületen felmerül a probléma, hogy hogyan segítsék elő a felhasználó és a nyilvános tudásstruktúrák kölcsönhatását. Az első lépésben nemzetközi szinten kell a szétszórt irodalmat rendezni és könnyen hozzáférhetővé tenni, ahogy azt *Henri La Fontaine* és *Paul Otlet* ajánlotta. A következő lépésben elemezni kell a bennük lévő szétszórt *információkat.* Mivel a tudományos irodalom a legrendezettebb és a legkevésbé szétszórt, úgy gondolom, hogy a második szakasz munkáját, akárcsak az első szakaszét, a tudományok irodalmán fogják elvégezni.

4. Az ismeretstruktúrák gépesítése

A nyilvános tudásstruktúrákat – azaz az ismeretstruktúrákat (a szerk.) – valakinek még ki kell dolgoznia. A dokumentáció korában ahhoz, hogy a nyilvános tudást tömör, hivatalos formában bemutassák, az enciklopédiák szerkesztése volt az egyértelmű út; ehhez sok szakértőt toboroztak, hogy a különféle tárgykörökben viszonylag rövid leírásokat készítsenek. Ezeknek az írá-

soknak az összegyűjtéséhez, egybevetéséhez, nyomtatásához és kiadásához azonban idő és szervezőképesség kell. Mire az enciklopédia eljutott az olvasóhoz, el is kezdhették a korszerűsítést. A korszerűsítés nemcsak az egyes cikkek felülvizsgálatát követelte meg, hanem egymás közötti megváltozott összefüggéseiket is át kellett értékelni. A tudás ismertetésének egy másik hatékony útja az, hogy hozzáértő szerzőket áttekintő vagy „állapotrögzítő" tanulmányok megírására ösztönöznek, melyekben összefüggő formában számolnak be egy-egy szűk szakterületen folyó munkáról, s válogatott bibliográfiát tartalmaznak. A korszerűsítés azonban ismét csak állandó erőfeszítést és nehezen fenntartható, következetes kritikai magatartást kíván. Milyen megoldást ajánlhat az „információ korszaka" ezekre a problémákra?

A problémák jelentősek, de szerintem legjobban empirikusan közelíthetők meg. Ma már az on-line rendszer jóvoltából folyóiratok adhatók közre a gépesített információs hálózatokban. Az ilyen folyóiratba benyújtott írásokat még mindig felül kell bíráltatni a szokásos módon a szóban forgó tárgykör két-három szakértőjével. Megismerve a bírálók véleményét a szerkesztő dönti el, hogy elfogadja-e a benyújtott írást úgy ahogy van, meghatározott módosításokat kér, vagy elutasítja.

A bírálónak két fő feladata van: először fel kell mérnie, hogy tartalmaz-e az írás új információt a témában és ha igen, milyent, másodszor pedig, hogy az új információ eléggé indokolja-e a tanulmány megjelentetését a folyóiratban. Az első feladat még úgy is megfogalmazható, hogy össze kell hasonlítani a tanulmány tudásstruktúráját a nyilvános tudásstruktúra megfelelő részével. Az összehasonlítást olyannak kell végeznie, aki feltehetően ismeri a nyilvános tudásstruktúra szóban forgó részeit és ennek időszerű bizonytalan pontjait és anomáliáit. Az első feladat tehát inkább tényszerű összehasonlításból áll mint ítéletből. Nem lehetne ezt az összehasonlítást gépesíteni? A második feladat inkább a személyes megítélés kérdése, mint a tényeké. Tudományos folyóiratban az új információ vonzza az olvasókat és a szerkesztő ösztönösen kisebb újdonsági szintet szab meg mérceként, amit az írásokban el kell érni ahhoz, hogy elfogadják őket. Az ilyen döntésekhez az kell, hogy néhány mennyiségileg nem megfogható tényező egyensúlyban legyen. Gépesíteni lehet-e az ilyen megítéléseket?

Az on-line folyóirathoz benyújtott írások elbírálása tartalmi feltáró és elemző műveletet képvisel a gépesítés második szakaszában. A gépesített tartalmi elbírálásból adódó következtetések eleinte összehasonlíthatók az emberi közreműködéssel végzett bírálatok következtetéseivel, míg a gépesítési feltárással kapott eredmények elég jók nem lesznek ahhoz, hogy teljesen rájuk hagyatkozhassunk.

A szóban forgó szakterületre érvényes nyilvános tudásstruktúra kezdettől fogva úgy bővül, ahogy az írásokat egyenként elfogadják kiadásra. Kezdetben a megjelent dokumentumok keresését és megtalálását biztosító természetes nyelven alapuló tartalmi feltárást – az indexelést – emberi erővel kell elvégez-

ni, mert először meg kell tanulnunk, mit is kell ehhez egyáltalán tenni. A tárgymutató-készítés jelenlegi módszerei ugyanis reménytelenül alkalmatlanok lennének; a Boole-algebrai ÉS, VAGY, NEM logikai operátorok még a jelenlegi finomításokkal is túlságosan durvák ahhoz, hogy a szükséges pontosságot biztosítsák. Már jó ideje foglalkozom az indexelés kérdésével, de valahogy mindig visszakanyarodom a relációs indexelés módszeréhez, amelyet először *Jason Farradane* javasolt 20 évvel ezelőtt. Farradane kilenc szintaktikai relációt használt, azt a kilencet, amellyel a leggyakrabban találkozott a tudományos értekezésekben. Ezt a kilenc relációt azonban még tovább finomíthatják. *Farradane* nekem azt mondta, hogy ő maga 80 vagy 90 relációt tud megkülönböztetni, bár legtöbbjük besorolható az általa mindennapi használatra kiválasztott kilencbe.

Amikor *Farradane* relációs indexelési módszerét az első szakasz információs rendszereiben összehasonlították a többi indexelési eljárással, úgy festett, hogy teljesítménye nem indokolja azt a szellemi többlet-erőfeszítést, amelyet módszere igényel, mégha szakképzett kezekben valóban jobb eredményeket nyújtott is. Most azonban úgy látom, hogy a második szakasz rendszereiben pontosan arra a mély elemzésre van szükség, amely az első szakasz nyers rendszereiben hatástalannak bizonyult. A relációs indexelés korát megelőzően született.

Elméletileg olyan relációkalkulusra vagy logikára van szükségünk, amely minden eddiginél rugalmasabb.

> A dolgokat nem önmagukban, hanem csak egymással való összefüggéseikben vizsgálhatjuk. Ezért a dolgok fogalmait a vizsgálatokban azok funkcióit képviselő fogalmaival helyettesítjük, melyek viszonyokat, egzakt értelemben relációkat fejeznek ki. Ezek a viszonyok persze nem mesterségesek, hanem a dolgok közötti összefüggéseket jelölik. E jelölés, melynek segítségével ezek az összefüggések kezelhetők, a relációkalkulus. A legegyszerűbb, általános formáját – amely két a és b fogalom közötti bármely viszonyt fejez ki – a következőképpen jelölik:
> aRb,
> és ez azt jelenti: a az R relációban áll b-vel.
> A relációkalkulus tehát az a formális, képletszerűen leírható logikai rendszer, amelyben meghatározott tartalmú logikai kérdést tanulmányozni lehet.

Mind a relációkat, mind a fogalmakat – legalábbis kezdetben – változónak, nem pedig állandónak kell tekintenünk. A nyelvi állandók holt nyelvekben fordulnak elő, olyan szövegekben, amelyeket eleven társadalmi összefüggéseikből kiemeltek és mint anatómiai példányokat elemzés céljából rögzítettek. Nekünk

azonban az információs munkában, ahol a fogalmak és a közöttük lévő összefüggések még bármikor visszavonhatók és helyesbíthetők, olyan tartalmak nyelvi megfogalmazásának feldolgozásával, tárolásával és közvetítésével kell foglalkoznunk, amelyekben az éppen érvényes tudás határai körvonalazódnak. Ha – mint javasoltam – a számunkra szükséges kalkulusok empirikusan alakulnak ki a működő on-line folyóiratok dinamikusan kezelt szövegkörnyezetében, akkor sokkal valószínűbb, hogy használhatók is lesznek az információs rendszerekben; használhatóbbak azoknál, melyeket logikával foglalkozó szakemberek és nyelvészek dolgoznak ki axiomatikusan, távol az információs rendszerek világától.

Ha az ismeretstruktúrát manuálisan – az emberi intelligencia közvetlen közreműködésével – dolgoztuk ki, meg kell találnunk a módját, hogyan építhető be az információs rendszerbe és hogyan közvetíthető ez a struktúra vagy ennek egy része a felhasználónak. Ekkorra már remélhetőleg a struktúra minden összetevőjét (nemzetközi egyezmény alapján) olyan mértékben formalizálták, hogy a struktúra leírása egyetlen élő nyelvtől sem fog függeni. (Könnyen lehet, hogy az, ami ily módon kialakul, nagyon hasonló lesz a Chomsky-féle generatív nyelvtan mélystruktúráihoz.)

A struktúrát első lépésben egyszerű, grafikus formában lehetne szemléltetni, amint arra már *Judge* is utalt; ezen a területen műszakilag még sokat lehet tenni. Nehéz lesz az elmélet és a bizonyíték konfliktusának a modellezése is. A következő lépésben a relációk helyét a fogalmakra ható szemantikai hatóerők veszik át, hogy az anomáliák feszültséget teremtsenek a struktúrában. Amint az anomáliák megoldódnak, enyhülnek a helyi feszültségek, a struktúra növekszik és változik, mialatt a periférián új feszültségek keletkeznek. Ezen a ponton a modell *majdnem* „emmorfotropikus" – információkereső –, abban az értelemben, hogy automatikusan ki kell, hogy válogassa a hozzátartozó információ inputból azokat az elemeket, amelyek optimális mértékben enyhítik a modell feszültségeit és elősegítik a növekedését. A további elmélkedés már a tudományos–fantasztikus irodalom világába tartozik.

Az én célom csak az volt, hogy felvázoljam az egyik lehetséges utat a gépesítés második szakaszához. Minden on-line folyóiratból kifejlődhet egy-egy önálló ismeretstruktúra, amely a hasonló adatbázisok szétszórt hálózatában komponens „adatbázissá" válik.

5. Az információ és a tudás mint alapkategóriák

A gépesítés második szakaszához szükséges technológia javarésze már rendelkezésre áll és felhasználható, amint a kívánt elméleteket és elképzeléseket elég részletesen kidolgozták. Az új elméleteknek szoros összhangban kell lenniük az információs folyamatok és a tudás kölcsönhatásáról szerzett növekvő ismeretekkel. Már jó néhány éve foglalkozom ezekkel a kérdésekkel,

és mára már – noha miránk szakmailag csak a kognitív jelenségek tartoznak – az a véleményem, hogy tulajdonképpen a különféle információs folyamatok léte különbözteti meg az emberi lényeket – az összes többi élőlénnyel egyetemben – a tehetetlen anyagtól. A minket elsősorban érdeklő kognitív jelenségek az információs folyamatok spektrumának csak az egyik végletét jelentik, a folyamatok maguk pedig nem különülnek el egymástól élesen. Minél fejlettebb egy állat, annál szélesebb azoknak az információs folyamatoknak a skálája, amelyek benne megvalósulhatnak. Nem szabad elfelejtenünk, hogy az emberi tudat elválaszthatatlan az élő emberi testtől.

A szervek anyagfejlődése a Földön úgy is leírható, mint az információs folyamatok spektrumának a bővülése, amelynek a pillanatnyi végpontja az ember nyelvvel összefüggő kognitív fogalomalkotó készsége. Az információs folyamatok azonban az élettel együtt kezdődtek. *Darwin* fogalma, a „létért való küzdelem" is újrafogalmazható: eszerint azok a fajok maradtak fenn, amelyek új információs forrásokra és velük együtt olyan eszközökre is szert tudtak tenni, amelyek révén inputjukat sikeresebben felhasználhatták arra, hogy a különböző környezeti körülmények között segítsék a faj fennmaradását. A legalkalmazkodóbb állat, az ember tovább folytatja ezt az evolúciós küzdelmet a létért és kezdi feltárni azokat a lehetőségeket, amelyekkel új, lakható bolygókra bukkanhat, ha valamilyen okból túl bizonytalanná válnak a feltételek a Földön az emberi élet folyamatosságának a biztosításához. Egy másik, jelentős evolúciós lépés azoknak az integrált tudásstruktúráknak a birtokolása lenne társadalmi szinten, amelyeknek a befogadására az egyes emberi elme nem képes. Az ilyen elképzelések összeegyeztethetetlenek a tudat és az anyag független voltáról szóló felfogással, amelyet a filozófus *Descartes* kényszerített az emberi gondolkodásra 350 évvel ezelőtt, a modern tudomány kezdetén. Abban az időben a kartéziánus dichotómia felbecsülhetetlen szerepet játszott. Feloszlatta a metafizikus ködöt és arra ösztönözte a tudósokat, hogy kizárólag a fizikai mechanizmusok alapján, korlátozó metafizikai megszorítások nélkül tárják fel a fizikai és biológiai világot. A tudósok éltek ezzel a gondolati szabadsággal és ennek eredményeképpen ma mind makro- mind mikroszinten sokkal jobban ismerjük fizikai környezetünket, mint saját magunkat vagy akár a környezetünkhöz fűződő viszonyunkat. A közvetlenül előttünk álló problémák azonban már nem a természettudományokhoz, hanem a társadalomtudományokhoz tartoznak. Ahhoz, hogy épp olyan kielégítően meg tudjuk oldani társadalmi problémáinkat, mint amilyen jól a tudományos problémáinkkal tettük azt, fel kell ismernünk, hogy *Descartes* komoly korlátokkal is megterhelte a gondolkodás szabadságát.

Descartes testetlen lelkekkel benépesített, őrült fizikai világot hagyott ránk örökül. Mivel a kognitív fogalmakon alapuló információ részben fizikai, részben testi jellegű, *Descartes* két világa alapján nem kezelhető az információ fogalma. Az információ mégis minden társadalomtudományban központi fogalom. Ahogyan már megfogalmaztam, az információ közvetíti mind az

egyének egymás közötti, mind az egyén és a környezet közötti összes kölcsönhatást. Ahhoz, hogy szilárd alapra helyezzük az információ és a tudás fogalmát, a *Descartes-énél* gazdagabb metafizikára van szükségünk.

A fizikai világ a tér/idő és az anyag/energia tovább nem elemezhető fogalmaival írható le, bár a modern fizika felismerései igencsak kikezdték ezeknek a fogalmaknak a klasszikus egyszerűségét. Azt azonban már nem értem, hogyan lehetne az információ fogalmát tudományosan *kizárólag* fizikai dolgokat képviselő fogalmakkal leírni. Ezért most azt javaslom, hogy mondjuk ki hangosan és nyíltan azt, ami hallgatólagosan már régen megfogalmazódott az információs munkáról és problémákról szóló vitákban. Azt ugyanis, hogy az *információ* és a *tudás,* úgy ahogy már leírtam őket, ugyancsak elemezhetetlen fogalmak; olyan alapvetőek a társadalomtudományok számára, akár a tér/idő és anyag/energia a természettudományok számára.

Ennek az elképzelésnek véleményem szerint már eleve megvolna az az előnye, hogy az alkalmazása révén az információ tudománya felsorakoztatva a többi természettudományhoz, önálló természettudománnyá válhatna. E nélkül az információ tudománya valahol a tudományos materializmus pokla (vagy mennyországa) és a tiszta szellemiség mennyországa (vagy pokla) közötti metafizikus börtönben található, és a természet- és a társadalomtudósok ezentúl is meglehetősen bizonytalan szellemi képződménynek fogják tekinteni.

Az Információt és a Tudást alapkategóriának tekintő elképzelés alapján az információ tudománya, az informatika a külső világ és az élő szervezetek kölcsönhatásának a tudománya lehetne, átfogva az információs folyamatok teljes spektrumát. Ennek a tudománynak az első gyakorlati eredményei azok a jól megalapozott információs rendszerek lennének, amelyek végül ki fogják szorítani a működő, rossz hatásfokú dokumentumkereső rendszereket.

A GÉPI INFORMÁCIÓKERESÉS KLASSZIKUSAI

A hatvanas évek elején megjelentek a második generációs nagyszámítógépek (a Kongresszusi Könyvtárban helyezték üzembe az első könyvtári célú gépet), 1963-ban adták ki először a Science Citation Indexet, az évtized közepére működni kezdett számos nagy gépi információkereső rendszer (pl. a MEDLAR, az ERIC), 1966-ban *Henriette D. Awram* elkészítette az első mágnesszalagos adatcsere formátumot (MARC), 1967-ben hozzákezdtek az OCLC off-line kötegelt feldolgozású gépi katalógusának kialakításához, az évtized végén pedig már folytak az első információs hálózatok tervezési munkái, 1970-ben működni kezdett az első távközlési hálózat, a TYMNET, 1971-ben már on-line elérhető az OCLC, a MEDLAR (MEDLINE) és az ERIC, elkészültek az első on-line tezauruszok. Elkezdődtek a távolsági on-line információkeresések.
Szinte ezzel egy időben publikálták az első, kizárólag a gépi információkereső rendszerekkel foglalkozó kézikönyveket és monográfiákat. Az 1959–1983 közötti termés gazdagságát a következő oldalon található táblázatban mutatjuk be. A felsorolásban nem szerepelnek az információkereséssel általában – tehát nemcsak gépi szempontból – foglalkozó szerzők (például *Brian C. Vickery, J. R. Sharp, R. Meetham, Peter Hermann*) művei.

A szerzők vagy a rendszerek tervezői és építői *(Kent, Lancaster, Salton)* vagy programok irányítói, csúcsmenedzserek, vállalkozók *(Becker, Hayes),* vagy kutatók, tudósok *(Spark Jones, van Rijsbergen, Maron, Meadow).* Szinte mindegyik tanít könyvtári és információtudományi főiskolán.

Gépi információkereső rendszerekkel általában foglalkozó művek		Automatikus indexeléssel/osztályozással foglalkozó művek
1959		Maron, M. E. [et al.]: Probabilistic indexing
1963	Becker, J., Hayes, R. M. Information storage and retrieval (2. kiad.: 1976)	
1966	Kent, A.: Textbook on machine information retrieval (német kiad.: 1966)	
1967	Meadow, Ch. T.: The analysis of information systems (2. kiad.: 1973)	
1968	Lancaster, W. F.: Information retrieval systems: characteristic, testing and evaluation (2. kiad.: 1979)	Salton, G. Automatic information organization on information and retrieval
1969	Lefkowitz, D.: File structures for on-line systems	
1970	Hayes, R., Becker, J.: Handbook of data processing for libraries	
1971		Spark Jones, K.: Automatic keyword classification for IR
		Salton, G.: The SMART retrieval system
1972	Campey, L.: Generating and printing indexes by computer	
	Lancaster, W. F.: Vocabulary control for information retrieval	Van Rijsbergen, C.: Inforrmation retrieval [magyarul **1987**]
1973	Lancaster,F. W., Fayen, E. G.: Information retrieval on-line	
1975	Wessel, A. E.: Computer-aided information retrieval	Salton, G.: Dynamic information and library Processing
1976		Spark Jones, K.: Linguistik und Informationswissenschaft
1978		Maron, M. E.: Theory and foundation of information retrieval
	Lockeman, P. C., Mayr, H. C.: Rechnergestützte Informationssysteme	
1981		Spark Jones, K.: Information retrieval experiment
		Harner, E. P.: An introduction to automatic literature searching
1982	Krause, J.: Mensch–Maschine–Interaktion in natürlicher Sprache	
1983		Knorz, G.: Automatisches Indexieren als Erkennen abstrakter Objekte
		Salton, G.: Introduction to modern information retrieval
1985	Dym, E.: Subject and information analysis	
1986		Lustig, G.: Automatische Indexierumg zwischem Forschung und Praxis

ROBERT MAYO HAYES (1926) ÉS JOSEPH BECKER (1923–1995)

A matematikus Robert M. Hayes a könyvtári kutatások intézetének igazgatója a California Egyetemen (Institute of Library Research, University of California, Los Angeles). Ő szerkeszti az információkeresés immár klasszikusnak számító folyóiratát, az Information storage and retrieval-t, és adja ki az Information Science Series-t.
Az informatikus Joseph Becker a Pittsburghi Egyetemen a könyvtártudomány professzora, a Becker and Hayes, Inc. vállalat elnöke, amely a John Wiley and Sons, Inc. kiadó leányvállalata. Becker fontos szerepet játszott az Egyesült Államok információs hálózatának szervezésében.

Már *Brian C. Vickerinél* is megfigyelhető volt, hogy az osztályozást és az információkeresést igyekezett általánosan leírni. Katalóguscédula helyett például az „egységnyi adathordozó" (tally), a jelzet, tárgyszó, bibliográfiai adat helyet az ismérv kifejezéseket használja, a helyeket azonosító elemeket (lelőhely adatokat, raktári jelzetet, oldalszámot) általánosítva „címeknek" (address) nevezi.
A gépi információkereső rendszerek leírásakor ez a tendencia csak fokozódott. Hayes és Becker könyveikben helyenként például csak „dolgokról" (things) beszélnek, hogy ne korlátozzák az információkereső rendszer leírását csak dokumentumokra, vagy csak információkra (pl. „Since an information system deals with descriptions of thing essentially external to it,..."). Ez az általánosító törekvés másfél–két évtized múlva az adatmodellezésben például az entitás (dolog, entity) fogalmához vezetett el. A terminológiai ellenőrzés alapján keletkezett mutatókat, tárgyszójegyzékeket és deszkriptorszótárakat (tezauruszokat) és a formai feltáráshoz használt egységesített besorolási adatokat is egyetlen, általánosított fogalommal (authority) ragadják meg, s beszélnek „tárgyi (szemantikai) szempontból egységesített adatokról" (subject authority). Mindezt magyarul sokszor nem is lehet egyetlen tömör kifejezéssel leírni.
Az alábbi szemelvényben az osztályozási rendszerek és információkereső nyelvek (illetve a jelzetek, tárgyszavak, deszkriptorok) kezelését a fájlszervezés szempontjából írják le.

A könyvtári adatfeldolgozás kézikönyve[1]

Könyvtárak és információs központok mint információs rendszerek[2]

Az 1. ábrán a tudományos információs rendszerben lejátszódó információfeldolgozás néhány szintje látható.

Ezek az információs folyamatok négy fő lépésben foglalhatók össze:

1. Adatok előállítása: rendszerint a tudományos kutatók feladata.
2. Az adatok terjesztése: különböző személyek végzik, kezdve az adat előállítójától a folyóirat referálóján keresztül a könyvszerkesztőig.
3. Az adatok válogatása, beszerzése, indexelése és katalogizálása: könyvtárosok, dokumentátorok, indexelők, kivonatkészítők és hasonló szakemberek feladata.
4. Az adatok lényegi elemzése: rendszerint a tudományos kutató végzi, de egyre inkább a „tudományos információs szakember" feladata lesz.

Az információs tevékenységek e spektrumában a könyvtár különösen fontos szerepet játszik és még hosszú ideig fog is játszani az anyagok kiválasztásában, beszerzésében, katalogizálásában, indexelésében, az ezeket reprezentáló fájlok tárolásában és karbantartásában, valamint elérésük biztosításában. Ha a könyvtárat úgy határozzák meg, mint a könyvek, a múlt feljegyzéseinek őrzőjét, akkor a könyvtár elsősorban az adatkiválasztás eszközének a szerepét játssza. Amikor pl. betűrendes katalógust hoznak létre, az adatátvitel eszközeként funkcionál. (A könyvtár történelmi szerepe ebben az esetben nem más, mint a gyűjtött adatok alapján végzett legegyszerűbb információfeldolgozás.) Amikor viszont szakkatalógust készítenek, akkor a könyvtár olyan eszköz, amellyel a birtokában lévő adatok kiválaszthatók.

A katalógusok azonban legfeljebb a tartalom durva megközelítésre alkalmasak. A könyvtárak használóinak erre van szükségük, de amikor ugyanezeket az osztályozási technikákat alkalmazzák nem könyv jellegű, részletes műszaki adatokat tartalmazó, szakosított gyűjteményekre, a tartalomra vonatkozó keresés e módszereinek pontatlansága nyilvánvalóvá válik. Még a mutatók is csak a potenciálisan releváns dokumentumokhoz vezetnek el és nem a tartalmukhoz. Ezért az adatok kódolásának és indexelésének olyan útjait és módszereit kezdték keresni, amelyek inkább a dokumentumtöredékek, semmint a teljes dokumentumok kiválasztására és kezelésére alkalmasak.

---

1 Handbook of data processing for libraries / Robert M. Hayes ; Joseph Becker. – New York [etc.] : John Wiley and Sons, 1970. 885 p. (A Wiley-Becker-Hayes publication.)

2 Libraries and information centers as information systems. In: Handbook of data processing for libraries, p. 752–770.

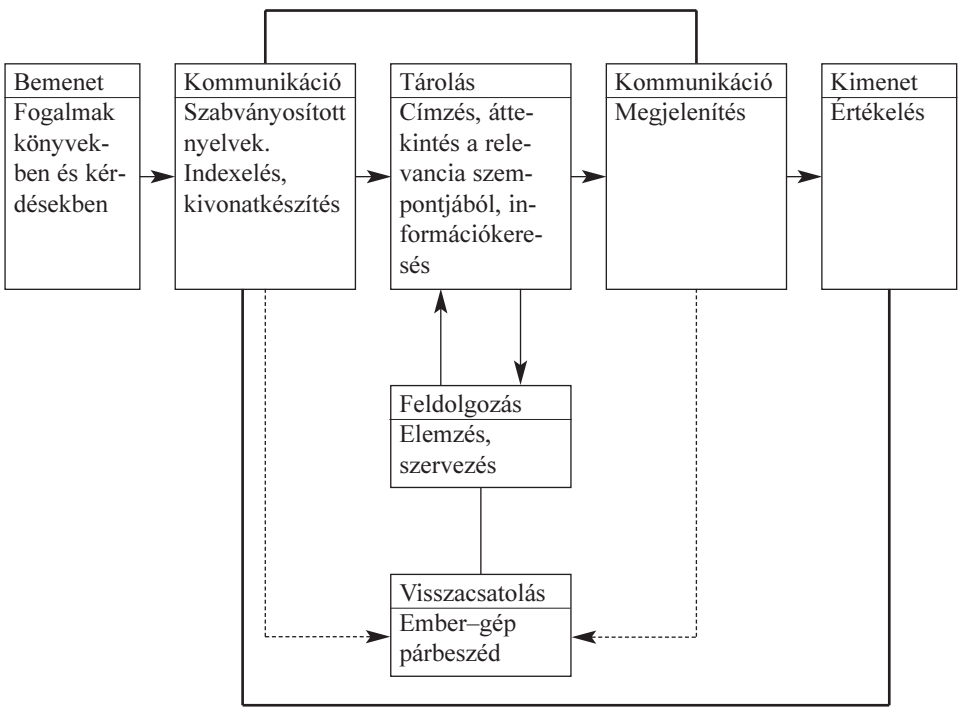
	Első szint (nyers adat)	Második szint (zárt adat)	Harmadik szint (közbenső adat)	Negyedik szint (nyílt adat)	Ötödik szint (nyílt adat)	Hatodik szint (nyílt adat)
A közlés célja	Időazonos figyelés és ellenőrzés	Belső, szervezeten belüli, team vagy személy számára	Személyes, szakmabeli kollégák közötti	Híradás a szakmabeli kollégáknak	Híradás a szakmabeli kollégáknak	Híradás a világnak, mindenkinek
A hallgatóság természete	Gép	Szorosan együttdolgozó munkacsoport	„láthatatlan kollégium”	Figyelő csoport	Szakmai csoport	Általános szakmai közvélemény
A hallgatóság nagyságrendje	–	1–10	10–200	100–1000	1000–10 000	10 000–100 000
A keletkezés és hozzáférhetőség közötti idő	10^{-2}–10^{2} sec	2 perc–1 óra	1 óra–2 hét	2 hét–1 év	2 év–3 év	3 év–9 év
Az áttekintés minősége	–	Semmi	Semmi	Gyenge	Elég jó	Jó
Kötet/év (csoporton belül)	–	3×10^{3}	3×10^{3}?	3×10^{3}?	3×10^{3}?	3×10^{3}?
A közlés formája	Jel	Személyes feljegyzések, belső mellékeztetők stb.	Telefon, levél	Beszámolók, konferenciaelőadások (egyedi teljesítmények)	Folyóiratok, konferenciakiadványok	Könyvek
A közlés rendeltetése	?	Vezetői feljegyzések	A kommunikáció megkönnyítése	Szakmai írás Feladatorientált indexek és kivonatok Elemzés, értékelés	Áttekintés, publikálás, terjesztés Diszciplínaorientált indexek és kivonatok Kiválasztás, katalogizálás, tárolás és hozzáférés	Szerkesztés, kiadás, terjesztés Nemzeti bibliográfiák Kiválasztás, katalogizálás és hozzáférés
Szakmai szerepek (a közlési folyamatban)	?	Tájékoztatási szakemberek	Tájékoztatási szakemberek	Könyvtárosok és dokumentátorok Tájékoztatási szakemberek?	Szerkesztés és előadók Könyvtárosok	Szerkesztők Könyvtárosok
Intézményesítés	?		Információcserélő csoportok, különleges érdeklődésű csoportok stb.	Könyvtárak Információs (elemző) központok	Szakmai társaságok Könyvtárak	Kiadók Könyvtárak
A gépesítés szerepe	Érzékelés, feldolgozás, átvitel, feljegyzés, ellenőrzés	Adatelemzés. Fájlkezelés	Konferencialátogatás, „on-line közösség”	Bibliográfiai keresés. Adatelemzés	Kivonatok és indexek nyomtatása. Bibliográfiai keresés	Katalógusok és bibliográfiák nyomtatása

1. ábra. Információfeldolgozás könyvtárakban és információs központokban

A kétféle – dokumentumokat válogató és kezelő, illetve a magasabb szintű információ-feldolgozást is végző – rendszer közötti különbségre többek között *Bar-Hillel* mutatott rá; szerinte a *dokumentumok* tárolása és keresése a bennük található adatok elemzésétől független kérdés. Az irodalomkutatás eredménye csak az, hogy mely dokumentumok, illetve könyvek relevánsak az adott témában. Az *információkeresést* ezzel szemben ennél általánosabban definiálja: válaszok gyűjtése a választott tárgyra vonatkozó kérdésekre.

Az információs folyamatok a könyvtárakban és az információs központokban

A kérdést olyan rendszer példáján tárgyaljuk, amelyben dokumentumokat reprezentáló hivatkozásokat (bibliográfiai tételeket) tárolnak és kereshetnek vissza. A folyamatok a 2. ábrán láthatók. Maguk a dokumentumok potenciálisan relevánsnak tekinthetők a bennük – többé-kevésbé jól – leírt elképzelések, gondolatmenetek és adatok miatt. Az irántuk támasztott igények is többé-kevésbé kifejeződnek a kérdésekben.



2. ábra. A dokumentumkeresés folyamatának vázlata

Az elkerülhetetlen nehézségek miatt, amelyekkel az emberek gondolataik kifejezésekor küzdenek, mindkét esetben bizonytalan az elképzelések leírásának pontossága. Az írott vagy beszélt nyelv a legjobb esetben is pontatlan közlést eredményez, de ebbe, akárhogy is van, bele kell törődnünk: az információs rendszerek input forrásainak ilyen a természete.

A rendszer első összetevőjének – a kommunikációs komponensnek – kell megfelelő módon kezelni ezt a dokumentumokból és kérdésekből álló, általában rendezetlen inputot, és ennek a komponensnek kell az inputot a rendszer nyelvén – leíró katalogizálással, indexeléssel, kivonatkészítéssel stb. – reprezentálnia. Az input eredeti formájában meglévő pontatlanság felerősödik a bemenetek és a rendszer nyelve közötti elkerülhetetlen összehangolatlanság következtében. A kommunikációs folyamatban – az adatok kiválasztásával, hasznosságuk értékelésével, a leírások szabványos formára fordításával – ezeket az eltéréseket meg kell szüntetni. Ez magyarázza, miért vizsgáljuk kitüntetetten az input és a rendszer közötti viszonyt. A rendszer annyira lesz hatékony, amennyire ez a párbeszéd hatékony, hiszen az összes további művelet azokra a tételekre irányul, melyek ennek alapján keletkeznek.

Ezek után azokat a helyeket kell a tárolóban meghatározni, ahová az új tételek kerülnek majd, vagy ahol a régieket fogjuk keresni. A megfelelő tárolóhelyi cím megtalálásának folyamata, bár formalizált – valójában talán éppen azért, mert formalizált – az a pont, amely a rendszer teljesítőképességét (szemben az egyszerű hatékonysággal) a leginkább meghatározza. A fájlok rendezése és szervezése, a címek elhelyezése, a tárolóhelyek kijelölése, az indexek strukturálása mind lényeges gépi módszerek, amelyekkel ésszerű költségek mellett gyors hozzáférés biztosítható a várhatóan releváns tételekhez. És mivel a „gyorsaság”, a „relevancia” és az „ésszerűség” minőségi jellemzők, a rendszertervezés ez irányú problémája éppen e minőségi jellemzők kvantifikálása, mérhetővé tétele, a teljesítőképesség – az összes kívánalmat tükröző – mértékének kialakítása, és a címzési folyamat e mértékének megfelelő optimalizálása.

Az azonosító lefordítása a tárolóhelyi címre viszont csupán a potenciálisan releváns tételek csoportjainak helyét határozza meg. Ezeket kell megtalálni és áttekinteni. A kívánt tárolóhelyi címek gyors pozicionálásából és tartalmuk gyors áttekintéséből adódó mechanikus problémák egyesítése a rendszer működésének az az oldala, amely a számítástechnikai berendezések gyártóit a leginkább érdekli. Ebből a célból rengeteg nagy kapacitású, közvetlen elérésű memóriát (például „asszociatív tárat”) fejlesztettek ki, amelyek az adatok rendkívül gyors keresésének, áttekintésének lehetőségét teremtették meg.

A tételek áttekintése közben mindegyiket értékelni kell abból a szempontból, hogy a bejövő kérdésre (vagy a tárolandó dokumentumra) vonatkoztatva releváns-e. A „relevancia” értelmezésében felmerülő logikai problémákat, va-

lamint a „relevancia” értékeléséhez használható kritériumok kialakításának gyakorlati problémáit különösen intenzíven tanulmányozták.

Ezek a folyamatok – a dokumentumok leírása, térbeli elhelyezése és értékelése – tipikusan könyvtári műveletek. A következő lépést – az adatok elemzését és redukálását – ritkán vonják be a gépesítés műveletébe, talán azért, mert a gépi rendszerekben tárolt részletek szintje e szempontból alacsony, de még inkább azért, mert az elemzés folyamata túlságosan bonyolult. Természetesen a tételek szervezettségének megjelenítése – egy bizonyos szintig – egyszerűen megoldható, és a könyvtárak alkalmazzák is ezeket a technikákat. Ezek közé tartoznak például a különböző formájú permutált címindexek. A relevancia mértéke szerinti besorolás és megjelenítés ugyanilyen egyszerű, de csak ritkán élnek vele. Ha a gépben a dokumentumok tartalmának nagyobb hányadát tárolnák, részletesebb elemzésre kerülhetne sor. Ez már elterjedt az olyan információs rendszerekben (például a leltárellenőrző és adattömörítő rendszerekben), amelyekben a teljes dokumentumtartalmat gépben tárolják. Az eredmények közlése a felhasználó számára, és a felhasználó által végrehajtott értelmezés (a rendszer működésének utolsó lépése) olyan mértékben könnyíthető meg, amilyen mértékben a dokumentumok tartalmának elemzése automatikusan elvégezhető. Ez az oka annak, hogy az informatikusok munkájának középpontjában azoknak az algoritmusoknak és heurisztikus módszereknek a kidolgozása áll, amelyekkel a természetes nyelvű szövegek feldolgozhatók, és a kérdésekre adandó válaszok, eredmények automatikusan generálhatók. Nem kétséges, hogy ez az információtudomány legnagyobb kihívása és intellektuálisan legizgatóbb kutatási területe. *Bar-Hillel* és mások ugyanakkor azt állítják, hogy lehetetlen ezt a feladatot elvégezni a természetes nyelvben rejlő logikátlanság miatt. Ennek ellenére az információtudomány számos kutatója továbbra is komolyan érdeklődik e téma iránt.

Mindezek a folyamatok következtében került az érdeklődés középpontjába a könyvtári tevékenységben az információtudomány. A következő fejezetekben a könyvtári tevékenységnek azokat az elméleti és gyakorlati problémáit összegezzük, amelyekkel az információtudomány művelői foglalkoznak. Célunk nem az, hogy mély és átfogó elemzést adjunk, hanem az, hogy olyan vázlattal szolgáljunk, amely jó alapot ad a szakirodalom további tanulmányozásához.

Még a fejezetek végén ajánlott irodalom sem meríti ki a tárgykört, inkább azokra a művekre szorítkoztunk, amelyek leginkább elvezetik az olvasót az alapvető kutatási jelentésekhez. A szakirodalomban a legfontosabbak közé tartozik a *Carlos Cuadra* által az American Society of Information Science (az Amerikai Információtudományi Társaság) számára szerkesztett „Annual Reviews of Information Science and Technology” című kiadványok.

Az adatbevitel problémái a kommunikációs folyamatban

A kommunikációs folyamatban felmerülő problémák első csoportja, amit tárgyalni fogunk, az adatbevitelből ered. Ezek egyike, a „szókincs szabványosítása", mely megfelel hagyományos könyvtári körülmények között az egységesített tárgyszójegyzékek karbantartásának és magában foglalja szótárak, tezauruszok és osztályozási rendszerek készítését. Egy másiknak, az „ember–gép párbeszédnek" nincs megfelelő párja a könyvtárban, de analóg az olyan, két ember között lezajló párbeszéddel, amely elfogadott kommunikációs modell nélkül folyik. A harmadik, az „indexelés, referátumkészítés és kódolás" a katalogizálás párja, sőt több mint puszta megfelelője, mivel csak a részletezés szintjében különböznek.

A szókincs szabványosítása. Minden – hagyományos vagy automatizált – információs rendszer többnyire tartalmaz többé-kevésbé szabványosított szótárat. Az ilyen szótárak kialakítása általában olyan problémákat vet föl, amelyekkel már hosszú ideje nyelvészek, filozófusok és pszichológusok foglalkoznak. A nyelvészek ezt a kérdést megfigyelési problémának tekintik: milyen az a nyelv, melyet az emberek sajátos „információs rendszerében" használnak? A nyelvészet eszközeivel feltárhatók az aktuális nyelvhasználat mögött meghúzódó szintaktikai és szemantikai struktúrák. A nyelvészek feladata, hogy meghatározzák e nyelv kifejezéseit, melyek kiejtett hangsorokban vagy vizuális jelekben egyaránt megjelenhetnek.

A nyelvfilozófusok viszont hajlamosak a nyelvet és szabványosítását logikai problémának tekinteni: melyek azok a struktúrák, amelyeken a nyelv alapszik? Megközelítésükhöz különböző szintű formális nyelveket használnak fel: az olyan legegyszerűbb logikai nyelvektől kezdve, amilyen például a kijelentéskalkulus, a véges állapotú nyelveken át a természetes nyelvek összetettségét megközelítő nyelvekig.

A pszichológusok megközelítése némileg eltér ettől: milyen hatással van a nyelv a használójára, és arra, akihez szólnak? A pszicholingvisztika gyökerei az általános szemantikából erednek.

Az adatfeldolgozó szakembereknek is foglalkozniuk kell a nyelv szabványosításával, mégpedig kétféleképpen: (1) a számítógéppel folytatott kommunikáció szintjén és (2) az információt reprezentáló kifejezések szintjén a számítógépen belül.

A cél – az adott megközelítési módtól függetlenül – az, hogy a nyelvben implicite meglévő összefüggéseket explicit formában fejezzük ki. Mivel bármit kezelünk is gépi eljárásokkal, azt előbb formalizálni kell, az alapvető kérdés az, hogy a nyelvfeldolgozás mekkora hányadát végezhetjük algoritmikusan. Valójában a szintaxis és a szemantika megkülönböztetésében egyszerűen az a különbség fejeződik ki, amely a nyelv formalizált – és ily módon explicitté tett – valamint még implicit, még nem formalizálható részei között van.

Mivel minket éppen a többé-kevésbé formalizált szókincs kialakításának a kérdése foglalkoztat, meg kell vizsgálnunk a szemantikai elemzés nagyon bonyolult kérdéskörét. A szemantikai elemzés első lépése olyan szótár elkészítése, amely a rendszerben használandó minden egyes kifejezést az információs rendszer működésének megfelelő formában tartalmazza. A definícióknak az adott (éppen meghatározott) szó és a rendszer többi kifejezése közötti kapcsolatok, relációk pontos megadására kell szorítkozniuk. Végül, a definíciók viszonylag kis számú „definiálatlan kifejezésre" támaszkodnak, amelyeknek értelme pusztán műveleti jellegű, és amelyeket „axiomatikusan" kell elfogadni.

A szótárakban általában gondoskodnak arról, hogy az egyes szavakhoz megadják a specifikusabb és az általánosabb jelentésű kifejezéseket is, továbbá az egész–rész és egyéb, asszociatív, rokonsági összefüggést. Szinonimák esetén feltüntetik a használandó kifejezést. A szavakat fogalomosztályokba (szakcsoportokba) is besorolják. Az egyik első, a gépesítést jóval megelőző világban született ilyen szótárat az angol *Roget* adta ki 1852-ben[3]. Ezeket a szótárakat nevezeték tezaurusznak, s ezt az elnevezést vették át az információkereső rendszerek a szabványosított szótárak megnevezéséhez.

Információfeldolgozás
Speciálisabb kifejezések: Információterjesztés, Információkeresés, Információtárolás, Információfelhasználás
Rokon kifejezés: Számítógép

Információkeresés
Átfogóbb kifejezés: Információfeldolgozás
Rokon kifejezés: Adatfeldolgozás

3. ábra. Deszkriptorszótár részlete

Az egyszerű adatbázis-rendszerekben a terminológiai ellenőrzés nem okoz különösebb gondot, és a kifejezések vagy implicit módon magukban a rekordokban szerepelnek, vagy legjobb esetben szó-, illetve kódjegyzék tartalmazza őket. A bibliográfiai (szakirodalmi) információkereső rendszerekben a terminológiai ellenőrzés sokkal összetettebb és a tezaurusz különösen fontos szerepet játszik. A dokumentum tartalmát a szótárból vett szakkifejezések-

3 Magyarul először *Póra Ferenc* készített ilyet 1912-ben: Póra Ferenc: A magyar rokon értelmű szók és szólások kézikönyve. Tartalmaz harmincezer szinonim szót és szólást nyolcszáz logikai csoportban. – 2. kiadás. (1907: 1. kiadás). – Budapest: Gondolat, 1991. 452 p. A tezauruszokkal részletesebben az első kötetben foglalkozunk (a szerk.).

kel kell leírni. Mivel a szavak sokféleképpen értelmezhetők, pontosan azokat a kifejezéseket kell megadni, amelyek a dokumentum tartalmát jellemzik. Ennek érdekében a hivatalosan elfogadott, szabványosított, kötött tárgyszójegyzékekben és a tezauruszokban feltüntetik a megengedett kifejezések közötti a *priori* (a generikus, a partitív) és az egyéb relációkat (3. ábra). Ebben az összefüggésben az osztályozási rendszerek olyan szótáraknak tekinthetők, amelyekben a kifejezéseket hierarchikus alá–fölérendelési relációk szerint rendezik.

A szövegfeldolgozó rendszerek esetében még bonyolultabb a helyzet, hiszen a természetes nyelv szavainak jelentései nemcsak az *a priori* meghatározásoktól függnek, hanem azoknak a mondatoknak a szintaktikai struktúrájától is, amelyben használták őket, továbbá a mondaton belüli többi szó jelentésétől is. A számítógépben tárolt szótárnak ezért le kell írnia mindazokat a mondatfajtákat, amelyekben egy szó valamennyi lehetséges értelemben előfordulhat. Az ilyen szótár mérete – még nagyon szűk témakör esetében is – meghaladja a kereskedelmi forgalomban beszerezhető számítógépek gyors elérésű memóriájának kapacitását. Ez az egyik oka annak, hogy a kifinomult szövegfeldolgozás még csak elméleti, legfeljebb kísérleti stádiumban van.

A szótárfejlesztés problémái mindig is a könyvtártudomány és a dokumentáció érdeklődésének középpontjában álltak. Hogyan állapíthatjuk meg a dokumentumok leírására alkalmas kifejezéseket? Hogyan definiáljuk a jelentésüket, különösen pedig az egymáshoz fűződő kapcsolataikat?

1. Megnézhetünk valamilyen szabványosított, kötött tárgyszójegyzéket, vagy az adott terület folyóiratainak mutatóit vagy kész szójegyzékeket. Ez az eljárás különösen a kiinduló szótár összeállításában segíthet. Néhányan azonban megkérdőjelezik ezt a módszert, ezért még ha el is ismerjük hasznosságát, teljesen megbízhatónak nem tekinthetjük.

2. Támaszkodhatunk magukra a dokumentumokra is a kifejezések megválasztásában. Amikor a szövegből kiválasztott szavakkal végzett információnyerés statisztikai (automatizált) és intellektuális módszereiről beszélünk, nem foglalkozunk a forrásdokumentum indexelésében betöltött szerepükkel. Ehelyett azt vizsgáljuk, hogyan használhatók fel első lépésben szabványosított szótárak kialakításához. Az a célunk, hogy eldöntsük, mely szavak használatosak adott szakterületen, és ehhez az egyes dokumentumok szolgáltatják a példákat. Ezek a szavak jól vagy rosszul írják le az adott dokumentum tartalmát. A permutált címindex a legkézenfekvőbb példa erre, de a statisztikai módszerekkel előállított kivonatok és bizonyos koordinált indexelési törekvések is idesorolhatók. A szervezettség hiánya, de még inkább a keresztutalások hiánya arra kényszeríti a felhasználót, hogy az egész mutatót átfussa vagy saját maga alakítson ki a kifejezések között kapcsolatokat.

3. Bevezethetünk olyan elemeket, amelyek révén az egyszerű szójegyzék szabványosított szótárrá alakul: definíciókat és magyarázatokat, utalókat és

keresztutalásokat, generikus és egyéb (osztály- és analitikus) relációkat. Milyen módszerekkel élhetünk? A témakör szakértőjét sok gépi eszköz és lehetőség támogatja. A permutált indexek és konkordanciák – azáltal, hogy összehozzák azokat a szövegkörnyezeteket, amelyekben egy szó vagy kifejezés előfordul – nagymértékben segítik a szakembert a relációk felismerésében. A „szemantikai mező térképe", amely a szavak asszociációjának mértékét mutatja, az asszociáció statisztikailag kimutatható erősségét használja fel arra, hogy a mögötte meghúzódó szemantikai kapcsolatot feltárja. Több egyedi jelzetrendszert fejlesztettek ki a szabványosított szótárakban előforduló kapcsolatok reprezentálására – ilyenek például a *Perry* és *Kent* szemantikai kódjai[4] és az Engineer's Jount Council jelzetei, illetve a kettőspontos osztályozások.

Sokan megállnak ennél a pontnál, és feltételezik, hogy a szabványosított szójegyzék vagy tezaurusz önmagában is elegendő. Mások azzal próbálkoznak, hogy a szabványosított szótárat valamilyen osztályozási rendszerrel kapcsolják össze.

Először – egyszerűen a skála egyik végén elindulva – az osztályozás tisztán intuitív megközelítésével találkozunk; eszerint ismeretterület – beleértve ennek szótárát is – rendszerezhető pusztán annak alapján, hogy mit találnak „racionálisnak". Ez a hagyományos osztályozási rendszerek és taxonómiák klasszikus világa.

Másodiknak vegyük a „fazettás elemzésnek" nevezett módszert, amelyben a tezaurusz kifejezései – rendezett sorok formájában – fazetták sorrendjének megfelelő helyzetükkel reprezentálhatók. Így a kifejezések közötti összefüggések egyszerű, könnyen áttekinthető formában jelennek meg.

Harmadszor ott van a hagyományos tárgyszókatalógus besorolási szabályaival példázható struktúra, a fő-, al- és melléktárgyszavas szerkezet. Ezek a szabályok olyan mértékben helyeznek a szótárra (azaz a szabványosított tárgyszójegyzékre) rá egy olyan sorrendet, amely összehozza az egymással kapcsolatban álló szavakat, bár ez csak lokális struktúrákat eredményez.[5]

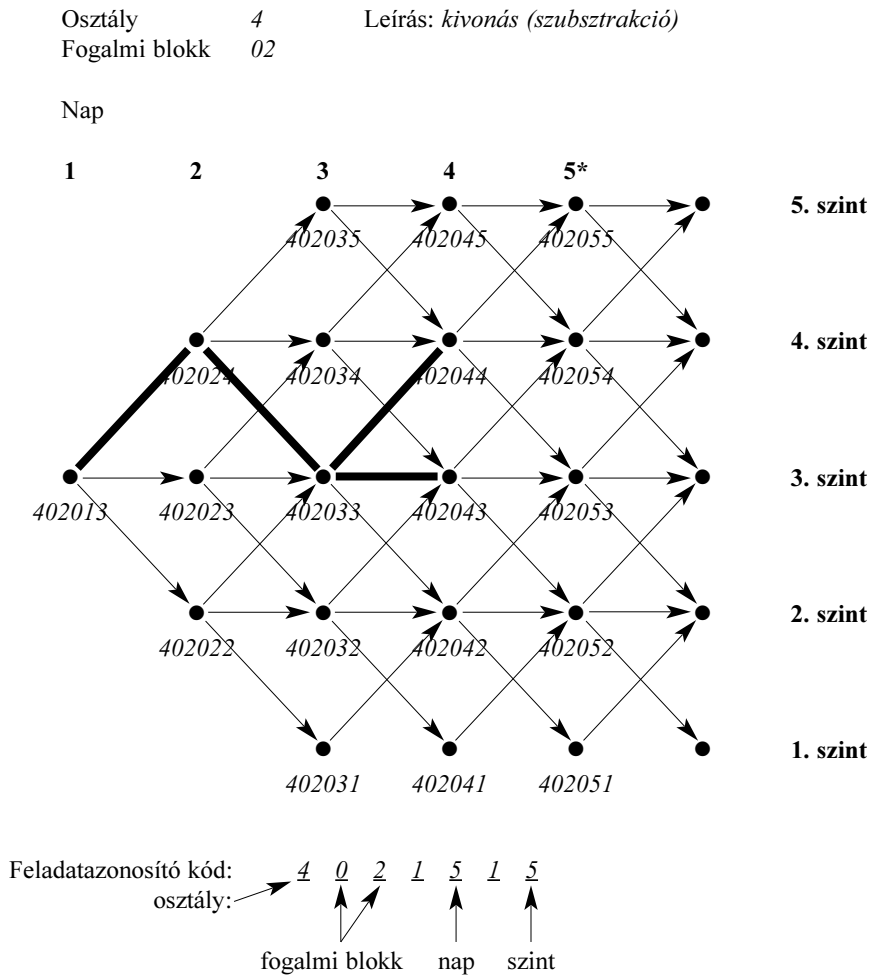
Negyedszer, számos olyan megközelítés van, amely abból a feltételezésből indul ki, hogy a szemantikai összefüggések levezethetők abból a statisztikai kapcsolatból, amely a szavak együttes előfordulásán alapszik. Megkísérlik – matematikai módszerekkel – felbontani a kifejezések közötti asszociáció erősségét leíró mátrixokat. Akár faktoranalízist, akár saját értékelemzést vagy klaszterelemzést alkalmaznak, a megközelítés közel azonos. Az igazi különbség az asszociáció mértékének megválasztásában rejlik. Ez az automatikus vagy numerikus osztályozás világa.

---

4 Lásd az első kötetben *Eric de Grolier* művének szemelvényét (a szerk.).

5 Cutter, Ch. A.: Szótárkatalógus-szabályzat. In: Szilágyi T.: A tárgyszó és a tárgyszókatalógus. (Szemle).

Ember–gép párbeszéd. Az ember–gép párbeszédhez négy dolog tartozik: (1) a megengedett beviteli jelek halmaza, amelynek segítségével az ember kommunikálhat, (2) a gép számára elérhető adatbázisok, (3) azok az üzenetformák vagy „sztereotípiák", amelyek segítségével a gép a beviteli jeleket össze tudja kapcsolni az adatbázisban tárolt jelekkel és ezáltal képes a hatékony kommunikációra és (4) azok a választási stratégiák vagy modellek, amelyekkel az ember, a gép vagy mind a kettő meghatározhatja a sorrendet. A 4. ábrán ilyen párbeszéd döntési folyamata látható (ahogy azt egy számítógéppel támogatott oktatási program szemlélteti).



* Nincs adat a jelentés dátumáról.

4. ábra Tipikus párbeszédes döntési folyamat

Mindez elsősorban a természetes nyelvet kezelő rendszerek számára fontosak. A kommunikációs probléma szinte leküzdhetetlen a gépben (a szótárban) tárolandó szókészlet mérete, a szavak különböző lehetséges jelentéseiből fakadó szemantikai bizonytalanságok, és a géppel felismerhető szintaktikai szerkezetek korlátozott száma miatt.

Indexelés, kivonatkészítés és kódolás. Mivel az információs rendszerek programjai a dolgokat (dokumentumokat, információkat) lényegében külsődleges szempontból (ahogy a dokumentum, az információ megjelenik) képesek leírni, gondoskodni kell olyan eszközökről is, amelyek az egyes dolgok belső, tartalmi leírásához az éppen odaillő jellemzők kiválasztását és a rekordon belül az egymással való összekapcsolását biztosítják. Ezt úgy oldják meg, hogy valamilyen szótárból kifejezéseket jelölnek ki, és ezek reprezentálják a megfelelő rekordokban a tartalmat. A kifejezések rekordokhoz rendelésének eredményét az 5. ábrán látható rekord–ismérv mátrix szemlélteti.

A rekord (dokumentum)–ismérv mátrixszal részletesen foglalkozik a kötet elején Vickery.

	Ismérvek					
	T_1	T_2	…	T_j	…	T_n
R_1	a_{11}	a_{12}		a_{ij}		a_{1n}
R_2	a_{21}	a_{22}		a_{21}		
...			…			
R_i	a_{i1}	a_{i2}		a_{ij}		a_{in}
...					…	
R_m	a_{m1}	a_{m2}				a_{mn}

Az a_{ij} elem értéke 1, ha a T_j ismérvet használják a rekordban, és 0, ha nem használják.

5. ábra. Rekord–ismérv mátrix
Rekordon akár dokumentum (dokumentációs egység) is érthető. *Vickery* például a kötet elején közölt szemelvényben dokumentum–ismérv mátrixról beszél.

Az adott dokumentumot jellemző ismérvek (kifejezések) kiválasztási eljárásai, legyenek azok statisztikai (automatikus), „szintaktikai" vagy „szemantikai" megoldások, mindig a beviteli adatok formáján alapulnak.

E leírások tárolásához elkerülhetetlenül szükséges a „formátumok" vagy adatszerkezetek kialakítása, amelyek elsősorban azt határozzák meg, hogy a kifejezések hogyan rendelhetők hozzá a leíráshoz (rekordhoz); ritkábban a kifejezések kölcsönös kapcsolatainak kimutatásáról is gondoskodnak szerepjelölők, kapcsolatjelölők és szabványosított szintaktikai szabályok segítségével.

A tárolás és a keresés problémái

Az általunk vizsgált második problémakör az adatok tárolásából és kereséséből következik. Az egyik probléma – a fájlszervezés – a katalógusszerkesztés besorolási szabályainak felel meg; a közvetlen adateltérés támogatására ki kell alakítani a mutatók és a keresztutalások rendszerét is. A második kérdéskör a „relevancia" értékelése, a harmadik a „keresési stratégia", amely a tájékoztató munka megfelelője, és a bibliográfiai adatok megtalálásával és értékelésével kapcsolatos.

Fájlszervezés

Különös, de a fájlszervezés kérdését alig tanulmányozzák a dokumentációs szakemberek. Valójában csak a tömegtároló eszközök hatékony felhasználásával foglalkozó adatfeldolgozó szakemberek szentelnek figyelmet ennek. Pedig ez a terület éppen olyan fontos a könyvtárak számára, mint a szótárszerkesztés és a terminológiai ellenőrzés.

Ez a viszonylagos elhanyagoltság talán abból fakad, hogy sokan hajlamosak a fájlszervezést a szótárszervezéssel azonosítani, és úgy vélik, hogy a fizikai szervezés szükségszerűen tükrözi az intellektuálist. Ez a szemlélet annyira felszínes, hogy miatta a fájlszervezést elhanyagolják, mert néhány alapvető problémája egyenesen érthetetlenné válik.

A fájlszervezés szükségességét mi sem bizonyítja jobban, mint az a tény, hogy a nagy fájlok esetén nem lehet a teljes fájl átnézésére vállalkozni a kívánt információ vagy tétel keresésekor. Ezért meghatározott egyszerű kritériumok alapján ki kell választani egy viszonylag kevés tételből álló halmazt, amely aztán részletesen vizsgálható. A feladat nyilvánvalóan olyan megfelelő struktúra kialakítása, amely a halmazok egymás utáni átvizsgálásának mechanizmusáról gondoskodik. A 6. ábrán fájlszerkezet vázlata látható.

A fájlszervezés valamennyi módszerének középpontjában a fájlon belüli sorrend szerepel, mivel a rekordokat a saját helyükön kell tárolni. A sorrend megválasztása rendkívül fontos, hiszen a rekord helye az a hely, ahová a számítógépnek később, a kereséskor vissza kell mennie. A legtöbb rendszerben a fájl rendezését a rekordformátum meghatározott mezője, az ún. Rekordazonosító (például a dokumentum azonosító száma) határozza meg, és a rekordok halmazát az ebben a mezőben található értékeknek megfelelően, numerikus vagy alfabetikus sorrendben tárolják.

A magától értetődő fájlszerkezet az, amely ténylegesen elmossa a szótár szerkezete és a fájlszerkezet közötti különbséget: használd a szótár szerkezetét – a tárgyszavakat, az osztályozási jelzeteket, a kulcsszavakat. Ennek a módszernek az előnye viszonylagos egyszerűsége: ismert megközelítési mód, ezért mindenki számára könnyen érthető. Ha nem nézünk a dolgok mélyére, úgy tűnhet, hogy minden könyvtárban ez a rendezés kritériuma.

Törzsfájl
(a dokumentum–ismérv mátrix sorai)

Cím	Rekord	A törzsfájl mutatója
1	D_1: T_{11}, T_{12} …, T_{1k1}	$D_1 – D_b$: 1. cím
2	D_2: T_{21}, T_{22} …, T_{2k1}	$D_1 – D_b$: *b*+1. cím
3	D_3: T_{31}, T_{32} …, T_{3k1}	$D_1 – D_b$: *2b*+1. cím
…	…	… ,
i	D_i: T_{i1}, T_{i2} …, T_{ik1}	$D_1 – D_b$: n–*b.* cím
…	…	
n	D_n: T_{n1}, T_{n2} …, T_{nk1}	

Keresztutalások mutatója

Cím	Rekord	A keresztutalások indexe
a	T_1: D_{11}, D_{12} …, D_{1p1}	$T_1–T_b$: a. cím
a+1	T_2: D_{21}, D_{22} …, D_{2p1}	$T_{c+1} – T_{2c}$: cím
...	…	…
a+j1	T_j D_{j1}, D_{j2} …, D_{jk1}	$T_{m–c} – T_m$: *a+m–1–c* cím
...	…	
a+m–1	T_m: D_{m1}, D_{m2} …, D_{mk1}	

6. ábra. Fájlszervezés és szerkezet

Létezik azonban másik fájlszervezési módszer, amelyet – szinte öntudatlanul – ugyanilyen mértékben használnak. Ez a „tevékenység” szerinti szervezés módszere, amely azon az elképzelésen alapszik, hogy a fájl azon tételeit tegyük könnyen elérhetővé, amelyeket nagy valószínűséggel használni fognak. Ez természetesen azt jelenti, hogy tudnunk kell, mit értünk azon, hogy „használni fognak” és hogyan mérhető ez? Nem feledkezve meg a problémákról, el kell ismerni, hogy sok kritérium van, amelyeknek lehetnek hiányosságai, de pragmatikusan megfelelőek. Az igazi kérdés az, hogyan oldjuk meg a fájl szervezését a kiválasztott kritériumnak megfelelően. Ennek a megközelítésnek az eredményeként egyetlen gyakran használt könyvet éppen olyan könnyű megtalálni – mind fizikailag, mind intellektuálisan – mint a sok könyvet tartalmazó nagy egységeket.

A harmadik eljárás sokkal távolabb áll a gyakorlattól, a tételek hasonlóságán alapul, és hasonló tételekből alkot csoportokat. Ehhez megint csak feltételezzük, hogy tisztában vagyunk azzal, hogyan határozható meg a hasonlóság. Éppen emiatt számít spekulatívnak ez a módszer. El kell azonban ismerni, hogy a tárgyszavak vagy osztályozási jelzetek szerinti szervezése valójában ugyancsak a hasonlóság egyszerű megállapításán alapszik; a hasonlóságot a közös téma képviseli, a így a relevancia intellektuális értelmezéseinek bővítésével párhuzamosan a fájlszervezéssel kapcsolatos elképzeléseinket is gazdagítanunk kell.

Akármilyen módszert használjunk is a rendezésre, a sorrend kialakítására, a számítógép csak akkor tudja a kívánt rekordot a törzsfájlban megtalálni, ha eleve tudja, hogy helyezkedik el, mi a helye (ez a „közvetlen elérés"), vagy ha végignézi az összes rekordot mindaddig, amíg a keresett rekordot meg nem találja (ez a „soros elérés"). Mivel nagy fájlok esetében a soros elérés nagyon időigényes, olyan mutatókat kell létrehozni, amelyek megmondják a számítógépnek, hogy a kívánt jellemzőjű rekordok milyen helyen találhatók. Jellemző példa erre az „invertált fájl" vagy „ kifejezés orientált fáj", amelyben a szótár minden egyes rekordja (pl. minden tezaurusz deszkriptor) számára egy „mutató rekordot" tárolnak (más szóval minden deszkriptor mutatónév az indexben). Az egyes mutató rekordok a törzsfájl összes olyan dokumentum rekordját felsorolják, amelyhez az adott kifejezést hozzárendelték (tehát ez tárolja a rekord–ismérv mátrix oszlopait, a törzsfájl pedig a mátrix sorait tartalmazza).

A „lista" fogalmán alapszik egy másik indexelési struktúra. A lista olyan rekordok egymásutánja, amelyben minden rekord tartalmazza a listában utána következő rekord címét (kivéve az utolsót, amely az első rekordra mutat, és így zárt lista keletkezik). Létrehozható lista úgy is, hogy adott ismérvet megfeleltetünk a listában azokkal a rekordokkal, amelyeket az ismérvvel jellemeztek. Az „ismérv fájlnak" csak a lista első rekordjának tárolóhelyi címét kell tartalmaznia, a következőket már rekordról rekordra címezik.

Relevancia. Az információs rendszerek feladata az igények szempontjából releváns adatok elemzése. Az igény szempontjából releváns adatok körének meghatározása meglehetősen problematikus. A gondok részben a felhasználói követelményekkel, részben az üzemeltetés módjával, részben az üzemeltetés hatékonyságával kapcsolatosak. Különböző, egymást követő „szűrőket" kell használni, amelyek valamilyen módon mérik a relevanciát vagy az egyezés mértékét. Az utolsó szűrő maga a felhasználó, aki saját kritériumait alkalmazza a tárolt adatok legbonyolultabb formáira. Mivel ezt nem teheti meg a fájl összes leírásával, előválogató szűrőkkel csökkentik a fájlt olyan méretűre, amelyet már át tud tekinteni.

A kérdések különböző formában érkeznek. Különböznek részletességükben, a leírás pontosságában, az érdeklődési körben és a kívánt információ for-

májában. Némelyek ismétlődők, és formanyomtatvánnyal elintézhetők. Mások „folyamatos kérések", amelyeket a szelektív információszolgáltatáshoz (SDI) használnak, amint az adatok beérkeznek. Van a kéréseknek olyan csoportja is, amelyek személyre szabott keresést igényelnek a fájlban. A kérdést minden esetben olyan formában kell kifejezni, amely összevethető a tárolt rekordokkal, azonosítva azokat az ismérveket, amelyeknek a kívánt rekordok meghatározott mezőjében meg kell jelenniük. Az adatbázisrendszerekben sok „ismérv" számérték formájában jelenik meg a mezőben. Ebben az esetben akár érték, akár értéktartomány megadható a kérésben az „egyenlő", „nagyobb mint", „kisebb mint" kapcsolatok formájában. A kifejezések ezután Boole-algebrai eszközökkel kapcsolhatók össze, a logikai ÉS, VAGY, NEM fölhasználásával. Például a kérdés előírhatja, hogy az 1. mező tartalmazza az A_1 vagy A_2 vagy A_3 kifejezéseket, a 2. mező tartalma legyen nagyobb mint *B* és a 3. mező ne tartalmazza a *C* ismérvet. A kérés formája:

1. mező: (A_1 VAGY A_2 VAGY A_3) ÉS 2. mező(:>*B*) ÉS 3. mező: (NEM *C*)

A bibliográfiai információkereső rendszerekben azoknak a kifejezéseknek a megállapításához, amelyek a legjobban kifejezik a kérést, rendszerint a kérdező és a tárolt tezaurusz közötti párbeszédre van szükség. A tezaurusz segít abban, hogy a kérdező által használt szavak elvezessenek a szótárban használt megfelelőkhöz. Időnként a kért kifejezést „fel kell robbantani", mert a tezaurusz azt jelzi, hogy a kifejezés több hasonló vagy specifikusabb értelmű szóhoz kapcsolódik, amellyel a dokumentumok szintén leírhatók. Ezeket a kifejezéseket a kérés leírásakor logikai VAGY kapcsolatba kell összefűzni. A felhasználó például az *A* témát kéri, és a tezaurusz azt mutatja, hogy A_1 és $\underline{A_2}$ specifikusabbak, *B* téma pedig hasonló. Az *A* kifejezés kiterjeszthető, a kérdés a következő lesz:

Tárgy (*A* VAGY A_1 VAGY A_2 VAGY *B*)

Szövegfeldolgozó rendszerekben a kérdést ugyanúgy lehet megfogalmazni, mint adatbázis vagy bibliográfiai információkereső rendszer esetében, vagyis nagyjából hasonló problémákkal küzdve. Másfelől viszont a szövegfeldolgozás egyik könnyebbsége, hogy a kérdéseket egyszerű természetes nyelvű mondatok formájában fogalmazhatjuk meg. Természetesen a fordításhoz és feldolgozáshoz szükséges programok bonyolultsága hasonló mértékű, mint a tárolt szöveg kezeléséhez szükséges programoké.

Ha a kérdést már megfogalmaztuk, meg kell vizsgálni a rekordokat, relevánsak-e. Mivel a relevancia fogalma rosszul definiált, a számítógépes program az „egyezés mértékét" méri egyszerűen; például azt, hány olyan kifejezés jelenik meg a tárolt rekordban, amely a kérdésben is szerepelt. A leginkább megfelelő rekordokat ezután relevánsnak tekintik, legalábbis a számítógépes feldolgozás szempontjából. A nagy fájlok esetében azonban rengeteg idő elmegy az-

zal, hogy minden rekordot összehasonlítanak a kérdéssel, ezért indexfájlokat használnak, hogy a megfelelő rekordokat megtalálják. A kérdés összehasonlítása az indexrekordokkal a fájlkeresés.

A keresési stratégia. A fájlok keresése rekordok és szurrogátumaik egymásután következő átnézéséből, a kérdésekkel való összehasonlításból, és a releváns rekordok kiválasztásából áll. A megfelelés bizonytalansága miatt azonban (aminek egyrészt a leírás bizonytalansága és a benne előforduló hibák az okai, másrészt az, hogy a szurrogátumok tételcsoportokat képviselnek), a keresési stratégia megválasztásának is komoly szerepe van. A fájlban való keresés-kor alapvető döntés, hogy a keresés irányát megtartsuk-e, vagy másik irányba induljunk el. A döntéskor figyelembe kell venni a válaszadás idejét, az eredmény megbízhatóságát, és magának a keresési folyamatnak a költségét.

A keresési folyamat eredményeként azoknak a rekordoknak a halmazát kapjuk vissza, amelyeket a rendszer a felhasználó szempontjából „releváns-nak" tart. Ekkor válik nyilvánvalóvá, mennyire hatékony a teljes rendszer. A felhasználó rendszerint meg akar bizonyosodni arról, hogy a releváns tételek nagy részét megtalálta. Egyébként nagyfokú bizonytalanságot érez, vajon nem hiányoznak-e azok a tételek, amelyekre igényt tartott?

Továbbá, arról is meg akar bizonyosodni, hogy a kapott anyag nagyobb részben nem irreleváns-e, különben saját idejét kell elfecsérelnie az irreleváns anyagok kiszórásával. Ezt a két igényt tükrözi a teljesség és a pontosság arány. Az előbbi mérése rendkívül nehéz, hiszen nem tudjuk, hány releváns dokumentum volt valójában a fájlban. Az utóbbit könnyű mérni, és ez a költség/hatékonyság egészen egyértelmű mértékéhez vezet.

Az adatfeldolgozás és -megjelenítés problémái

A problémák harmadik csoportja, amelyről beszélünk, az adatok elemzéséből és megjelenítéséből ered. Az egyik, az „adatszervezés", viszonylag egyszerű, az adatok előírt sorrend szerinti elrendezését jelenti. A másik, a „kérdések megválaszolása" olyan összetett, hogy egyes kutatók érzése szerint nem is gépesíthető.

Adatszervezés. A fájlkezelő rendszerek általában alkalmasak kimutatások készítésére. Ez azt jelenti, hogy elő lehet írni az output formákat, rendezési sorrendeket és statisztikai összefoglalásokat. A keresés eredményeként megtalált bibliográfiai tételek rendezhetők például a relevancia sorrendjében, időrendben, szerző szerinti betűrendben stb. A KWIC index tipikus forma, amely a címeket a bennük előforduló kulcsszavak sorrendjében adja meg. Az előírt sorrendeket nem nehéz előállítani.

Kérdések megválaszolása. Ezzel szemben a kérdések megválaszolásához szükséges logikai következtetések levonása a tárolt adatokból rendkívül sok problémát vet fel. Programok készültek a szillogizmusok levezetésére, amelyek elemzik a következtetéseket, valamint a predikátum-kalkulus alapján a rekordok közötti kölcsönös kapcsolatokat. A 7. ábrán néhány kísérleti kérdés–válasz rendszert soroltunk föl. Ezek a rendszerek meglehetősen kísérleti jellegűek, több szempontból is korlátozott a felhasználásuk.

Rendszer	Folyóirat/év	Nyelvi adatfeldolgozás	Keresés	Adatredukció
Phillips ORACLE	1960	Szintaktikai modell (SVO, idő, hely)	A szintaktikai modellben szereplő szavak egyszerű összevetése	
Sable IDL	C–ACM 1962	Generikus relációk		
Cooper Tényadat-keresés	J–ACM 1964	Angol alnyelvű. Fordítás nyelvtani osztályok segítségével „logikai" megfelelőkre. Specifikus algoritmus az összetevő-struktúrájú nyelvekre, amelyek áthágják a nyelvtani szabályokat	Három tárolt adatmondatig az összes lehetséges alhalmaz, amelyek közül mindegyik tartalmazza a kérdés legalább egy alapvető kifejezését	Arisztotelészi logika
Salton SMART	1964	„Szemantikai" osztályokat tartalmazó specifikus szintaktikai szerkezetek	A kifejezések összevetése a klaszterált értékekkel	Statisztikailag elemzett fastruktúra
Black SQA	1964	Állandó formátumok	A szavak és szerkezetek pontos megfelelése	Következési szabályok

C–ACM = Communications of the Association for Computing Machinery;
J–ACM = Journal of the Association for Computing Machinery

7. ábra. Néhány kérdés–válasz rendszer

FREDRICK WILFRID LANCASTER (1933) ÉS EMILY GALLUP FAYEN

F. Wilfrid Lancaster a könyvtártudomány professzora az illinoisi egyetemen (Graduate School of Library Science, University of Illinois). A hatvanas években igazgatója volt az orvosbiológiai könyvtári fejlesztési programnak (Program in biomedical librarianship), később kurzusokat szervezett az információtárolás és -keresés, az információszolgáltatás tesztelése, az automatizált szövegelemzés és a terminológiai ellenőrzés tárgykörében. Több állami és kereskedelmi ügynökség tanácsadója. Érdeklődésének középpontjában az információkereső rendszerek jellemzésének, vizsgálati módszereinek és értékelésének kérdései állnak. Tagja a brit könyvtáros egyesületnek (British Library Association), az információtudományi társaságnak (American Society of Information Science) és az osztályozáskutatási csoportnak (CRG).
Szakmai hírnevét a hatvanas évek második felében a MEDLARS fejlesztésében való közreműködésével alapozta meg. Az erről írt beszámolója 1969-ben a legjobb amerikai dokumentációs tanulmány díját nyerte el. *Robert M. Hayes, Allen Kent* és *Charles Meadow* mellett egyike volt az elsőknek, aki az információkereső rendszerekről kézikönyvet[6] adott ki, melyet 1970-ben az amerikai információtudományi társaság (American Society for Information Science; ASIS) „az év legjobb könyve" díjjal jutalmazott. Az információkereső nyelvekről 1972-ben írt könyvéből az első kötetben szerepel részlet. Emily Gallup Fayennel együtt elsőként írt kézikönyvet 1973-ban az on-line információkeresésről, melyből az alábbi részlet származik.

Emily Gallup Fayen számítástechnikai szakember, a Computer Sciences Corporation munkatársa.
Újabban a szakértői rendszerek információkeresési célokra való alkalmazásával foglalkozott.[7]

6 Lancaster, F. W.: Information retrieval systems: characteristics, testing and evaluation. – New York [etc.]: Wiley,. 1968. (Information sciences series.)

7 F. W. Lancaster and Lina C. Smith [Ed.]: Artificial intelligence and expert systems. Will they change the libraries. Papers presented at the 1990 clinic on library applications of data processing. March 25–27. 1990.]. Urbana-Champaign: University of Illinois, 1992. 291 p. – (Clinic on library applications of data processing) (Ism. Koltai Tibor, Könyvtári Figyelő, 1994, 4. (40.) köt., 3. sz. p. 426–430.).

On-line információkeresés[8]

Az on-line információkereső rendszerek jellemzői[9]

Megjegyzés: Ebben a könyvben az információkeresés kifejezést úgy használjuk, ahogy azt általában a szakirodalomban, *vagyis* olyan rendszer jellemzésére, amely dokumentumokra vonatkozó információk – rendszerint a dokumentum bibliográfiai vagy tartalmi leírásának, ún. szurrogátumának (például a cím, a kivonat, a referátum) – visszanyerésére képes. Ezen kívül elsősorban a tartalom, a téma és nem a szerző vagy más ismérv szerinti keresést tárgyaljuk.

Vegyük egyszerűen úgy, hogy on-line információkereső rendszer az, amelyben a felhasználó közvetlen kapcsolatot teremthet számítógép segítségével a dokumentumok vagy dokumentumleírások géppel olvasható adatbázisával. Az on-line információkereső rendszerben a számítógép és a felhasználó között kétirányú a kommunikáció, amely az input/output berendezéseken (ilyen berendezés például az írógép), kommunikációs csatornán, – mint amilyen a közönséges telefonvonal is – keresztül valósul meg, a számítógéphez kötött képernyős terminál segítségével. Bár az on-line rendszer „dedikált" (kijelölt) módon – egy használó számára fenntartva is működhet, gyakran kivitelezik időosztásos környezetben. Az „időosztás" kifejezést többféleképpen határozzák meg. Lényegében azt az üzemmódot jelenti, amikor a számítógépen egy időben két vagy több, egymástól teljesen független tevékenység folyik, azaz megosztoznak a feldolgozási időn. Az időosztásos rendszerben különböző felhasználók egyidejűleg férhetnek hozzá a számítógéphez. Az on-line időosztásos rendszer számos, egymástól független, egyidejűleg használható terminálon keresztül működik. Eközben mindegyik terminál használója akkor kap feldolgozási időt, amikor szüksége van rá, s így az az illúziója (az idő nagy részében), hogy ő a számítógép egyetlen felhasználója. Egy másik, az on-line fogalmához társuló kifejezés az időazonosság. Az időazonos működés azt jelenti, hogy a számítógép megkapja és feldolgozza az adatokat, az eredményeket pedig elég gyorsan küldi vissza ahhoz, hogy ezeket az eredményeket az éppen folyamatban lévő feladat folytatásához felhasználhassák. Az információkeresés esetében az időazonosság azt jelenti, hogy a számítógép

8 Information retrieval on-line / F. Wilfrid Lancaster, Emily G. Fayen. – Los Angeles : Melville, 1973. (Information sciences series).

9 Some characteristics of on-line retrieval systems. In: Information retrieval on-line, p. 1–4.

elég gyorsan válaszol ahhoz, hogy párbeszédben maradhasson a felhasználó heurisztikus keresési folyamatával. Ezzel szemben a késleltetett idejű feldolgozás a kötegelt feldolgozású rendszereket jellemzi. Az időazonos információkereső rendszerek lényegében olyan gyorsan tudnak válaszolni kérdésekre, hogy válaszadásuk azonnalinak vagy majdnem azonnalinak tekinthető. Az online, időazonos rendszerben a felhasználó és a számítógép között párbeszéd lehetséges.

A digitális számítógépek körülbelül 1948 óta léteznek, de felhasználásuk az információkeresésben a hatvanas évektől kezdődött. Az USA-ban a legtöbb nagy számítógépet felhasználó információkereső rendszer 1965 után keletkezett. A hardver fejlődése az információkeresés szempontjából a következőképp foglalható össze:

1. Az 1940-es évek előtt: főleg kártyakatalógusok és nyomtatott, könyv alakban közreadott mutatók.
2. 1940 és 1949 között: a félautomatikus megoldások első változatai, beleértve a peremlyukkártyákat és az optikai egybeesés elvén alapuló találatképzést (a fénylyukkártyákat); az első mikrofilmes táron alapuló információkereső rendszerek (gyorsválogatók).
3. 1950 és 1959 között: az első szélesebb körben használt lyukkártyás adatfeldolgozó berendezések, néhány korai számítógépes rendszer, fejlettebb mikrofilmes információkereső rendszerek voltak forgalomban.
4. 1960 és 1969 között: egyre gyakrabban alkalmazták a digitális számítógépeket az információkereséshez off-line kötegelt feldolgozású üzemmódban; elszórt kísérletek folytak on-line párbeszédes üzemmódú rendszerekkel; továbbfejlődtek a mikrofilmes információkereső rendszerek.
5. 1970-től mostanáig: az on-line rendszerek tervezése és kötegelt rendszerek átalakítása on-line üzemmódra mindinkább meghatározó jellemzői a fejlődésnek. A kézi információkereső rendszereket – mint amilyenek például a nyomtatott mutatók (indexek) és a kártyakatalógusok – a következők jellemzik:

    1. Véletlen elérés. Közvetlenül a fájlhoz lehet fordulni és csak azt a részt kell megnézni, amely szükséges.
    2. Interaktív (párbeszédes) üzemmód: a tételek között átfutó, böngésző, a legjobb eredmény elérésének rendszerére rávezető, ún. heurisztikus keresés lehetséges.
    3. Egy kereső egyszerre csak egy tárgy alapján kereshet. Egyszerre általában csak egyetlen témakörben kereshető információ. Ugyanabban az időben viszont egyszerre több felhasználó kereshet ugyanabban a fájlban.

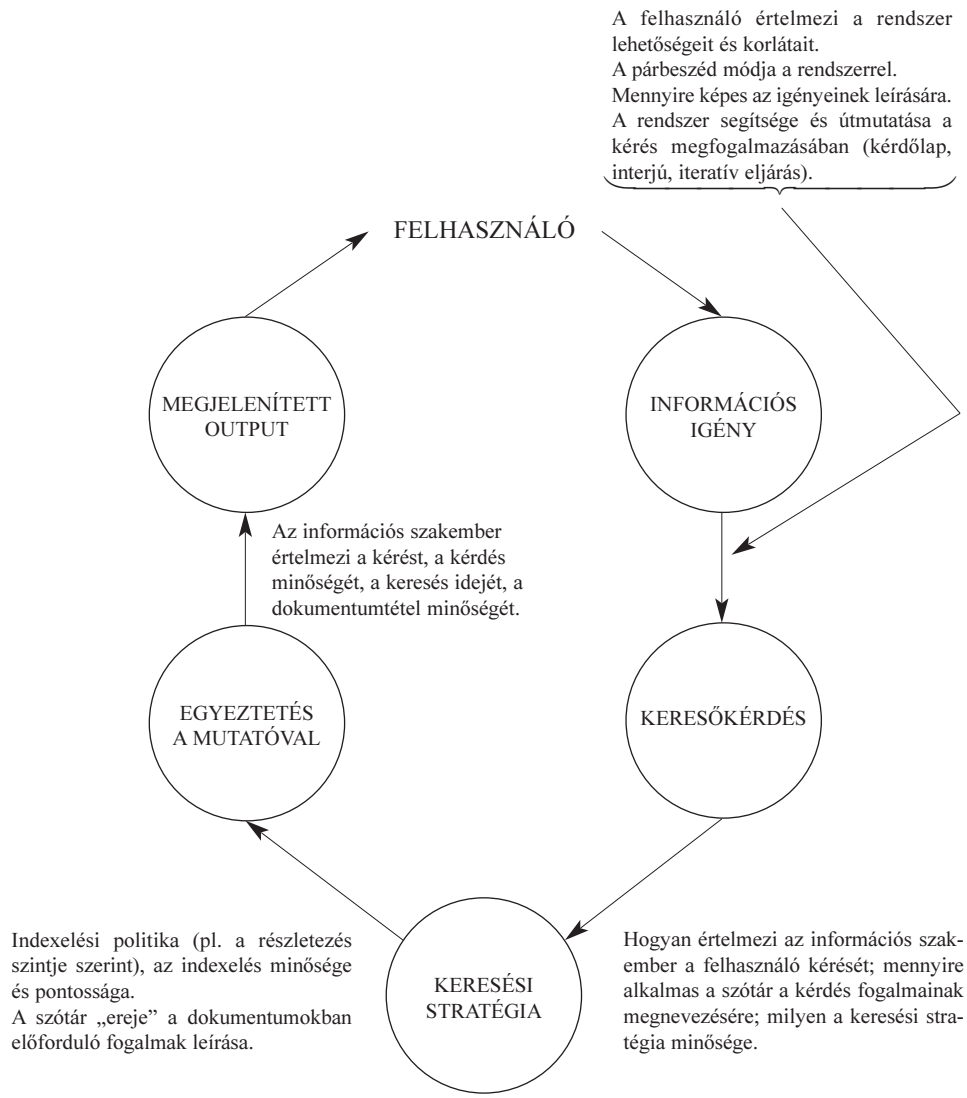
4. Az információ igénylője (többnyire a téma szakértője) saját maga végezheti a keresést a fájlban, ha úgy kívánja; nem kell az irodalomkutatást könyvtárosra vagy más információs szakemberre (közvetítőre) bíznia.
5. Semmiféle időkésleltetés nincs. Mivel fizikailag bárki hozzáférhet a fájlhoz, ezért a keresés akkor végezhető, amikor az információ iránti igény fölmerül.

A lyukkártyák és később a számítógépes rendszerek alkalmazása a keresés technikáját teljesen megváltoztatta. Az első gépesített rendszereket off-line, kötegelt üzemmódra tervezték. Ennek az üzemmódnak a hátrányai a következők:

1. Kevés a lehetőség a böngészésre.
2. A keresési stratégia nem alakítható ki heurisztikusan. Ha a kereső kérdését a tárnak „feltette", közbenső eredmények alapján már nem avatkozhat be a keresési folyamatba. A sikeres keresés érdekében ezért előre át kell gondolnia a keresés összes lehetséges megközelítését.
3. A keresést információs szakemberre kell bízni. Az információs szolgáltatás vevője nem képes a keresést saját maga elvégezni. A keresés átruházása azonban problémákat okoz. A felhasználók néha nagyon nehezen tudják csak leírni, mit akarnak keresni, a keresést végzők viszont félreérthetik a felhasználók kéréseit. Ezek a problémák az összes megbízáson alapuló keresőrendszerben megvannak.
4. A kötegelt üzemű rendszerekben a keresés és a keresési eredmények kézbevétele között elkerülhetetlen az időkülönbség.

Az on-line üzemmódban használható információkereső rendszerekre a fenti hátrányok egyike sem jellemző. Csak az on-line üzemmód rendelkezik a gyors válasz, az interaktivitás, a böngészés és a heurisztikus keresés előnyeivel, és ezt nem ellensúlyozza, hogy az off-line üzemmódban a keresést képzett információs szakemberre bízzák („delegálják"). Végül pedig, mint arra a fentiekben utaltunk, az on-line rendszereket „nem delegált" (nem átruházott) keresési módban is lehet használni, azaz a téma szakembere saját maga végezheti el a keresést információigényeinek kielégítésére. Így hatékonyabb a keresés és elkerülhetők a félreértések.

Az 1. ábra a „delegált" (szakemberre bízott) információkeresés folyamatában előforduló lépéseket mutatja be attól kezdve, hogy a felhasználó először lép kapcsolatba a rendszerrel, addig, hogy a rendszer megadja a választ. Az ábrán szerepelnek a keresés sikerét, illetve sikertelenségét a ciklus egyes lépéseiben leginkább befolyásoló tényezők is.



1. ábra. A kijelölt keresési módban használt információkereső rendszer teljesítményét befolyásoló tényezők

Az ábrából világosan kitűnik, hogy a kérdező felhasználó és a keresést végző információs szakember közötti kapcsolatrendszer (az ún. interfész) a teljes folyamat szempontjából kritikus. Bármennyire kitűnő az információkereső nyelv szótára és az ezen alapuló indexelés, aligha számíthatunk sikeres keresésre, ha a felhasználó gépre – pontosabban: a gép által kezelhető infor-

mációkereső nyelvre – fordított kérése nem tükrözi pontosan a felhasználó tényleges igényeit, vagy ha az információs szakember alapvetően félreérti a felhasználó kérdéseit.

Az on-line rendszerben a felhasználó maga kerül közvetlen kapcsolatba az adatbázissal, s ezért ezek a problémák elkerülhetők. Ebben a helyzetben viszont nagy valószínűséggel más problémának lesz nagyobb jelentősége. Ha a felhasználó nem információs szakember, valószínűleg kevésbé járatos a rendszer szótárának használatában, az indexelési stratégiákban és eljárásokban. Keresési stratégiájának minősége feltehetően ennek megfelelő lesz. A gyakorlati szakemberek közvetlen használatára szánt on-line rendszerekben a tervezés és az implementálás – a használatbavétel – problémái teljesen mások, mint az off-line rendszerekben, amelyeket „delegált" keresésre szántak.

AZ INFORMÁCIÓKERESÉS ÉRTÉKELÉSE

Az információkeresés és e tevékenység értékelésére tett kísérletek szinte egyidősek. Az elsőre 1953-ban az Egyesült Államokban (az ASTIA tezaurusz használatával kapcsolatban) került sor, egy évvel később pedig Angliában *Cyrill W. Cleverdon* végzett vizsgálatokat. Az értékelés két alapvető premisszáját, a visszahívás (recall) és a pontosság (precision) fogalmát először *William Perry* használta a szemantikai kódok használhatóságának elemzésével kapcsolatban (ők még pertinenciatényezőnek, *Cleverdon* 1963-ban relevancia-tényezőnek nevezte; mai nevét *Gerard Salton* adta 1965-ben).

Noha az értékelés premisszái mind az intellektuális keresés, mind pedig az automatizált információkereső rendszerekkel végzett keresés esetén lényegében azonosak, meglehetős különbség van a kétfajta keresés és értékelésének tartalmában. Ma az automatizált információkereső rendszerek értékelése áll inkább a középpontban. E téren a helyzetet az értékelés szakavatott művelője, *Cornelis van Rijsbergen* így jellemzi:

„Sok erőfeszítésre, kutatásra került már sor az információkereső rendszerek értékelési problémájának megoldása érdekében. Mégis, alighanem joggal állíthatjuk, hogy nagyon sokan, akik az információtárolás és -keresés területén dolgoznak, még mindig úgy érzik, hogy messze vagyunk a megoldástól. Az erőfeszítések méreteiről képet kaphatunk, ha megnézzük azt a nagyszámú áttekintő cikket, amelyet e témában publikáltak (köztük az Annual Review of Information Science an Technology állandó fejezetét az értékelésről)." [p. 142][1]

1 Az összes szögletes zárójelben megadott oldalszámú van Rijsbergen-idézet forrása a szerző magyar fordításban megjelent, 1979-ben írt könyve: Információ-visszakeresés. – Budapest: Múzsák Közművelődési kiadó, 1987.

Az intellektuális keresések értékelésében, ha lehet, még nagyobb a bizonytalanság. Egy 1991-ben a DIALOG rendszerben végzett on-line keresésre vonatkozó vizsgálat a következő összegzéssel zárul:

„Az eredmények igen kevés törvényszerűséget mutattak. A szakma valószínűleg nincs is tudatában, hogy a keresésekben mennyire kevés a megegyezés… Az on-line keresés messze van attól, hogy tudománynak lehessen nevezni…. Nincs semmiféle elfogadható módszer, átfogó és következetes irányelv, mely az on-line információkeresés útmutatójaként szolgálhatna. A keresés még mindig művészet, mégpedig meglehetősen kevéssé megfogható és meghatározható művészet."[2]

Ha jobban meggondoljuk, ez nem is olyan nagy baj: a művészet segítségével a legnehezebb fajta emberi problémák oldhatók meg. (A keresésben alkalmazható „művészi" – pontosabban heurisztikus – módszerekről kötetünk előző részében *Marcia John Bates* szemelvényében olvashatunk.) Megint *van Rijsbergent idézzük.*

„Az egzakt szemlélet tükrében – talán éppen a keresés »művészi« jellege miatt – a keresőrendszerek értékelése különösen nehéznek bizonyult…

Cleverdon és társainak úttörő munkája és az azóta javasolt számos hatékonysági mérőeszköz ellenére az értékelés általános elmélete még várat magára…" [p. 15]

„Arra a kérdésre, hogy mit értékeljünk, mi az, amit mérni lehet, már 1966-ban *Cleverdon* választ adott. Munkájában hat fő mérhető mennyiséget sorolt fel:

(1) a gyűjtemény *lefedő* volta, vagyis az, hogy milyen mértékben tartalmaz a rendszer releváns dokumentumokat;
(2) az *időtényező,* vagyis az, hogy mekkora az átlagos időeltérés a keresési kérdés feltétele és a válasz megadása között;
(3) az output *megjelenítési* módja;
(4) az *erőfeszítés* a használó részéről annak érdekében, hogy keresőkérdésére a választ megkapja;
(5) a rendszer *teljessége,* vagyis a megtalált releváns dokumentumok aránya az összes (akár talált, akár nem) releváns dokumentumhoz viszonyítva…;
(6) a rendszer *pontossága,* vagyis a talált dokumentumok összességén belül a releváns dokumentumok aránya." [143]

2 Kaular, Paul: On-line searching. Still and imprecise art. In: Library Journal, 1991, 116. vol, No. 16, p. 47–51.

Az első négy könnyen becsülhető. Mérni igazán a teljességet és a pontosságot kellene, e kettő fejezi ki a leginkább az információkereső rendszer hatékonyságát. Más szóval feltételezik, hogy ez a mértéke a rendszer azon képességének, hogy megtalálja a releváns dokumentumokat, s ugyanakkor visszatartja az irrelevánsakat. *Cleverdon* óta se szeri, se száma a legkülönfélébb ilyen jellegű vizsgálatnak.
A baj csak az, hogy e két paraméter alapja nem kezelhető ugyanolyan egzakt módon, mint maga a két paraméter. *Van Rijsbergen* szerint

„[A teljesség és a pontosság] valamilyen módon még mindig igényli a relevancia meghatározását… A relevancia [azonban] szubjektív fogalom. Különböző használók különböző álláspontot foglalhatnak el adott dokumentumnak egy kérdésre vonatkozó relevanciáját vagy irrevelanciáját illetően… a kísérleti eredményeket… rendszerint jóhiszemű használókból csalták ki, olyanoktól, akik valamilyen szakterületen dolgoznak és információs igényeik vannak… Olyan helyzetben vagyunk tehát, hogy nagyszámú kérdéshez ismerjük a 'helyes' válaszokat. Az információkeresés területén általános feltételezés, hogy ha egy információkereső stratégia elegendően jól működik nagyszámú *kísérleti* feltétel mellett, akkor vélhetően jól fog működni *gyakorlati* szituációban is, amikor a relevanciát nem ismerjük előre". [144]

Más szóval mindannak, ami objektív – a pontosság és a teljesség –, olyasmi az alapja, ami szubjektív. Ez minden értékelés alapvető ellentmondása az információkeresés világában. Egy jól nevelt természettudós vagy matematikus ezért egy kicsit olyan viszonyban van a relevanciával, akár az ördög a szenteltvízzel.
Még súlyosabb a helyzet a pertinenciával. A relevancia azt fejezi ki, hogy mekkora a közelség a felhasználói kérdés és a talált dokumentumok tartalma között (azaz azok a dokumentumok, melyek a kérdésnek megfelelnek, relevánsak); a vele szorosan összefüggő pertinencia nem más, mint a felhasználói információszükséglet és a talált dokumentumok tartalma közötti megfelelés (tehát nem pusztán a kérdést, hanem azt, ami – irgalom atyja ne hagyj el – a „tudaton belül van", kellene összevetni az eredménnyel). Ezért aztán az olyan szerzőnek a könyvében, mint például *van Rijsbergen*, a pertinencia még mutatóban sem fordul elő.
Az *M. E. Maron* és *J. L. Kuhns* által 1960-ban megalkotott valószínűségi relevanciamodell segítségével bevezetett revelancia-valószínűség a matematikailag megalapozatlan relevanciával szembeni kritika méregfogát hivatott kihúzni.
Figyelemre méltó, hogy *Mortimer Taube* 1965-ben lándzsát tört a nyilvánvalóan szubjektív relevancia fogalom használata mellett, és élénken tiltakozott mindenfajta „matematizált" relevancia bevezetése ellen. Az

utóbbin alapuló képleteket, melyekkel az információkereső rendszerek hatékonyságát számszerűen értékelhették, pszeudo-matematikai konstrukcióknak tekintette.[3]
Van Rijsbergen beszámol arról, hogy elvileg ugyan van olyan relevancia-fogalom, amelyet objektívnek tekinthetünk, s amelyet „logikai relevanciának" nevezhetünk.

„…ez a relevancia-fogalom jelenleg nagyon korlátozottan használható csupán. Ennek fő oka, hogy olyanfajta rendszert, amely ahhoz kellene, hogy olyan információkereső stratégiát alkalmazzunk, amely csak a logikailag releváns dokumentumokat keresi vissza, még nem hoztak létre… Lehetséges, hogy egy ilyen típusú rendszer mérete túlzott a mai számítógépek számára; de a végső szót majd a jövő mondja ki." [145–146]

Azóta, hogy a fenti sorokat leírták, több mint tíz év telt el, feltehetően túl rövid idő ama bizonyos jövendő bekövetkeztéig.

A pontosság, a teljesség és a relevancia értékelésének kezdetei a hatvanas évek elejére nyúlnak vissza és a már említett angliai könyvtároshoz, *Cyrill W. Cleverdonhoz* köthetők: *Jack Millsszel* és *Michael Keennel* együtt az Aslib megbízásával végzett ún. I. (1962) és II. (1966) cranfieldi vizsgálatokban a relevancia fogalmán alapuló keresési teljesség és a pontosság alapján értékelték a keresési eredményeket. Fontos szerepet játszottak ebben az időben *Hans Peter Luhn* eszméi is a szavak gyakoriságának felhasználhatóságáról az információkeresésben. A matematikai logika két művelője, *M. E. Maron* és *J. L. Kuhns* feltehetően először használták 1960-ban a relevancia fogalmát. A kötetünkben nem szereplő *Tefko Saracevic* írta később a legismertebb monográfiát erről a fogalomról, 1975-ben pedig megfogalmazta a „relevancia filozófiáját"[4]. *Cornelis van Rijsbergen* magyarul megjelent művében részletesen összegezi az automatizált információkereső rendszerekre vonatkozó, 1985-ig végzett kutatásokat.

3 Taube, M.: A note on the pseudo-mathematic of relevance. In: American Documentation, 1965, Vol. 16, p. 69–72.

4 Saracevic, T.: Introduction to information science. – New York; London: P. R. Bowker, 1970.

Saracevic, T.: Relevance. A review of a framework for the thinking on the Notion in information science. In: Journal of the ASIS, 1975, Vol. 26, p. 321–343.

CYRILL W. CLEVERDON

Az értékelés egyik legjelentősebb úttörője, aki azóta is mindig hallatta a szavát, ha pontosságról és teljességről volt szó, Cyrill W. Cleverdon, a hatvanas években az angliai Cranfield aeronautikai főiskoláján a könyvtár igazgatója volt és az Aslib támogatásával 1962-ben (Cranfield I)[5] és 1966-ban (Cranfield II)[6] az addigi legnagyobb szabású két összehasonlító vizsgálatot végezte el az információkereső rendszerek használatával kapcsolatban. Vizsgálata klasszikus példája lett az információkeresés kísérleti megközelítésének, számos olyan jó ötletet tartalmaz, melyet később a szakirodalomban mások részletesebben is kifejtettek. Az első jelentést a kezdeti lelkesedés és az eredmények ellenőrzése után kiábrándulás fogadta. Mivel a legtöbb kifogás arra vonatkozott, hogy a túlságosan életszerű vizsgálati körülmények túl sok ellenőrizhetetlen változót eredményeztek, a második vizsgálatot „laboratórium körülmények" között hajtották végre, hogy mindig csak egyetlen jellemző változását mérhessék. A vizsgálatokba Cleverdon bevonta *Jack Millst*, *Bliss* osztályozási rendszerének gondozóját és az információkereső rendszerek tesztelésével professzionálisan foglalkozó *Michael Keent*, a wales-i főiskola tanárát. A két „jelentésről" és az egyik követő vizsgálatról Horváth Tibor írt 1967-ben részletes, értő beszámolókat. *Keen* a későbbiekben az automatikus indexelés vizsgálatával is foglalkozott.

A felsorolt szerzőktől, illetve a tárgykörben magyarul megjelent:

Horváth Tibor: A második cranfieldi jelentés

In: Babiczky Béla: Szöveggyűjtemény az osztályozás és indexelés kérdéseinek tanulmányozására. – Budapest : Tankönyvkiadó, 1970. p. 213–230.

Az első cranfieldi jelentés legfontosabb és máig bíztató hatású megállapítása, hogy a gyakorlott osztályozó specialisták a szakterület közelebbi ismerete nélkül is ugyanolyan jól osztályoznak, mint adott esetben az aerodinamikai szakemberek. (Ebben feltehetően – akár-

5 Cleverdon, C. W.: Aslib cranfield research project. Report on the testing and analysis of an investigation in the comparative efficiency of indexing systems. Cranfield: College of Aeronautics, 1962. 305 p.

6 Cleverdon, C. W., Mills, J, Keen, M.: Factors determining the performace of indexing systems. Vol. I.: Design. Vol. II.: Test results. Cranfield: Aslib, 1966.

csak mondjuk az íróknál – a nyelvi kompetenciával való jobb bánni tudás játszik szerepet, mely kiegyensúlyozza az osztályozó szakemberek konkrét ismereteinek hiányát.
A második vizsgálat célkitűzése az volt, hogy megállapítsák, hogyan befolyásolják a keresést az osztályozási rendszer/dokumentációs nyelv szerkezeti elemei, illetve ezek szerkezete az egész információkereső rendszer működését. A legfontosabb eredmények:
– a legjobb eredményt a teljesség vonatkozásában az egyszerű szavakból álló dokumentációs nyelvekkel érhetők el;
– a fazettás osztályozáson alapuló alapfogalmak eredményezik a legkisebb pontosságot, összehasonlítva az egyszerű szavas, illetve deszkriptoros nyelvekkel;
– a keresési szabályok befolyásolják a teljesség és pontosság viszonyát;
– a pontosság fokozására használt segédeszközök (szintaktikai eszközök, szerep- és kapcsolatjelölők) hatástalanok az eredményre.

Általánosítva a legfőbb tapasztalat, hogy az osztályozást nem szabad túlszabályozni. Optimálisnak tekinthető, ha a szabályozás túlnyomórészt a szinonimákra és homonimákra terjed ki, azaz főleg morfológiai.

Horváth Tibor: Az aberyswyth-i jelentés

In: Könyvtári Figyelő, 1975, 21. évf., 3. sz., p. 564–569.
Eredeti: Keen, E. M., Digger, J. A.: Report of an information science index language test. Part 1.: Text. Part 2.: Tables. – Aberystwyth: Department of Information Retrieval Studies. College of Librarianship Wales, 1972.

„…az eredmények nem annyira az általánosítható tapasztalatokban, mint inkább a részletekben vannak… A vizsgált nyelvek nem mutattak szignifikáns különbségeket a keresés hatékonyságát illetően.” Általános tapasztalatok:
– A szabályozatlan nyelvek mindazt nyújtották, amit a szabályozott nyelvek nyújtottak.
– A szabályozatlan nyelv sohasem volt olyan rossz, mint a legroszszabb szabályozott nyelv, sem pedig olyan jó, mint a legjobb szabályozott nyelv.
– A nagy specifikusság nem eredményez rosszabb teljességet, mint a kis specifikusság.

„A jelentés ugyanakkor egy sor negatív megállapítást tartalmaz. Kb. oda jutottak el, hogy mi nem befolyásolja a keresés alapvető mutatóit, ti. a teljességet és a pontosságot."

Michael Keen: A rangsorolásos információkeresési eljárások hatékonysága különböző súlyozási eljárások esetén

In: Tudományos és Műszaki Tájékoztatás, 1995, 40. évf., 4–5. sz., p. 199–203.
Eredeti: Keen, E. M.: Query term weightig schemes for effective ranked output retrieval. In: Proceedings (of) 15th International Online Information Meeting, 10–12. December 1991, London. – Oxford; Nootering: Learned Information, 1991. p. 135–142.

Noha közel 30 éve folynak kísérletek olyan információkeresési eljárások kidolgozására, melyek használata esetén nincs szükség arra, hogy a kereső Boole-algebrai kifejezéseket állítson össze, kevés ilyen fajta keresőrendszer működik üzemszerűen, és egyikük sem terjedt el szélesebb körben. *Michael Keen,* a cranfieldi vizsgálatok egyik résztvevője több rangsorolásos eljárás összehasonlító elemzését végezte el. Az egyes eljárások az összesített gyakoriság, a rekordonkénti gyakoriság, a páronkénti gyakoriság, illetve a kereső által szubjektíven adott súlyok módszere szerint működtek. A leghatékonyabbnak a páronkénti távolságok módszere bizonyult, és az összesített gyakoriság és a rekordonkénti gyakoriság módszerével kombinálva a leginkább látszik alkalmasnak a rangsorolásos keresésre.

Jack Mills: Lépcsőzetes mutatózás és a szakkatalógus

In: Könyvtári Figyelő, 1956, 2. évf., 10. sz., p. 33–43.
Eredeti: Mills, J.: Chain indexing and the classified catalogue. In: Librarian Association Rec, 1955, Vol. 57, p. 141–148.

A lépcsőzetes mutatót (láncindexet, hierarchikus permutált mutatót) általában osztályozási rendszerekhez használják. A lehető legtöbb információt adja a szakkatalógushoz használt rendszer indexelhető kifejezéseiről.

Az ETO alábbi részletéből például a következő lépcsőzetes, hierarchikus mutató képezhető:

64 Háztartás
641 Élelmiszer. Feldolgozás
641.7 Feldolgozási eljárások
641.74 Párolás

Párolás: Feldolgozási eljárások: Élelmiszer: Háztartás	641.74
Feldolgozási eljárások: Élelmiszer: Háztartás	641.7
Élelmiszer: Háztartás	641
Háztartás	64

Az ötvenes és hatvanas években sok angol könyvtáros az egyik legjobb indexelési eljárásnak tartotta. Noha első nyomai már *Dewey-nál* (relatíve index) és *Cutternál* is megfigyelhetők, szellemi atyja *Ranganathan* volt, az ő terminológiája alapján született a neve is (*Ranganathan* az egymás fölé rendelt fogalmak sorát nevezte láncnak). 1950 és 1970 között, a PRECIS bevezetéséig a Brit Nemzeti Bibliográfia mutatóját készítették ezzel a módszerrel, ennek következtében rendkívül gazdag tapasztalatokat szereztek a használatáról. Mivel mereven a mindenkori osztályozási rendszerhez tapad, annak összefüggéseit és terminológiáját követi, elkerülhetetlenül hiányzik belőle minden terminológiai szabványosítás. Alkalmatlan továbbá arra, hogy az adatcsere-formátumokba mint a rekordhoz tartozó mutatóneveket (indexkifejezéseket) fölvegyék.

(A lépcsőzetes mutatót a „Tárgyszó és a tárgyszókatalógus" című kötetben *Eric J. Coates* szemelvénye részletesen tárgyalja.)

AZ INFORMÁCIÓK CSERESZABATOSSÁGA

Ugyanazt a dokumentumot nagyon sok könyvtárban, illetve dokumentációs intézményben is tárolhatják, ezért ugyanennyi helyen kell ugyanazt a dokumentumot feldolgozni. A feldolgozás eredménye a dokumentum leírása, amely (i) meglehetősen költséges munka, és (ii) országonként csak kevés olyan könyvtár vagy dokumentációs intézmény, ügynökség létezik, mely feladatának tekinti, hogy a szabványoknak tökéletesen megfelelő leírásokat készítsen. Ezért a könyvtáraknak komoly érdeke fűződik ahhoz, hogy elkerüljék a párhuzamos feldolgozást, és a leírásokat ettől a kevés „profi" könyvtártól vagy ügynökségtől szerezzék be.

A hatvanas években a nagyobb könyvtárak és dokumentációs intézmények és ügynökségek elkezdték a bibliográfiai leírások számítógépes feldolgozását és tárolását; már az évtized közepén fölmerült az igény, hogy az így elkészült bibliográfiai rekordokat más könyvtárak is átvegyék. Ahhoz azonban, hogy ez lehetséges legyen, olyan logikai rekordformátumot kellett kialakítani, melyet az adásvételben vagy a cserében résztvevő intézmények egyaránt használhattak, mert csak így lehet az adatok cseréjét a leggazdaságosabban megszervezni (hiszen ebben az esetben minden, a cserében résztvevő intézmény ugyanahhoz a rekordformátumhoz készítheti el a saját konvertáló programját). Ezt, a géppel olvasható adatcsere céljára kialakított logikai formátumot nevezték el az angol rövidítés alapján MARC (Machine Readable Cataloguing) formátumnak. Éllovasa az első ilyen formátum elkészítésének a Kongresszusi Könyvtár volt, és a fejlődés első szakaszának kulcsembere a könyvtár információs rendszereket koordináló osztályának vezetője, *Henriette D. Awram*.

Awram Frederick Kilgourral és *Richard de Gennaroval* közösen az 1966-ban a londoni Brasenore-Intézetben rendezett konferencián számolt be először erről a formátumról. A továbbfejlesztett, 1968-tól használt változatot nevezték a Kongresszusi Könyvtár első kísérletétől (MARC I) való megkülönböztetés érdekében MARC II formátumnak.

A MARC formátum létrehozói nemcsak a Kongresszusi Könyvtár bibliográfiai leírási gyakorlatához kötődtek rendkívül erősen, hanem a korabeli logikai adatbázis-kezelés formájához is. Ennek máig ható következményei lettek:

– A MARC formátumok alapvetően a bibliográfiai leírás és besorolás adatelemeire, tehát a formai feltárásra épülnek. Ez a magyarázata annak, hogy a jellegzetesen dokumentációs–információs adatelemek, melyek előírt tárgyi (tartalmi) vagy dokumentumtipológiai értékkészlettel rendelkeznek, osztályozáselméleti, tipológiai szempontból nemcsak rendkívül kezdetlegesek, hanem ellentmondásosak is.
– Ennek tulajdonítható, hogy a formátumban a tárgyi–tartalmi feltárás adatelemei (osztályozási jelzet, tárgyszó stb.) rendkívül szegényesen, szinte strukturálatlanul jelentek meg, összehasonlítva a formai feltárás egységesített besorolási adatelemeivel. (Az első kísérletekre, hogy a tárgyi–tartalmi feltárás adataihoz is megfelelően részletezett formátum készüljön, a nyolcvanas évek végén került sor.)
– A formátum alapvetően a könyvekre vonatkozott; ami még jobban beszűkítette az alkalmazhatóságát. Szinte „programozva" volt, hogy az analitikus feltárás igényeihez igazodó dokumentációs célú leírások adatcseréjéhez eleve nem lehet majd alkalmassá tenni, legyen szó akár folyóiratcikkekről, akár dokumentumok részletéről, de akár arról, hogy az adatcserébe a referátumot vagy annotációt is bevonják.
– A korai időszak adatbázis-kezelési gyakorlatához való kötődés további következménye, hogy az adatcsere-formátum hierarchikus (mező–almező) felépítésű. A rekordon belüli nem hierarchikus összefüggések csak körülményesen adhatók meg.

Ezek az érthető, természetes és ugyanakkor sajnálatos kötődések máig éreztetik a hatásukat; az egyes országokban és nemzetközi szinten is hosszú ideig minden későbbi fejlődés ebből a MARC II formátumból indult ki, ezt tekintette szerkesztésfilozófiai mintának. Ezen csak az 1978–84 között kidolgozott „közös" adatcsere-formátum (a Common Communication Format; CCF) próbált változtatni. A fejleményekről *Mirna Willer* számol be, kiegészítve *Vajda Erik* kommentárjával.

Mára, az on-line környezetben az adatcsere-formátumok korábbi jelentősége kissé ugyan csökkent, de fokozatosan olyan szerepe értékelődött föl, melyre eredeti megalkotói még nem gondoltak. A formátumban a dokumentumleírás logikai szerkezetét a legkisebb részletekig pontosan és egyértelműen rögzítik, ezért valóságos kincsestára a leírásra vonatkozó ismereteknek. A formátumokat a dokumentumleírásra és mindenfajta reprezentációjára vonatkozó információgazdagsága miatt az adatbázis-tervezők rájöttek arra, hogy nagyon jól használhatók a dokumentációs és bibliográfiai

adatbázis-tervezéshez. Éppen azt a logikai rekordstruktúrát képviselik ugyanis, amelynek teljes körű kezelésére kell kialakítani a mindenkori katalogizálási és információkereső rendszert.

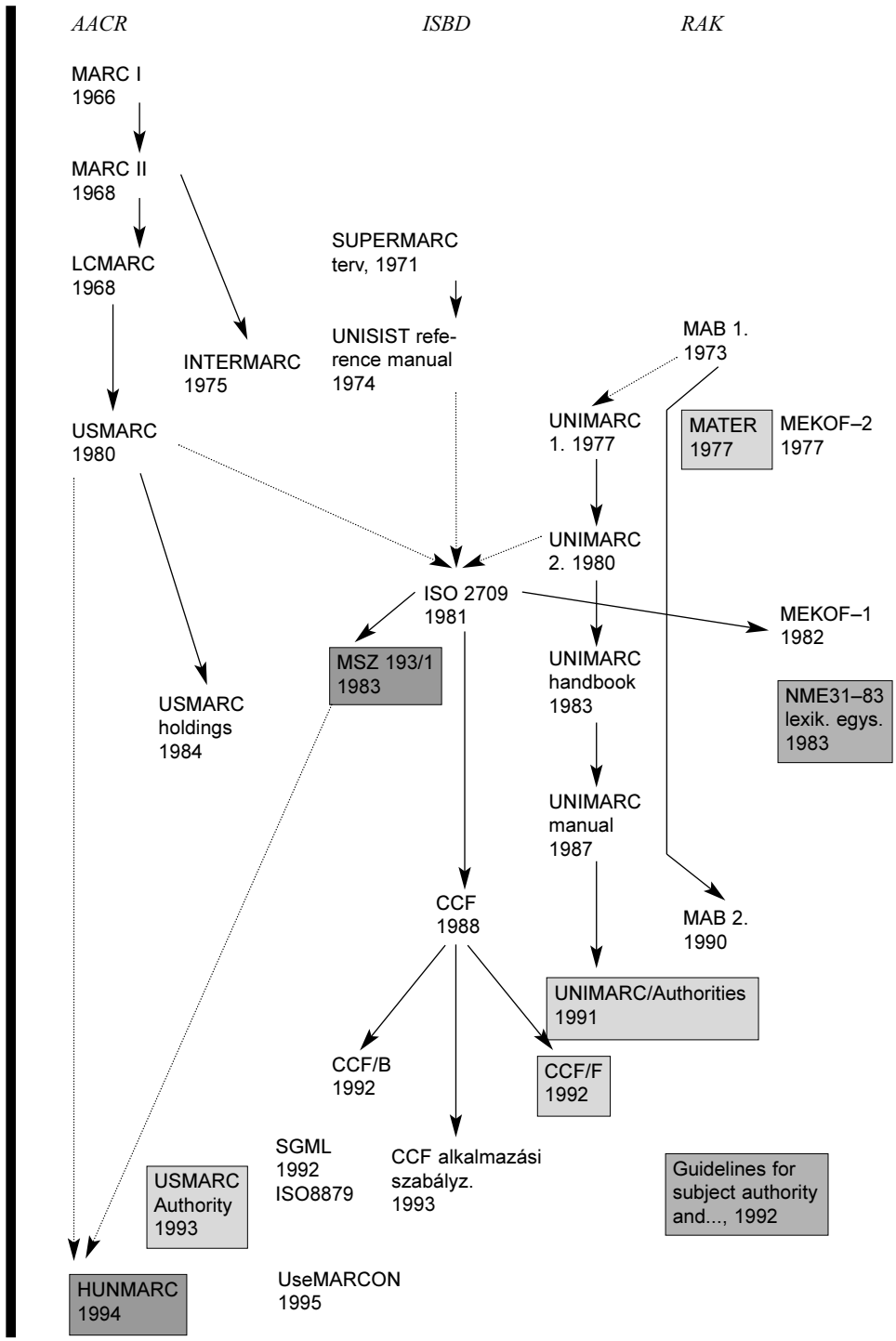
Meg kell még említeni egy másik fejleményt, mely nem közvetlenül az adatcsere-formátumokhoz, de mégis csak a dokumentumleírások egységes – csereszabatos – felhasználásához kapcsolódik. Idővel számos országban megszerveződött a központi katalóguscédula ellátás is. Leghíresebb intézménye az ohioi főiskolán 1951-ben alapított non-profit könyvtári központ (Ohio College Library Center; OCLC), melyet a Kongresszusi Könyvtár bábáskodásával 1961-ben újjászerveztek, 1967-ben pedig *Frederick Kilgour* irányításával indultak el azok a munkák, melyek eredményeként 1970-től kötegelt üzemmódban gyártották és szolgáltatták a katalóguscédulákat. 1971-től a OCLC katalógusa on-line hozzáférésű lett, s immár a bibliográfiai rekordokat szolgáltatják, 1981-ben pedig az On-line Computer Library Center (OCLC) Inc. néven megújult társaság megnyitotta első irodáját Angliában, s ezzel elkezdődött az OCLC-szolgáltatások igénybevétele Európában is.[1]

A szabványos bibliográfiai leírások egységes hasznosításával függ össze az a fejlemény is, hogy idővel több országban, például az Egyesült Államokban, Németországban a könyvkiadók és terjesztők a kiadványuk impresszumában feltüntetik a kiadvány szabványos bibliográfiai leírását. Ez nevezik kiadványba nyomtatott katalogizálásnak (Cataloguing in Publication; CIP). Elterjedése azzal függött össze, hogy a kiadók és terjesztők felismerték, az egységes, mindenki számára hozzáférhető leírás feltüntetése megkönnyíti a kiadvány megrendelését, tehát forgalomnövelő hatású.

Az alábbi ábrán a közölt szemelvények alapján összefoglaltuk az adatcsere-formátumok időbeli fejlődését. A folytonos vonalú nyilak az adott formátum fejlődését, a pontozott vonalú nyilak az egyes formátumok, illetve szabványok legfontosabb hatásait fejezik ki. Az ábrán feltüntetett fejlődésvonalak és hatások természetesen csak közelítő érvényűek, és az ábra arra való, hogy durva időbeli áttekintés formájában segítséget nyújtson az adatcsere-formátumok történeti fejlődéséhez.

1 Mitchell, J.: Az OCLC – múlt, jelen, jövő. In: Tudományos és Műszaki Tájékoztatás, 1993, 40. évf., 9–10. sz., p. 387–390.

Mitchell, J.: OCLC – nemzetközi bibliográfiai forrásmegosztó hálózat könyvtáraknak. In: Tudományos és Műszaki Tájékoztatás, 1994, 40. évf., 4. sz., p. 567–573.


AACR
ISBD
RAK
MARC I
1966
MARC II
1968
LCMARC
1968
INTERMARC
1975
SUPERMARC
terv, 1971
UNISIST refe-
rence manual
1974
MAB 1.
1973
USMARC
1980
UNIMARC
1. 1977
MATER
1977
MEKOF–2
1977
UNIMARC
2. 1980
ISO 2709
1981
MEKOF–1
1982
MSZ 193/1
1983
UNIMARC
handbook
1983
NME31–83
lexik. egys.
1983
USMARC
holdings
1984
UNIMARC
manual
1987
CCF
1988
MAB 2.
1990
UNIMARC/Authorities
1991
CCF/B
1992
CCF/F
1992
SGML
1992
ISO8879
CCF alkalmazási
szabályz.
1993
Guidelines for
subject authority
and..., 1992
USMARC
Authority
1993
HUNMARC
1994
UseMARCON
1995

Az első sorban dőlt betűkkel feltüntettük a három legfontosabb katalogizálási, illetve bibliográfiai leírási szabványt, melyek ennek a fejlődésnek a hátterében álltak (és amelyeket akár hagyományos, akár számítógépes feldolgozás esetén alkalmazni kell. Az angol-amerikai katalogizálási szabvány (Anglo-American Cataloguing Rules; AACR), a bibliográfiai leírás nemzetközi szabványa (International Standard Bibliographic Description; ISBD), és a német katalogizálási irányelvek (Richtlinien für den Allgemeinen Katalog; RAK) természetesen erősen hatottak az adott nyelvterületen, illetve környezetben kialakított MARC formátumokra, és persze a nemzetközi (ISO) szabványosításra is. Mindez csak nagyon durva megközelítés, a valóságos összefüggések és összefonódások természetesen ennél bonyolultabbak, hiszen szinte minden hatott mindenre.
Említést érdemel, hogy már megjelentek az egységesített besorolási adatokra, lexikai egységekre vonatkozó adatcsere-formátumok. Az elsőt Németországban készítették[2], ezt követte a KGST-államok nemzetközi információs rendszerében (NTMIR) a lexikai egységekre vonatkozó normatív–műszaki irányelv[3]. Napjainkban pedig az figyelhető meg, hogy az egységesített besorolási adatokra vonatkozóan formátumok leválnak a bibliográfiai, illetve dokumentációs adatelemeket tartalmazó formátumokról (UNIMARC/Authorities, CCF/F, USMARC Authority) (halványan árnyékolt keretben).

A nyolcvanas évek végétől az IFLA egyik munkabizottságában kezdtek dolgozni a tartalmi feltárás lexikai egységeire vonatkozó szabványtervezen is, melynek „végleges" tervezete (final draft) 1992-ben készült el[4]. Ebben részletesen szabályozzák, hogy milyen szerkezetű tételt alkossanak az állományba szervezett tárgyszavak, illetve a deszkriptorok (pl. milyen mezőkben szerepeljenek a lexikai egység forrására, kezelésére stb. vonatkozó adatok). Az MSZ 193/1 magyar szabvány és a HUNMARC mellett[5]

2 MATER. Magnetic Tape Exchange Format for Terminological/lexicographical Records. Magnetband-Austauschformat für terminologische/lexikographische Daten. ISO/DP 6156. 1977.

3 NTMIK NME 31–83 Adatok tartalma és közlésmódja a mágnesszalagos információcserére szánt rekordokban. Az információkereső nyelvek lexikai egységei és terminológiai adatok. 1983. december. In: Az NTMIR normatív–műszaki dokumentumai. X. rész. 28. köt. [közr. az] OMFB NTMIR Magyar Tanács Titkárság. – Budapest: OMIKK, 1988. p. 57–70.

4 Guidelines for subject authority and reference entries. [ed by] Working Group of the Section Classification and Indexing of the IFLA Divisiin of Bibliographic Control. – [Frankfurt]: [IFLA], 1992 [Final draft].

5 MSZ 193/1. Mágnesszalagos bibliográfiai adatcsere formátuma. A rekordok szerkezete. Budapest: Magyar Szabványügyi Hivatal, 198. 7 p. [Tartalma megegyezik az ISO 2709 nemzetközi szabvánnyal.]

HUNMARC: a bibliográfiai rekordok adatcsere-formátuma. [Összeáll. Sípos Márta; kész. az Országos Széchényi Könyvtár Fejlesztési Osztályán] – Budapest: OSZK, 1994. 129 p.

(mindkettő sötéten árnyékolt keretben) az időközben megszűnt NTMIR normatív–műszaki dokumentumai is megjelentek magyarul.[6]
Az Országos Széchényi Könyvtár hosszas, országos egyeztetés után 1994-ben megjelentette a magyar HUNMARC adatcsere-formátum első kiadását.

A HUNMARC-ról az alábbi közlemények jelentek meg:

OSZK Fejlesztési Csoport: A nemzeti adatcsere formátuma és az összevont adatelemek

In: Könyvtári Figyelő, 1995. 5. évf., 2. sz., p. 257–263.

Részletes példákkal mutatják be az adatelemek konverziójának problémáit, különös tekintettel az adatelemek összevonásából keletkező nehézségekre. A nemzeti adatcsere-formátumnak az adatelemek maximális szétválasztását és azonosítását kell biztosítania. Az ilyen formátumból egyszerűbb célokra bármikor könnyen elvégezhető a konverzió. Fordítva, összevont, leegyszerűsített adatelemeket tartalmazó formátumból már nem lehet differenciáltabb adatszerkezetre konvertálni.

Sipos Márta: USMARC–UseMARCON–HUNMARC. A bibliográfiai rekordok adatcsere-formátuma és a konverzió

In: Könyvtári Figyelő, 1997. 7. évf., 1. sz., p.

Az adatcsere-formátumoknak történetileg három, az adott nemzeti formátum szellemét meghatározó típusa alakult ki: az USMARC, az UNIMARC és a CCF típusú formátumok. Új fejlemény az UseMARCON általános konvertáló program megjelenése, melyet a felhasználó para-

6 NTP 2–74 Mágnesszalagon rögzített adatrekordok .... 1974 május. In: Az NTMIR normatív–műszaki dokumentumai. I. rész. 2. köt. [közr. az] OMFB Szakmai Információs Tárcaközi Bizottság. – Budapest: OMIKK, 1976. p. 34–40.

NTP 1–74 Bibliográfiai információcsere mágnesszalagos formátuma. 1974 május. In: Az NTMIR normatív–műszaki dokumentumai. I. rész. 2. köt. [Közr. az] OMFB Szakmai Információs Tárcaközi Bizottság. – Budapest: OMIKK, 1976. p. 27–40.

NTP MCNTI 19–77 Adatok tartalma és közlésmódja a mágnesszalagos információcserére szánt rekordokban. 1977 december. In: Az NTMIR normatív–műszaki dokumentumai. VI. rész. 14. köt. [közr. az] OMFB Szakmai Információs Tárcaközi Bizottság. – Budapest: OMIKK, 1979. p. 10–279.

NTP MCNTI 30–82 Bibliográfiai információk tartalma és közlésmódja a mágnesszalagos információcserére szánt rekordokban (MEKOF–1 nemzetközi csereformátum). 1982 május. In: Az NTMIR normatív–műszaki dokumentumai. VII. rész. 23. köt. [közr. az] OMFB NTMIR Magyar Tanács Titkárság. – Budapest: OMIKK, 1983. p. 45–128.

méterezhet, és bármely adatcsere-formátumok közötti konverzióra használhat. A tanulmány második felében az adatcsere-formátum speciális magyar jellemvonásainak kérdéseiről (pl. a párhuzamos címek, a magyar személynevek kódolásáról) van szó.

Az adatcsere-formátumokkal kapcsolatban magyarul az alábbi fordítások, illetve ismertetések jelentek meg:

Harold Dierickx: Egységes nemzetközi bibliográfiai adatcsere-formátum felé. A bibliográfiai leírás nemzetközi UNISIST programja

In: Könyvtári Figyelő, 1980, 26. évf., 1. sz., p. 88–90.
Eredeti: Toward a common international bibliographic exchange format? (In: International cataloguing, 1978, Vol. 7, No. 2, p. 19–24.) és International Centre for Bibliographic description (UNIBID): A review of objectives, activites for date an future development (In: Unesco Bulletin Libr., 1978, Vol. 32, No. 3, p. 157–185.) E cikkek alapján összeállította és tömörítette Bobokné Belányi Beáta.

A Bibliográfiai Leírás Nemzetközi Központját (UNISIST International Centre for Bibliographic Description; UNIBID) az angol nemzeti könyvtárban hozták létre. Egyik feladatuk az UNISIST Refrerence Manual gondozása. 1978-ban konferenciát rendeztek azzal a céllal, hogy a könyvtári/bibliográfiai és a tájékoztatási/dokumentációs adatcseréhez mindkét területen egyformán alkalmazható formátumot dolgozzanak ki. A nemzetközi adatcsere-formátumnak teljesen alkalmazásfüggetlennek kell lennie. Ilyen formátum 1978-ig nem készült.
Az ISBD-k és a meglévő nemzetközi adatcsere-formátumok a könyvtári és tájékoztatási szolgáltatások legfontosabb követelményeit testesítik meg; fontos hivatkozási pontok a jövő, valóban egyetemes formátumai számára.

Fredrick G. Kilgour: Az „on-line" katalógus forradalma

In: Könyvtári Figyelő, 1980, 26. évf.
The on-line catalogue revolution és Simonds, M. J.: Database limiation and on-line catalogs (In: Library Journal, 1984, Vol. 109, No. 3, p. 319–330) c. tanulmányok alapján Varga Ildikó szemléje.

A kritikai észrevételekben gazdag beszámoló legfontosabb megállapítása, hogy a MARC formátumok létrehozói túlságosan a Kongresszusi Könyvtár céduláihoz kötődtek. Mindennek a középpontjában a bibliográfiai adatelemek álltak. Az információs/dokumentációs célú adatelemek értékeinek rendszerei osztályozási szempontból átgondolatlanok.

Gerő Péter–Vladimir A. Skripkin: Az NTMIR mágnesszalagos bibliográfiai adatcsere-formátumáról

In: Tudományos és Műszaki Tájékoztatás, 1984, 31. évf., 5. sz., p. 176–187.

Az NTMIR-ben két rekordszerkezetet fogadtak el. Az NTP MCNTI 1–82 teljesen megfelel az ISO által elfogadott szerkezetnek, az NTP MCNTI 2–74 a Csehszlovák Tudományos és Műszaki Gazdasági Információs Központban készült el a Chemical Abstracts Service által alkalmazott szerkezet alapján. Az elsőnek a MEKOF–1 (= NTMIK NME 30–82) adatcsere-formátum felel meg, a másodiknak a MEKOF–2 (= NTMIK NME 19–82). Táblázatban:

Rekordszerkezet szabványa	Adattartalom normatív–műszaki előírása	Rendszer neve	KGST szabványa
NTP MCNTI 2–74	NTMIK NME 19–82	MEKOF–2	SZT SEV 4269–83
NTP MCNTI 1–82	NTMIK NME 30–82	MEKOF–1	SZT SEV .....–84

Alan Hopkinson: A bibliográfiai adatok nemzetközi használhatósága: a MARC és a MARC-okat összefogó munkák

In: Tudományos és Műszaki Tájékoztatás, 1985, 32. évf., 8–9. sz., p. 383–391.

A bibliográfiai adatcsere-formátumok az ISO 2709 szerinti, a Library of Congress MARC formátumának szerkezetéből létrejött rekordszerkezetet alkalmazzák. AZ UNISIST Reference Manualt az ICSU–AB és az Unesco hozta létre. E korai nemzetközi adatcsere-formátum a referáló és indexelő szolgáltatások céljainak megfelelő

katalogizálási adatokat tartalmaz. A nemzeti könyvtárak az IFLA égisze alatt az UNIMARC-ot dolgozták ki adatcsere-formátumként, mivel nemzeti formátumaik nem voltak teljesen kompatibilisek. Az UNESCO a nagy nemzetközi adatcsere-formátumok közötti kompatibilitás hiányából adódó problémák megoldására szimpóziumot szervezett, melynek ajánlásaiból kiindulva kidolgozták a Közös Adatcsere-formátumot (Common Communication Format; CCF), az ISO-ban pedig megkezdték a bibliográfiai adatelemek szabványos gyűjteményének az összeállítását. Nemzetközi formátumok léteznek még a nukleáris és a mezőgazdasági, valamint az időszaki kiadványok adatainak cseréjére. Más – egymástól és az említettektől eltérő szerkezetű, nemzetközileg hasznos – adatcsere-formátumok is léteznek. A nemzetközi MARC-hálózat lehetőségét vizsgáló bizottság a bibliográfiai adatok cseréje érdekében megszervezte az UNIMARC tesztelését és ellenőrizte a formátum egyértelmű gyakorlati használatát célzó UNIMARC kézikönyv tartalmát. A nemzetközi adatcsere-formátumok fejlődése arra utal, hogy növekvő mértékben látják el őket katalogizálási szabályokkal. Akkor lesznek ezek a formátumok igazán hasznosak, ha léteznek majd nemzetközileg elfogadott katalogizálási szabályok és szabványos besorolási adatgyűjtemények.

Alan Hopkinson: Információátvitel és adatcsere-formátumok

In: Tudományos és Műszaki Tájékoztatás, 1992, 39. évf., 6. sz., p. 269–274.

Az elsősorban könyvtári célokra megfelelő UNIMARC, a dokumentációs szolgáltatásokban alkalmazott és az UNISIST Reference Manualban leírt formátum, valamint az UNESCO mindkét területen használható Közös Kommunikációs Formátumának (Common Communication Format; CCF) közös és eltérő vonásainak kritikai elemzése olvasható a tanulmányban.

Az alábbiakban Vajda Erik bevezetőjével olvasható Mirna Willernek az adatcsere-formátumokról írt tanulmánya.
Vajda Erik bevezetőjében az alapfogalmakat tisztázza és magyarázza meg, és néhány olyan kérdést érint, amit Willer tanulmánya – amely elsősorban a MARC formátumokról szól – szükségképpen mellőzött.

Vajda Erik: Az adatcsere-formátumokról[7]

(Alapozás és kiegészítések Mirna Willmer: A szabványosítás szükségessége a géppel olvasható katalogizálásban c. cikkéhez)

1. Mi az, hogy formátum?

A formátum – esetünkben, vagyis bibliográfiai adatrekordok esetében, a bibliográfiai adatközlés formátuma – megnevezéseket, definíciókat, paramétereket (megengedett, illetve előírt értékeket) tartalmazó *szabályrendszer,* amely lehetővé teszi, hogy a számítógép szoftverje felismerhesse a bevitt, feldolgozandó, tárolandó és megjelenítendő adatokat. A formátum lehet adatbeviteli, feldolgozási, megjelenítési és adatcsere-formátum.

Az *adatbeviteli* (input) formátumot az teszi szükségessé, hogy az adatokat már bevitelükkor is a szoftver által felismerhető formában kell rögzíteni, bár – és ezt itt jegyezzük meg, de minden formátumra vonatkozik – a formátum olyan szabályokat is tartalmaz, amelyek valójában a közlendő adatok körére és közlésük módjára vonatkoznak, tehát – bibliográfiai rekordok esetében – voltaképpen katalogizálási előírások. Az adatbeviteli formátum lényegében egy űrlap, vagy munkalap, a hozzá tartozó magyarázatokkal együtt, jóllehet az űrlap szót nem kell fizikailag konkrétan felfogni, mert ezt a bevivendő adatok listája és beviteli előírásai is helyettesíthetik. Az adatbeviteli formátum tartalmában a rendszer céljához – tehát ha úgy tetszik a feldolgozási formátumhoz – igazodik, formájában (legalábbis bevallott elvek szerint) igyekszik az adatokat bevivő könyvtáros, vagy egyéb „rabszolga" munkáját könnyíteni és pontosabbá tenni. Az alkalmazói szoftver többféle adatbeviteli formátum alkalmazását is lehetővé teheti.

A *feldolgozási* (vagy belső) formátum a gépen belül tárolt és kezelt adatok körének meghatározására, egyértelmű megjelölésére, kezelésük módjának rögzítésére szolgál. Az alkalmazói szoftvertől függ, hogy a használónak milyen „beleszólása" van a feldolgozási formátumba, jóllehet ma már ritkák a kulcsrakész, de merev, minden paramétert előre meghatározó, zárt jellegű formátumok. A felhasználónak tehát különböző mértékű lehetőségei, és egyben kötelezettségei vannak a feldolgozási formátum meghatározásában, aminek gyakorlati minimuma az adatelemek körének meghatározása. Aki azonban már hozott létre adatbázist, az tudja, hogy – pl. a CDS/ISIS mezőmeghatározó táblájának (field definition table) elkészítésekor – ennél lényegesen többet határoz meg.

A *megjelenítési formátum* (a szoftverek készítői gyakran csak ezt nevezik formátumnak, pl. CDS/ISIS, ALEPH) a tárolt adatok közül megjeleníten-

7 In: Könyvtári Figyelő, 1994, 4. évf., 1. sz. , p. 35–41.

dő adatok körét, sorrendjét, elhelyezését, központozását, esetleges állandó, vagy feltételes kísérő szövegeit, jeleit, vagyis a képernyőn láthatóvá tett, illetve kinyomtatható rekordok (tételek) képét határozza meg. Egy-egy adatbázishoz többféle megjelenítési formátum tartozhat (és többnyire tartozik is) a különféle használók különféle igényeinek kielégítése végett.

Végül az *adatcsere-formátumokról* csak annyit (hiszen erről szól e cikk), hogy – a megjelenítési formátumokhoz hasonlóan – ezek is kommunikációs (output) formátumok, de egyfelől esetükben nem képernyő, vagy nyomtatás útján, hanem géppel olvasható formában történik a kommunikáció, másfelől e kommunikáció címzettei más, számítógépes könyvtári és rokonterületi számítógépes rendszerek.

2. A formátumok elemei

Minden formátumnak – kérem fogadják el a felosztás önkényét, hiszen ezek az „elemek" is alárendelhető elemekből állnak – két, eltérő rendeltetésű és jellegű eleme (és további egy kiegészítő eleme) van.

Mielőtt ezek ismertetéséhez fognánk, a rend kedvéért tisztázzunk egy általánosan ismert, de gyakran pontatlanul értelmezett fogalmat, az *adatelem* fogalmát. E témakör szövegkörnyezetében az adatelem a rendszer (a szoftver) által egységként kezelt legkisebb karaktersor. (A szoftver az ezen az egységen belüli karaktereket már csak speciális célokra „bontja" tovább, és kezeli elkülönítetten, mint pl. akkor, amikor szavakból álló keresőfájlokhoz készít ott adatelemként viselkedő kisebb elemeket.) Az adatelemeket természetesen a formátum is elkülöníti, méghozzá hívójelekkel megjelölt *mezőkben* és – esetleg – almező azonosítókkal megjelölt *almezőkben*, amelyek egymástól még a sajátos alkalmazási célt jelző *indikátorokkal* megjelölve is eltérhetnek. Más szóval, egy adott hívójellel megjelölt mező és – esetleg – almező azonosítóval megjelölt almező, adott indikátorral megjelölt változatának tartalma egy adatelem. Ezért nevezik a hívójel, az esetleges almező azonosító és az esetleges indikátor együttesét *tartalmi azonosítónak* (content designator). A mezők – mint ismeretes – rekordokat alkotnak, egy-egy rekord pedig – a leggyakrabban használt bibliográfiai adatcsere-formátumok esetében – a rekordfejből, a mutatóból és az adatmezőkből, valamint a rekordhatároló jelből áll.

A formátumnak felismerhetővé kell tennie a szoftver számára, hogy az egyes rekordok, ezen belül pedig az egyes adatelemek hogyan különböztethetők meg és hogy találhatók meg. Erre különféle eszközök szolgálhatnak. A legkezdetlegesebb megoldás szerint a formátum egységesen rögzíti a rekordok és azon belül a mezők hosszát. Ezáltal a mezők elkülöníthetővé válnak, tartalmukat pedig sorrendjükkel lehet meghatározni. A korszerű formátumok ettől eltérnek, és *változó (előre meg nem határozott) hosszúságú rekordokat,*

illetve mezőket tartalmaznak. Ennek megfelelően bonyolultabb „megjelölő eszközöket" (egyebek között a fent említett tartalmi azonosítókat) kell alkalmazniuk.

Azok az információk, amelyek lehetővé teszik a rekord és a mezők, illetve almezők és indikátorok felismerését a *rekordfejben* (más néven rekordcímkében) helyezkednek el. Ezek az információk – amelyek a rekordfej fix hosszúságú részeinek pozíciójához kötöttek – pl. megmutatják, hogy a rekordban milyen határolójeleket használnak, milyen hosszúságúak az almező-azonosítók és indikátorok, milyen a mutató elemeinek szerkezete. A rekordfej tartalmaz emellett a rekord jellegére, tartalmára és konkrét szerkezetére vonatkozó információkat is. A rekord *mutatója* már az adott rekord adatmezőinek azonosítását, keresését segíti, az *adatmezők* – határoló elemeik, almező azonosítóik és indikátoruk mellett – a bibliográfiai információkat tartalmazzák.

A formátumnak az az eleme, amely a rekordfej, a mutató és az adatmezők szerkezetét, részeit és ezek rendeltetését határozza meg, a *formátum szintaxisa* (struktúrája). Ez a szintaxis – mint említettük – független a rekord tartalmától, így természetesen azoktól a szabályoktól is, amelyek szerint a rekord tartalmát meg kell választani és amilyen formában e tartalmat le kell írni, vagyis a struktúra független – esetünkben – a katalogizálási és kapcsolódó bibliográfiai adatközlési szabályoktól.

Az *ISO 2709–1981 sz. nemzetközi szabvány*[8] és *magyar megfelelője, az MSZ 193/1-1983*[9] a bibliográfiai információcsere formátumaként a formátum szintaxisát határozza meg. A szintaxisnak a tartalomtól független használhatósága magyarázza az ISO 2709 sikerét: az adatcsere-formátumok többsége erre épül (mégpedig nemcsak azért, mert – lásd M. W. cikkét – a formátumok jórészt MARC-alapúak, az ISO 2709 „nagyanyja" pedig az USMARC szintaxisa volt). E siker következménye, hogy a legtöbb könyvtári alkalmazói szoftver képes ISO 2709 szintaxisú rekordokat importálni.

A formátumok másik eleme – amit a *formátum szemantikájának* nevezhetünk és neveztünk is a néhai szocialista országok néhai szabványosítási együttműködése keretében – a *formátum tartalmi kereteit* (nem tartalmát, hiszen ez a mindenkori leírt dokumentumtól függ) határozza meg, azáltal, hogy felsorolja a kötelezően közlendő, vagy a közölhető adatelemeket, megállapítva természetes nyelvű nevüket, „gépi nevüket", vagyis tartalmi azonosítójukat, valamint fogalmukat, kötelezőségi státusukat, ismételhető, vagy nem ismételhető voltukat, továbbá – közvetlenül, vagy katalogizálási szabályokra hivatkozva – közlésük módját (ide értve azon adatelemek esetében, amelyeket kódolva kell közölni, a kódkészletet és feloldását is). E „szemantikát" nevezhetjük az adatelemek adat-

8 ISO 2709–1981. Format for bibliographic information inter-change on magnetic tape.

9 MSZ 193/1–1983. A mágnesszalagos bibliográfiai adatcsere formátuma. A rekordok szerkezete.

tárának is. Aligha nehéz belátni, hogy itt már nehezebb a szabványosítás, mert – nemzetek, kultúrák és iskolák szerint, de a könyvtárak és tájékoztatási intézmények funkció-adta igényei miatt is – majdhogynem ahány rendszer, annyi formátum. Ezért aztán a több rendszert érintő adatcsere-formátumok mindenképpen kompromisszumokra kényszerülnek a szemantikát illetően.

A formátum szintaxisa és szemantikája önmagában még nem elégséges az adott formátumban közölt rekordok kezeléséhez. Ehhez még ismerni kell a kódolt jelkészletet is, amit az adott formátumban rögzített rekordokban használnak. Magát a jelkészlet-azonosító megnevezését a rekord tartalma, vagy a fájl címkéje megadhatja, de ez még nem elégséges. Ezért tekinthető – feltételesen – a formátumokhoz nem tartozó *jelkészlet a formátum harmadik elemének.*

3. Adatcsere-formátumok szükségessége

E cikk keretében nem szorul bizonyításra, hogy a korszerű információs világ az információk cseréjének és kölcsönös hasznosításának, a rendszerek összekapcsolásának, a hálózatoknak a világa. Másfelől a fentiekből következő, és egyébként is ismert tény, hogy az egyes könyvtárak, információs rendszerek meghatározó belső formátumai elütőek. Ehhez járul még, hogy a különböző könyvtárak eltérő alkalmazói szoftvereket használnak (ha nem is mindenütt érik el az egy könyvtárra jutó alkalmazói szoftverek számának magyarországi, kiemelkedő értékét, vagyis – legalábbis szoftver-lobbyjaink által – dédelgetett szoftver-heterogenitásunkat, finomabban sokszínűségünket). A szoftverek eltérése szintén a formátumok különbségéhez vezet. Ez azt jelenti, hogy még optimális esetben is (az abszolút gátló tényezőkről utóbb lesz szó) *a rendszerek közötti adatcsere konverziót igényel,* kedvező esetekben számítástechnikai eszközökkel megvalósítható konverziót. Az adatcsere gyakorlata aránylag gyorsan vezetett két felismerésre, nevezetesen arra, hogy

- a belső, és ebből adódóan az input-formátumok olyan közelítése, amely a konverziót szükségtelenné tenné, nem oldható és nem is oldandó meg, de
- a konverziós erőfeszítések és ráfordítások csökkentendők lennének, adatcsere-formátumok útján.

Ha feltételezzük, hogy egy adatcserét lebonyolító könyvtári együttesnek öt tagja van, és mindegyiküknek eltérő saját formátuma van, akkor a teljes körű adatcseréhez mindegyik könyvtárnak négy konverziós programot kell készítenie, a másik négy könyvtár által adott adatok használatához (vagy ha az ügy iránti őszinte odaadást tételezünk fel, akkor saját formátumának mind a négy cserepartnerének formátumára való konverziójához, ami ugyanannyi programot jelent). Öt könyvtár esetében ez 20 konverziós program, általánosítva, X könyvtár esetében X(X–1) konverziós program. Ha azonban létezne egy közö-

sen elfogadott adatcsere-formátum, akkor mindenkinek csak két konverziós programra (sajátjáról az adatcsere-formátumra és az adatcsere-formátumról a sajátjára) lenne szüksége, tehát az öt könyvtárnak 10 programra (X×2). Minél nagyobb a könyvtárak száma, annál inkább nő ez a különbség: nyolc könyvtár esetében az első variáns maximálisan 56, a második maximálisan 16 konverziós programot igényel.

Ilyen egyszerűen, a konverziós programok számának csökkenthetőségével *bizonyítható az adatcsere-formátum (elvileg optimálisan egyetlen adatcsere-formátum) szükségessége.* Emellett egy általános adatcsere-formátum (legalábbis szemantikájában) befolyásolhatja, közelebb hozhatja, adott körön belül akár egységesítheti is az érdekeltek saját formátumait.

Az *adatcsere-formátumnak* – a fentiekből adódóan – *a leggyakoribb alkalmazásokhoz kell alkalmazkodnia, illetve lehetővé kell tennie egymást nem szükségképpen kizáró variánsok rögzítését.* Az első cél viszonylag jól elérhető, ha a szemantika a legelterjedtebben használt katalogizálási szabályzat(ok)ra épül (mint teszi pl. az USMARC, vagy a MAB), az utóbbi – kissé ellentmondó – cél viszont minél többféle megoldás befogadását követeli meg (mint a CCF és – korlátozottabb mértékben – az UNIMARC esetében).

Az adatcsere-formátum valójában szabvány. Következésképpen elvileg *egyetlen adatcsere-formátum* országos, vagy világszerte elterjedt alkalmazása lenne célszerű. Az előbbiek és a gyakorlat utalnak arra, hogy ez a cél legfeljebb megközelíthető.

Ide tartozik még egy félreérthető és egy vitás kérdés lehetőség szerinti tisztázása:

a) Az adatcsere-formátum nem csodaszer! Semmilyen *adatcsere-formátum sem oldhatja meg az adatcserét,* vagyis a kapott adatok gépi konverzióját és hasznosítását, ha az adatelemek adattára úgy tér el, hogy
   - a fogadónak olyan adat kellene, amit a küldő nem dolgoz fel,
   - a küldő (és az adatcsere-formátum) adatai összevontabbak mint a fogadóé,
   - ha az adatelemek közlésmód-szabályai eltérőek a küldőnél és a fogadónál.

   Ilyenkor csak a megalkuvás és/vagy a kapott rekordok intellektuális/manuális szerkesztése segít!

b) Sokan úgy vélik, hogy az on-line adatcsere, pontosabban egymás adataihoz való on-line hozzáférés korszakában az adatcsere problémája elavult. Ebben van némi igazság, de csak igen korlátozottan, mert a sok adatbázisban on-line kereső használónak a különböző körű, és igen korlátozottan szabványosított adatokban való eligazodás nehézségeket jelent, vagyis az adatcsere, vagy legalább az elterjedt formátumok egységesítő hatása nem mellőzhető. Maradéktalanul igaz vi-

szont ez a feltételezés (mert vagy feleslegessé válik, vagy igen egyszerű az adatcsere) olyan zárt hálózatokon belül, ahol
– azonos az alkalmazói szoftver,
– azonos az adatelemek adattára,
– azonosak a közlésmódra vonatkozó és egyéb szabályok.

4. Kiegészítő történelmi és alkalmazói áttekintés a CCF-ről és az „adatcsere-formátum politikáról"

Olvasóm már gazdagabb a MARC-formátumok és fejlődésük számos ismeretével, és röviden tájékozódott a Common Communication Format-ról (Közös Adatcsere-formátumról, CCF-ről) is, M. W. cikkéből. Az alábbi kiegészítések részben a CCF kialakulását, fejlődését és helyzetét, részben ennek tanulságait illetik.

Az 1978-ban a szépséges és felejthetetlen Taorminában tartott szimpózium, amelynek célja az volt, hogy tanulmányozza, vajon mennyire kívánatos és lehetséges a létező bibliográfiai adatcsere-formátumok maximális kompatibilitásának megvalósítása, olyan helyzetben került sorra, amikor (a hivatkozásokat lásd M. W. cikkében)

– egy évvel korábban már publikálták az egységes nemzetközi adatcsere-formátum igényével fellépő *UNIMARC*-ot,
– 1974 óta létezett a szintén nemzetközi használatra szánt *UNISIST Reference Manual*,
– *további* olyan *adatcsere-formátumok* készültek a közelmúlt években, amelyeket *nemzetközi használatra* szántak, de
– a géppel olvasható *bibliográfiai adatok tényleges kommunikációja* változatlanul elsősorban nemzetinek hívott, valójában inkább nemzeti könyvtári/nemzeti bibliográfiai, illetve egyes szolgáltatások által kifejlesztett formátumok használatával ment végbe,
– mind nemzeti, mind nemzetközi keretekben egyre nyilvánvalóbb lett, hogy a *szükségletek nem ismerik* a nemzeti bibliográfiák közötti, vagy általában a könyvtárközi adatcsere, vagy éppenséggel a referáló és indexelő szakbibliográfiai szolgáltatások közötti *adatcsere korlátait*, hanem e „táborok" között is szükséges az adatcsere.

Nem lehet csodálni, hogy lelkesen fogadták a szimpózium ajánlásait[10], így azt is, hogy *létre kell hozni egyetlen, közösnek nevezett adatcsere-formátumot,*

10 Towards a common bibliographic exchange format? Proceedings of the International Symposium on Bibliographic Exchange Formats, Taormina, Sicily, 27–29 April 1978. Budapest, OMKDK–Technoinform, London, UNIBID, 1978.

amely formátumok, vagy akár adatcsere-formátumok közötti hídként szolgálhat, és amelynek alapelvei, hogy

- szintaxisa az ISO 2709 szerinti legyen,
- kevés számú kötelező adateleme legyen alkalmas a legfontosabb, szabványosan meghatározott bibliográfiai adatokból álló rekordok létrehozására,
- ezek az adatelemek egészüljenek ki szabványosan meghatározott fakultatív adatelemekkel,
- szabványos módszert használjon a bibliográfiai szintek és a bibliográfiai tételek, közötti kapcsolatok kifejezésére.

A munka hat kiválasztott nemzetközi formátum elemzésével és egy nagy terjedelmű *adatelem-gyűjtemény* összeállításával indult. Ennek további elemzésére és az említett szabványos technikák meghatározására épült fel a CCF 1984-ben megjelent első kiadása, amelyet *Peter Simmons* és *Alan Hopkinson* szerkesztő munkája hozott létre, de amelynek minden adatelemét és technikáját az egész munkát szervező és finanszírozó Unesco Általános Információs Programja keretében működő CCF munkacsoport (Ad hoc Group on the Establishment of a Common Communication Format) vitatta meg és hagyta jóvá *Nathalie Dusoulier* elnöklete mellett. Bár nem volt jó jel, hogy a munkacsoportból (amelynek – más szakértők mellett – tagjai voltak számos nagy formátum reprezentánsai is) munka közben az utóbbiak közül többen kiváltak, mégis *remélni lehetett, hogy a CCF betölti szerepét mint egységes nemzetközi formátum, amely nem szorítja ki a többi formátumot, de azok között hidat képezhet és mindazok rendelkezésére áll, akik nem kötelezték el magukat egyetlen formátum mellett.* E szerepre a CCF rugalmassága folytán is alkalmasnak látszik, ugyanis csak minimálisan tartalmaz közlésmód-szabályokat, és változatokat kínál ugyanazon adatelem különböző formájú közléseinek elhelyezésére és megkülönböztetésére. Hasonlóan hat a kötelező adatelemek viszonylag csekély száma és a fakultatív adatelemek széles köre.

Tíz évvel első kiadásának[11] megjelenése után be kell látni, hogy a várakozások túlzottak voltak. Ennek egyik oka, hogy valamiféle *rivalizálás alakult ki az UNIMARC és a CCF között.* Az UNIMARC összeállítói – hiába volt az UNIMARC a maga nemzeti könyvtári és nemzeti bibliográfiai indíttatásával érzéketlen a szakkönyvtárak, szakbibliográfiák és a referáló/indexelő szolgáltatások követelményei iránt, mind összetételével, mind a bibliográfiai szintek kezelésével (így az analitikus-cikkrekordok problémáinak megoldatlanságával) – egyetlen nemzetközi formátumként kívánták látni az UNIMARC-ot. Másfelől

11 CCF: The Common Communication Format. Edited by Peter Simmons and Alan Hopkinson for the General Information Program and UNISIST. Paris, Unesco, 1984. (PGI–84/WS/4)

a nemzeti formátumok használói és a *különféle bibliográfiai szolgáltató központok nem mutattak különösebb hajlandóságot egyik nagy nemzetközi formátum használatára sem*, eltekintve az UNIMARC M. W. által is említett alternatív használatától a nemzeti MARC formátumok mellett.

A CCF elsősorban a fejlődő országokban és nemzetközi szervezetek körében aratta sikereit, de – jó egy néhány kivétellel – nem adatcsere-formátumként, hanem *olyan mintaként és referenciagyűjteményként, amire adatbázisok létrehozói építették adatelem-specifikációikat,* vagy éppenséggel technikáikat is, más szóval a CCF belső formátumok alapjává vált.

Időközben a CCF a további években több irányban is fejlődött. 1988-ban megjelent második, a tapasztalatok alapján javított kiadása, 1989-ben sor került a CCF-használók első nemzetközi tanácskozására, amelynek ajánlásai nyomán a CCF legújabb kiadása két teljesen kompatibilis részre oszlott, és 1992-ben mint *CCF/B, a bibliográfiai információk közlésére* szolgáló CCF[12] és *CCF/F, a faktografikus információk közlésére* szolgáló CCF[13] jelent meg. E változást a faktografikus információk és a bibliográfiai információk integrált adatbázisokban és ennek nyomán az adatcserében végbemenő ötvöződése tette lehetővé. Az immár két részre oszlott CCF sem csak adatcsere-formátumként, hanem adatbázis-építéshez is kitűnően használható – ezt e sorok írója saját tapasztalatával is igazolhatja. Megjelent a CCF használati tanácsadója (Implementation notes for users of the Common Communication Format) is, *Alan Hopkinson* tollából[14].

Enyhült az UNIMARC/CCF feszültség is. 1992 júniusában Firenzében *UNIMARC/CCF munkaértekezletet* (workshop-ot) tartottak a problémák tisztázására (lásd még alább). Az ott elhangzottaknál is nagyobb mértékben járult hozzá az együttműködéshez a *Peter Simmons* által készített UNIMARC/CCF konvertáló program.

Hogy áll ezek után az adatcsere-formátum politika?

a) Az *adatcsere-formátumok eredeti szerepe és jelentősége vitathatatlanul csökkent* az „on-line környezetben", de amíg off-line adatcsere lesz – és még belátható ideig lesz –, addig adatcsere-formátumokra szükség lesz.

---

12 CCF/B: The Common Communication Format for Bibliographic Information. Edited by Peter Simmons and Alan Hopkinson for the General Information Program. Paris, Unesco, 1992. (PGI–92/WS/9)

13 CCF/F: The Common Communication Format for Factual Information. Edited by Peter Simmons and Alan Hopkinson for the General Information Program. Paris, Unesco, 1992. (PGI–92–WS/8)

14 Hopkinson, Alan: Implementation Notes for Users of the Common Communication Format (CCF). Paris, Unesco, 1993. (PGI–90/WS/3)

b) Sem az *UNIMARC*, sem a *CCF* nem tudott a másik helyére lépni, sem kiszorítani a nemzeti és szolgáltatói kommunikációs formátumokat. Mégis, *vitathatatlan vezető szerepük* kialakult, és ma már jellemző egymás kiegészítése, és nem az, hogy ki akarnák szorítani egymást (még akkor is, ha a fent említett UNIMARC/CCF Workshop-on olyan előadás is elhangzott, nem is akárki, hanem egy, mindkét formátumban érdekelt személy szájából, aminek – egyszerűsített – mondandója az volt, hogy legjobb lenne mindkettőt abbahagyni, és egységesen az USMARC-ot használni). A két formátum különösen az újonnan nemzeti, vagy szolgáltatói formátumot megalkotók számára lehet vonzó, ha Isten és az alkalmazói szoftver is úgy akarja.

c) Előtérbe került a két formátum – főként a CCF – „melléktermék"-hatása, nevezetesen az, hogy *belső formátum alapjául* alkalmazzák kisebb, vagy nagyobb mértékben adaptált változatát.

5. Hazai berkeinkből

Éber olvasóm figyelmét aligha kerülte el, hogy az ISO 2907 magyar változatának szabványjelzete MSZ 193/**1**–1983 (kiemelés tőlem, V .E.) Ez a /1 azt jelzi, hogy a /2, illetve esetleg a /x még várat magára, ugyanis ez(ek) a teljes formátumhoz még szükséges „szemantika" szabványa(i) lett(ek) volna. Nem nehéz kitalálni, hogy a folytatás miért nem született meg: nincs megegyezés abban, hogy *mit adaptáljunk.*

Időközben az *Országos Széchényi Könyvtárnak* létre kellett hoznia saját *kommunikációs formátumát.* A HUNMARC az USMARC-hoz (tehát korántsem mindenben az UNIMARC-hoz és alig valamiben a CCF-hez) hasonlít, több reális indokkal (a belső formátum adottságai, az alkalmazói szoftver „gusztusa" stb.). Úgy vélem, hogy ez nem kifogásolható, akár az sem, hogy az újszülött a *HUNMARC* nevet kapta. Születésének mindenki örül, és várja a HUNMARC-ban terjesztett központi szolgáltatásokat (hiszen a konverzió szempontjából a legfontosabb mégis csak az ISO 2907-en nyugvó szintaxis).

Mindez azonban nem jelenti azt, hogy e HUNMARC legyen az MSZ 193/2. Személy szerint úgy vélem, hogy *formátum-szabványra szükség* van. Semmi baj nem történik azonban, ha több változatot enged(nek) a szabvány(ok). Jelöltjeim: a CCF (mind a kettő), az UNIMARC és – esetleg – a HUNMARC, vagy ezek egyszerűsített változatai. A formális szabványt más ajánlások is pótolhatják, a jelzett tartalommal. Ha lehet, mielőbb.

MIRNA WILMER

Mirna Wilmer a zágrábi egyetemi könyvtár tudományos munkatársa és a dokumentációs adatcsere-formátum (Common Communication Format; CCF) nemzetközi munkacsoportjának tagja.

A szabványosítás szükségessége a géppel olvasható katalogizálásban[15]

1. Bevezetés

A MARC formátumok megjelenése szorosan összefügg a szabványosítással. A MARC formátumok gyakorlatilag nem jöhettek volna létre magas szintű szabványosítás nélkül.

A MARC rövidítés a MAchine Readable Cataloguing (géppel olvasható katalogizálás) kifejezésből származik. Lényegét legjobban a következő meghatározás tartalmazza: „Formátumok csoportja, amely különleges konvenciók (megállapodásos szabályok) halmazát alkalmazza a bibliográfiai adatok azonosítására és rendszerezésére, azok gépi kezelésekor."[16]

A fenti fogalommeghatározás a gépi katalogizálás két lényeges fogalmát – a „formátumok csoportja" és a „konvenciók halmaza" – tartalmazza, melyekkel jelen írásomban foglalkozom.

2. A MARC formátum történeti háttere

A számítógépek könyvtárakban való elterjedése felvetette a bibliográfiai adatok egységes leírásának, azonosításának és szerkesztésének szükségességét, vagyis azt, hogy az adatokhoz szabványos kódokat rendeljenek annak érdekében, hogy a számítógép kezelni tudja őket.

A hatvanas évek küszöbén a Library of Congress fejlesztette ki az első MARC formátumot, amelyből azután a későbbi többi MARC formátum származott. Az Egyesült Államokban végzett fejlesztést közvetlenül követte a hasonló célú nagy-britanniai fejlesztés, és az együttes munka eredményeként létrejött két formátum az LCMARC, későbbi nevén USMARC, illetve az UKMARC lett. Ez volt az az alap, amelyre építve azután több nemzeti MARC

15 A szerző tanulmányát a Könyvtári Figyelő számára írta Machine readable cataloguing the necesity for standardization címen. In: Könyvtári Figyelő, 1994, 4. évf., 1. sz., p. 42–48.

16 Gredley, Ellen, Hopkinson, Alan: Exchanging bibliographic data : MARC and other international formats. – Ottawa [etc.]: Canadian Library Association [etc.], 1990. – 70 p.

formátumot fejlesztettek ki. E formátumok széles körű elterjedése a bibliográfiai adatok cseréjére bátorított, és új fejlődési szakaszt jelentett az Egységes Bibliográfiai Számbavételhez (Universal Bibliographic Control) vezető úton.

Kanada kétnyelvűsége miatt a MARC gondozására itt létrehozott munkacsoport az LCMARC, az UKMARC, valamint a francia, az olasz és a német MARC formátumok messzemenő figyelembevételével alkotta meg a kanadai MARC formátumot. A kanadai MARC formátum az LCMARC-on alapszik, csak bővített mezőkkel és mezőazonosítókkal a sajátos kanadai igények kielégítése érdekében. Az ausztráliai MARC formátum, az AUSMARC alapja az USMARC és az UKMARC közösen, csekély módosításokkal. Az LCMARC, UKMARC, CANMARC és AUSMARC továbbfejlesztését és karbantartását az ABACUS (Association of Bibliographic Agencies of Britain, Australia, Canada, US: Nagy-Britannia, Ausztrália, Kanada és az Egyesült Államok bibliográfiai központjainak Szövetsége) keretében végzik.

A Német Szövetségi Köztársaság 1973-ban adta közre saját formátumát. Ez volt a MAB (Maschinelles Austauschformat für Bibliotheken). Legfőbb jellemzője, nem véletlenül az, hogy létrehozták a különböző bibliográfiai leírási szintű rekordok – így a többkötetes monográfiák, a monográfiák és analitikus leírásai – rekordjainak kapcsolására szolgáló eszközöket. Nem véletlenül, mert a német katalogizálási szabályzat, a RAK előírja, hogy e sajátos bibliográfiai körülmények között többlépcsős bibliográfiai leírásokat kell alkalmazni. Az egyes országok bibliográfiai leírási szabályzataiban rögzített feltételek és előírások nagymértékben hatnak a nemzeti formátumok kialakítására. A MAB formátum kapcsolási technikája, amely e bibliográfiai feltétel (mármint a többlépcsős leírás) szükségességét felismerő határozott európai gyakorlatra támaszkodik, nagy hatással volt az UNIMARC formátum fejlesztésére.

A hetvenes években különböző más nemzeti formátumokat is kialakítottak, így Olaszországban az ANNAMARC-ot, Spanyolországban az IBERMARC-ot, Dániában a DANMARC-ot, Katalóniában a CATMARC-ot, amelyek mindegyike vagy az USMARC-on, vagy az UKMARC-on nyugszik. Azt is meg kell jegyezni, hogy egyes könyvtári információs szervezetek vagy intézetek, amelyek saját szoftvert fejlesztettek ki, megalkották saját MARC formátumukat is, melyeket később, a szoftverek elterjedésével együtt, széles körben használtak. Ezek közül a legjelentősebbek a DOBIS/LIBIS és a Pica formátumai.

A legelső nemzetközi MARC formátumnak az 1975-ben megjelent INTERMARC (International MARC) formátum tekinthető. A nyugat-európai nemzeti könyvtárak képviselőinek egy csoportja fejlesztette ki, és végül ez lett Belgium és Franciaország nemzeti formátuma is, jóllehet Franciaország nem sokkal az UNIMARC közreadása után az UNIMARC-ot fogadta el belföldi használatra is, belső adatbázis- és adatcsere-formátumként. Az INTERMARC-ot belső célokra azonban még mindig használja a francia nemzeti könyvtár, bár minden cserére szánt terméke lényegében UNIMARC formátumú.

A nyugat-európai országokon kívül a létező formátumok különféle formái és adaptációi terjedtek el. Több egyéb formátum mellett Kelet-Európában a MEKOF-ot használják; a különféle latin-amerikai nemzeti formátumok többnyire USMARC alapúak, míg Indiában és Délkelet-Ázsiában a használatban lévő formátumok vagy USMARC, vagy UKMARC alapúak mint az INDIMARC, THAIMARC, MALMARC, INDOMARC stb.

A hetvenes évek vége felé és a nyolcvanas évek elején az UNIMARC formátum – melynek első kiadása 1977-ben jelent meg – erősen hatott a nemzeti formátumok kialakítására. Ez volt a helyzet az ázsiai és az afrikai országokban és később Európában is. Az, hogy az UNIMARC-ot vették a nemzeti formátumok alapjául, vagy magát az UNIMARC-ot használták mint nemzeti formátumot, a következőkkel magyarázható:

- az UNIMARC-ot sokkal általánosabb elvek szerint tervezték, mint az USMARC-ot, vagy az UKMARC-ot;
- az UNIMARC nem ír elő, illetve nem feltételez különleges katalogizálási szabályokat, mint például az Anglo-American Cataloguing Rules, melynek használatát az USMARC vagy az UKMARC előírja;
- miután az UNIMARC maga is IFLA szabvány, az IFLA által javasolt szabványokat és irányelveket alkalmazza;
- az UNIMARC-ot a szakterület nemzetközi szakértői együttesen tervezték;
- lévén, hogy IFLA formátum, elvárható, hogy az IFLA biztosítja annak szakértői szintű fenntartását és továbbfejlesztését;
- az UNIMARC integrált formátum, ami azt jelenti, hogy különböző típusú könyvtári anyagok esetében használható;
- az UNIMARC lehetőséget biztosít a nem latin betűs írás és a többnyelvű rekordok kezelésére;
- az UNIMARC biztosítja a rekordkapcsolás mechanizmusát;
- az UNIMARC nemzetközi szintű adatcsere-formátum, így minden azt alkalmazó intézmény közvetlenül átveheti az adatokat más országok intézményeitől.

A hetvenes évek vége felé Japán, Dél-Afrika, Tajvan, majd később Kína is az UNIMARC-ot adaptálta nemzeti formátumként. Európában 1980-ban Horvátország Nemzeti és Egyetemi Könyvtára adaptálta belső adatbázis- és adatcsere-formátumként. A Maribori Egyetemi Könyvtár 1984-ben kezdte alkalmazni adatbázisának építéséhez, a későbbiekben pedig a formátumot adaptálták, hogy kielégítse a szlovén közösség használati igényeit; ma ez a formátum COMARC néven ismert. 1986-ban Portugália is az UNIMARC-ot fogadta el nemzeti és adatcsere-formátumnak. A nyolcvanas évek derekén néhány nemzeti könyvtár UNIMARC formátumban is rendelkezésre bocsátotta mágnesszalagos csereszolgáltatásait, így a Library of Congress, a Deutsche Bibliothek, az

olasz, a francia nemzeti könyvtárak stb. Érdekességként megjegyzem, hogy Olaszországban és Franciaországban az UNIMARC egyben az országon belüli adatcsere formátuma is.

A kilencvenes években az UNIMARC használata újabb lökést kapott. Az Európai Közösségben lévő könyvtárak az UNIMARC-ot vették figyelembe szabványos adatcsere-formátumként projektjeik adatcseréjéhez. Ilyen folyamatban lévő projekt az EROMM, European Register of Microform Masters (a mikrofilmek alappéldányainak európai regisztere), valamint a Consortium of European Research Libraries-nek (Európai Tudományos Könyvtárak Konzorciuma) a régi nyomtatványok (a nyomtatás kezdetétől 1830-ig megjelent művek) számbavételére irányuló projektje.

3. Az UNIMARC: nemzetközi MARC formátum

A hetvenes években úgy tűnt, hogy az eltérő nemzeti MARC formátum, azok egymással való gyakori inkompatibilitása és az eltérő katalogizálási szabályzatok miatt a nemzetközi bibliográfiai számbavétel (UBC) és a nemzeti könyvtáraknak a bibliográfiai adatok géppel olvasható formájú cseréjén nyugvó együttműködése távol áll a megvalósulástól. Több találkozó, megbeszélés, vita és tanulmány után a nemzeti könyvtárak igazgatóinak konferenciája tervbe vette a Nemzetközi MARC Hálózat kialakítását. A munkák irányát meghatározó tervtanulmányokat *Stephen Green* (British Library) és *R. M. Duchesne* (National Library of Canada) dolgozták ki. Duchesne volt az, aki már 1971-ben felvetette a valóban nemzetközi kommunikációs formátumnak a gondolatát, amely a nemzeti változatok szuperhalmaza volna.[17] Ez a SUPERMARC formátum kettőre csökkentené le a könyvtárak konverziós programjainak számát: a nemzeti formátumról a SUPERMARC-ra és a SUPERMARC-ról a nemzeti formátumra. 1977-ben adták közre *A. G. Wells International MARC Network : A study for an international bibliographic data network* című művét.[18] E tanulmány a szabványosítást az alábbi területeken javasolja:

- a leíró adatok ISBD alapú szabályozása;
- a tárgyköri adatok szabályozása;
- a nemzeti központok közötti kommunikációs hálózat, a bibliográfiai adatcsere, a rekord-tulajdon politikájának szabályozása;

17 Duchesne, R. M.: MARC and SUPERMARC in the exchange of bibliographic data and the MARC format. Proceedings of the International Seminar on the MARC Format and the Exchange of Bibliographic Data in Machine-readable Form, Berlin, 1971. – 2. ed. – Berlin : Verlag Dokumentation, 1973. pp. 37–57.

18 Wells, A. G.: The International MARC Network. A study for an international bibliographic data network. – London : IFLA International Office for UBC, 1977. – (Occasional paper : no. 3.)

– együttműködésben végzett tervezés és fejlesztés az egyes országokban létrehozott, regionális MARC hálózati csoportok, valamint az International MARC Network Tanácsadó Bizottsága, az IFLA-testületek, az ISO, az Unesco, az ISDS Központ és minden más, a szabványosításban érintett intézmény részvételével.

A fenti javaslatok egyik eredménye az IFLA International MARC Network Committee létrehozása volt, mely a MARC fejlesztése nemzetközi koordinátoraként volt hivatva működni. Az IFLA Katalogizálási Bizottsága és a Gépesítési Bizottság által fenntartott IFLA Group on Content Designators (tartalmi azonosítókkal foglalkozó csoport; tartalmi azonosítók alatt itt a nemzetközi formátum mezőit és almezőit meghatározó hívójelek, almező-azonosítók és indikátorok értendők; ford. megj.) által ajánlott UNIMARC első kiadása 1977-ben jelent meg. Elsődleges célja az volt, hogy megkönnyítse a nemzeti bibliográfiai központok között a géppel olvasható bibliográfiai adatok cseréjét. Azt is állították, hogy a formátum mintaként szolgálhat új, géppel olvasható bibliográfiai formátumok kifejlesztéséhez. 1980-ban megjelent az UNIMARC második kiadása. Az UNIMARC Handbook 1983-ban jelent meg. E kézikönyv UNIMARC Manual című, 1987-es átdolgozott kiadása tartalmazza az ISBD legújabb kiadásaiban található változásokat. Így ez tekinthető az UNIMARC formátum harmadik kiadásának.[19]

Az IFLA 1986-os tokiói konferenciáján elhatározták két IFLA program összevonását: a Nemzetközi MARC Program (International MARC Programme) és az Egyetemes Bibliográfiai Számbavétel (Universal Bibliographic Control) összevonását UBCIM Programme (Universal Bibliographic Control and International MARC Programme) néven.

1991-ben jelent meg az UNIMARC/Authorities (egységes besorolási adatok formátuma).[20] E formátum az IFLA Working Group on an International

19 UNIMARC : Universal MARC Format. Recommended by the IFLA Working Group on Content Designators set by the IFLA Section on Cataloguing and the IFLA Section on Mechanization. – London : IFLA International Office for UBC, 1977.

UNIMARC : Universal MARC Format. Recommended by the IFLA Working Group on Content Designators set by the IFLA Section on Cataloguing and the IFLA Section on Mechanization. – 2. rev. ed. – London : IFLA International Office for UBC, 1980.

UNIMARC handbook. Compiled and edited by Alan Hopkinson with the assistance of Sally McCallum and Stephen Davies. – London : IFLA International Office for UBC, 1983. – ISBN 0–903043–40–8

UNIMARC manual. Edited by Brian Holt with the assistance of Sally McCallum and A. B. Long. – London : IFLA Universal Bibliographic Control and International MARC Programme : British Library Bibliographic Services, 1987. – ISBN 0–9030043–44–0

20 UNIMARC/Authorities : Universal format for authorities. Recommended by the IFLA Steering Group on a UNIMARC Format for Authorities ; approved by the Standing Committees of the IFLA Section on Cataloguing and Information Technology. – München [etc] : K.G. Saur, 1991. – ISBN 3–598–10986–5

Authority System (Egységes Besorolási Adatok Nemzetközi Munkacsoportja) ajánlásait tartalmazó, 1984-ben kiadott *Guidelines for Authority and Reference Entries* alapjaira épül. Az UNIMARC/Authorities besorolási, utaló- és általános, magyarázatos utaló-tételek mezőinek hívójeleit határozza meg. Ez a formátum kiegészíti az UNIMARC Manualban definiált, a bibliográfiai adatok közlésének szerkezetét, tartalmi azonosítóit, kódolási rendszerét és kapcsolási technikáját szabályozó formátumot. A két formátum összefüggését az UNIMARC/Authorities speciálisan meghatározza.

Az UNIMARC/Authorities-t még vizsgálni kell. A formátumát 1990-ben már adaptálták és használatba vették a Horvát Nemzeti és Egyetemi Könyvtárban, amikor az még csak tervezet formájában állt rendelkezésre. A hirtelen jött használatbavételnek az volt az oka, hogy éppen akkor vált hozzáférhetővé, amikor a könyvtár az új számítógépes rendszerét tervezte.

Az UNIMARC formátumokkal kapcsolatos munka azonban ezzel a legújabb kiadással nem zárult le. Éppen ellenkezőleg: új impulzust kapott az Állandó IFLA UNIMARC Bizottság (Permanent UNIMARC COMMITTEE, PUC) 1991-es létrehozásával.[21] A PUC fő célja az UNIMARC formátum ellenőrzése az egységes bibliográfiai számbavétel elveinek szempontjából. A bizottság most dolgozik az UNIMARC Manual új kiadásán; megjelentetését 1994-re tervezik. A bizottság népszerűsítő és marketing kampányt is folytat, tanácskozások és projektek útján. A projektek egyike az UNIMARC használók és szakértők névjegyzékének összeállítására szolgáló kérdőív elkészítése; brosúra közreadása az UNIMARC-ról és más akciók. A bizottság kapcsolatban áll és együttműködik az IFLA más testületeivel, az Unesco CCF munkacsoportjával, az ISDS központtal (új nevén: ISSN központtal; ford. megj.), az ISO-val és különféle, UNIMARC alapú projektek végrehajtóival.

4. A MARC formátum szerkezete

A formátum három szintből áll, amelyek mindegyikének alkalmazása magas fokú szabványosítást igényel.

a) A formátum struktúrája: a géppel olvasható rekordok fizikai megjelenítése. Mindegyik MARC formátum az ISO 2709 szabvány alkalmazásának speciális formája, amely szabvány meghatározza a bibliográfiai adatokat tartalmazó rekordok szerkezetét.

21 Permanent UNIMARC Committee : Terms of reference and procedures. In: International Cataloguing and Bibliographic Control. vol. 21. no. 4. (October/December 1992), pp. 51–52.

A könyvtáros-társadalom számára közvetlen hasznot jelent ez a szabvány, mert következetes alkalmazásával különböző formátumok között lényegében egyértelmű konverzió lehetséges, vagyis lehetővé teszi a géppel olvasható bibliográfiai adatok cseréjét.
E szabvány szerint minden cserére szánt bibliográfiai rekordnak tartalmaznia kell:

– a rekordfejet (record label), amely a rekord szerkezetével kapcsolatos, az ISO 2709 szabványban és egyes, a formátum sajátos alkalmazásaihoz meghatározott adatokat tartalmaz;
– a mutatót, amely (helyesen: amelynek egy-egy eleme; ford. megj.) három részből áll, a mező hívójelét, az illető mezőben található karakterek számát és a mező induló (első) karakterének a rekordon belüli pozícióját tartalmazza;
– változó hosszúságú adatmezőket. Ezek indikátorokat, almező azonosítókat és adatokat tartalmaznak.

b) A tartalmi azonosító kódok – a mezők hívójelei, az indikátorok és az almező azonosítók –, melyek az adatelemeket, vagyis az információk legkisebb, egyértelműen megkülönböztetett egységeit azonosítják, illetve kiegészítő információkat közölnek az adatelemekről.

A különböző intézmények különféle kódrendszereket, vagyis gyakorlatuknak és szabályaiknak megfelelő formátumokat fejlesztettek ki. A könyvtáros-társadalomban ezek a formátumok az USMARC, UKMARC, az INTERMARC... és az UNIMARC.

c) A rekord tartalma: bibliográfiai adatok, például cím, szerző neve, a kiadás helye, éve, megjegyzések, osztályozási jelzetek, tárgyszavak, de olyan adatok is, mint a mű nyelve és írásrendszere, a megjelenési ország neve is.

A rekordok tartalmának szabványosítása nemzetközileg a Párizsi Alapelveken, a különböző típusú dokumentumok bibliográfiai leírásainak szabványain, az ISBD-ken, az egységes besorolási adatok alakjára vonatkozó irányelveken, a személynevek alakjairól és az egységesített címekről készült listákon és egyebeken nyugszik, nemzeti keretekben pedig a katalogizálási szabályzatokon, osztályozó rendszereken, a tárgyszavazást szabályozó szabályokon és szabványokon nyugszik. A szabványok másik csoportjába – amely szintén a bibliográfiai rekordok tartalmához kapcsolódik, jóllehet nem közvetlenül a bibliográfiai adatok köréhez és alakjához – tartoznak azok a szabványok, amelyek a transzliterációt, a jelkészletet, a megjelenési országoknak,

a címek nyelvének, a katalogizálási szabályzatoknak és a tárgyköri osztályozási rendszereknek a kódjait határozzák meg.

Nem-MARC formátumok: a szolgáltatások környezete

A könyvtárakon kívüli környezetben számos intézmény fejlesztett ki sokféle formátumot saját bibliográfiai adatbázisainak, illetve szakosított számítógépes hálózatokban végzett adatcseréjének szükségleteire. Ezeknek a formátumoknak a tervei a nemzeti könyvtárak MARC formátumainak korai munkálataira, az ISO szabványokra és a referáló és indexelő (szakbibliográfiai) szolgáltatások körének sajátos gyakorlatára épültek. Ennek a fejlesztésnek azután az lett az eredménye, hogy e formátumok olyan mértékben tértek el a MARC formátumoktól, hogy az adatkonverzió ezek között és a MARC formátum között rendkívül nehézzé vált. Bár a bibliográfiai feltételek alapjában azonosak, a hangsúlyt ezeknél a formátumoknál a kiadványok analitikus szintű feldolgozására helyezik, különös figyelmet szentelve a tárgyköri indexelésnek és a referátumok rögzítésének.

Ennek a környezetnek a meghatározó formátuma az UNISIST–ICSU/AB Working Group on Bibliographic Description (Bibliográfiai Leírási Munkacsoport) által kifejlesztett UNISIST Reference Manual. A munkát az Unesco és az International Council of Scientific Unions (ISCU) égisze alatt végezték, a Világméretű Tudományos Információs Rendszer (World Science Informations System; UNISIST) létrehozása végett.

Az UNISIST Reference Manual jellemzője, hogy táblázatokat tartalmaz a különféle bibliográfiai szintek minden egyes kombinációjának rögzítésére szolgáló mezőkkel, így időszaki kiadványok cikkeinek, monográfiák részeinek, sorozatokba tartozó monográfiáknak és monográfiáknak a feldolgozására. Nem készült tábla az időszaki kiadványok feldolgozására. A formátum legmeghatározóbb felhasználói voltak: az ottawai International Development Research Centre, ahol a MINISIS szoftverre alkalmazták, az American Geological Institute, amelyik a GeoRef referáló szolgáltatásához használta, a Comisión Económica para América Latina (CEPAL), és még jó néhány intézmény, szerte a világban.

Az INSPEC és a Chemical Abstracts (CAS) formátumok még a MARC formátumok előtt alakultak ki, amelyek nem az ISO 2709-re épülnek. A referáló szolgáltatások környezetében kifejlesztett további formátumok az AGRIS (Agricultural Information System); az INIS (International Nuclear Information System) formátumai stb.

Az ISDS formátumot az International Serials Data System, az UNISIST program részeként fejlesztette ki, és az International Centre for the Registration of Serial Publications (CIEPS; neve időközben ISSN Inter-

national Centre-re változott) gondozza. A CIEPS felelős az ISSN (International Standard Serial Number) számok kiadásáért, és ők kezelik az időszaki kiadványok központi adatbázisát. Az ISDS adatbázisban az ISDS katalogizálási szabályzata alapján katalogizálják a dokumentumokat. Növekedett azonban annak jelentősége, hogy fokozzák a kompatibilitást az ISDS formátum és az ISBD(S) között, mivel a nemzeti központok által katalogizált és rögzített rekordok kettős célt szolgálnak: az inputot az ISDS-be és ISBD(S) alapú bibliográfiai tételek létrehozását.

6. A formátumok közötti híd: CCF és az SGML

A kommunikáció szükségessége, vagyis az az igény, hogy az adatok a különféle közösségek határain keresztül cserélhetők legyenek, vezetett a Common Communication Format (CCF) kidolgozásához. A CCF az ISCU/AB, az ISDS, az IFLA, az ISO, az UNIBID (az UNISIST Reference Manual-t gondozó, azóta megszűnt nemzetközi központ) képviselői által végzett fejlesztés eredménye. Az UNIMARC, az UNISIST Reference Manual, az ISDS, a MEKOF (és néhány más nemzetközi formátum), valamint az ISO 2709-es szabvány elemzésén alapul. A formátumot a CCF ad hoc munkacsoportja kezeli, és az Unesco PGI (Általános Információs Program Osztálya) felügyeli. A formátum alapeszméje az, hogy olyan adatcsere-formátum legyen, melyet hídként használnak a könyvtári közösség és a referáló és indexelő szolgáltatási környezete között. Jellemzői lehetővé teszik azonban, hogy sok intézet saját belső formátumaként is felhasználja. A formátum két fő jellemzőjét úgy lehet összefoglalni, hogy:

- a formátum nem kötődik katalogizálási szabályokhoz;
- a CCF rekord struktúráját az ISO 2709–1983 szabvány legfrissebb kiadása határozza meg.

A rekordszerkezet szegmenseken nyugszik: minden egyes rekordban annyi szegmens alkalmazható, amennyit a különböző dokumentumok különböző bibliográfiai leírási szintjei megkövetelnek. Például, a rekordban egy szegmens az időszaki kiadvány adatai céljára szolgál, egy másik az időszaki kiadványban található cikkek adatainak feldolgozására. Másik példaként említhető a sorozatban megjelenő többkötetes monográfia többszintű leírása, ugyanabban a rekordban, de annak egymástól elválasztott szegmenseiben. Mindegyik szegmens tartalmazza a bibliográfiai szint kódját és kapcsolatot a többi szegmenshez. A kapcsolási technika mind szegmensek, mind rekordok közötti kapcsolat kifejezését lehetővé teszi. A formátum egy további jellemzője az ISO 2709–1983 használata, vagyis a szegmenseken belül ismétlődő mezők kódolt megjelölése.

A CCF 1984-es, első megjelenése óta újabb kiadások láttak napvilágot, ami a használók elismerését tükrözi: a CCF/B (bibliográfiai adatokra), ami a CCF második kiadása; a CCF/F (faktografikus adatokra), és az Implementation Notes (Használói segédlet), amelyek mind 1992-ben jelentek meg.

A SGML (Standardized General Mark-up Language; Szabványos, Általános Szövegfeldolgozási Nyelv) ISO 8879 formátum a kiadói környezetből jött, ahol kezdetben a szövegek szedés és nyomtatás előtti megjelölésének szabványa volt. Mára a géppel szerkeszthető dokumentumok kódolt logikai struktúrájának szabványául fogadták el. Ez azt jelenti, hogy az SGML lehetővé teszi az elektronikus dokumentumok cseréjét, azok logikai struktúrájára, formátumára és az információk fizikai struktúrájára vonatkozó információkkal együtt, nyomtatás vagy képernyős megjelenítés céljából.

Az SGML a DTD-ből (Document Type Definition – dokumentumtípus-meghatározóból), a dokumentum logikai struktúrájából és magából a dokumentumból áll. A DTD szerinti SGML szabályok szintaxisát egy parser (nyelvészeti alapú szintaktikai elemző program) ellenőrzi.

Hogyan kapcsolódik az SGML a MARC formátumokhoz? A kapcsolódási lehetőségek sokfélék. Minden SGML szerint készült elektronikus dokumentum DTD-jében szerepelnek azok a kódolt bibliográfiai információk, amelyek leírják a dokumentumot. Tehát minden kiadó kódolja a bibliográfiai adatokat a DTD-ben, saját szabványai szerint. Ha mármost a könyvtári környezetben ezeket a mezőket áttehetik a saját szabványaik szerinti MARC típusú formátumokba, vagy MARC formátumba, ezáltal már a közeljövőben lehetővé válik, hogy az elektronikus publikációkat könyvtárak is felismerhető módon kezelhessék, új szolgáltatásokat nyújtva használóiknak.

FERNANDA M. CAMPOS; M. INES LOPES; ROSA M. GALVAO

Tanulmányukban leltározták a már létező nemzeti adatcsere-formátumokat és ezek eredetét. Részletes statisztikai adatokat közölnek arról, milyen mértékben játszanak szerepet a nemzetközi adatcsere-formátumok az egyes nemzeti formátumok kidolgozásában. Az alábbiakban néhány részletet közlünk a statisztikából.

MARC adatcsere-formátumok és alkalmazásuk[22]

Az USMARC és UKMARC adatcsere-formátumokon alapuló nemzeti adatcsere-formátumok

MAB – Maschinelles Austauschformat für Bibliotheken. 1973-ban készült el a Német Szövetségi Köztársaságban, második kiadása 1990-ben jelent meg. Legfontosabb jellemzője, hogy részletes rekordkapcsolatokat tartalmaz a különféle bibliográfiai szintek között (pl. az analitikus és monografikus, illetve kötet és monografikus közös leírás között). Ezzel számos későbbi formátumra, többek között az UNIMARCra is hatott.

PICA-MARC – Hollandia adatcsere-formátuma; különféle MARC formátumok jellegzetes hatásai érződnek rajta, mint INTERMARC, USMARC, UKMARC és UNIMARC.

CANMARC – Canadian MARC Communication Format. Első kiadása 1974-ben, második kiadása 1979-ben jelent meg. Elsősorban az USMAR-on alapszik. Számos mezőt vettek át az UKMARC és a MONOCLE formátumból (a MONOCLE formátum a korai hetvenes években Franciaországban készült).

ANNAMARC – Specifichje relativi ani mastri magnetici comntenenti i record della Bibliografia Nazionale Italiana nel formato ANNAMARC. A firenzei olasz nemzeti könyvtárban dolgozták ki 1978-ban. 1985-ig használták; ettől kezdve áttértek az UNIMARC formátumra.

DANMARC – DanMARC1. Udgave omfattende monogafier. Túlnyomórészt az UKMARC formátumból nőtt ki, első kiadása 1975-ben, második kiadása 1979-ben jelent meg Dániában.

SWEMARC – SWEMARC format specificatioin. 1980-tól alkalmazzák Svédországban; későbbi revízióját LIBRISMARC (LIBRary Information System MARC format) néven használták, és az UKMARC formátumon alapszik.

FINMARC – UKMARC formátumon alapuló finn nemzeti formátum, mely a svéd MARC formátum kidolgozásának nyomán készült el.

NORMARC – USMARC formátumon alapuló norvég nemzeti formátum.

---

22 MARC formats and their use / Fernanda M. Campos; M. Ines Lopes; Rosa M. Galvao. In: Program, 1995, Vol. 29, No. 10, p. 443–459.

IBERMARC – USMARC formátumon alapuló, 1976-ban elkészült első spanyolországi adatcsere-formátum.

CATMARC – Az UKMARC formátumhoz szorosan kapcsolódó, Spanyolországban 1987-ben elkészült adatcsere-formátum a katalóniai nemzeti könyvtár számára.

INTERMARC – Miközben az IFLA támogatásával már dolgoztak a nemzetközi szintű UNIMARC formátumon, 1975-ben a francia és a belga nemzeti könyvtár nemzetközi célú felhasználásra is alkalmas formátumként jelentette meg az *INTERMARC(M)* adatcsere-formátumot (*Format bibliographique d`échange pour les monographies*). Lényegében az UKMARC formátumon alapszik, számos önálló vonással és indikátorral. A belga és francia nemzeti könyvtárak használják. Az INTERMARC nem vált nemzetközi adatcsere-formátummá, mivel a közművelődési és egyetemi könyvtárak többsége az UNIMARC formátumot választotta.

HUNMARC – Az USMARC formátumon alapuló magyar nemzeti adatcsere-formátum, melyen 1990-től dolgoztak és 1996-ban jelent meg.

AUSMARC – Australian MARC Specification. Az UKMARC formátumhoz rendkívül közel álló ausztráliai nemzeti formátumot 1973-ban adták ki először. Második kiadására 1979-ben került sor, követve az UKMARC második kiadását és az angol–amerikai katalogizálási szabályzat (AACR2) változását. Az egységesített besorolási adatok dolgában szorosan kapcsolódik az USMARC formátumhoz.

MALMARC (Malajzia) – Az IMARC formátumon alapuló, 1977-ben kiadott adatcsere-formátum.

INDOMARC (Indonézia) – Az alapváltozata (1981) a SEAMARC formátumon alapult, amelynek viszont az USMARC formátum volt az alapja (a SEAPRINT – South-East Asean Imprints Project – adatcsere formátuma).

KORMARC (Korea) – mind az UKMARC, mind az USMARC formátumon alapuló, 1981-ben megjelent adatcsere-formátum.

SINGMARC (Szingapúr) – az UKMARC formátumon alapuló, 1979–80-ban elkészült adatcsere-formátum.

THAIMARC (Thaiföld) – az UKMARC formátumon alapuló, 1976-ban elkészült adatcsere-formátum.

MARCAL (Latin-Amerika) – az USMARC formátumon alapuló, 1991-ben elkészült adatcsere-formátum, melynek tényleges használatáról azonban nincs információ.

Az UNIMARC adatcsere-formátumon alapuló nemzeti formátumok

A nemzeti adatcsere-formátumok kidolgozásának második szakaszát 1977-től az UNIMARC megjelenése és hatása jellemezte. Számos államban egyszerűen átvették az UNIMARC formátumot, másutt pedig a nemzeti formátum kidolgozásának alapjául választották. A már létező nemzeti adatcsere-formátumok revízióját is befolyásolta az UNIMARC. A következő államokban alkalmazzák nemzeti adatcsere-formátumként az UNIMARC formátumot:

Horvátország és Szlovénia (1981), India (noha 1985-ben kidolgoztak és kiadtak egy INDIMARC nevű, az UKMARC formátumon alapuló nemzeti adatcsere-formátumot is), Görögország (1991), Portugália (1987) és Olaszország (megszüntetve 1985-ben az ANNAMARC használatát).

Az alábbi nemzeti adatcsere-formátumok alapszanak az UNIMARC formátumon:

SAMARC (Dél-Afrika) – Első kiadása 1977-ben jelent meg.

JAPANMARC (Japán) – Első kiadása 1981-ben, 3. kiadása 1988-ban jelent meg.

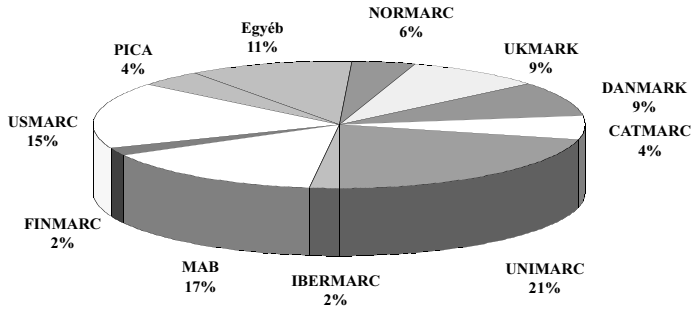
CNMARC (Taivan) – Első kiadása 1980-ban, 2. kiadása 1984-ben jelent meg.

Léteznek olyan nemzeti adatcsere-formátumok, melyek ugyan az UKMARC és/vagy az USMARC formátumon alapulnak, az UNIMARC mégis erősen hatott rájuk:

CALCO (Brazília) – Első kiadása 1981-ben jelent meg és az USMARC formátumon alapult, revíziójának eredményeként átkeresztelték IBICT névre.

KORMARC (Korea) – Megjelent 1981-ben.

CSMRC (Csehország) – Jelenleg folyik a revíziója.



Egyéb: CNN, Minisis, DobisMARC, helyi MARC formátumok, ISO 2709

3.1 A nemzeti bibliográfiákhoz használt adatcsere-formátumok

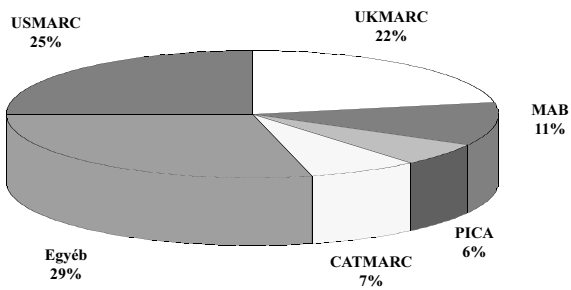
Európai államok

Európai állam	Nemzeti formátuma	ISO 2709	Alapformátum				Nyomtatott kiadás	Mágneses hordozón	Milyen formátumban szolgáltat
			US	UK	UNI	Egyéb			
Csehország	CSMARC	i							
Dánia	DANMARC	i		i			i	i	DANMARC
Finnország	FINNMARC	i		i			i	i	FINMARC
Franciaország	INTERMARC	i	i	i	i		i	i	UNIMARC, INTERMARC
Görögország	UNIMARC	i					i	i	
Hollandia	PICAMARC	i						i	USMARC
Horvátország	UNIMARC	i					i	i	UNIMARC
Lengyelország	POLMARC	i			i				
Magyarország	HUNMARC	i	i				i	i	HUNMARC, USMARC
Nagy-Britannia	UKMARC	i					i	i	UKMARC
Németország	MAB	kompat.		i			i	i	MAB, UNMARC
Norvégia	NORMARC	i	i				i	i	NORMARC
Olaszország	ANNAMARC, UNIMARC	i					i	i	ANNAMARC a), UNIMARC b)
Oroszország	MEKOF	kompat.						i	MEKOF, UNIMARC
Portugália	UNIMARC	i					i	i	UNIMARC
Spany. Katal.	CATMARC	i		i			i	i	
Spanyolország	IBERMARC	i	i		i		i	i	IBERMARC, UNIMARC
Svájc	USMARC	i						i	UNIMARC
Svédország	LIBRISMARC	i		i			i	i	LIBRISMARC

Európán kívüli államok

Európai állam	Nemzeti formátuma	ISO 2709	Alapformátum				Nyomtatott kiadás	Mágneses hordozón	Milyen formátumban szolgáltat
			US	UK	UNI	Egyéb			
Ausztrália	AUSMARC	i	i	i			i	i	AUSMARC, USMARC
Brazília	CALCO	i	i		i		i	i	
Dél-Afrika	SAMARC	i		i					
Fülöp-szigetek	PHILMARC	i					i	i	PHILMARC
India	UNIMARC	i							
Indonézia	INDOMARC	i	i				i		
Japán	JAPANMARC	i			i		i	i	JAPANMARC
Kanada	CANMARC	i					i	i	CANMARC
Kína	CNMARC	i			i	CCF			CNMARC, CCD
Korea	KORMARC	i	i	i	i				
Malajzia	MALMARC	i		i			i	i	ISO 2709
Szingapúr	SINGMARC	i		i				i	USMARC
Thaiföld	THAIMARC	i		i					
Új-Zéland	USMARC	i						i	USMARC
USA	USMARC						i	i	USMARC, UKMARC, UNIMARC

Belső és/vagy katalogizálási formátumként használt adatcsere-formátumok



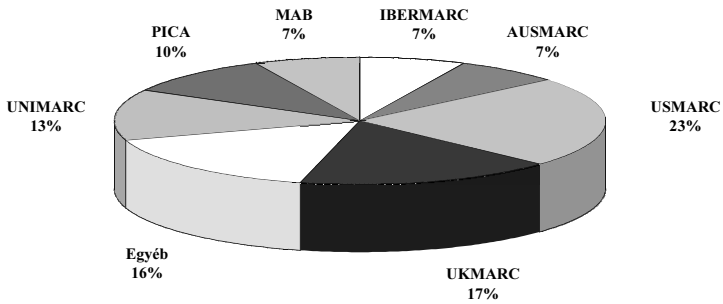
Importformátumként használt adatcsere-formátumok

Azok közül a MARC formátumon alapuló rendszerek közül, melyek rekordokat képesek importálni külső adatbázisokból

– 40% egynél több speciális MARC formátum alapján képes importálni rekordokat,

– 10% csak egyetlen speciális MARC formátumú rekordokat képes importálni rekordokat,
– 50% képes MARC formátumú rekordokat importálni, de nem állnak rendelkezésre adatok arról, hogy pontosan melyik ez a formátum.

Az importformátumként használt adatcsere-formátumok megoszlása



II. RÉSZ
ON-LINE ÉS INTERNET

A TELJES AUTOMATIZÁLÁS FELÉ: AUTOMATIKUS INDEXELÉS, AUTOMATIKUS OSZTÁLYOZÁS

Az automatikus indexelés

Automatikus indexelésen olyan eljárásokat értünk, melyekben a szöveg, referátum vagy a cím szavait gépi segítséggel kivonatolják, extrahálják, akár úgy, ahogy a szavak az eredeti dokumentumban szerepelnek, akár konflálva (szótövek alapján összevetve stb.), akár súlyozva, statisztikai vagy valószínűségi alapon, és az így kivonatolt szavakból mutatót, invertált fájlt készítenek.

A természetes nyelvű szövegnek azonban szintaktikai és szemantikai szerkezete van. Ezekkel a tartalmi, minőségi jellemzőkkel az automatikus indexelés és osztályozás művelői nem foglalkoznak, és a nyelv problémáit kizárólag mint statisztikai–kvantitatív kérdést kezelik. A nyelv szempontjából ezeket az eljárásokat kvantitatív nyelvészeti eljárásoknak nevezik.

Az automatitikus indexelés kezdetei az ötvenes évek közepére nyúlnak vissza. *Hans Peter Luhn* 1958-ban publikálta elképzeléseit a súlyozott kulcsszavas leíráson alapuló automatikus indexelésről (lásd a kötet elején a *Luhnnal* foglalkozó részt).

Ezekhez a korai munkákhoz kapcsolódott *Gerard Salton*, aki 1966-ban készítette el az egyik első automatikus információkereső rendszert (System for the Mechanical Analysis and Retrieval of Text; SMART; egy másik vélemény szerint a helyes feloldás: Saltons Magic Automatic Retrieval Technique, *van Rijsbergen* szerint pedig Salton's Magic Automatic Retriever of Texts) a Harward egyetemen. A SMART talán a legfontosabb alátámasztást adott az automatikus indexelésnek.

1960-ban a matematikai logikus *M. E. Maron* és *J. L. Kuhns* a valószínűségi logika kalkulusát alkalmazták az indexelés elemzésére, és elsőként használták a relevancia fogalmát. Céljuk az volt, hogy a súlyozott kulcsszavas feldolgozáshoz képest megbízhatóbb indexelést dolgozzanak ki. Az eljárást 1974–75 körül többek között *Stephen P. Harter* tökéletesítet-

te a szavak statisztikai eloszlásának figyelembevételével. *Cornelis Van Rijsbergen* a következőképpen foglalja össze a lényeget:

„Úgy találták, hogy a funkcionális szavak minden dokumentumban jól modellezhetők a Poisson-eloszlással, a speciális szavak viszont nem követik ezt az eloszlást. Pontosabban, ha egy *w* funkcionális szó eloszlását vizsgáljuk egy szöveghalmazon, akkor *f(n)* annak a valószínűsége, hogy a szövegben a *w* funkcionális szó *n*-szer fordul elő,

$$f(n) = \frac{e^{-x}x^n}{n!}$$

Általában az *x* paraméter szavanként változik, és adott szóra a szöveg hosszúságával arányos, *x*-re azt is mondhatjuk, hogy egy szöveghalmazban a *w* szó átlagos előfordulási számát jelenti.

A Bookstein–Swanson–Harter modell feltételezi, hogy a speciális szavak »tartalomhordozók«, a funkcionális szavak viszont nem azok. Ez azt jelenti, hogy ha egy szó véletlen eloszlása a Poisson-eloszlásnak felel meg, akkor ez a szó nem informatív arra a dokumentumra vonatkozóan, amelyben előfordul. Ugyanakkor az a tény, hogy egy szó nem követi a Poisson-eloszlást, a feltevések szerint azt jelzi, hogy az információt tartalmaz a dokumentum tárgyára vonatkozóan. Ez nem tekinthető megalapozatlan álláspontnak: ismerve azt, hogy a »háború« speciális szó előfordul a gyűjteményben, elvárható, hogy csak abban a viszonylag kevés dokumentumban forduljon elő, amely valóban a háborúról szól. Másrészt viszont elvárható az is, hogy az olyan tipikusan funkcionális szó, mint az »előtt« vagy az »ilyen« véletlenszerűen forduljon elő.”[1]

Mind a súlyozott kulcsszavas, mind pedig a valószínűségi indexeléskor tezaurusz alkalmazható a kérdések és a dokumentumreprezentációk módosítására. Az ilyen tezauruszok általában egyszerűbb szerkezetűek, mint a manuálisan használt információkereső tezauruszok, többnyire csak az egymással helyettesíthető (szinonim vagy annak tekintett) szavakat kapcsolják össze ekvivalenciaosztályokba. (A manuális használat nyelvén: csak a nem-deszkriptorok és deszkriptorok közötti „Lásd” kapcsolatot és annak inverzét alkalmazzák.)

Nem nehéz belátni, hogy a fenti elv alapján a dokumentumokban szereplő kulcsszavak osztályozása automatizálható, s az osztályok analóg módon használhatók, mint a manuális tezauruszoknál. Pontosabban az automatikus indexe-

1 Van Rijsbergen, C.: Információ-visszakeresés. Budapest: Múzsák Közművelődési Kiadó, 1987. p. 32. A további idézetek is ugyanebből a műből származnak.

léskor a kulcsszó-osztályozás használatának két fő megközelítési módját fedezhetjük föl:

(1) a dokumentum (és kérés) reprezentáció minden egyes kulcsszavát helyettesítjük annak az osztálynak a megnevezésével, amelyben szerepel;
(2) minden egyes kulcsszót helyettesítünk mindazon kulcsszavakkal, amelyek vele egy osztályban szerepelnek.

Azaz a szöveg gépben tárolt specifikus szövegszavait helyettesítik a tezaurusz megfelelő deszkriptorával, vagy minden szövegszót helyettesítenek minden nem-deszkriptorral és a deszkriptorral is. Az utóbbi eset a manuális tezauruszhasználatban nem szokásos.
Az indexelésnek létezik félautomatikus eljárása is, amikor például az intellektuális indexelés eredményeit gépi segítséggel extrapolálják (ahogy ez például a PRECIS kiadvány-indexelő rendszer esetén történik, ahol az indexelők által megadott operátorok alapján a gép jeleníti meg szintaktikailag helyesen a mutatót).

Az automatikus osztályozás

Az automatikus dokumentumosztályozáskor a szöveg szavait arra használják föl, hogy a klaszterelemzéssel (Cluster-analízissel, „klaszterálással") hasonlóságokat állapítsanak meg a dokumentumok között. A hasonlóság megállapításának az alapja pedig az, hogy milyen mértékben fordulnak elő a dokumentumokban vagy azok reprezentációiban (pl. referátumaiban, deszkriptorláncaiban) közös szavak. E mérték, „asszociáció" alapján a dokumentumok csoportosíthatók, és e csoportosítás, klaszterálás az osztályozás eredménye.
Robert A. Fairthorne már az ötvenes években utalt arra, hogy az automatikus osztályozás hasznosnak bizonyulhat a dokumentumok keresésében. A matematikai ***osztályozási*** eljárás vonzerejét nem(csak) az adatok növekvő mennyisége adta, hanem az a törekvés is, hogy az intellektuális osztályozáshoz képest objektívebb eredményhez jussanak. Emellett még az a remény is hatott, hogy matematikai úton az információk, dokumentumok új, intellektuálisan nem könnyen felismerhető, innovatív csoportosítási összefüggései állapíthatók meg. A kötetünk elején bemutatott *Fairthorne* az ún. extenzionális, szigorúan a halmazelméleti megalapozású automatikus osztályozás egyik előfutára volt. Idővel a fogalmak tartalmi, ún. intenzionális elemzésének matematikai módszereivel is foglalkozni kezdtek; ezeknek a ritka kutatásoknak egyik figyelemreméltó képviselője az orosz *Jurij A. Šrejder*. (Az extenzionális és intenzionális megközelítés fogalmára alább, az automatikus és intellektuális osztályozás összehasonlításakor még visszatérünk.)

A dokumentumklaszterálásban (ezt a szót az automatikus dokumentumosztályozás szinonimájaként használhatjuk) a legkorábbi munkák a SMART kísérlet keretében folytak; a SMART rendszerben mind az automatikus indexelés, mind az automatikus osztályozás végezhető. (Az automatikus osztályozással a kötetünkben található szemelvényeken kívül magyar nyelven részletesen foglalkozik Horváth Tibor és könyvében *van Rijsbergen*.)[2]
A klaszterelemzés mindenekelőtt dokumentumok csoportjainak (füzéreinek, nyalábjainak, klasztereinek–osztályainak) a meghatározására irányul. E csoportok kialakításának alapja a dokumentumok közötti hasonlóság. Feltesszük, hogy egy csoporton belül egy dolog jobban hasonlít a csoport többi tagjához, mint bármelyik csoporton kívüli dokumentumhoz. Magyarán: a hasonlóság annál nagyobb, mennél nagyobb a közös jellemzők (a közös szövegszavak, vagy közös deszkriptorok, összefoglalóan kulcsszavak, indexkifejezések) száma. Ahhoz azonban, hogy az eredmény ne legyen megtévesztő, azaz a dokumentumok a hasonlósághoz ne a méretükkel arányosan járuljanak hozzá, ügyelni kell arra, hogy minden egyes dokumentumhoz ugyanannyi kulcsszó tartozzék. Ezt nevezik normalizálásnak. Ez a fajta automatikus osztályozás nem-determinisztikus eljárás, azaz még akkor is valószínűségi eljárásnak tekinthető, ha a statisztikai ingadozásokkal eleve nem törődünk. Mint ilyen az ún. nem-dimenzionális matematikai módszerekhez sorolható, amilyen a strukturális elemzés. A klaszterálás eredményeként ezért beszélnek „taxonómiai struktúrákról" is. A klaszterelemzés a matematika sokváltozós (multivariáns) eljárásai közé tartozik, mint amilyen a faktor- és diszkriminancia-analízis, de ezekkel ellentétben nem érzékeny az adatok minőségére. Vele alapvetően nem hipotéziseket igazolnak, hanem adathalmazok szerkezetét írják le.
Az automatikus osztályozásnak az információkeresésben két fő alkalmazási területe van:

(1) Dokumentumok klaszterálása

Ilyenkor hasonló dokumentumokat állapítanak meg a dokumentumokban előforduló kulcsszavak előfordulásainak elemzése alapján, és ezeket klaszterekbe vonják össze, azaz a dokumentumokat a bennük előforduló kulcsszavak előfordulásának gyakorisága alapján dokumentumosztályokba vonják össze.

Minden dokumentum leírható dokumentumvektor formájában, azaz

$$D_i = (g_{ij})_{j=1}^{N} \quad \text{és} \quad D = (g_{ij})_{pxN}$$

2 Horváth Tibor: Automatikus osztályozás. In: Könyvtári figyelő, 1978, 24. évf., 5. sz., p. 528–542.
Van Rijsbergen, C.: Információ-visszakeresés, p. 41–69.

A (g_{ij}) (i = 1, ..., p; j = 1, ..., N) értékek a D_i (i = 1, ..., p; j = 1, ..., N) dokumentumokban szereplő T_{ij} (j = 1, ..., N) kulcsszavak valamilyen statisztikailag súlyozott előfordulásai lehetnek. Egy s korrelációs vagy hasonlósági mérték alapján mérhető két ilyen dokumentumvektor hasonlósága. Az

$$s_{ij} = s(D_i, D_j)$$

hasonlósági koefficiensek az alábbi dokumentum–dokumentum mátrixba írhatók be:

$$D^* = (s_{ij})_{p \times p},$$

Egyszerű nyelven szólva a dokumentumok annak alapján lesznek hasonlóak, és kerülnek e hasonlóság alapján egy osztályba, hogy bennük előre megadott értéknél többször fordulnak elő közös kulcsszavak.

(2) Kulcsszavak klaszterálása

Ilyenkor összetartozó kulcsszavakat állapítanak meg a dokumentumban előforduló kulcsszavak együttes előfordulásainak elemzése alapján, és ezeket klaszterekbe vonják össze, azaz a kulcsszavakat a dokumentumban való előfordulásuk gyakorisága alapján kulcsszóosztályokba vonják össze.

A kulcsszóosztályozás alapja az előbbi D mátrix transzponált T mátrixa:

$$T = D' \; (g_{ji})_{n \times p} \quad \text{és} \quad T_i = (g_{ji})_i \, \underline{p}_{=1}$$

Az

$$s = s(T_i, T_j)$$

hasonlósági koefficiensek az alábbi kulcsszó-hasonlósági mátrixba írhatók be:

$$T^* = (s_{ij})_{N \times N}$$

Egyszerű nyelven szólva a kulcsszavak annak alapján kerülnek egymással kapcsolatba (függnek össze „hasonlóság” alapján), hogy bennük a dokumentumokban előre megadott értéknél többször fordulnak elő együttesen.

(3) Egyidejű kulcsszó- és dokumentumklaszterálás

Ilyenkor mind a kulcsszavak, mind pedig a dokumentumok csoportjait (klasztereit) kialakítják, azaz az osztályozás mind a tárgyi, mind pedig az ismérvtérben lejátszódik. Ennek egyik propagálója *Jiri Panyr*.

Salton elsősorban dokumentumok klaszterálásával foglalkozott. Az eljárás indoklásaként *Salton* a hatékonyságra hivatkozik: ha az információkeresést automatikusan akarják elvégezni, nem lehet minden elemzett dokumentumot minden keresőkérdéssel összevetni, mivel ennek túl nagy az időigénye. Olyan megoldást kellett keresni, mellyel az összehasonlító műveletek száma lényegesen csökkenthető. Erre való a dokumentumok hasonlóság alapján végzett előzetes automatikus csoportosítása. A rokon dokumentumok egyes csoportjait egy-egy jellemző ismérvvektor reprezentálja. Ennek alapján elég, ha a keresőkérdést először csak ezzel a vektorral hasonlítják össze.

A dokumentumosztályozás homlokterébe tehát az információkeresés hatékonysága áll. Ezzel szemben a kulcsszóosztályozás révén az információkereső rendszer teljesítményét, használhatóságát kívánják növelni. A kulcsszóosztályozás egyik legismertebb kutatója a következő szemelvényben bemutatott *Karen Sparck Jones.*

A dokumentumosztályozás legalábbis elméletileg teljesítette a hozzáfűzött várakozásokat. A kulcsszóosztályozás alkalmazását sokan vitatják. Amikor ugyanis a keresést az automatikusan klaszterált kulcsszóosztályokkal is kibővítik, ez gyakran teljesen váratlan és nehezen értelmezhető eredményekkel jár. Ha például olyan dokumentumokat osztályoznak automatikusan, melyekben higanymérgezett halakkal foglalkoznak, akkor a kulcsszóosztályozás eredményeként az derül ki, hogy a Hal és a Higany között szignifikáns szemantikai összefüggés áll fenn, holott ez az összefüggés nem szemantikai, kontextus független (analitikus), hanem esetleges. Ilyen alapon a Hal kulcsszó elvileg minden elképzelhető elem kulcsszavával szemantikai összefüggésben állhatna, hiszen bármelyik elem feldúsulása a tengervízben okozhatja a halak mérgezését. E példából is látható az automatikus eljárások korlátozott érvényessége.

Automatikus és intellektuális osztályozás

Az automatikus – vagy más néven numerikus – osztályozással kapcsolatban meg kell említeni, hogy az információkeresésen kívül még rendkívül sok, egymástól távol eső szakterületen alkalmazzák, mint a biológia, a szociológia, az alakfelismerés. Az első lépéseket ezen a téren 1957–62 között a biológiai taxonómia területén tették, és 1963-ban jelent meg az automatikus osztályozási eljárások máig ható első, 1971-ben pedig a második legfontosabb kézikönyve.[3]

3 Sneath, R. H., Sokal, R. R.: Numerical taxonomy. The principles and practice of numerical classification. – San Francisco: W. H. Freeman Co., 1973.
Jardine, N., Sibson, R.: Numerical taxonomy. – New York: J. Wiley & Sons, 1971.

A matematikai eljárásokon alapuló osztályozásnak 1985-ben megalakult a nemzetközi szervezete is (International Federation of Classification Societies; IFCS). Mivel mind a tartalmi–fogalmi alapon dolgozó „intellektuális" osztályozók, mind pedig a matematikai–statisztikai eljárások művelői egyaránt csak „osztályozásról" beszélnek, messzebbről nézve a két szakmai kör összemosódik – noha tagjainak szemlélete között ég és föld a különbség (ami például azzal járt, hogy az első kötetben tárgyalt Gesellschaft für Klassifikationból az intellektuális osztályozás művelőinek egy része 1991-ben kivált, és az Ismeretszervezés Nemzetközi Társasága [International Society of Knowledge Organization, ISKO] néven önállósult.)

Mind *Salton*, mind pedig *Sparck Jones* alább következő szemelvényében jól felismerhető, mennyire csak a numerikus, matematikai módszerek szempontjából tárgyalják az osztályozást; az a benyomása az olvasónak, mintha az intellektuális módszerek legfeljebb a matematikai módszerek valamiféle elhanyagolható részhalmazát alkotnák. (Az igazság az, hogy ez fordítva is így van: az intellektuális, „fogalmi" osztályozás művelői sem szoktak gyakran utalni arra, hogy az osztályozás felfogható úgy is, mint matematikailag algoritmizálható, numerikus folyamat, illetve rendszer.) Ez időnként a süketek párbeszédére emlékeztet, amit jól példáz az Osztályozási Társaság (Gesellschaft für Klassifikation) 1978-ban rendezett konferenciáján a *Salton* előadását követő, szemelvényben kötetünkben közölt vita, továbbá az, ahogy *Sparck Jones* az általa felállított – és elsősorban az automatikus osztályozás szempontjából releváns – osztályozási-rendszer-tipológáról azt állítja, hogy teljes körű.

A félreértések egyik magyarázata abban rejlik, hogy az automatikus osztályozás művelői szerint az osztályozás objektumok osztályozása, melyek tulajdonságokkal rendelkeznek, ezért ismertetőjegyekkel, az őket reprezentáló deszkriptorokkal leírhatók. Az így kapott leírásokat dolgozzák föl matematikailag (például a klaszterelemzéssel), és ennek eredményeként keletkezik az osztályozási rendszer, azaz a konkréttól az elvont irányában, „alulról fölfelé" indulnak el. A logika nyelvén azt mondhatjuk, hogy ez az osztályozás extenzionális, a fogalmak terjedelmén alapszik.

Az intellektuális, mondhatni intuitív osztályozás művelői ezzel szemben fogalmak közötti összefüggések alapján alakítják ki osztályozási rendszereiket, az elvonttól a konkrétabb felé, „felülről lefelé" indulnak el, és az objektumokat sorolják be az így kapott osztályokba (természetesen egyfajta folytonos gondolati visszacsatolással figyelembe véve az objektumok tulajdonságait). A logika nyelvén azt mondhatjuk, hogy ez az osztályozás intenzionális, a fogalmak tartalmán alapul.

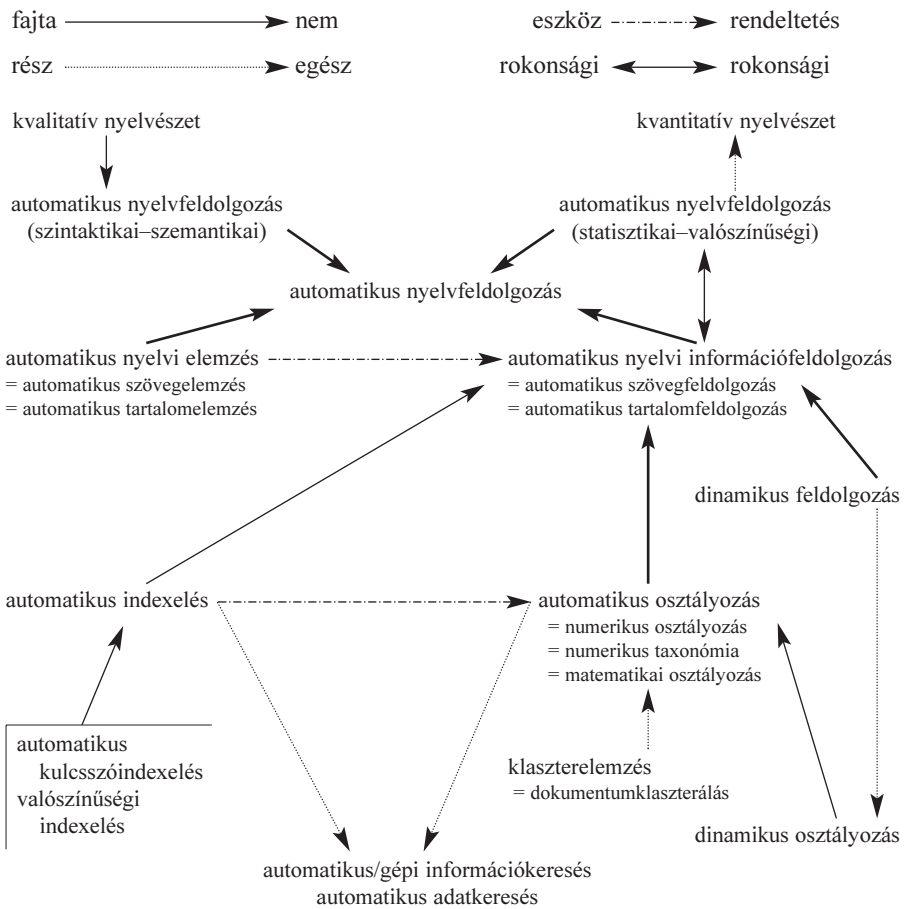
A félreértések másik magyarázata bizonyos mértékben a fentiekből következik, de inkább lélektani jellegű. Az extenzionális alapú automatikus osztályozás művelői úgy járnak el, hogy teljesen kikapcsolják a fogalmi

összefüggéseket a kiindulásaikból, sőt egyenesen hibának tekintik a fogalmi összefüggések előzetes, intuitív figyelembevételét, és kizárólag az objektumok tulajdonságaira támaszkodnak.
Az intenzionális, intuitív osztályozás művelői ezzel szemben nem képesek figyelmen kívül hagyni az objektumok tulajdonságait, amikor osztályoznak, sőt az osztályozási rendszereik szerkesztésekor is támaszkodnak az ilyen ismereteikre. A fogalmi összefüggések (pl. az olyan relációk, mint Asztal–Bútor, Fiók–Asztal, Ég–Kék, Kutya–Ugatás, Fizika–Természettudomány) nem írhatók le numerikusan, nem algoritmizálhatók teljes körűen a formális logika eszközeivel, és ezért az intenzionális osztályozás művelői még ha akarnák is, nem alkalmazhatnak teljes körűen matematikai módszereket, noha mindazt, amiből az extenzionális osztályozás művelői kiindulnak, figyelembe veszik, de csak intuitíven.
A két felfogás közötti eltérés bizonyára örökké megmarad, amiből azonban nem következik, hogy csak az egyik, vagy csak a másik a helyes. Mindkét eljárásra mindig szükség lesz.

Az automatikus információkeresés (indexelés és osztályozás) folyamatai

Az alábbiakban összefoglaljuk a különféle automatikus eljárásokat és azok részfolyamatait:
Az *automatikus (vagy gépi) információkeresést* egyes szerzők a másodlagos (szakirodalmi) információkra korlátozzák, megkülönböztetésül az elsődleges információkra vonatkozó *automatikus adatkereséstől.*
Az automatikus információkeresésnek két területe van: az *automatikus indexelés* (amikor a szövegből kiválasztott kulcsszavak alkotják az ismérvet [az indexkifejezést vagy keresőkifejezést]), és *automatikus osztályozás* (amikor a dokumentum szövegében szereplő szavak gyakorisági elemzése alapján a dokumentumokat hasonlósági osztályokba sorolják).
Az *automatikus nyelvfeldolgozás* egyik fajtája a *nyelvi elemzés* (szövegelemzés, tartalomelemzés), a másik fajtája a *nyelvi információfeldolgozás* (szövegfeldolgozás, tartalomfeldolgozás).
Az automatikus indexelés legegyszerűbb formája a szókiválasztás (amikor a szöveg szavaiból többnyire konflációval ismérvül szolgáló kulcsszavakat emelnek ki), ilyenkor *kulcsszóindexelés*ről beszélnek. Alkalmazhatnak statisztikai elemzéseket (például a szógyakoriságot is vizsgálják a kulcsszavak súlyozott megállapítása érdekében). A legösszetettebb eljárások egyike, amikor valószínűség számítási módszereket is felhasználnak, ilyenkor beszélünk *valószínűségi indexelés*ről.
Ha az elemzés/feldolgozás a szintaktikai/szemantikai összefüggésekre is kitér, *kvalitatív nyelvészet*ről beszélnek. Ha csak statisztikai és valószínűségi

az elemzés/feldolgozás, *kvantitatív nyelvészet*ről beszélnek. Az automatikus indexelési eljárások túlnyomó többsége, az automatikus osztályozási eljárások mindegyike kvantitatív jellegű. (A gépi fordítás, és az automatikus strukturális nyelvi elemzés szükségszerűen mindig kvalitatív.)
Az *automatikus osztályozás*kor a többnyire az automatikus indexelés eredményeként kiemelt kulcsszavakat *klaszterelemzés*nek vetik alá, és segítségével dokumentumok hasonlóságának mértékét állapítják meg. E hasonlóság alapján a dokumentumok halmazai osztályokat alkotnak. Ha az automatikus osztályozási rendszer viszonylag gyorsan képes átrendeződni az újabb dokumentumok felvételekor és feldolgozásakor új klaszterált állapotba, *dinamikus feldolgozás*ról, illetve *dinamikus osztályozás*ról beszélnek.
Az összefüggések az alábbi címkézett, irányított gráfon láthatók:



(A gráf egyszerűbb formában Bar-Hillel egyik szemelvényében is tanulmányozható.)

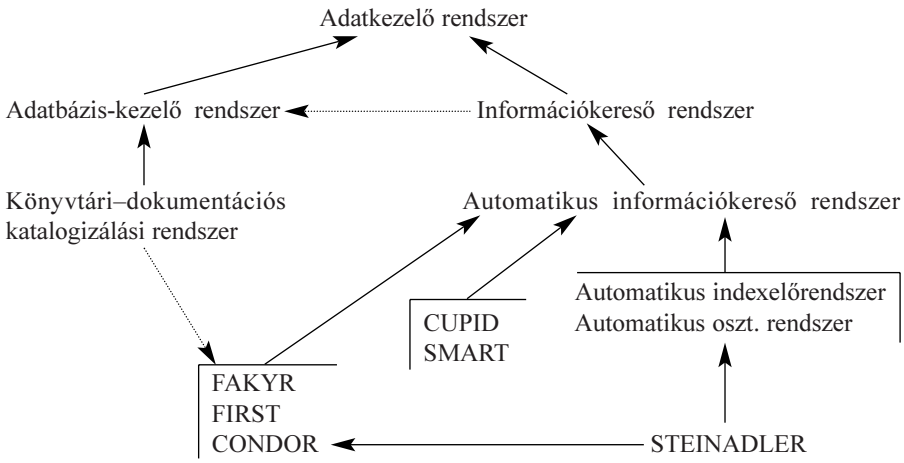
Automatikus információkereső rendszerek és az adatbázis-kezelő rendszerek

A könyvtári–dokumentációs katalogizálási és dokumentum-nyilvántartási rendszerek az adatbázis-kezelő rendszerek (ABKR; Data Management System; DBMS) osztályába tartoznak.
Az első adatbázis-kezelő rendszerek hierarchikus szerkezetűek voltak. Az ilyen szerkezetű rendszerek korlátainak felismerése hamarosan a hálós adatszerkezeteken alapuló adatbázis-kezelő rendszerekre irányította a figyelmet. Az adatrendszer-nyelvekről 1969-ben tartott konferencia (Conference on Data System Languages; CODASYL) e hálós rendszerek propagálásának jegyében szerveződött. A konferencia javaslatára az adatbázis munkacsoport (Data Base Tast Group; DBTG) a hálós adatbázis-kezelő rendszerek szerkezeti kialakítására vonatkozó nemzetközi ajánlások kidolgozásával kezdett foglalkozni. Az elkövetkező évtizedben sorra születtek meg ajánlásai, melyeket CODASYL ajánlásoknak neveztek. A hetvenes évek közepén megjelentek az első publikációk a hálósnál még rugalmasabb, attól teljesen eltérő logikai szerkezettel rendelkező relációs adatbázis-kezelő rendszerekről, és a nyolcvanas évek közepétől már csak ilyen szerkezet alapján működő rendszereket terveztek. Ma a kereskedelmi forgalomban jelen vannak a korai, hierarchikus és hálós, valamint a relációs adatbázis-kezelő rendszerek. Az időszerűségüket lassan elveszítő CODASYL ajánlások a hálós rendszerekre vonatkoznak.[4]

A könyvtári–katalogizálási rendszerek jelentős csoportja (pl. a DOBIS/LIBIS, ALPEHP, BRS SEARCH, ISIS) tartozik a már elavultnak számító hierarchikus és hálós adatbázis-kezelő rendszerek csoportjába. A korszerűbb termékek (pl. ORACLE, TYNLIB) már relációs szerkezetűek. Mint adatbázis-kezelő rendszerek, jellemző rájuk az adatelem–rekord–fájl szerkezet és a mezők közötti, rekordok közötti és fájlok közötti kapcsolat, valamint az adatelemek és rekordok azonosított kezelése (az adatelemet például az adatelemnév/mezőnév, a rekordot a rekordazonosító adatelem azonosítja). Minden ilyen rendszer tartalmaz információkereső komponenst, melyre az jellemző, hogy a keresés alapja maga a karakterlánc (például a szövegszó, a jelzet vagy a deszkriptor), nem pedig valamilyen azonosító, melyek mindig csak egyetlen tételt határoznak meg. Az információkereső elemeket, mint például a deszkriptorokat, mivel általában egyszerre több tételt is jellemeznek, ezért nevezik másodlagos azonosítóknak is. Jellem-

4 Részletesebben lásd Halassy Béla: Adatbázisok kezelésének alapvető problémái. – Budapest, KSH SZÁMOK, 1978, p. 165 és Halassy Béla: Az adatbázis-tervezés alapjai és titkai. Budapest: IDB, 1994, p. 61–62.

ző ezekre a rendszerekre még az invertált fájlok és a Boole-algebrai műveletek kitüntetett szerepe is.[5]
A „tiszta" automatikus indexelő- és osztályozórendszerek (összefoglalóan automatikus információkereső rendszerek, pl. SMART) alapvetően tehát nem adatbázis-kezelő, hanem csak adatkezelő rendszerek (ahogy az adatbázis-kezelő rendszerek is azok). Előfordul – és feltehetően ez a „jövő zenéje" – hogy az automatikus információkereső rendszert összekapcsolják a könyvtári–dokumentációs katalogizáló (tehát adatbázis-kezelő rendszerrel); ennek nyomai fedezhetők fel például a FAKYR, a FIRST és a CONDOR rendszerben. (Lásd még a szerkesztői kommentárt *Jiri Panyr* szemelvényében a CODASYL ajánlásokkal összefüggésben.) A mondottakat az alábbi ábrán foglaltuk össze:



A címkézett irányított gráf jelölései:

fajta ——▶ nem　　rész ┈┈▶ egész

A *Jiri Panyr* által ismertetett FAKYR, FIRST és CONDOR automatikus indexelésre és osztályozásra alkalmas információkereső rendszereknek van némi adatbázis-kezelő képessége (összetevője) is, a CUPID és a SMART tisztán automatikus információkereső rendszerek. A STEINADLER automatikus osztályozási rendszer a CONDOR része. Az automatikus információkereső és az adatbázis-kezelő rendszerek az adatkezelő rendszerek fajtái.

5 Egyes szerzők – például a kötetünkben szereplő *Richard J. Hartley,* vagy a magyar nyelvű szakirodalomban *Halassy Béla* – szöveges információkereső rendszernek nevezik azokat az adatbázis-kezelő rendszereket, melyek kitüntetett jellemzője az invertált fájlok és a Boole-algebrai műveletek használata.

Az alább következő első szemelvény a SMART készítőjétől, *Gerard Saltontól* származik. Kiegészítésképpen olvashatjuk annak a vitának a részletét is, melyet az intellektuális osztályozók híveivel folytatott az egyik konferencián.
A következő második szemelvény szerzője *Karen Sparck Jones,* aki a kulcsszavak klaszterálásának vizsgálatával vált híressé.
A harmadik szemelvényben a SIEMENS rendszerfejlesztője, *Jiri Panyr* világít rá az automatikus eljárások mai helyzetére és ismerteti a fenti gráfon látható kísérleti rendszereket.

M. E. MARON

A matematikai logikával foglalkozó *M. E. Maron* és *J. L. Kuhns a* dokumentumon belüli szóeloszlás *Hans Peter Luhn* által elkezdett statisztikai elemzését fejlesztették tovább és a relevancia valószínűségi modelljének egyik megalapozói voltak, abban a reményben, hogy ezzel a stochasztikus módszerrel a szubjektív, intuitív relevanciát matematikailag megalapozottabb relevanciafogalommal válthatják föl. Az e modell alapján működő információkereső rendszereket képesnek tartják arra, hogy a talált dokumentumokat a felhasználó pozitív relevancia megállapításai alapján rendezzék. A valószínűségi eljárásban a keresés bővítését és élesítését a kulcsszavak statisztikai együttes előfordulásának a vizsgálata alapján oldják meg. E modellből fejlődött ki a hetvenes évek közepén a valószínűségi indexelés.

Az információkeresés logikai analízise

In: Varga Dénes: A dokumentáció nyelvészeti kérdései I. (Szemelvénygyűjtemény) – Budapest: Országos Műszaki Könyvtár és Dokumentációs Központ, 1966. – (A tudományos tájékoztatás elmélete és gyakorlata; 10 sz.) p. 61–67.
Eredeti: A logician's view of language-data processing. In: Natural Language and the Computer, McGraw-Hill, 1963. p. 144–150.

A javasolt valószínűségi indexelés a logika mellett az aritmetikát veszi igénybe, amennyiben a számítógéppel kiszámítják a dokumentum várható relevanciájának a mértékét valamely információkéréssel összefüggésben. A gépi indexkészítés hagyományos formájának ismertetése után a súlyozott jelzetek elvét írja le, majd kifejti a relevancia valószínűségi fogalmát és kvantitatív mérését. Az információkereséskor e modell alapján az olvasók kéréseit tekintik irány-

mutatónak. Ennek az iránymutatásnak és a statisztikai elemzésnek az alapján a rendszer heurisztikus eljárások segítségével állíthatja össze az igényeket kielégítő jegyzéket, mely a kérésre vonatkoztatott valószínűségi jegyzék szerint rangsorolja a megfelelő dokumentumtételeket.

GERARD SALTON (1927–1995)

A németországi születésű Gerard Salton az egyik legismertebb kísérleti jellegű automatikus információkereső rendszer, a SMART (System for the Mechanical Analysis and Retrieval of Text) kialakítója. Mióta 1966-ban rendszerével színre lépett, keresési kísérletek tömegét végezte el, egyre finomítva mind az automatikus indexelés, mint pedig az osztályozás folyamatát. Indexelési felfogása tömör és gyakorlati: „...általában a legegyszerűbb indexelési eljárás, amely az adott dokumentumot vagy kérdést a dokumentum vagy a kérdés szövegéből kiválasztott, súlyozott vagy súlyozatlan kulcsszóhalmazzal azonosítja, egyúttal a legjobb hatásfokú." A súlyozott kulcsszavakat pedig olyan dokmentumkivonatokból kell választani, melyek hossza legalább egy referátumének megfelelő.
Rendszerének automatikus osztályozó komponense mintegy kiegészíti az indexelőrészt, ezzel is bizonyítva, hogy az automatikus osztályozás nem önmagában álló és felhasználható valami a dokumentumfeldolgozás és keresés világában, hanem olyan kereső eszköz, mely a többivel kölcsönhatásban alkalmazható optimálisan.
Salton „dinamikusnak" nevezi rendszerét. Ennek megértéséhez a következőket kell tudni: Ha egy dokumentumállományon a klaszterelemzést elvégzik, a dokumentumok osztályozási struktúrába rendeződnek. Ha aztán új dokumentumok kerülnek a rendszerbe, ezeket is be kell illeszteni az osztályozási szerkezetbe. A „nem dinamikusan" működő klaszterálási rendszereknek ilyen esetben az egész osztályozási szerkezetet újra kell strukturálniuk. Azokat az automatikus osztályozási eljárásokat, melyek a bővítést, felújítást (updating) viszonylag gyorsan végre tudják hajtani, azaz megfelelő sebességgel képesek a környezet változásaira reagálni, „dinamikusnak" nevezik. Ez az igény triviálisnak tűnik; a valóság az, hogy a legtöbb eljárás nem elégíti ki a dinamikusság követelményét.
A SMART 1961–64 között készült a Harvard Egyetemen; 1968-ig IBM 7094-es, utána pedig IBM 360/65-ös gépen működött. Később a Cornell egyetemre került át a rendszer.
Salton a számítógép-tudomány professzora volt a Cornell Egyetemen (Department of Computer Science, Cornell University) és az ACM Communications és az ACM Journal szerkesztője.

Dinamikus információ- és könyvtári feldolgozás[6]

1–5 A dinamikus könyvtár alapelvei[7]

A könyvtári adminisztrációs – feldolgozási – műveletek bonyolultabbak, mint azt általában feltételezik, és számos olyan problémát vetnek fel, amelyeket az egyes könyvtárak önállóan nem egykönnyen oldanak meg. Ennek ellenére nem valószínű, hogy a közeljövőben kialakul olyan hatékony, centralizált hálózati szervezet, amely tevékenységét teljes mértékben helyettesítené.

Felmerül azonban a kérdés, vajon az önálló könyvtári szervezet valójában tud-e valami értelmeset tenni a jelenlegi kényelmetlen helyzet könnyítéséért. Nem egyszerű erre válaszolni. A korábbiakat figyelembe véve azonban valószínűleg jó stratégia, ha erőforrásainkat nem vesztegetjük olyan műveletekre, amelyek szabványosítás vagy közös, illetve centralizált kivitelezés előtt állnak. Ez különösen érvényes a könyvtári világra, ahol a főbb könyvtári feldolgozási műveleteknek ilyen ellenszenves adatfeldolgozási jellemzőik vannak.

Ehelyett a könyvtáraknak azoknak az eljárásoknak a korszerűsítésére kellene erőiket összpontosítani, amelyek felhasználói környezetenként változnak, így különösen a könyvek és a dokumentumok tartalmának azonosításához használt szótárakra, a tartalomazonosítás tényleges módszerére, a tárolt információs fájlok szervezésére, az információkereséshez és -visszanyeréshez használt módszerekre, és az állomány növekedését és apasztását kezelő eljárásokra. E műveletek végrehajtására *alkalmas* módszertan jelenleg vagy nem létezik – ez a helyzet például az állományapasztással – vagy (a legjobb esetben) tökéletlen, és egyáltalán nincs összhangban a különböző felhasználói csoportok igényeivel.

Semmi sem indokolja azonban annak a statikus környezetnek a fenntartását a jövő könyvtárában, amely ma még a feldolgozás legtöbb lépését jellemzi. Számtalan nyilvánvaló előny származik ugyanis olyan dinamikus környezet megtervezéséből, amelyben a felhasználók közvetlenül befolyásolhatják az intellektuális feldolgozási sorrendeket. Először, a könyvtári műveletek hatékonysága növelhető, ha bizonyos kritikus pontokon a felhasználók gondoskodnak a megfelelő inputokról; másodszor, sok könyvtári folyamatot félig automatikusan – azaz a számítógépes műveleteket felhasználói inputokkal kiegészítve – a legjobb elvégezni, és nem teljesen kézi, illetve automatikus eljárásokkal.

---

6 Dynamic information and library processing / Gerard Salton. – New Jersey; Englewood Cliffs, N. J.: Prentice Hall, 1975. 523 p.

7 1–5 The dynamic library – basic principles. In: Dynamic information and library processing. p. 14–16.

Az általunk javasolt dinamikus környezet így három fő elven alapul:

(a) A „totális" (teljes) könyvtári rendszeren, amelyben alapinput indítja el az egymást követő feldolgozási lépések láncolatát, miközben – a feldolgozási lánc különböző lépéseinek következtében – ez az alapinput folyamatosan módosul.
(b) Az együttműködéssel vagy megosztottan végrehajtott műveletek lehető legszélesebb körű igénybevételén, beleértve az állománygyarapítási együttműködést és centralizációt, az osztott, „közös" katalogizálást és a szabványosított könyvtári feldolgozási műveleteket.
(c) Adaptív környezeten, amelyben a felhasználói kör befolyásolja a főbb intellektuális folyamatokat, például az indexelőszótárt és az indexelési gyakorlatot, a tárolás szerkezetét, a keresési és visszanyerési műveleteket, végül a gyűjtemény növeléséhez és apasztásához szükséges állomány-ellenőrzést.

Mivel az alapvető input-műveletek és a feldolgozási műveletek gépesítése nem indokolt, ha az egyes szervezetek külön-külön, egymástól függetlenül végzik ezeket a műveleteket, a gépesítésnek olyan közös vállalkozásnak kell lennie, amely a lehető legnagyobb mértékben támaszkodik a szabványosításra és a munkamegosztásra. Ideális esetben ezt a munkát on-line, időazonos eljárásokat felhasználva lehetne elvégezni, bár a MARC formátumú adatcsereszalagok szétosztásából és az ezzel kapcsolatos adatcsere-szolgáltatásokból következő központosított off-line feldolgozástípus is nagyon biztató. Ma már egyre kevesebb szervezet tud meggyőzően érvelni amellett, hogy az ő körülményeik annyira speciálisak és felhasználói körük annyira szokatlan, hogy nekik nem felelnek meg a szabványosított feldolgozási eljárások. Ezek után remélhető, hogy a könyvtári feldolgozás területén a fejlesztések a nagyobb mérvű szabványosítás és együttműködés irányába haladnak majd, mivel ezektől várható a költség–haszon arány végső javulása.

Másrészt, bár az ügyviteli input műveleteket szabványosítani kell, magának a gyűjteménynek a kezelését mindegyik könyvtár a saját környezetének megfelelően kell, hogy megoldja. Nem kell mindenkinek feltétlenül azonos fájlelrendezést fenntartania, sem ugyanazt az indexelőszótárat[8] használnia, a felhasználói kör tájékozódását figyelmen kívül hagyva. Az is bizonyos, hogy amikor eljön az ideje a kevéssé használt anyagok kivonásának és kisegítő raktárba költöztetésének, mindegyik szervezetnek a saját tapasztalata alapján kell eldöntenie, hogy a gyűjtemény mely része távolítható el a leginkább.

8 Indexelőszótáron információkereső nyelv (információleíró nyelv, dokumentációs nyelv) szótára – tárgyszójegyzék, tezaurusz, de akár osztályozási rendszer is – értendő. Az információkereső nyelv terminológiai kérdésével az első kötetben *Gernot Wersig* szemelvényével összefüggésben foglalkozunk (a szerk.).

A következőkben bemutatandó dinamikus rendszer lehetővé teszi minden könyvtár számára, hogy közvetlenül reagáljon a felhasználói körére, és elkerülje azokat a szokványos kudarcokat, amelyek a statikus indexelési és osztályozási rendszerek használatából, a specializált, szabványosított indexelőnyelvek megtanulásának szükségességéből, és az információkeresés során a felhasználó–rendszer párbeszéd hiányából fakadnak.

Az alapötlet az, hogy a könyvtári rendszert az állandó változás állapotában kell tartani. A dokumentumokat először – remélhetőleg automatikusan – ideiglenesen indexelik úgy, hogy mindegyik egységhez a dokumentum tartalmát reprezentáló, súlyozott kulcsszavak (ismérvek) csoportját rendelik hozzá. Ezeket a kulcsszó- (ismérv) vektorokat azután az egységek automatikus osztályozására, rokonsági csoportokba sorolására használják oly módon, hogy a hasonló vektorú egységek egy osztályba kerülnek. A feldolgozás során mind a kérdés-, mind a dokumentumvektorok kisebb változásokon mennek keresztül, például bizonyos kulcsszavak súlyának csökkentésével vagy növelésével, új kulcsszavak hozzáadásával vagy régiek elvételével. Ahogy az egyes kulcsszó- (ismérv) vektorok változnak, úgy változik a dokumentumok osztályozása is. Végeredményként egy olyan fájlt kapunk, amelyben az „érdekes" egységek a fájl közepére kerülnek, a nem kívánt egységek pedig a perifériára.

A fájlszervezést és a főbb feldolgozási lépéseket ennek a fejezetnek a további részében írjuk le.[9]

[...]

1–7 Dinamikus keresés és találatképzés[10]

A dinamikus könyvtár alapgondolata, hogy a szolgáltatás javításában és a fájlszervezés karbantartásában, felfrissítésében a felhasználó tapasztalataira kell támaszkodni. *A rendszer kialakításának sarkalatos mozzanata, hogy a felhasználói körtől érkező kéréseket a lehető legpontosabban fogalmazzák meg, és ekként építsék be a rendszerbe a felhasználói inputokat.* A SMART rendszerben a kérés módosításának módszerét relevancia-visszacsatolásnak

9 Nagyon fontos, hogy megértsük: a dinamikus műveletek a gépesített katalógusfájlok szervezésére és feldolgozására, valamint a tárolt tételek eléréséhez használt szoftver segédletekre alkalmazhatók. A könyvek fizikai elrendezését a polcokon nem érinti, kivéve persze az új tételek elhelyezését és a kivont tételek távolabbi raktárba való elszállítását.

10 1–7 Dynamic search and retrieval. In: Dynamic information and library processing. p. 22–25.

Az eredetiben a „search" és a „retrieval" kifejezések szerepelnek, amelyek a keresést, illetve az eredményes keresést, azaz a találatképzést (használják a „visszakeresés" kifejezést is) jelentik. Ezt próbáltuk a fenti módon tükrözni (a szerk.).

nevezik, mivel a kéréseket a felhasználótól származó, a korábban talált dokumentumok értékeléséből fakadó relevancia információ alapján automatikusan aktualizálják. A relevancia-visszacsatolási eljárás feltételezi, hogy a rendszerhez intézett valamennyi kérésre előzetes (próba) keresést végeznek. Az output egy kis része, a legmagasabb értékű dokumentumok közül néhány kerül azután a felhasználóhoz, akit arra kérnek, állapítsa meg relevánsak-e (R) ezek a dokumentumok az ő információs igénye szempontjából vagy sem (S). Ezek az *ítéletek* kerülnek vissza a rendszerbe, amelyeket az úgy használ fel a keresőkérdések automatikus módosítására, hogy a releváns dokumentumokban jelenlevő, a kérdés kulcsszavainak megfelelő kulcsszavaknak nagyobb hangsúlyt ad (növeli a súlyukat), míg az irreleváns dokumentumokban előforduló kulcsszavakat lefokozza.

A korábbi azonosításkor relevánsnak (R), illetve irrelevánsnak (S) minősített dokumentumok olyan új kérdés (q') megfogalmazását szolgálják, amely várhatóan hasonlóbb lesz a releváns, és kevésbé hasonló a nem releváns dokumentumokhoz, mint az eredeti kérdés (q). Ha a releváns egységekből származó kulcsszavakat a keresőkérdésekhez adjuk (az irrelevánsakból származókat pedig elvesszük), a kérdés aktualizálása, finomítása, amely a relevancia-visszacsatoláson alapszik, a következő egyenlettel írható le:

$$q' = q + \alpha \sum_{i \in R} r_i - \beta \sum_{j \in S} s_i$$

ahol r_i az R releváns halmazban lévő i-edik dokumentum, az s_j pedig az S nem releváns halmaz j-edik dokumentuma, α és β pedig állandók.

A relevancia-visszacsatolási folyamat értékelése azt mutatja, hogy a különböző interaktív információkeresési módszerek közül ez vezet a legjobb eredményekhez, és a felhasználóra is ez rója a legkisebb terhet. A relevancia-visszacsatolásról a SMART és a Medlars rendszereket összehasonlítva megállapították, hogy két visszacsatolt keresés után a SMART-output legalább 10%-kal jobb, mint a szótőelemzést alkalmazó Medlars, és 20-30%-kal jobb, amikor az elemzés céljaira tezauruszt használtak (1. táblázat).

A fent ismertetett kérdésmódosítási eljárás a normál információkeresési folyamat során a felhasználóktól kapott információn alapul. *A felhasználók intelligenciáját* igénybe vehetjük maguknak a dokumentumvektoroknak a javítására is oly módon, hogy azokat a dokumentumokat, amelyeket a felhasználók fontosnak ítéltek „előléptetjük", míg a többieket lefokozzuk.

Különösen akkor, amikor egy adott kérdésre kapott válaszban a felhasználó sok dokumentumot jelöl „relevánsnak", e dokumentumok a jövőben könnyen megtalálhatóvá tehetők úgy, hogy mindegyiket hasonlóvá tesszük az információkeresésükhöz felhasznált kérdéshez. Ugyanígy, a talált dokumentumok közül az irrelevánsnak minősítettek megtalálását megnehezíthetjük, ha a kérdés-

ből kivonjuk ezeket. Remélhetőleg egy sor ilyen „párbeszéd" után a felhasználók által igényelt dokumentumok lassanként a dokumentumtér aktív részébe kerülnek, vagyis abba a részbe, amelyre a felhasználók kérdései koncentrálódnak, míg a szükségtelen egységek a perifériára szorulnak, ahonnan végül akár ki is dobhatók.

Az elemzés módszere	Teljesség	Pontosság
Medlars (ellenőrzött kulcsszavak)	0,3117	0,6110
SMART szótő		
0 – kezdő keresés	0,2622 (-16%)	0,4901 (-19%)
1 – ismételt visszacsatolás	0,3235 (+4%)	0,6385 (+5%)
2 – ismételt visszacsatolás	0,3433 (+10%)	0,6892 (+13%)
SMART tezaurusz		
0 – kezdő keresés	0,3232 (+4%)	0,6106 (0%)
1 – ismételt visszacsatolás	0,3915 (+25%)	0,7427 (+18%)
2 – ismételt visszacsatolás	0,4029 (+25%)	0,7438 (+18%)

(450 dokumentum, 29 kérdés)

1. táblázat. A visszacsatolást felhasználó rendszerek, a SMART és a Medlars összehasonlítása

Brauen olyan dokumentumtér-módosító folyamatot vezetett be és ellenőrzött, amely az alábbi stratégiát használta fel:

(a) A visszacsatolási folyamat során relevánsnak minősített egységhez tartozó dokumentumvektort a kérdés kulcsszavainak hozzáadásával vagy a dokumentum- és kérdésvektorban egyaránt szereplő kulcsszavak súlyának növelésével változtatta meg; másrészt, a kérdésekből hiányzó dokumentum-kulcsszavak jelentőségét kisebb súlyozással csökkentette.

(b) Hasonlóan, az irrelevánsnak minősített dokumentumoknál a dokumentum- és kérdésvektorban együttesen jelentkező dokumentum-kulcsszavak súlyát csökkentik, míg a kérdésből hiányzó kulcsszavak súlyát növelik.

Mindkét eljárás azon a feltételezésen alapul, hogy a kérdésvektorok aktív tárgyköröket képviselnek. Így az aktív kérdésekhez közelebb vitt dokumentumok „előlépnek" az által, hogy közelebb kerülnek az érdeklődés középpontjához. A fő kérdésterületektől eltávolított dokumentumokra ennek a fordítottja igaz.

Az eljárást először 125 felhasználói kérdés dokumentumtér-módosításra való felhasználásával vizsgálták. Ezután harminc kérdésből álló halmazt dolgoztak fel, a kérdéseket első lépésben az eredeti (dokumentumvektor-módosítás előtti) dokumentumgyűjteménnyel, második lépésben a 125 kérdés feldolgozása után kapott végső, módosított dokumentumtérrel összevetve. Az eredmények azt mutatják, hogy a harminc új felhasználó hasznát látta a korábbi párbeszédnek, mivel az új, módosított térben végzett keresés eredményei a normalizált teljesség szempontjából 3%-kal, a normalizált pontosság tekintetében pedig 8%-kal javultak az eredeti térben elért eredményekhez képest.

Az igazi könyvtárban a dokumentumtér módosításának a folyamata a rendszer állandó jellemzőjének tekinthető. Ha a felhasználói kör elég homogén, gyorsan kialakul egy egyensúlyi állapot, amelyben a legrelevánsabb egységek a megfelelő kérdéskörök köré csoportosulnak („klaszterálódnak"). Ahogy a felhasználók érdeklődése változik, vagy új témakörök válnak népszerűvé, a dokumentumszervezés is megváltozik, az új állapotoknak megfelelően.

Természetesen, amikor sok dokumentumvektor változik meg a beérkező felhasználói kérdések függvényében, a megfelelő dokumentumosztály- (centroid) vektorok *már nem képviselik* a megváltoztatott dokumentumokat. A dokumentumok centroid vektorainak megváltoztatására, korszerűsítésére számos stratégia alkalmazható. Idetartozik például az új kulcsszavak hozzáadása és a régiek törlése a centroidból, a meglévő kulcsszavak súlyának megváltoztatása. Végül, ahogy a dokumentumtér-változtatások mértéke növekszik, lehetővé kell tenni dokumentumok átcsúszását az egyik klaszterből a másikba. Az újra klaszterálás műveletét a következő fejezetben tárgyaljuk.

[...]

4–1 Információs rendszerek[11]

Az *információs rendszer* géppel olvasható formában tárolt rekordok (információegységek) gyűjteménye, melynek célja adott felhasználói kör ellátása információs szolgáltatásokkal. A közönséges fájlrendszerektől eltérően az információs rendszert általában nem egyetlen meghatározott célra tervezik, és

11 Salton e tanulmányát 1975-ben publikálta, ezért beszél fájlkezelő rendszerekről. Az eltelt időben a fájlkezelő rendszerekből kifejlődtek az adatbázis-kezelő rendszerek, amelyek technikailag, a szoftvert illetően és szervezési szempontból is lényegesen meghaladják a fájlkezelés szintjét. Ezt a szövegrészt mégis érdekesnek tartjuk, nem számítástechnikai, hanem könyvtári informatikai szempontból, mert amit az adat, illetve szövegkezelés problémáiról és viszonyáról ír, az lényegében ma is érvényes (a szerk.).

4–1 Information systems. In: Dynamic information and library processing. p. 115–117.

nem feltétlenül végezne benne mindig periodikus, rendszeres feldolgozást. Ehelyett sokféle, eltérő célra használható, és gyakran a központi fájloktól távoli felhasználói csoportok veszik igénybe. Ehhez természetesen olyan számítógépes programokról kell gondoskodni, amelyek lekérdezik az adatbázist, kikeresik az információt, kinyomtatják az outputot, és más szükséges feladatokat oldanak meg, amelyek lehetővé teszik a tárolt fájlok hasznosítását.

Célszerű az információs rendszerek két típusát megkülönböztetni: az *adatszolgáltató rendszereket*[12] és a *referenciaszolgáltató rendszereket.* Az előbbiek adatokra irányuló egyedi kérdésekre adnak olyan specifikus válaszokat, amelyek lehetőleg csak a ténylegesen kívánt adatokat tartalmazzák. A referenciaszolgáltató rendszerek viszont olyan felhasználókat szolgálnak ki, akiket nem meghatározott tények, hanem meghatározott szakterület állapota, szemle jellegű beszámolói érdekelnek. E rendszer outputja – találathalmaza – rendszerint valamilyen dokumentumegyüttes.

Bár ez a kétfajta információs rendszer kapcsolatban áll egymással, mivel az adatszolgáltató rendszerben talált tényadatok megtalálhatók a referenciaszolgáltató rendszerből kiválasztott dokumentumokban is, feldolgozási szempontból valójában teljesen eltérőek. A társzervezés, a struktúra-kezelés, a különböző műveletek végrehajtásához szükséges feldolgozási lépések meglehetősen különböznek az adat- és referenciaszolgáltató környezetben, és nem építhetők jó rendszerek egyformán könnyen e két típusban.

Nagy általánosságban az adatszolgáltató rendszerek a kérdések széles skálájára adhatnak választ, például „Írt-e Smith könyvet?", „Smithnek minden publikációja könyv?", „Kik írtak tanulmányt a szintaktikai elemzésről?" és így tovább. A gyakorlatban az ilyen általános *kérdés–felelet* vagy *tényadat-kereső* rendszerek valójában nem építhetők meg, mivel a dekódoláshoz és a tárolt információ kezeléséhez szükséges műveletek összetettsége és bonyolultsága meghaladja a jelenlegi képességeket, lehetőségeket. A működő adatszolgáltató rendszerek egyetlen, általában elérhető típusa a *fájlkezelő* vagy *információkezelő* rendszer, amely képes strukturált és formalizált adatbázisokhoz alkalmas fájlok feldolgozására. A fájlkezelő rendszerek tervezésekor a hangsúlyt rendszerint az igen eltérő fájl- és adatszerkezetekre alkalmazható általánosított eljárásokra, folyamatokra, valamint a kérdésbevezetés és a válaszgenerálás egyszerűsítését célzó rugalmas felhasználói interfész („illesztési") módszerekre helyezik. Valójában a működő rendszerek koncepcionálisan elég korlátozottak, és olyan formalizált fájlokon alapulnak, amelyek – specifikus szervezettségük miatt – az előzetesen definiált kérdéstípusoknak csak egy kis halmazára képesek válaszokat generálni.

---

12 Az adatszolgáltató rendszerek a faktografikus (tényadat-szolgáltató) rendszerek, a renferenciaszolgáltató rendszerek pedig a másodlagos információkat szolgáltató rendszerek. Másodlagos információk például a dokumentumleírások adatelemei (a szerk.).

A jelenlegi adatszolgáltató rendszerek tehát elsősorban fájlkezelési célokra – üzleti fájlok létrehozására, kezelésére, korszerűsítésére, aktualizálására és összefoglalására – alkalmasak, fájllekérdezési és információkeresési képességeik azonban rendszerint korlátozottak. Ezen túlmenően az általánosításra való törekvés – például a következetességi követelmények alkalmazása eltérő összetevőkre, új adatok vagy gyűjtemények ráhelyezése új feldolgozási követelmények által – gyakran rendkívül költséges és időrabló. A jelenleg használatban lévő adatszolgáltató rendszerek képességei ezért korlátozottak, és nem tekinthetők az általánosabb kérdés–felelet rendszerek előfutárainak, amelyek képesek lesznek egy adott témakörhöz tartozó kérdések nyitott halmazára válaszolni.

Referenciaszolgáltató rendszereket viszont elvileg sokkal könnyebb tervezni és üzemeltetni, és igen sokféle ilyen rendszert helyeztek már üzembe. Az elsődleges irodalomhoz (folyóirat vagy könyv) hozzáférést nyújtó rendszereket *másodlagos információs szolgáltatásoknak* nevezzük. Közéjük soroljuk hagyományosan a szemléket tartalmazó folyóiratokat, referálólapokat, a különböző indexeket, bibliográfiákat, a felhasználóra szabott referenciaszolgáltatásokat és bizonyos információkereső rendszereket. Kézenfekvő, hogy ezeknek a rendszereknek a jelentősége az elsődleges irodalom növekedésével és elérésének nehezedésével egyenes arányban nő. Az, hogy az utóbbi években olyan nagy hangsúlyt fektettek az információkereső rendszerek tervezésére és működtetésére, egyértelműen az információrobbanásnak és mellékhatásainak a következménye.

A másodlagos információs szolgáltatások általában három fő funkció teljesítésére alkalmasak:

(a) *Kivonatolás és referálás,* az előállított kivonatok és referátumok indexelésével és osztályozásával együtt.
(b) *Tárolás és információkeresés* olyan egységek esetében, amelyek az elsődleges irodalomra vonatkoznak.
(c) *Jeladó és folyamatos tájékoztatás* (az ún. „current awareness") a már rendelkezésre bocsátható újonnan publikált egységekről szóló információ nyújtása.

Vegyük az első funkciót, a kivonatkészítést és referálást. A kritikai jellegű és átfogó áttekintések nyilvánvalóan óriási potenciális segítséget nyújtanak azoknak, akik lépést akarnak tartani valamelyik szakterület eredményeivel. A jó szemletanulmány elkészítése bonyolult és időigényes feladat, és a jelenlegi ösztönzési rendszer nem kedvez a szemleírásnak annyira, mint a kutatási jelentések és beszámolók írásának. Ez lehet a magyarázata, hogy a jó szemlék hiányoznak.

A szemlék írásával kapcsolatos helyzethez hasonló, bár kevésbé tragikus a referálás állapota. A referátumok készítőitől rendszerint elvárják, hogy szak-

emberei legyenek a referált témakörnek, ugyanakkor a publikálási hierarchiában a beszámolónak nem tulajdonítanak jelentőséget, így a referátumok többségének színvonala kívánnivalókat hagy maga után.

Bár a referálást és szemlézést komolyan igénylik, sajnos nem valószínű, hogy a megoldást számítógépes automatikus technikáktól várhatjuk. Az automatikusan készített referátum a dokumentum szövegéből kivont mondatokból áll, és a szemlék, összefoglalások automatikus készítése meghaladja a jelenlegi szövegfeldolgozó rendszerek lehetőségeit. Így – gyakorlati célokra – az automatikus technikát az információkereső és folyamatos tájékoztató rendszerek tervezésére kell korlátozni. E referenciaszolgáltató területek és rendszerek jelenlegi állapotát írjuk le a fejezet további részében.

[...]

8–1 Klaszterált fájlok[13]

8–1–A A fő jellemzők

Az információkeresés szempontjából számunkra kétféle osztályozás érdekes: a *kulcsszó*osztályozás, amelyet annak reményében hoztak létre, hogy a kulcsszavakat (szinonima) osztályokba sorolva nagyobb megfelelést lehessen elérni a kérdés és a dokumentum-kulcsszavak között; valamint a *dokumentum*osztályozás, amely jobb, gyorsabb keresési outputot eredményez oly módon, hogy a keresést a fájlnak meghatározott részére korlátozzák. A két osztályozás nem független egymástól, hiszen a dokumentumokhoz rendelt kulcsszavaknak kell megalapozniuk a dokumentumcsoportosítási eljárással létrehozott osztályokat.

> Salton mindvégig csak „osztályozásról" beszél, valójában azonban klaszterelemzéssel végzett automatikus (numerikus) osztályozást ért rajta.

A jó kulcsszavas osztályozásnak általában sikerül a különböző, egymáshoz kapcsolódó, alacsony előfordulású kulcsszavakat közös deszkriptorosztályokba tömöríteni. A közös osztályba tartozó kulcsszavak ezután információkereséskor helyettesíthetők egymással, és az ilyen eljárással kapott teljességi eredmény várhatóan tovább tökéletesíthető. A dokumentumosztályozások esetén a kereséseket csak a legérdekesebb dokumentumosztályokra korlátozzák, és ezáltal nagyon pontos outputot kapnak. Amikor az osztályozott dokumentumfájlokat egy

13 8–1 Clustered files. In: Dynamic informatian and library processing, 1975. p. 323–333.

jó tezaurusszal együtt használják, nagy teljességű és pontosságú keresési eredményekre számíthatunk.

Egyébiránt kétféle osztályozási törekvést célszerű megkülönböztetni: az egyik esetben az osztályozás teljesen elvont folyamat, melyben jól körülhatárolt kritériumon alapuló formális eljárást alkalmaznak a formálisan meghatározott dokumentumok halmazára; ugyanakkor az osztályozás lehet empirikus folyamat is, amelyet konkrét cél elérése érdekében hajtanak végre.

Létezik néhány olyan terület is – a taxonómiában például –, ahol az osztályozás formális elmélete használható a csoportosítási módszerek definiálására, és az osztályozandó adatok formális leírása is rendelkezésre áll. Nem ez a helyzet az információkereséssel, mivel az osztályozandó objektumok – kulcsszavak vagy dokumentumok – gyakran rosszul definiáltak, és nem állnak rendelkezésre a priori kritériumok az optimális osztályozás leírására. A dokumentációban a fő cél valójában az, hogy a felhasználói kérdésekre adott válaszban a hasznos anyagot visszanyerjük, a nem hasznosat pedig elvessük. Ilyen feltételek között a legnagyobb erőket a jó kulcsszó- és dokumentumosztályozások készítésére kell összpontosítani.

Amikor az objektumok halmazát az osztályok meghatározott halmazához kell rendelni, a létrejövő osztályozástól rendszerint a következő jellemzőket várjuk:

(a) Az osztályozás legyen annyira jól definiált, hogy bármilyen adategyüttesre egyetlen eredményt kapjunk.
(b) Az osztályozás eredményét nem befolyásolhatja az a sorrend, amelyben az objektumok az osztályozási folyamatba lépnek (sorrendfüggetlenség), vagyis az objektumok újracímkézése hagyja érintetlenül az osztályozást.
(c) Az osztályozás legyen stabil, hogy az adatokban beálló kis változások csak kis változásokat *idézzenek elő* a kapott osztályozásban.
(d) Az osztályozás legyen léptékfüggetlen, hogy az objektumokat azonosító tulajdonságállandóval való szorzás ne befolyásolja az osztályozást.
(e) Az erős hasonlóságokat mutató objektumokat nem szabad szétválasztani úgy, hogy más-más osztályba soroljuk őket.

Az első két tulajdonság – a jól definiáltság és a sorrendfüggetlenség – összekapcsolódik, mivel meglétüket csak akkor érjük el, ha az objektumok valamennyi, az osztályozási kritériumokat kielégítő alkalmazását az osztályok tényleges definiálása előtt megvizsgáljuk. Az ilyen alapos döntés azonban túlságosan időigényes ahhoz, hogy a gyakorlatban akkor is végrehajtsuk, ha az osztályozási folyamat rendkívül sok objektumra terjed ki. Ha az első két kritérium nem teljesül, a stabilitás válik különösen fontossá, mivel ez biztosítja, hogy a tulajdonságok és objektumok egymáshoz rendelésében előforduló kisebb hi-

bák korrekciója csak kis módosulásokat hozzon létre az osztályokban. Ugyanez áll fenn arra az esetre is, amikor az objektumokat tulajdonságok hozzáadásával azonosítjuk, vagy amikor meglévő tulajdonságokat távolítunk el. A léptékfüggetlenség (a [d] tulajdonság) szintén természetes követelmény lehet, mivel az objektumok közötti hasonlóság mérésére használt lépték általában önkényes.

Az információkeresésben mindig kívánatos, hogy a kulcsszavak és a dokumentumok osztályai stabilak legyenek, elsősorban azért, mert az objektumokat jellemző tulajdonságvektorok nem mindig pontosak és megbízhatóak. A sorrendfüggetlenség elvileg kívánatos, de a gyakorlatban sokszor egyenértékű keresési eredményeket lehet elérni olyan kulcsszó- és/vagy dokumentumosztályozásokkal, amelyek lényeges eltéréseket mutatnak. Ilyen körülmények között a formális osztályozási követelmények jelentősége a vártnál kisebb lehet.

8–1–B Osztályozási típusok

Az osztályozási rendszerek különböző formai jellemzőkkel írhatók le. Az osztályozás lehet monotetikus vagy politetikus, ami azt jelenti, hogy az osztály valamennyi tagja rendelkezik egy jellemző tulajdonsággal (monotetikus), vagy ellenkezőleg: nem tehető ilyen megkötés. Az osztályok lehetnek kizáróak, ha az objektumokat legfeljebb egy osztályba sorolják be, vagy átfedőek. Végül az osztályozások lehetnek rendezettek, ha megadják az eltérő osztályok közötti szisztematikus kapcsolatokat, illetve vannak nem rendezett osztályozások.

A numerikus osztályozási rendszerek tipológiájával részletesen *Karen Sparck Jones* foglalkozik a következő szemelvényben.

Az információkeresésben minden esetben a legkevésbé korlátozó követelményt részesítik előnyben. Általában sem a dokumentumokat, sem a kulcsszavakat nem definiálják olyan részletesen, hogy érdemes legyen monotetikus kulcsszó- vagy dokumentumosztályozásokat építeni. Ugyanezen okból a legjobb osztályok átfedőek, éppen azért, hogy az egységek egynél több osztályban jelenhessenek meg. Néha célszerű lehet rendezett kulcsszó-osztályozásokat (kulcsszó-hierarchiákat), illetve rendezett dokumentumosztályokat létrehozni. Általában azonban, ha semmilyen speciális követelmény nincs, a nemrendszerező, rendezetlen osztályozás a valóságot jobban megközelítő osztályokat hoz létre. Így többnyire politetikus, átfedő, nem rendező osztályozási rendszerekre van szükség.

Az osztályozási folyamattal generált osztályok leírására számos módszer áll rendelkezésre:

(a) A *faktoranalízissel* az adatokban fellépő maximális változások irányának meghatározását kísérlik meg a mérési térben, tengelyek for-

gatásával. A különböző kategóriákat azután olyan vektorokkal írják le, amelyek egymástól annyira *elkülönültek,* amennyire csak lehet, és azokat a dimenziókat választják, amelyekben a középértékek maximális változásokat mutatnak.

(b) A *felhalmozás* (nyalábosítás, füzéresítés, bogosítás „clumping")[14] vagy a legközelebbi szomszéd technikája az osztályleíráson alapul: a leírást az ugyanabba az osztályba tartozó elemek listája képviseli. Az osztályokat úgy nyerik, hogy az objektumok párjai közötti hasonlóságokat kiszámítják, és olyan klaszterpárokból álló rendszerbe egyesítik őket, amelyek elegendő közelséget mutatnak.

(c) *Particionálás* (felosztás) vagy alakkeresés esetén az osztályt átlagos pozíciójú vagy alakú kulcsszavakkal írják le, amelyek képesek az osztály elemeit reprezentálni. Az osztályok egy vagy több (a térben pontokkal vagy egyenesekkel képviselt) klaszterközpont kiválasztásával, és a centroid(ok) adott küszöbértékén belüli összes objektumnak a megfelelő osztály(ok)ba helyezésével alakíthatók ki. Másik alternatívának megfelelően az osztályok az adathalmaz többé-kevésbé önkényes felosztásával nyerhetők, amit a homogén osztályok kialakítását szolgáló finomítási eljárás követ.

(d) Bizonyos *döntéselméleti* és *felbontó* módszerekben valószínűségi eloszlással dolgoznak a kategóriák és osztályok leírásakor. A becsült valószínűségi eloszlások viselkedésén alapuló döntést használják azután az osztályok közötti határok kialakítására.

(e) A *lépcsős* vagy *lineárisan adaptív* rendszerek az egyenesekkel vagy síkokkal felosztott osztályok közötti határok definícióján alapszanak oly módon, hogy az eredményként kapott osztályok maximálisan távoliak legyenek.

(f) A különböző felsoroló (enumeratív) technikák akkor használhatók, amikor az adott adathalmazok mérete korlátozott. Az osztályokat úgy határozzák meg, hogy az összes lehetséges adatfelosztást kipróbálják, és megállapítják a legjobb halmazokat az osztályok definíciójához.

(g) A klaszterek (csoportok) különböző cél- vagy szereporientált technikákkal szintén definiálhatók, például a térsűrűségi mérések számításaival, a gráfelmélet felhasználásával (a tárgyak maximálisan teljes alhalmazaira), vagy azáltal, hogy a teret a tömegvonzás vagy más kapcsolódó eljárások segítségével összeomlasztják.

(h) Az elmúlt években az interaktív klaszter (csoport) módszerek voltak használatosak, amikor az operátor a klaszterszerkezetre vonatkozóan ad hoc döntéseket hozhat a képernyőn megjelenő, és az előző klaszterműveletekből nyert adatoknak a megtekintése után.

---

14 A clump (= bog, csomó, nyaláb) egyfajta klaszter (a szerk.).

Függetlenül attól, hogy milyen stratégiát használnak az ideális osztályok leírására vagy a tényleges csoportosítási eljárások végrehajtására, az eredményül kapott osztályozás hasznossága különböző megfontolásoktól függhet. Egyrészt vizsgálhatjuk a csoportosítási eljárás hatékonyságát a legkívánatosabb tényezők, például a felhasznált tároló terület, a futási idő vagy a végrehajtott műveletek száma alapján, a legkisebb költséget jelentő felosztás megadásával.

Másrészről viszont kívánhatjuk az osztályozási kritériumok optimalizálását vagy egyetlen osztályba tartozó egységek közötti erős asszociációt, vagy a különböző osztályokban lévő egységek közötti gyenge kötődést. Vegyük az egységek A osztályát, B pedig reprezentálja a nem A-ba tartozó egységeket. Ekkor egy osztályt úgy definiálhatunk, mint A halmazt, amely minimalizálja az összetartozási (kohéziós) funkciót:

$$C = \frac{S_{AB}^2}{S_{AA} \times S_{BB}}$$

Az S_{AB} úgy definiálható mint az A-ba tartozó N_A egységek és a B-be tartozó N_B egységek páronkénti hasonlóságainak átlagértéke, azaz

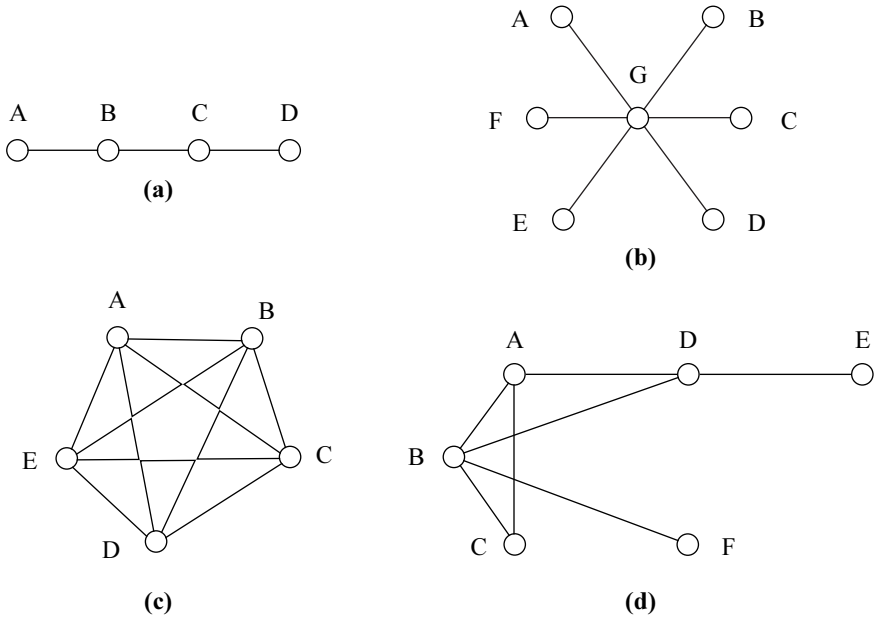
$$S_{AB} = \frac{1}{N_A N_B} \sum_{i \in A} \sum_{j \in B} S_{ij}$$

ahol S_{ij} az i és j egységek közötti hasonlóságot fejezi ki.

A másik lehetőség, hogy S_{AB} az összes S_{ij} hasonlóság maximuma lehet i### A és j### B esetére.

A számítási költségeken és az osztályok kohézióján kívül az osztályozási folyamat jellemezhető a használt paraméterekkel és az eredményül kapott osztályok típusával is. Az osztályok a bennük lévő objektumok hasonlósági tulajdonságaival jellemezhetők: vagy az 1. egység hasonló a 2.-hoz, amely viszont a 3. egységhez hasonlít és így tovább (lánctípus); vagy minden egyes egység egy központi helyzetű egységhez hasonló (csillagtípus); vagy minden egyes egység hasonló az osztály összes többi eleméhez (összekapcsolt, „klikkelt" típus); vagy az egyes egységek és a többi egységek gyűjteménye közötti hasonlóság meghalad adott küszöbértéket. Ezeket az osztálytípusokat mutatja be grafikusan az 1. ábra.

Az osztályok azonosíthatók a használt hasonlósági függvénnyel és a hasonlósági együttható ama értékével, amely az osztályok szétválasztásához szükséges; az osztályok profiljának, centroidjának meghatározásával, amelyet az osztályok azonosítására használnak; és azokkal a kulcsszavakkal vagy tulajdonságokkal, amelyekkel az osztályozandó egységeket azonosítják.



1. ábra. Jellegzetes osztálytípusok: (a) lánc, (b) csillag, (c) klikk, (d) clump (a klaszter egyik fajtája).

Egy dolog közös a különböző elméletekben: kevés szót ejtenek arról, milyen hatása van a különböző osztályozási kritériumoknak az osztályok három fontos jellemzőjére. Nevezetesen az adott eljárással várhatóan nyerhető osztályok számára, eme osztályok méretére és az osztályok közötti átfedésre. Ez a három paraméter részben meghatározza, hogy mennyire használható az osztályozási folyamat információkeresés céljaira. Sajnos, nem tudjuk, melyik eljárás a leghatékonyabb a paraméterek értékeinek optimalizálására, vagy milyen értéktartományok alapján születnek jó keresési eredmények.

A továbbiakban számos, az információkeresés szempontjából hasznos osztályozási módszert részletesebben elemzünk.

8–1–C Az osztályozás módszertana

Az információkeresésben szoros kapcsolat van az indexelés és az osztályozás között. Az automatikus kulcsszó-osztályozások – amelyek rendeltetése, hogy minden kulcsszóhoz a keresésben használható „helyettesítőket” hozzanak létre – nyilvánvalóan az információs egységek azonosításához használt

kulcsszavaktól függenek. Sőt, mivel az adott osztályozás generálásához szükséges számítógépidő bizonyos mértékig az osztályozandó egységek számától függ, fontos, hogy a lehető legjobb kulcsszavak álljanak rendelkezésre.

Vegyük most a dokumentumreprezentáció problémáját. Az információkeresésben adott dokumentumot rendszerint a kulcsszavak valamelyik halmaza azonosítja. A ténylegesen használt kulcsszavak így mind a kulcszóosztályozást, mind pedig a dokumentumosztályozást közvetlenül befolyásolják. Pontosabban: a két dokumentum között mutatkozó hasonlóság gyakran a dokumentumvektorban szereplő kulcsszavak függvényeként számítható ki. A dokumentumosztályok tehát közvetlenül a rendelkezésre álló indexeléstől függenek.

A későbbiekben látni fogjuk, hogy az információkeresésben használható legvonzóbb osztályozási rendszerek némelyike a redukálhatóság elvén nyugszik, vagyis a egységek – kérdések vagy dokumentumok – ama részgyűjteményeinek kiválasztásán, amelyek együttesen tágabb felhasználói környezetet képviselnek és amelyeknek az osztályozása lényegesen olcsóbb, mint az egész gyűjtemény osztályozása. Természetesen az ilyen részgyűjtemények kiválasztása és elérhetősége szintén az eredeti indexelés függvénye.

A klaszterálási módszereknek általában három meghatározott részfolyamata van: az *indító eljárás,* amelyet adott osztály generálására alkalmaznak; a *hozzárendelési eljárás,* melynek során az egységeket az osztályokhoz rendelik; és a *befejező eljárás,* amely azokkal a feltételekkel foglalkozik, amelyek alapján adott osztály véglegesnek és teljesnek tekinthető.

Az indítási eljárás gyakran abból áll, hogy az egységek által alkotott tér sűrű területein elhelyezkedő, egymással szomszédos, hasonló egységek közül kiindulási osztályközpontokat választanak. A hozzárendelési eljárás is megoldható úgy, hogy ay egységeket azokba az osztályokba helyezik, amelyekkel elég nagy a hasonlósági együtthatójuk. Végül a befejező eljárás mutatja meg, hogy az eredményül kapott osztályok kielégítik-e az előzőleg megállapított feltételt, például a korábban említett összetartozási (kohéziós) feltételt, vagy sem.

Az üzemeltetés szempontjából két főbb osztályozási módszert célszerű megkülönböztetni: az *alulról felfelé* vagy *generáló* eljárásokat, amelyek egy klaszterálatlan egységhalmazból indulnak ki, és először egységpárok, majd egységhármasok stb. kialakításával egyre nagyobb csoportokat hoznak létre. A *felülről lefelé* vagy *felosztó* eljárásoknál viszont feltételezzük, hogy valamennyi egység kezdetben egy osztályba tartozik, amelyet azután több részre bontanak, majd e részeket is további kisebb csoportokra osztják, egészen addig, míg a végleges osztályok ki nem alakulnak. A gyakorlatban időnként kevert felülről lefelé eljárást alkalmaznak, amelyben több kiindulási osztályt tételeznek fel, és az egységek kezdeti felosztását a klaszter minőségének javítására ismételt hozzárendelési folyamatokkal finomítják úgy, hogy az összetartozást (kohéziót) maximalizálják, vagy, hogy a távoli klaszterekben lévő egységek közötti hasonlóságot a legkisebbre csökkentsék.

A legtöbb felülről lefelé módszert úgy tervezik, hogy *hierarchikus* osztályelrendezéseket kapjanak, abban az értelemben, hogy a szintenkénti eljárás olyan osztályokat generáljon, amelyek valamilyen magasabb szintű osztály alhalmazai vagy beletartoznak ebbe a magasabb szintű osztályba. Az eredmény fastruktúra, amelynek a gyökere (a legfelsőbb szintje) a teljes teret reprezentáló szuperosztály, a levelek (alsó szint) pedig egyedi dokumentumok vagy kulcsszavak csoportjai. Néhány alulról felfelé építkező osztályozási módszer alkalmazásával is létrejönnek hierarchikus csoportosítások. Nem-hierarchikus osztályozások azok, amelyekben a generált osztályok között nem állnak fenn formális belefoglalási tulajdonságok. Amikor osztályhierarchiát fastruktúrába rendeznek, általában igyekeznek a szűken definiált, specifikus egységeket a struktúra aljára, az általánosabb jellegű egységeket pedig a struktúra tetejéhez közel elhelyezni.

Nehéz az osztályozási algoritmusok számítási hatékonyságáról általános érvényű megállapításokat tenni. A páronkénti hasonlósági mátrix kialakításán alapuló jól ismert eljárások például, amelyeknél a mátrix minden egységpár hasonlósági együtthatóját megadja, n egység esetén legalább $R(n^2)$ nagyságrendű műveletet igényelnek. [Csupán a párok hasonlósági együtthatójának kiszámítása $R(n^2)$ vektor-összehasonlítást jelent, mivel n egység esetén $n(n–1)/2$ külön pár lesz.] Ha feltételezzük, hogy a legalább n^2 műveletet igénylő eljárások a gyakorlati kivitelezés szempontjából túl költségesek, a választás azokra a módszerekre szűkül le, amelyekben a tárgyi tér kiindulási felosztása adott, vagy amelyekben az objektumok alkotta tér kisebb egységhalmazra redukálható. Ez utóbbi esetben az osztályozási eljárás viszonylag gyorsan elvégezhető a jól definiált kívánatos tulajdonságok, a sorrendfüggetlenség és a stabilitás megtartásával.

8–1–D Profilmeghatározás

Ha azt akarjuk, hogy az osztályok rendszere a gyakorlatban használható legyen, az egyes osztályokat általában egy vagy több azonosítóval kell reprezentálni, amelyeket általában *profil*nak vagy *centroid*nak neveznek. Az osztály profilja lehet az osztály középpontjában elhelyezett „hamis” egység, vagy bármelyik olyan reprezentatív objektum, mely az osztályban lévő többi objektumot jellemezni tudja. Mivel elvárjuk, hogy az egységek egy osztályon belül jobban hasonlítsanak egymáshoz, mint a többi osztály egységeihez, az egyes egységeknek inkább kell saját profiljukhoz hasonlítani, mint a többi osztály profiljaihoz. Továbbá, az információkeresésben általában úgy jutnak el a releváns osztályokhoz, hogy a kérdéseket összehasonlítják az osztályok profiljaival. A jó profilmegfogalmazás tehát igen lényeges feltétele a klaszterált fájlok sikeres működésének.

Vegyünk egy C osztályt, amely D_n objektumból áll, azaz

$$C = \{D_1, D_2, \ldots , D_n\}$$

A következő természetes profildefiníciók adódnak:

(a) P_1 logikai profil. Ha $D_i = (d_{i1}, d_{i2}, \ldots , d_{it})$ a kulcsszavak vektora, és $d_{ii} = 1$, ha a j kulcsszót D_i-hez rendeljük, és egyébként j = 1, 2, ... , t-re $d_{ij} = 0$, akkor a P_1 profil definíciója a következő lehet: $(P_1 = p_{11}, p_{12}, \ldots , p_{1t}) = D_1 \cup D_2 \cup \ldots \cup D_n$. Azaz, $p_{1j} = 1$, akkor és csak akkor, ha legalább az osztály egy eleme tartalmazza a j kulcsszót. Egyébként $p_{1j} = 0$.

(b) P_2 dokumentumgyakoriság profil. Ha minden egyes egységre ugyanazt a leírást használjuk, olyan súlyozott profil használható, amelyben az egyes p_{2j} profil-kulcsszavak súlya egyenlő a j kulcsszót tartalmazó klaszterben lévő egységek számával. Pontosabban, $P_2 = (p_{21}, p_{22}, \ldots , p_{2t}) = D_1+D_2+ \ldots +D_n$, ahol $d_{ij} = 1$, ha a j kulcsszót D_i -hez rendelték, minden más esetben $d_{ij} = 0$.

(c) P_3 kulcsszógyakoriság profil. Azokban a rendszerekben, amelyekben a kulcsszavakat súlyozzák, azaz, ahol d_{ij} a D_i j-edik kulcsszavához rendelt fontosság (súly), a profildefiníció a következőképpen terjeszthető ki a P_3-ra: $P_3 = (p_{31}, p_{32}, \ldots , p_{3t}) = D_1+D_2+ \ldots +D_n$, ahol p_{3j} jelenleg a j kulcsszó teljes súlya a gyűjtemény valamennyi egysége között.

Rendszerint itt is bevezethetők különböző normalizációs tényezők. Így a P_2-ben és P_3-ban lévő kulcsszavak úgy normalizálhatók, hogy elosztjuk őket az osztályba tartozó egységek számával (n). Ez a tömegközépponttal analóg definíciót eredményez, nevezetesen:

$$P_{cm} = \frac{1}{n} \sum_{i=1}^{n} D_i$$

Másik lehetőség, hogy az egységekben a kulcsszavak súlyértékeit az egyes dokumentumok $|D_i|$ hosszával elosztva normalizáljuk. Ekkor

$$P_{cv} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i}{|D_i|}$$

A különböző profildefiníciók példáit a 2. táblázat tartalmazza.

Ha bizonyos kulcsszavak hangsúlyozására gyakorisági súlyozást használnak, gyakran előfordul, hogy tekintélyes súlytartományokat kapnak. Ilyen esetekben a nagyon nagy súlyértékű tényezők elnyomják az átlagos vagy ala-

csony súlyértékűeket. Ennek az az eredménye, hogy a kis előfordulási gyakorisággal rendelkező kulcsszavaknak nincs befolyásuk a keresési folyamatra. Ennek a nemkívánatos helyzetnek a megelőzésére azt javasolják, hogy a tényleges gyakorisági súlyok helyett *rangértékeket* használjanak.

(A) EREDETI EGYSÉGVEKTROROK

Súlyozatlan kulcsszavak	*Súlyozott kulcsszavak*
$D_1 = (1, 0, 1, 0, 0, 1, 0, 0, 0, 0)$	$D_1 = (1, 0, 3, 0, 0, 1, 0, 0, 0, 0)$
$D_2 = (0, 0, 1, 0, 0, 1, 1, 0, 0, 0)$	$D_2 = (0, 0, 2, 0, 0, 3, 1, 0, 0, 0)$
$D_3 = (0, 0, 0, 0, 0, 1, 1, 0, 0, 1)$	$D_3 = (0, 0, 0, 0, 0, 1, 2, 0, 0, 3)$
$D_4 = (0, 0, 0, 0, 0, 1, 0, 0, 0, 1)$	$D_4 = (0, 0, 0, 0, 0, 3, 0, 0, 0, 2)$
$D_5 = (0, 0, 0, 0, 0, 1, 1, 0, 0, 0)$	$D_5 = (0, 0, 0, 0, 0, 2, 1, 0, 0, 0)$
$\lvert D_1 \rvert = \sqrt{3}$, $\lvert D_2 \rvert = \sqrt{3}$, $\lvert D_3 \rvert = \sqrt{3}$,	$\lvert D_1 \rvert = \sqrt{11}$, $\lvert D_2 \rvert = \sqrt{14}$, $\lvert D_3 \rvert = \sqrt{14}$,
$\lvert D_4 \rvert = \sqrt{2}$, $\lvert D_5 \rvert = \sqrt{2}$,	$\lvert D_4 \rvert = \sqrt{13}$, $\lvert D_5 \rvert = \sqrt{5}$

(B) A MEGFELELŐ OSZTÁLYPROFILOK

Osztályprofilok (súlyozatlan)	*Osztályprofilok (súlyozottak)*
$P_1 = (1, 0, 1, 0, 0, 1, 1, 0, 0, 1)$	$P_3 = (1, 0, 5, 0, 0, 10, 4, 0, 0, 5)$
$P_2 = (1, 0, 2, 0, 0, 5, 3, 0, 0, 2)$	$P_{cm} = (0{,}20, 0, 1, 0, 0, 2, 0{,}80, 0, 0, 1)$
$P_{cm} = (0{,}20, 0, 40, 0, 0, 1, 0{,}60, 0, 0, 0{,}40)$	$P_{cv} = (0{,}06, 0, 0{,}28\ 0, 0, 0{,}62, 0{,}25\ 0, 0, 0{,}27)$
$P_{cv} = (0{,}12, 0, 0{,}23, 0, 0, 0{,}63, 0{,}37, 0, 0, 0{,}26)$	
$\lvert P_1 \rvert = \sqrt{5}$, $\lvert P_2 \rvert = \sqrt{43}$, $\lvert P_{cm} \rvert = \sqrt{1{,}72}$,	$\lvert P_3 \rvert = \sqrt{167}$, $\lvert P_{cm} \rvert = \sqrt{6{,}68}$,
$\lvert P_{cv} \rvert = \sqrt{0{,}67}$,	$\lvert P_{cv} \rvert = \sqrt{0{,}60}$,

2. táblázat. Tipikus profildefiníciók

A rangérték az alapérték és a kulcsszóhoz rendelt rang közötti különbség. A rangot úgy állapítják meg, hogy minden kulcsszót a gyakoriság csökkenő sorrendjébe rendeznek. A rangérték megállapítására a következő eljárás használható:

(a) A vektor összes kulcsszavát a meglévő kulcsszósúlyoknak megfelelő csökkenő rendbe sorolják be, vagyis a legnagyobb súlyú kulcsszó vagy kulcsszavak az 1-es rangot kapják, a következő legnagyobb súlyhoz 2-es rangot rendelnek stb. a legkisebb súlyú kulcsszóig.

(b) Az i-edik kulcsszót a $v_i = b–r_i$ súlyhoz rendelik, ahol b egy állandó alapérték, és r_i az 1. lépésben az i-edik kulcsszónak adott rang.

Az alapérték előre kiválasztott állandó, rendszerint elég nagy ahhoz, hogy valamennyi v_i rangérték pozitív maradjon.

A rangértékek számításához használt módszerből világosan kitűnik, hogy a kulcsszósúlyokra kapott súlytartományok lényegesen csökkennek, amikor a teljes gyakoriság helyett rangsorokat használunk. Az alapérték bevezetése ezen kívül még azt is biztosítja, hogy az összes vektor azonosan, maximális súllyal terhelődik. A kulcsszósúlyok relatív nagysága pedig ugyanakkora marad, mint szabályos súlyok esetén.

A 2. táblázatban található P_2 és P_3 profilok transzformációjának rangsor szerinti súlyozása a 3. táblázaton látható.

Eredeti profil	$P_2 =$	(1,	0,	2,	0,	0,	5,	3,	0,	0,	2)
		|		|			|	|			|
Gyakorisági rangok		4		3			1	3			2)
		|		|			|	|			|
Alapérték 5	$P_r =$	(1,	0,	2,	0,	0,	4,	3,	0,	0,	2)
Alapérték 10	$P`_r =$	(1,	0,	2,	0,	0,	4,	3,	0,	0,	2)
Nagyságok	$\lvert P_2 \rvert = \sqrt{43}$,			$\lvert P_r \rvert = \sqrt{34}$,			$\lvert P`_r \rvert = \sqrt{279}$				
Eredeti profil	$P_3 =$	(1,	0,	5,	0,	0,	10,	4,	0,	0,	5)
		|		|			|	|			|
Gyakorisági rangsorok		4		2			1	3			2
		|		|			|	|			
Alapérték 5	$P_r =$	(6,	0,	3,	0,	0,	4,	2,	0,	0,	3)
Alapérték 10	$P`_r =$	(6,	0,	8,	0,	0,	9,	7,	0,	0,	8)
Nagyságok	$\lvert P_3 \rvert = \sqrt{167}$,			$\lvert P_r \rvert = \sqrt{39}$,			$\lvert P`_r \rvert = \sqrt{294}$				

3. táblázat. Rangérték-transzformációk

Másik, az információkeresés szempontjából érdekes profiltranszformáció a profilvektor hosszának lehetséges csökkentése. A 2. táblázatban bemutatott profilkialakítási módszerek alapján világos, hogy a tipikus osztályprofilban a nem-zérus kulcsszavak száma sokkal nagyobb, mint a közönséges egységvektorban. Ennek az a következménye, hogy a kérdés megfogalmazása és a profilvektor közötti összehasonlítások sokkal drágábbak, mint a szabályos objektumvektorokkal való összevetések.

Éppen ezért csökkenteni szokták az osztályprofil-vektorok hosszát (a nem-zérus kulcsszavak számát), hogy egybevágjanak az egy adott osztály objektumaiban szereplő, nem-zérus kulcsszavak átlagos számával úgy, hogy egy sor kisebb súlyú kulcsszót kirekesztenek. Mivel a klaszteres keresési folyamat visszanyerési hatékonysága nagy mértékben a választott keresési stratégiától függ, vagyis attól, hogy a keresés széles vagy szűk körű-e (az aktuálisan összehasonlított osztályok száma a kérdésekkel), valamint a klaszter struktúrán való előre és/vagy hátra irányuló mozgástól, nem meglepő, hogy a keresésben használt osztályprofil-vektorok hossza viszonylagosan elhanyagolható a keresés hatékonyságának meghatározásakor.

Vita az „Automatikus információkereséhez használható gyors dokumentum osztályozás" című előadás után[15]

Bollmann: Hány dokumentum osztályozását végezték el ezzel a módszerrel? Készült összehasonlító vizsgálat más eljárások hatékonyságával?

Salton: Az első kérdéshez: laboratóriumunkban legutóbb 5000 informatikai tárgyú dokumentumot osztályoztunk. Általában nem dolgozunk 5000-nél több dokumentummal, de a módszer gyors, akár több tízezer dokumentumra is alkalmazható. A második kérdésre nehezebb válaszolni, mivel az értékelés kérdése eleve problematikus. Szívem szerint a következőt mondhatom: nem törődőm azzal, hogyan fest az osztályozás, hanem csak azzal, hogy mire szolgál. Információkereső célú osztályozási módszerek értékelése számomra azt jelenti, hogy ne járjon kisebb teljességgel, mint más módszerek. Persze a számítógépidő, a tárolószükséglet, az összehasonlítások száma, és hasonlók fontos kritériumok. Alapvetően azonban nem az osztályozási rendszert értékelem, hanem az információkereső rendszert.

Schmitz-Esser: Fogalmi rendszerek osztályozására is alkalmasak az ön által előadott módszerek?

Salton: Az osztályozás tárgyak osztályozása, melyek tulajdonságokkal vagy deszkriptorokkal írhatók le. Ha a fogalmak is leírhatók ilyen tulajdonságokkal, akkor a módszerek alkalmazhatók.

Kuhlen: A nehézségeket nyilván az okozza, hogy a fogalmaknak nem tulajdonságaik, hanem összefüggéseik vannak, ezeket meg hogyan lehetne algoritmikusan megragadni?

---

15 Gerard Salton: Fast document classification in automatic information retrieval. In: Kooperation in der Klassifikation I. Proceedings der Sekt. 1–3 der 2. Fachtagung der Gesellschaft für Klassifikation e. V. Frankfurt–Höchst, 6–7. April 1978. Red. Wolfgang Dahlberg. – Frankfurt: Gesellschaft für Klassifikation, 1978. p. 129–146. (pp. 145–146)

Salton: Nem ismerem az ön problémájának az összetettségét, de úgy vélem, hogy a relációkat is beépíthetjük azokba a vektorokba, melyekkel az objektumokat leírjuk. Arra a kettősségre is gondolni kell, hogy nemcsak az objektumokat jellemzik az adatmátrixban szereplő tulajdonságok, hanem a tulajdonságokat is jellemzik az objektumok. Különösen dokumentumok esetén foglalkozunk ezzel.

Schulze: Hogyan oldja meg az objektumok és a jellemzők összekapcsolását: mennyiségileg vagy minőségileg? Úgy vélem továbbá, hogy az ön eljárása az objektumok sorrendjétől függ: nem okoz ez lényeges eltéréseket a végeredményben?

Salton: Mindenekelőtt ami a második kérdést illeti: kísérleti eredményeink az objektumok különféle mennyisége esetén arra utalnak, hogy a sorrend soha sem befolyásolta sem a teljességet, sem a pontosságot. A csoportképződések ugyan különbözhetnek a sorrend függvényében, de ez nem olyan fontos, mivel ezeknek a csoportoknak többnyire csak esztétikai és nem reálisan fontos a szerepük. Engem nem érdekel az osztályozás egyértelműsége.

Schulze: A felhasználónak azonban lényeges, hogy mindenekelőtt jó osztályozás jöjjön létre.

Salton: Mit jelent az, hogy „jó”? A „jó” azt jelenti, hogy az osztályozás használójának hasznos. És ha a felhasználó különféle osztályozásokkal ugyanazt éri el, akkor ezek egyformán jók a számára. Ami a második kérdést illeti: mi automatikus indexeléssel foglalkozunk. Azt kérdezzük, mi a jó „tér” és azt válaszoljuk: az a jó ismértér, melyben a dokumentumok jól megkülönböztethetők. Azokat a jellemzőket használjuk, melyek a teret a lehető legjobban kitágítják. Numerikus (nem bináris) adatokkal dolgozunk.

KAREN SPARCK JONES (1935)

A cambridge-i egyetem matematikai–nyelvészeti kutatócsoportjának (Cambridge Language Research Unit, CLRU) munkatársa, ahol – kezdetben *Paul Needham* irányításával – 1960 óta elsősorban a kulcsszó-osztályozással foglalkozott. Sparck Jones az automatikus szövegelemzés egyik legismertebb és publikációk dolgában legtermékenyebb kutatója. A szövegelemzés célja, hogy a dokumentum szövegét átalakítsa olyan formájúra, amely a gépi feldolgozásnak megfelel. E téren a kötetünk elején bemutatott *Fairthorne* mellett az 50-es években *Hans Peter Luhn* végzett úttörő munkát: *Luhn* a dokumentum szövegében szereplő szavak gyakoriságát vizsgálta, hogy meghatározza, mely szavak eléggé szignifikánsak ahhoz, hogy a dokumentumot a számítógépben reprezentálják. Az ilyen, a szövegből ki-

választott kulcsszavakból álló jegyzék minden dokumentumhoz elkészíthető. A szavak előfordulási gyakorisága a szövegben felhasználható a szignifikancia fokának jelzésére is, ez pedig egyszerű eszköz a kulcsszavak súlyozásához, és lehetővé teszi, hogy a dokumentumot „súlyozott kulcsszavas leírással" reprezentálják. A dokumentumon belüli szóeloszlás statisztikai elemzését többek között *Maron* fejlesztette tovább, aki a kulcsszavak között valószínűségi kapcsolatokat állapított meg. Ezeknek a kapcsolatoknak a tárolására születtek meg az automatikusan készített tezauruszok a 60-as évek elején.
Sparck Jones tovább ment, és kulcsszavak együttes előfordulásának gyakoriságát mérte (azaz azt vizsgálta, hogy bármely két kulcsszó ugyanazon dokumentumban hányszor fordul elő együtt). Kimutatta, hogy az így kapcsolt szavak jelentősen javíthatják a keresés teljességét. E kutatásai nyomán dolgozta ki a kulcsszóosztályozás elméletét.
A nyelvészet és az informatika összefüggéseivel foglalkozó, Marti Kayjal közösen írt monográfiája, a „Linguistics and Information Science" e két szakterület együttes művelésének egyik legfontosabb műve (egy másik ilyen jelentős könyvvel, William J. Hutchins nyelvészeti tanulmányával első kötetünkben foglalkozunk).[16]
Az alábbi tanulmányában közérthető formában foglalja össze a szövegelemzés és a kulcsszóosztályozás kérdéseit.

Karen Sparck Jonestől magyarul megjelent:

Gépi indexek

In: Könyvtári Figyelő, 1975, 21. köt., 5. sz., p. 545–552.
Eredeti: Karen Sparck Jones: Automatic indexes. In: Journal of Documentation, 1974, 4. sz., p. 393–432.

A történeti áttekintés és *Luhn* munkásságának ismertetése után a részleges, közepes és teljes (statisztikai, kvantitatív nyelvészeti értelemben vett) szintaktikai elemzés kérdéseivel foglalkozik. A teljesen automatikus indexelésre a SMART a példa. A szemantikai elemzés alapja a szavak statisztikus előfordulása. Beszámol a klaszterelemzésről és részletesen ismerteti az indexelés működő rendszereit, valamint az értékelési kísérleteket.

16 Sparck Jones, K. and Kay, M. Linguistics and Information Science. – New York: Academic Press, 1973.

Gondolatok az automatikus információkereséshez használt osztályozásról[17]

A dokumentáció és az információkeresés körén belül közhely, hogy a dokumentum leírásának, tárolásának és keresésének folyamatában az osztályozásnak is szerepet kell kapnia.

Bár az osztályozás fogalma ebben az összefüggésben nem új, szokatlan feladat, hogy dokumentumok osztályozását vagy szótárak készítését automatikusan lehessen elvégezni. Merőben gyakorlati problémákhoz vezet a nagyságrend, de az a tény, hogy az osztályozás megalkotása során a számítógép „fekete dobozként" működik, felvet néhány további érdekes kérdést is. Ezek egyike az, hogy milyen típusú osztályozást keressünk. Előfordulhat, hogy az osztályozó nem vizsgálja meg elég kritikusan azokat az elveket, amelyeken a rendszere alapszik, vagy talán meg sem fogalmazza azokat megfelelően, vagy nem alkalmazza következetesen. Az automatikus osztályozás durva sokkhatást okozhat, amikor váratlan eredményekhez vezet; a formális eljárás ugyanis a csoportosítási folyamat során már kizárja a módosítás vagy az elhagyás lehetőségét. A másik érdekes kérdés, hogyan aknázzuk ki az osztályozáselmélet lehetőségeit a keresőképhez releváns dokumentumok keresésének érdekében. Cikkünk válaszol ezekre a kérdésekre; de már az is nyereség, ha megvitatjuk a problémákat.

Tételezzük fel tehát, hogy az információkeresés céljából automatikusan működő tezauruszt kívánunk szerkeszteni. Azaz a kulcsszavakat úgy kívánjuk csoportosítani, hogy ha adott szó a keresőképben előfordul, az akkor is egyezéshez vezessen, ha a dokumentumban e szó valamelyik „megfelelője" fordul elő. Feltevésünk az, hogy amennyiben az osztályokat pontosan alakítják ki, ez azoknak a dokumentumoknak a megtalálásához vezet, amelyek ugyanarról szólnak, mint a keresőkép, ha másként is fejezik azt ki. Ez persze meglehetősen kézenfekvő feltételezés. Csak azért említjük, hogy legyen mihez kapcsolnunk az erről szóló általános gondolatmenetet. A továbbiakban ezért a szavak osztályozását példaként használjuk, de valamennyi fontosabb megállapításunk érvényes a dokumentumok osztályozására is, talán a részletekben némi módosítással.

Ha azt mondjuk, hogy kulcsszóosztályozást kívánunk kialakítani, ez szükségképpen vezet a kívánt osztályozás típusára vonatkozó alábbi gondolatokhoz. A kulcsszavakat úgy csoportosítjuk, hogy egyazon halmazba az egymást helyettesíthető szavak kerüljenek. A szinonimák nyilvánvaló példák a helyettesíthető kifejezésekre. Ha azonban osztályozásunkat automatikusan kívánjuk elvégez-

17 Some thoughts on classification for retrieval. In: Journal of Documentation. 26 (1970) 2. p. 89–101.

ni, nem vizsgálhatjuk a szavak jelentését annak eldöntésére, hogy helyettesíthetőek-e. Kénytelenek vagyunk a verbális kapcsolatot illetően valamilyen más információt keresni, amely a helyettesíthetőség jelzéseként elfogadható, s ugyanakkor alkalmas arra is, hogy a gép kezelje. Ilyen a szavaknak a dokumentumon belüli előfordulása és együttes előfordulása. Ha ugyanis két szó mindig együtt fordul elő, akkor bizonyosan helyettesíthetőek, hiszen bármelyikük ugyanazon dokumentumok megtalálásához vezet. Az ebből fakadó általánosítással olyan szavakból álló halmazokat kapunk, amelyek gyakran szerepelnek együtt. Fogalmazhatunk úgy, hogy olyan tematikus szóosztályok iránt érdeklődünk, amely szavak gyakran fordulnak elő azonos összefüggésben, és ahol ezt, ha előzetesen ismerjük a szavak előfordulását egy dokumentumgyűjteményben, remélhetőleg meg tudjuk határozni. A lényeges ebben a megközelítésben az, hogy az ilyen típusú ismeret tisztán automatizált manipulációk révén is egyszerűen megkapható.

Nem kívánok további részletekbe bocsátkozni, megállapításaim igazolására sem törekszem. Mostani célunkhoz elegendő annyi, hogy az automatikus tezauruszkészítéshez alapként rendszerint ezt szokták javasolni. Fontosabb ennél, hogy a fenti érvelésből meghatározott, politetikus (tulajdonságátfedő) és többszörös (dokumentumátfedő) osztályozás következik. Azaz olyan osztályozás, amelyben egyazon osztály valamennyi tagja nem rendelkezik szükségképpen egy vagy több közös tulajdonsággal, és amelyben az egyes elemek egynél több osztályba is tartozhatnak. E végkövetkeztetés okai eléggé nyilvánvalóak. Valószínűtlen ugyanis, hogy olyan szavak halmazaihoz jussunk, amely szavak valamennyien előfordulnak ugyanazon dokumentumban vagy dokumentumokban. Az, hogy olyan szavakat keresünk, amelyek hajlamosak az együttes előfordulásra azt jelenti, hogy olyan szavakból álló halmazokat keresünk, amelyek „osztoznak" a dokumentumok halmazán. Ilyenek például az a, b, c szavak, ha a, előfordul az 1. és a 2. dokumentumban, b, az 1.-ben és a 3.-ban, c pedig a 2.-ban és 3.-ban. Ez természetes következménye annak, hogy egy gyűjtemény dokumentumai, és, noha tartalmilag igen közel állhatnak egymáshoz, tárgyukat és szóhasználatukat tekintve nem azonosak. Az pedig, hogy megengedjük: a szavak egynél több osztályban szerepeljenek, nemcsak azt tükrözi, hogy még a műszaki kifejezések is rendelkezhetnek eltérő jelentésekkel egy specializált gyűjteményben, de azt is, hogy egy szó, azonos jelentéssel különböző összefüggésekben használható.

Továbbmenve úgy véljük, hogy semmi sem indokolja egy adott kifejezés helyettesi körének nagyfokú kiterjesztését – nem törekszünk ***rendezett*** osztályozásra. Rendezettnek az olyan osztályozást nevezzük, amelyben az osztályok között szisztematikus kapcsolatok állnak fenn, s ennek következményeként egy adott kifejezés lehetséges helyettesítőinek a száma növelhető. Esetünkben azonban úgy tűnik, hogy noha meg kívánunk engedni némi helyettesítést, nem akarunk egyazon szóhoz túl sok alternatívát. De akár ki is jelenthetjük, hogy

rendezetlen osztályozásra van szükségünk, minthogy abban az esetben sem jutnánk kielégítő eredményre, ha szerkesztenénk egy rendezett osztályozást, majd eltekintenénk ennek szerkezetétől.

Ily módon a kívánt osztályozás típusát, általános formális tulajdonságait illetően eléggé általános megállapításhoz jutottunk el. Azt mondhatjuk, hogy politetikus, többszörös, rendezetlen osztályozást keresünk. Másként fogalmazva, olyan osztályozást, amely jellegében igen összetett, s ugyanakkor szeretnénk azt nagy tárgyhalmazokra alkalmazni. Mégsem a nagyságrendi probléma a legfontosabb, bár a gyakorlatban elég kellemetlen lehet, s nem hagyható figyelmen kívül az osztályozási algoritmus megválasztásakor sem. Az igazi nehézségek azonban (a) az általunk keresett osztályozási típussal és (b) általánosságban magával az osztályozással kapcsolatosak. Ami az első problémát illeti, minden olyan kísérlet során, amikor meghatározott típusú osztályozást akarunk létrehozni, szembetaláljuk magunkat a megfelelő osztályozási módszerek, mi több, olyan programozható algoritmusok hiányával is, amelyek meghatározott osztályokhoz vezetnek. A második problémakörben az értékelés nehézségével találkozunk, amely ugyan bármilyen jellegű osztályozáshoz kapcsolódik, de különös hangsúlyt kap számítógép használatakor. Általa döntjük el, hogy egy osztályozás jó-e, vagy több osztályozás közül melyik a legjobb.

Kijelenthetjük viszont, hogy noha hiányoznak a megfelelő osztályozási módszerek, amelyek közül választhatnánk, nem igaz, hogy egyáltalán nincsenek ilyenek. *Needham* „clumpokra" (a klaszterek egyik fajtájára) vonatkozó elmélete például eredetileg információkeresési kívánalmaink kielégítése érdekében született meg. Az alternatívák köre azonban nem nagy, s a szóban forgó módszerek helyzete, egymáshoz való viszonya sem világos. Így végül is nincs nyilvánvaló választásunk. Hasonlóképpen mondhatjuk, hogy az információkeresésben rendelkezünk az osztályozás értékelését célzó eszközzel, amelynek révén megtudhatjuk, hogy megfelelőek-e a megszokott információkeresési teljesítménymértékek, a teljesség és a pontosság értékei. A szükséges kulcsszóosztályozás célja és az ennek kialakítása érdekében alkalmazott technika közötti kapcsolat azonban egyáltalában nem tisztázott. Ugyanez áll a dokumentumok osztályozására is. Az információkeresésre szolgáló osztályozás automatikus megszerkesztése ily módon meglehetősen általános kérdéseket vet fel magával az osztályozással kapcsolatban. A következőkben annak reményében foglalkozunk velük, hogy speciális feladatunk megoldásához kaphatunk valamiféle segítséget.

Vizsgáljuk tehát először röviden az osztályozást szélesebb összefüggésében, szembeállítva másfajta adatfeldolgozás-típusokkal. Majd a számunkra érdekes típusú osztályozást más típusúakkal állítjuk szembe.

Az osztályozást általában olyan folyamatként írhatjuk le, amely csoportokba sorolja a tulajdonságaikat tekintve egymásra hasonlító tárgyakat. Ez az igaz

megállapítás azonban túlságosan tág ahhoz, hogy használhassuk: aprópénzre kell váltani. Ugyanakkor sokkal használhatóbb és némileg pontosabb állítás, hogy az osztályozás egyidejűleg információvesztő és információnyerő folyamat. Másként fogalmazva, kiindulunk bizonyos, tárgyakra és tulajdonságaikra vonatkozó tapasztalati tényekből, például, hogy a, tárgy rendelkezik az 1, 2 és 4 tulajdonságokkal, b, tárgy a 2, 3, 4 és 5 tulajdonságokkal, hogy a- nak az 1 és 2 közös tulajdonsága c-vel, 2 és 4 meg d-vel, hogy b, a 2 és 3 tulajdonságokban megegyezik c-vel, míg az 1, 2 és 3 tulajdonságokban d-vel, és így tovább. Feladatunk az, hogy ezeket a részletező információkat becseréljük arra az általánosabb megállapításra, hogy ezek a tárgyak valamennyien hasonlóak, vagyis az a, b, c és d tárgyak ugyanabba az osztályba tartoznak, mert vannak közös tulajdonságaik, tekintet nélkül arra, hogy ezek milyen módon közösek. Ily módon veszítettünk, vagy inkább elvetettünk információt, mert az egyes tárgyak közötti sajátos tulajdonsági viszonyokat elhanyagoltuk. Ugyanakkor nyertünk is, mert explicit módon fejezzük ki azt a tényt, hogy egyes tárgyak hasonlóak. Ez a tény persze benne volt az eredeti adatokban is. Az osztályozás célja azonban az, hogy ezt akkor is bemutassa, ha egyébként nem nyilvánvaló. Elérhetünk azonban lényegesen fontosabb információnyereséget is. Mégpedig, hogy egy osztály minden egyes tagját ezentúl úgy kezelhetjük, mint ami bizonyos, az osztályra jellemző tulajdonságokkal rendelkezik, még akkor is, ha nem tudjuk, hogy ez eredetileg így volt-e. Az a tény, hogy egy tárgy egy osztály tagja, lehetővé teszi számunkra e következtetés levonását.

A tezaurusz esetével illusztrálhatjuk is ezt. Kiindulásként csak azt tudjuk, hogy meghatározott szavak előfordulnak meghatározott dokumentumokban. Azután azt mondjuk, hogy bizonyos szavak hasonlóak, egyazon osztály tagjai, mert hajlamosak arra, hogy ugyanazokban a dokumentumokban együtt forduljanak elő. Végezetül pedig az osztály tagjait egymással helyettesíthetőként tekintve úgy tehetünk, mintha valamennyi vizsgált szó előfordulna minden olyan dokumentumban, amelyben az osztály bármelyik tagja előfordult. A szavak csoportosítása tehát azt jelenti, hogy feltételezzük: valamennyi egyformán előfordulhatott volna mindazokban a dokumentumokban, amelyeken az osztályt létrehoztuk.

Ez a gondolatmenet kicsit többet mond arról, mi is az osztályozás, balszerencsés módon azonban azok az adatkeresési módok, amelyeket osztályozóknak nevezhetünk, még mindig nagyon változatosak. A nehézség pontosan ezek egymáshoz való viszonyításában, s az osztályozás általánosított, használható és tartalmi jellemzésében rejlik. Mindazonáltal azt mondhatjuk, hogy egyes tevékenységfajták nem eredményeznek osztályozást abban az értelemben, amelyet megpróbálunk kifejteni. Az egyik ilyen tevékenységi területre jó példa az olyan arányosítási technika mint *Shepherd-Kruskal* többdimenziós eljárása. Az arányosítási technikákat az adattípusú osztályozási módszerek egyszerűsítésére hozták létre, de ezek nem szétválasztóak. Mondhatjuk tehát, hogy az osztályo-

zás elvégzi azt a szétválasztást, amelyet az arányosítás óvatosan elkerül. Egy másik, nyilvánvalóan összevethető területet képvisel a faktoranalízis. Itt már közelebb kerülünk az osztályozáshoz, mint az arányosításnál, de a faktoranalízis eredményeiben az osztályozás sokkal inkább implicit, semmint explicit. Az arányosításnál és a faktoranalízisnél egyaránt elvetünk információkat, de helyettük nem nyerünk annyit, mint az osztályozás esetében.

Mindez azonban az osztályozás jellemzésének csak negatív megközelítése. Ha pozitívabbra törekszünk, azt látjuk, hogy az osztályokat bizonyos általános tulajdonságok vagy jellemzők (fenntartva a „tulajdonság" szót a tárgy lényeges jegyeire) szerint kategorizálhatjuk. Ezek a jellemzők három kérdésre adható válaszok segítségével határozhatók meg, amely kérdések viszont minden osztályozást illetően feltehetők. Az első kérdés a tárgyak (objektumok, dolgok) tulajdonságai és az osztályok közötti viszonyt érinti. A válasz szerint osztályaink lehetnek monotetikusak vagy politetikusak. Ha az osztály monotetikus, ez azt jelenti, hogy valamennyi tagja rendelkezik ugyanazon közös tulajdonsággal vagy tulajdonságokkal, ami politetikus osztályoknál nem igaz. A második kérdés a tárgyak és az osztályok közti kapcsolatra vonatkozik, s e válasz szerint osztályaink lehetnek kizáróak vagy átfedők. Ha a tárgyak csak egyetlen osztályba tartozhatnak, kizáró osztályaink vannak, ha viszont egynél több osztálynak is tagjai lehetnek, osztályaink átfedők. A harmadik kérdés az osztályok közötti kapcsolatokat érinti, s ennek révén rendezett vagy rendezetlen osztályozást kaphatunk. Az, hogy egy osztályozás rendezett, azt jelenti, hogy az osztályok szisztematikus kapcsolatban állnak egymással, ami a rendezetlen osztályozásra nem érvényes.

Tehát három szempont szerint szétválasztott alternatívapárt nyerünk, amelyek az alábbi módon írhatók le:

1. Tulajdonságok és osztályok közötti kapcsolat szempontjából létezik
   (a) monotetikus (nincs tulajdonságátfedés)
   (b) politetikus (tulajdonságátfedő)

2. Tárgyak és osztályok közötti kapcsolat szempontjából létezik
   (c) kizáró (egyszeresen besoroló)
   (d) átfedő (többszörösen besoroló)

3. Osztályok és osztályok közötti kapcsolat szempontjából létezik
   (e) rendezett (osztályokat egymáshoz képest valamilyen elv szerint elhelyező)
   (f) rendezetlen (az osztályok felsorolása nincs valamilyen elvhez kötve).

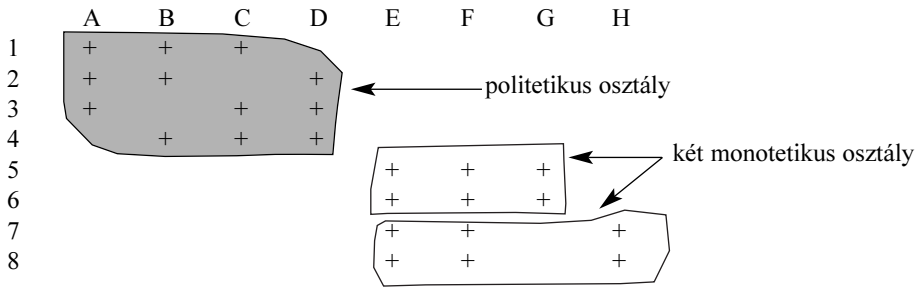
Úgy gondolom, hogy az osztályozás általános jellemzését illetően ez a hat jellemző kimeríti a lehetőségeket. (Ám ez a megjegyzésünk azokra az elvekre vonatkozik, amelyeken a csoportosítás alapul, s nem a tényleges csoportokra: Egy olyan eljárás, amely átfedő osztályok kialakítását célozza, adott tárgyhalmaz esetén létrehozhat kizárókat is.)

Ez a hat jellemző nem meríti ki a lehetőségeket. *Karen Sparck Jones* csak a numerikus osztályozási rendszerek tipológiáját adta meg. Más szóval kizárólag olyan szempontokat vett figyelembe, melyek a matematikai eszközöket használó osztályozó számára relevánsak. Ezek közül vannak, melyek az intellektuális, „fogalmi" osztályozási rendszerek tipológiájában is használhatók (mint például a 2. és a 3. csoport), és vannak, melyek gyakorlatilag nem jönnek számításba az utóbbi esetben (az 1. csoport).

A politetikus osztályozás megértéséhez vegyünk egy olyan K csoportot, melyet az f_1, f_2,... f_n tulajdonságok G halmaza jellemez a következőképpen:

(1) minden egyed rendelkezik a G-beli tulajdonságok közül sokkal (de nem meghatározott számúval);
(2) minden G-beli tulajdonsággal sok egyed rendelkezik;
(3) egyetlen olyan G-beli tulajdonság sincs, amellyel a csoport valamennyi tagja rendelkeznék.

Mindezt táblázatban szemléltethetjük:



Ha például megengedett, hogy a „kutya" mind a RAGADOZÓK, mind pedig a HÁZIÁLLATOK közé besorolható, akkor átfedő osztályozásról beszélünk. Ha nem megengedett, akkor kizáró az osztályozás. Ez a két típus lényegében megfelel a fogalmi osztályozás poli- és monohierarchiájának (monohierarchikus osztályozási rendszerre az ETO, polihierarchikusra általában a deszkriptornyelv a példa).

A rendezett osztályozásra példa minden hierarchikus osztályozás (pl. ETO). A rendezetlen osztályok nem kapcsolódnak egymáshoz, rendszerint az automatikus tezaurusz-szerkesztés során bukkannak föl (intellektuálisan készített osztályozási rendszer esetében a rendezetlen formájú rendszernek nincs sok értelme).

Fordítva is igaz: vannak szempontok, melyek alapján meghatározható osztályozási típusok az automatikus osztályozás szempontjából irrelevánsak:

[4] Az elemzés tárgya szempontjából létezik:
  [h] Természetes osztályozás, mely az elemzett dolgok tulajdonságain alapul (pl. az elemek periódusos rendszere) (ebben az értelemben minden automatikus osztályozás természetes).
  [i] Absztrakt osztályozás, mely gondolati koncepción alapul (pl. ETO).
[5] A származtatott nyelv szempontjából létezik:
  [h] Mesterséges nyelven alapuló (pl. ETO).
  [i] Természetes nyelven alapuló (pl. deszkriptornyelv).
[6] Az alkalmazás szempontjából létezik:
  [j] Prekoordinált, a várható osztályokat előre meghatározó (pl. ETO).
  [k] Posztkoordinált, az osztályt az osztályozáskor elemi összetevőkből (elemi osztályokból) megszerkesztő (pl. Uniterm, deszkriptornyelv).
[7] Az elemzés módszere szempontjából létezik:
  [l] Generalizáló, az osztályozási rendszer által rendezendő univerzum egészére irányuló, annak rendszerét általánosító, szintetizáló (pl. ETO).
  [m] Individualizáló, az osztályozási rendszer által rendezendő univerzum részleteit számba vevő, annak rendszerét elemi szinten analizáló (pl. deszkriptornyelv).

Ugyanakkor a fenti lehetőségek bármely kombinációja osztályozást eredményez. Eszerint az osztályozások a fenti jellemzők lehetséges kombinációinak megfelelően az 1. táblázatban látható nyolc típusba sorolhatók.

	1		2		3	
	a	b	c	d	e	f
I	x		x		x	
II	x		x			x
III	x			x	x	
IV	x			x		x
V		x	x		x	
VI		x	x			x
VII		x		x	x	
VIII		x		x		x

1. táblázat

Természetesen azt szeretnénk, ha minden, ami megérdemli az osztályozás nevet, megtalálható lenne a fenti táblázatban, és hogy semmilyen más

tárgyhalmaz kezelés ne legyen osztályozás. Ezt azonban nem könnyű bizonyítani. Az egész kategorizálás a tárgy, a tulajdonság és különösen az osztály előzetes fogalmán alapul, amit viszont nem definiáltunk, csak feltettünk.

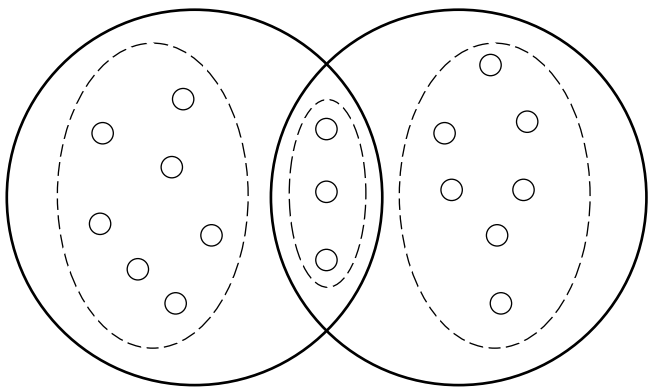
Megint más kérdés, hogy még ha tudnánk is, mik azok a tárgyak, tulajdonságok és osztályok, s ha abban a helyzetben lennénk is, hogy bizonyíthassuk, az osztályozásokat a fenti kategóriák lefedik, mindez nem volna túlságosan informatív, mert még mindig az általánosság túlságosan magas szintjén vagyunk. Mind a nyolc osztályozástípushoz több konkrét osztályozási módszer vagy osztálymeghatározás tartozik, és ez az, amire most figyelmünket összpontosítani kell. Osztályozásról beszélve éppen három vonatkoztatási szintet különböztethetünk meg. Általánosságban beszélhetünk különféle osztályozásokról az előbb bemutatott séma szerint. Ahhoz azonban, hogy meghatározott típusú osztályozáshoz jussunk, kell valami, amit osztályozási módszernek vagy osztálymeghatározásnak nevezhetünk, ami pontosan meghatározza azt az alapot, amelyre támaszkodva a tárgyakat csoportosítjuk. Végezetül pedig kell olyan algoritmus, amellyel meghatározott adatokhoz adott meghatározáson alapuló osztályokat találhatunk. Itt lényeges az, hogy különféle módszerek vagy meghatározások azonos fajta osztályozást állíthatnak elő, s továbbmenve, különféle algoritmusokkal lehet azonos meghatározáson alapuló osztályozásokat létrehozni.

Ezeket a megkülönböztetéseket példával illusztrálhatjuk. Tegyük föl, hogy politetikus, többszörös (átfedő), rendezetlen osztályokat akarunk. Ha ezt kiindulásként leszögezzük, definiálhatunk osztályt akár egymással kapcsolatban álló tárgyak halmazaként is, abban az értelemben, hogy minden tárgypárnak legalább egy közös tulajdonsága van; vagy meghatározhatunk osztályt úgy, mint olyan halmazt, amelyben több a tagok közötti (belső), mint a tagok és nem-tagok közötti (külső) kapcsolat. Kapcsolaton ismét egy vagy több közös tulajdonságot értünk a szóban forgó tárgypár tagjai között. Végül, ha ez utóbbit választjuk, megkereshetjük az osztályokat úgy, hogy egyetlen tárgyból kiinduló halmazhoz adunk további tárgyakat, vagy olyan folyamatot alkalmazhatunk, amely a tárgyak összességének kezdeti, véletlenszerű szétválasztását korrigálja mindaddig, míg a választott oldalon lévő tárgyak ki nem elégítik kívánalmainkat.

Ezen a ponton hangsúlyoznunk kell: az, hogy adott tárgyhalmazhoz ilyen vagy olyan típusú osztályozást alkotunk, semmiféle kapcsolatban sincs a szóban forgó tárgyak természetével. A tárgyak összességének nincs egyetlen egy helyes vagy természetes osztályozási módja. Azért szükséges ezt hangsúlyozni, mert sokszor hallani olyan megjegyzéseket, mintha bizony az osztályozás a tárgyak halmazának belső lényegéből tükrözne vissza valamit, holott sokkal inkább az osztályozást igénylő személy viszonyulásait tükrözi. Teljesen jogosan mondhatjuk, hogy adott adattömegre elvileg bármilyen osztályozási módszert alkalmazhatunk, és ennélfogva bármilyen típusú osztályozáshoz eljutha-

tunk. Így az 1. ábrán, ahol a tárgyak közti hasonlóságot kétdimenziós távolság reprezentálja, a szaggatott és a folytonos vonalak két, egyaránt ésszerű, ám eltérő alapokon álló osztályozást körvonalaznak. A szaggatott vonalas osztályozás abból indul ki, hogy az osztályoknak kizáróknak kell lenniük, míg a folytonos vonal az osztályok közötti átfedés megengedésének következménye. Az alternatívák azonban kizárólag az osztályozó eltérő érdekeihez kapcsolódnak: a tárgyakban semmi olyan nincs, ami indokolná az egyik osztályozás kiválasztását a másikkal szemben.

Az a tény, hogy ugyanazt a tárgyhalmazt többféleképpen lehet osztályozni, különösen fontos, ha az értékelés kérdéséhez érkezünk. Ha kialakítunk egy osztályozási eljárást, természetesen feltesszük a kérdést, hogy jó-e. A gyakorlatban pedig többször alakítunk ki alternatív osztályozásokat ugyanahhoz az adattömeghez, hogy megkérdezhessük, melyik a legjobb. Hogyan egyeztethető össze azonban az, hogy egyfelől ugyanazt a tárgyhalmazt többféleképp is tökéletesen jól osztályozhatjuk, másfelől valamelyik eredményről azt állítjuk, hogy az a legjobb?



1. ábra

Az osztályozásokat lényegében két különböző alapon értékelhetjük. Egyrészt teljesen formalizált osztályozási elmélet alapján: ha például adott tárgyhalmazt bizonyos formai jellegzetességek segítségével azonosítunk, olyan következtetésre juthatunk, miszerint ezeket az adatokat ilyen és ilyen osztályozás sérti meg a legkevésbé. Másrészt az osztályozás céljának teljesen formalizált leírásával, olyan megfogalmazásban, amely elvezet a megfelelő módszer kiválasztásához. Az első esetben torzítási mérték társul az osztályozáselmélethez; ha célunk csupán olyan osztályozás megszerkesztése, amely a tárgyhalmaz statikus leírását adja, akkor e torzítási mérték alapján a csoportosítási módszert összevethetjük adatainkkal. Az eredmény szükségképpen az adatok legjobb osztályozó leírása lesz. A második esetben azt mondhatjuk, hogy az a cél,

amelyhez az osztályozást keressük, önmagában tartalmaz vagy maga után von torzítást, így azután olyan osztályozási módszert keresünk, amely ezt a torzítást minimálisra csökkenti. Az eredmény ebben az esetben a meghatározott célnak leginkább megfelelő osztályozás.

Lényeges azonban, hogy megkülönböztessük a kétfajta osztályozást: az elvont folyamat alapján keletkező, csupán bizonyos belső feltételeknek eleget tevő osztályozást attól az osztályzástól, mely külső követelményeket kielégítő, célra adott választ képvisel. Mindkét esetben csak akkor lehetünk biztosak eredményünkben, ha rendelkezésünkre áll az egész leírás formális apparátusa. A probléma viszont éppen az elvont osztályozási kívánalmak megfogalmazása az első, illetve a feladat kívánalmainak újrafogalmazása osztályozási kifejezésekkel a második esetben. (Feltételezve, hogy egyáltalán meg tudjuk határozni a feladat kívánalmait, ami önmagában is probléma.) Valóban, mert ha elvileg igaz is, hogy adott feladat kívánalmai fölébe kerülhetnek a formalizált osztályozáselméleti követelményeknek, nagyon is valószínűtlen, hogy a feladatnak eleget tegyen olyan osztályozási módszer, amely nem elégíti ki egyúttal az osztályozáselmélethez kapcsolódó elvont követelményeket is. Ezt persze a teljességre törekvés mondatja. A gyakorlatban hasznosítható eredmények születnek elméletileg meglehetősen hibás csoportosítási eljárásokkal is. Ám céljainkat tekintve bizonyosan jobbak lesznek eredményeink, ha elméletileg kielégítő módszereket alkalmazunk.

Sajnálatos módon azonban nem rendelkezünk teljesen formalizált adat- és osztályozáselmélettel, s ugyanígy igen sok esetben nem elegendően pontos a feladat-meghatározás sem. Sokan álmodtak meg sajátos csoportosítási módszereket, főként biológiai tárgyak kizáró és rendezett osztályozásainak létrehozására, de nagyon kevés történt e módszerek értékelésére, vagy éppen az általános osztályozáselméleten belül a szigorú és hatékony összehasonlításra. Az e névre valóban méltó osztályozáselméletben legalább olyan kritériumokra van szükség, amelyek minden osztályozáselméletnek eleget kell hogy tegyenek, és amelyek azután felhasználhatók arra, hogy a különféle javasolt módszerekben válogassunk. Nehéz azonban alkalmas kritériumokat találni, s azt is nehéz lenne kimutatni, hogy valamilyen adott módszer ezeknek eleget tesz. A valóban hatékony osztályozáselméletnek azonban ezen is túl kellene mennie, megmutatva, hogy különböző osztályozási módszerek hogyan viszonyulnak egymáshoz, hogy azután a kiválasztott módszer következményei előre jelezhetők legyenek.

Ez láthatóan óriási feladat, különösen ha arra gondolunk, hogy az osztályozáselméletnek tartalmaznia kellene a formalizált adatelméletet is, továbbá a hasonlósági vagy különbözőségi mértékek alkalmazását. Utóbbiakat kellene ugyanis alkalmazni a kiinduló tárgy–tulajdonság információkra, hogy a tárgypárok közötti kapcsolatok tényleges megállapításával a csoportosítási eljárás inputját megkapjuk. Valamelyes kutatás folyt már e problémával kapcsolatban, de nagyrészt rendezett osztályozásokat szerkesztő módszerekhez illeszkedve, ami viszont aránylag könnyen kezelhető részterület. Így például *Jardine* és *Sibson*

megfogalmaztak egy sor osztályozáselméleti kritériumot, és ezek tükrében vizsgálták a rendezett csoportosítási technikákat. Valamit tudunk a hasonlósági és különbözőségi mértékekről, együtthatókról és ezek rokonairól is. De azoknak, akik a rendezetlen meghatározások iránt érdeklődnek, mint mi az információkeresés esetében, a mostani helyzet semmiképp sem kielégítő. Nem nehéz belátni, hogy egyes javasolt módszerek lehetnek elméletileg nagyon elégtelenek – például a „clumpok" (bogok, nyalábok)[18] meghatározásai –, nincs azonban olyan útirány, amelyről előre tudhatnák, hogy vélhetően valami jobbhoz vezet.

Én mindenesetre nem tudok ezúttal semmi elfogadhatóbbat adni. Ezért úgy vélem, hogy a megfelelő módszerek megtalálására irányuló kísérletek problémájának legjobb megközelítése az, ha kicsit előbbre nézünk. Arra, hogy mit is várunk el egy osztályozáselmélettől, majd pedig újra információkeresési céljainkat vesszük szemügyre.

Ha például a megfelelőségi kritériumok kérdéséről van szó, milyen követelményeket támaszthatunk egy osztályozási eljárással szemben? *Jardine* és *Sibson* jegyzéke – a formalizációból kissé nehézkesen kibontva és általánosítva – az alábbi kritériumokat tartalmazza:

1. adott adattömeghez egyértelmű eredményt adjon;
2. ha az adatok már osztályozottak, ezt az osztályozást őrizze meg;
3. az eljárás legyen független az adatok megnevezésének módjától;
4. a módszer legyen skálafüggetlen;
5. a módszer csökkentse minimálisra a torzítást;
6. az eredményt az adatok kis változtatása ne érintse radikálisan; és
7. maximálisan összefüggő tárgyhalmazok ne válhassanak szét.

Ezek a kívánalmak nagyon ésszerűeknek tűnnek, bár meg kell jegyeznünk, hogy feltételezik például a kielégítő torzulási mérték létét. Emellett bizonyos további kívánalmaktól sem tekinthetünk el, közvetlenül a hasonlósági vagy különbözőségi együtthatókat illetően, végső soron pedig az adatokkal kapcsolatban. Így a hasonlósági együtthatóknak az alábbi kívánalmakat kell teljesíteniük:

1. az azonos tárgyak között maximális legyen a hasonlóság; és
2. az egymást kiegészítő tulajdonságeloszlású tárgyak között minimális legyen a hasonlóság.

Ami az alapadatokat illeti, olyan követelményeket határozhatunk meg, amelyeket ki kell elégíteni ahhoz, hogy az adatok elfogadhatók legyenek az osztályozási folyamat inputjaként, például:

1. a tulajdonságmegjelölések legyenek egyértelműek;
2. a tulajdonságok legyenek függetlenek egymástól.

18 A klaszterek egyik fajtája (a szerk.).

Mindemellett kell használati útmutató a torzítási mértékhez, amely a hasonlóságszámítással és a csoportosítási folyamattal egyaránt összefügg. Ez adatosztályozhatósági mértékként volna kezelhető. Torzítási mértékünk ugyanis azt mondja meg, hogy az osztályok mennyire torzítják az adatokat, de mi azt is tudni szeretnénk, hogy a csoportok megtalálására irányuló kísérletünk nem hibás-e eleve, ha ugyanis véletlenszerű adatokkal vagy egymástól idegen tárgyak halmazával állunk szemben, akkor nincs mit megtalálnunk.

Bár a csoportosítási eljárás egészének mindezeket a kívánalmakat ki kell elégítenie, ebből még nem következik, hogy minden ezeket kielégítő eljárás egyúttal osztályozási folyamat is. Ezek az osztályozásnak csak szükséges, de nem elégséges feltételei. A fenti kívánalmak szavatolják, hogy minden megtalált osztály megfelelő lesz, de nekünk azt is biztosítani kell, hogy egyáltalán találjunk osztályokat. Más szavakkal: mitől lesz egy adatfeldolgozási technika osztályozási módszer, és nem – teszem azt – szortírozó eljárás? Itt jutunk el ahhoz az igazi problémához, mely az osztályozás fogalmának kiterjesztésével jár, és erről nagyon nehéz bármi használhatót mondani. Azt kimondhatjuk, hogy az osztályozás három, egymástól elhatárolható gondolatot tartalmaz. A tárgyak összességét fel kell osztanunk; ezt pedig úgy kell megtennünk, hogy azokat a részhalmazokat, amelyekbe a tárgyak kerülnek, a tagjaik közti hasonlóság tartsa össze, és hogy a tárgyak így keletkező leírása, az osztályhoz való tartozással kifejezve, egyszerűbb legyen, mintha eredeti tulajdonságaikkal írnánk le őket. Rendkívül nehéz azonban a fenti megállapításokat pontosítani. Azt érzékeljük például, hogy tíz tárgy olyan osztályozása, amely tíz, egyenként egy-egy elemből álló részhalmazhoz mint osztályhoz vezet, nem osztályozás a szó valódi értelmében. Általánosságban is, ha ugyanannyi osztály van, ahány tárgy, ez nem kielégítő. De vajon hogyan határozzunk meg valamilyen egyszerűségi vagy szétválasztási kritériumot, hogy ez ne tiltson valamely, az adattömeghez már illeszkedő osztályozást? Végül is lehet tíz olyan tárgyból álló halmaz, amelyet az említett osztályozás pontosan reprezentál. Továbbá, bár a csoportosítási eljárásokat úgy szeretnénk meghatározni, hogy a megszerkesztendő osztályozás a korábban említett típusok valamelyikéhez tartozzék, ezzel meg nincs szükségképpen olyan módszer a kezünkben, amely kielégíti az éppen most tárgyalt kritériumokat.

Az önálló osztályozási elmélet hiánya, s még inkább az a gyanú, hogy az eléggé átfogó elmélet voltaképp tartalmatlanná válna, meglehetősen lehangoló. Az se valami szívderítő, ha meghatározott – s különösen az információkeresési – célú osztályozásokkal kapcsolatos problémákkal foglalkozunk. Ilyenkor az a feladatunk, hogy elsősorban pontosan megfogalmazzuk a célunkat; azután pedig le kell fordítanunk mindezt osztályozási kifejezésekre. Néhány feladat azonban nem alkalmas arra, hogy pontos kifejezésekkel közelítsük meg, más feladatok viszont, bár sokkal pontosabban meghatározhatók, na-

gyon nehezen fordíthatók le. Ha például osztályozni próbáljuk a könyveket méretük szerint, hogy a raktári férőhelyet gazdaságosan használhassuk ki, megtartva ugyanakkor valamilyen mértékig a tartalmi csoportosítást is: ez sokkal meghatározottabb kívánalom, mint az előbbi, mégis nagyon nehéz formalizált osztályozási kifejezésekre lefordítani.

Ebben az összefüggésben az információkeresésre való alkalmazás különösen érdekes, mert ez minden szinten problémákat vet fel. Elsőként az információkeresést optimalizáló kulcsszavak osztályait kutatjuk, vagyis azokat a kifejezéshalmazokat, amelyek helyettesíthető elemekből állnak, mert ezek segítik a releváns dokumentumok megtalálását. A relevancia azonban szubjektív fogalom: amelyet az egyes használók ítélnek meg. Megfigyelhető viszont, hogy a használók statisztikai konzisztenciával viselkednek. Változatlanul igaz, hogy a relevancia a dokumentum tartalmához, tárgyához kapcsolódik, s ez számunkra hozzáférhetetlen. Nekünk, mint korábban láttuk, a helyettesíthető kifejezések osztályait azokra az információra alapozva kell létrehoznunk, melyek a dokumentumon belüli kulcsszó-előfordulásokról a rendelkezésünkre állnak, mert ez az, amivel rendelkezünk, a szavak relevanciatulajdonságainak megközelíthetetlen jellemzői helyett. Így abban a helyzetben vagyunk, hogy nem fejezhetjük ki célunkat közvetlenül, csak indirekt megállapítást tehetünk. Csakhogy az egész folyamat rendkívül bizonytalan, mert ha osztályozásunkat az információkeresés hatékonyságát illetően hiányosnak találjuk is, nem világos, hogyan korrigáljuk azt. További nehézség lép fel amiatt, hogy pusztán az a kijelentés, mely szerint a dokumentumok szövegén alapuló helyettesíthető kifejezések osztályaira van szükségünk, nem túlzottan nagy segítség a csoportosítási eljárás megtalálásához. Nem elég pontos. Nem mondja meg például, hogy szorosan vagy lazán összefüggő osztályokat akarunk. Kellemetlen helyzetben vagyunk, hiszen még ha létezik is az osztályozási módszerek olyan jól szervezett halmaza, amelyhez célkövetelményeinket viszonyíthatjuk, nem áll rendelkezésünkre kielégítően specifikus kívánalom. Így azután ha azt is figyelembe vesszük, hogy az osztályozási rendszerek jól szervezett halmaza is hiányzik, s csupán rosszul kipróbált javaslatok csoportjával rendelkezünk, világossá válik, hogy információkeresés céljára automatikus osztályozásba fogni ugyanolyan, mint sötétben mezítláb mocsárba gázolni.

Mi legyen hát munkamódszerünk, ha feltételezzük, hogy az automatikus információkeresés céljából végzett osztályozás iránti igény megmarad, sőt, növekszik? Egyetlen lehetőség a kompromisszum. Egyfelől törekedni kell az osztályozáselmélet fejlesztésére, legalább azon a részterületen, amely a számunkra különösen érdekes típussal foglalkozik; ugyanakkor a lehető legtöbb kísérletet kell elvégezni, a lehető legrendszeresebb módon, hogy megpróbáljuk tisztázni igényeinket. Ennek az a haszna, hogy felismerhetjük, milyen legyen az információkereső célú osztályozás, és – remélhetően – fogalmat alkothatunk a jobb osztályozási eljárásokról is.

CORNELIS VAN RIJSBERGEN (1943)

Cornelis van Rijsbergen a dublini University College, majd a Glasgow-i Egyetem számítástechnikai tanszékének professzora; elméleti szakemberként maga is alapkutatásokat végzett a tárgykörben, és a matematikai mértékelmélet felhasználásával kidolgozott egy, a teljességen és pontosságon alapuló értékelési elméletet. Monográfiája – melyben (hangsúlyozzuk) ***csak az automatikus információkereső rendszerekkel*** foglalkozik – magyarul is megjelent:

Információ-visszakeresés

[közr. az] Országos Széchényi Könyvtár Könyvtártudományi és Módszertani Központ. – Budapest: Múzsák Közművelődési Kiadó, 1987. 187 p.
Eredeti: Information retrieval. – London: Butterworts, 1979. 201 p.
A magyar kiadást ismerteti és fordítást kritikailag értékeli Roboz Péter. In: Könyvtári Figyelő, 1986, 34. köt., 2–3 sz., p. 221–222.

Az információkereső rendszereknek[19] a gyakorlati könyvtárost, dokumentátort és keresőt érintő részéről a könyvben nem sok szó esik; annál alaposabban tárgyalja az automatikus indexelés és osztályozás elméleti kérdéseit: az automatikus szövegelemzést, az automatikus osztályozást, a fájlstruktúrákat, a keresési stratégiák logikai mechanizmusát (pl. összevetési függvényeket, klaszter-reprezentatívokat, visszacsatolást), a valószínűségi keresést (pl. a legjobb összefüggési fa kiválasztását, az egy indexkifejezés szétválasztó erejét), az értékelést, végül pedig megpróbál a jövőbe is tekinteni.
A gyakorlati kereső, és a kérdésnek inkább a tartalmi–értelmi része iránt érdeklődő a konkrét levezetésekből feltehetően csak nehézségekkel profitál. De minden fejezet tartalmaz átfogó és értékelő összegezést a megjelent szakirodalomról, a fejezetek elején pedig mindig található utalás a történeti előzményekre, ami a nem „numerikus" érdeklődésű olvasó számára is fontos információkat tartalmaz. Különösen a bevezetés és az automatikus szövegelemzésről szóló első fejezet tarthat általános érdeklődésre igényt.

19 A kereséssel összefüggő teljes folyamatot jelentő „retrieval" kifejezésnek – melybe beleértjük a keresőkérdés elemzését, a keresőprofil és a keresési stratégia és taktika kialakítását, a tárolóban végzett keresőműveleteket, és a találatok képzését és kiadását – magyarul az „információkeresés" kifejezés felel meg. Korábban elterjedt a „visszakeresés" és az „információ-visszakeresés" kifejezés is.

JIRI PANYR (1942)

A prágai születésű matematikus Jiri Panyr a Siemens tudományos munkatársa, a CONDOR integrált információkereső rendszer és a STEINADLER automatikus osztályozási rendszer egyik tervezője. Azok közé tartozik, aki az egyidejű dokumentum- és kulcsszóosztályozás alkalmazásának lehetőségeit vizsgálják.
Az alábbiakban az ismertebb automatikus információkereső rendszereket tekinti át. Ezek csak kísérleti körülmények között működnek vagy működtek. Ahhoz, hogy az automatikus osztályozási rendszerek kereskedelmi forgalmazásra éretté váljanak, rengeteg adatra és nagyon sok számítógépidőre van szükség. Ezzel szemben a tényleges kísérleteket legfeljebb néhány ezer dokumentumon hajtották végre, és ez nem szolgáltatott elég bizonyítékot arra, hogy az egyelőre csak „laboratóriumi körülmények" között működő rendszerek a valóságos környezetben fellépő hatalmas beviteli adatmennyiség esetén is állni fogják a sarat. Ahhoz, hogy átfogó kísérleteket hajtsanak végre, például rendkívül sok indexkifejezést (kulcsszót) tartalmazó szótárakra, tezauruszokra lenne szükség a megfelelő szemantikai összefüggésekkel a kulcsszavak között. A mesterséges intelligencia rendszerekhez képest ugyan sokkal egyszerűbb szótári struktúrákról van szó, de a lexikai egységek több tízezerre tehető száma akkora ráfordítást igényelne, amire a megszokott intellektuális módszerek nem elegendőek. Mindez minden jel szerint lehűtötte a hatvanas évek lelkesedését, és a nagy számítógépgyártók fokozatosan visszakoztak a kísérleti rendszerek további finanszírozásától. Annál is inkább, mert – ahogy erre szemelvényében Jiri Panyr rámutatott – a „primitívebb" automatikus indexelési eljárások egyelőre még megfelelnek a célnak.

Automatikus osztályozás és információkeresés[20]

1.6 Kísérleti és labormodellek[21]

Az információtechnológiai kutatások eredményei (különösen ami az információk feltárását, tárolását és keresését illeti) és ama módszerek között, melyek alapján a kereskedelmi forgalmazásra készített információkereső rendszerek működnek, szinte alig van kapcsolat. A kereskedelmi forgalmazásra készített infor-

20 Automatische Klassifikation und Information Retrieval / Jiri Panyr – Saarbrücken: [s. n.], 1985. – 418.

21 1.6 Forschungs- und Labormodelle zu IR-Systemen. In: Automatische Klassifikation und Information Retrieval, p. 37–43.

mációkereső rendszereken azok a szoftvertermékek értendők, melyeket – ellentétben a kísérleti vagy labormodellekkel – széles körben alkalmaznak a gyakorlatban. A legtöbb ezek közül a késői hatvanas vagy a korai hetvenes években készült, és ennek következtében olyan jellemzőik és korlátaik vannak, melyek részben az akkori számítógépes technológiával, részben a központosított információközvetítés hagyományos szervezési adottságaival függnek össze.

Az ilyen rendszerek használója és a rendszer közötti kapcsolat az információszolgáltatás centralisztikus felfogásán alapszik, azaz szükség van olyan információközvetítőre, például szakemberre, aki a számítógépes rendszert kezelni, és lehetőségeit maximálisan kihasználni képes. Az első ilyen dokumentumkereső rendszerek intellektuális dokumentum feldolgozáson (indexelésen) alapultak. Az automatikus indexelőeljárások valamivel később, de még ebben a korai időszakban születtek meg, és ezért nem fedezhetők föl bennük az automatikus osztályozás és információkeresés újabb eredményei.

A dokumentumkereső rendszerek kísérleti vagy labormodelljeit elsősorban kutatási célokra készítették, hogy az újabb elméleti felismerések alkalmazhatóságát a feltárási és keresési technológia területén kipróbálják. A CONDOR kivételével egyetemi kutatóközpontokban készültek, és mindeddig csak nagyon korlátozottan használták őket a laboratórium falain kívül. Ezek a rendszerek egyébként a kereskedelemben forgalmazott rendszerek hagyományos feladatait is képesek ellátni, de többnyire csak az összehasonlító vizsgálatok elvégzése érdekében.

Ma az információkereső rendszerek fejlődésében bizonyos fokú stagnáció tapasztalható. A nagy számítógépgyártók, melyek egyben a leggyakrabban használt kereskedelmi rendszerek kifejlesztői és gazdái is, minden jel szerint azon a véleményen vannak, hogy uralkodó piaci helyzetük megengedi, hogy elöregedett rendszerkoncepcióikkal elégítsék ki a felhasználókat, mivel manapság amúgy sem kínálnak jobb rendszereket a meglévőknél a kereskedelemben. A japán 5. generációs számítógépek következtében ugyan fellendült a tudásbázisú rendszerek kutatása, de ez senkit se tévesszen meg: az információtechnológia problémái (különösen a szélesebb körben használható automatikus információkereső rendszerek továbbfejlesztése) a háttérbe szorult. A tudásbázisú rendszerek semmiképpen sem helyettesíthetik az automatikus információkereső rendszereket. Mindezt igazolja az a terjedő kiábrándulás, mely a mesterséges intelligencia rendszereinek és technológiájának jobb megismerésével lassan mindenütt megjelenik.

Az átfogó automatikus információkereső rendszerek területén a három legismertebb kutatási program és termék:

– a Harward Egyetemen (1966 közepéig) majd a Cornell Egyetemen (1966 közepétől mind a mai napig) működő SMART rendszer;
– a Siemens cég CONDOR rendszere (1973 és 1981 között)
– a berlini műszaki egyetem FAKYR rendszere (1972-től a mai napig).

Rajtuk kívül röviden megemlíthetjük a Rank Xerox cég FIRST rendszerét és a Syracuse-i Egyetem SIRE rendszerét. Mindegyikre jellemző, hogy központi funkciójuk az automatikus osztályozás.

Az eddig legjelentősebb kísérleti rendszer a SMART. Az információk feldolgozása és keresése terén számos megoldás, mint például a dokumentumklaszterálás (az ún. Vektortér modellel), a relevancia-visszacsatolás a SMART-(vagy Salton-) iskolából származnak. A rendszer jelentőségét növeli a nagy számú kísérlet, melyet vele elvégeztek. A SMART-ról egész sor publikáció is megjelent, ezért nem foglakozunk vele részletesebben.

A STEINADLER-eljárást (**S**tatistische **Te**xt**in**dexierung und **a**utomatische **D**okumentklassifikation unter Einbeziehung der **L**inguistischen **Er**gebnisse) legnagyobbrészt a CONDOR-program keretében alakították ki, mely utóbbi a legfontosabb és nemzetközileg is legismertebb automatikus német információkereső rendszernek számít. Ahogy *Reinhard Kuhlen* fogalmazta: „...a Siemens CONDOR rendszere egészében, vagy legalábbis néhány kiérlelt összetevőjével a nyolcvanas években meghatározó szerepet játszott az információkeresés történetében." A CONDOR fejlesztését 1981-ben a zárójelentést követően abbahagyták anélkül, hogy eljárásainak és felismeréseinek későbbi adaptációjáról egy majdani kereskedelmi változat számára gondoskodtak volna.

A CONDOR (**Co**mmunication in **N**atural language with **D**ialogue **O**riented **R**etrieval systems) eredeti változata csupán a természetes nyelvű ember–gép párbeszédre épült. Csak miután kialakították a STEINADLER 1. részváltozatát, és a hozzá való keresőkomponenst (kb. 1975 közepétől), irányult a továbbfejlesztés az átfogó automatikus információkereső rendszerre. Később (kb. 1978 óta) a CONDOR már integrált adatbázis-kezelő/információkereső rendszerré vált. Magát az integrációt azonban nem hajtották végre következetesen; ennek oka, hogy a szöveg– adatfeldolgozó komponenseket kezdetben teljesen külön fejlesztették. A kutatás súlypontjai a következők voltak:

- Szövegek és keresőkérdések nyelvészeti elemzése. A CONDOR jellemzője, hogy a szintaktikai kategóriák felismeréséhez nem használ szótárat.
- Formális dokumentumelemzés minden olyan formális információ interpretációja érdekében, melyek a leírásban és a jelsorozatokban előfordulnak, beleértve a jelek és szövegrészek osztályozását; eredményeként a keletkezett dokumentumleírás a dokumentum szerkezetét reprezentálja.
- Vonalas rajzok felismerése és elemzése (folyamatábrák, gráfok stb.) a képi információk feltárása és keresése érdekében.
- Durva keresési stratégiák, különösen az alábbi esetekben:
  - lineárisan súlyozott pszeudo-boole-algebrai keresési logika (a kulcsszavak súlyozása prioritási osztályaiknak felel meg);

  - teljesen vagy tökéletlenül feltárt osztályozási hálón alapuló keresési stratégiák (a STEINADLER segítségével);
  - a keresési eredmények rangsorolása.
- Interaktív keresési stratégia teljesen vagy részlegesen feltárt osztályozási háló alapján, azaz a relevancia-visszacsatolás alkalmazása (a CONDOR terminológiában pontosításnak is nevezik).
- Finomkeresés a (lineárisan súlyozott pszeudo-boole-algebrai keresési logika alapján végzett) durvakereséssel kapott dokumentumtételek között.
- Adatbázis-modellezés, különösen az adatbázis-kezelő és információkereső rendszerek integrációja.
- Ember–gép interakció és megjelenési formátumok tervezése párbeszédes rendszerekben.
- Automatikus tezauruszgenerálás.
- Automatikus indexelés (a teljes szövegek hozzárendelő indexelését[22] azonban elhanyagolták).
- További problémák, mint pl. továbbfejlesztés, az újabb rendszerváltozatok adaptációja, automatikus szemantikai elemzés, adatvédelem és adatbiztonság.

A keresési eredményeket ugyan értékelték, de nem követtek különösebb értékelési stratégiát. Többnyire a felhasználóra jellemző szubjektív alapon értékelték a rendszert, azaz a pertinenciát értékelték. A fejlesztési munkákat 1981-ben úgy zárták le, hogy a tervezett javításokat és új elképzeléseket nem lehetett már realizálni.[23]

A FAKYR (**Fa**chbereich **Ky**bernetik **R**etrievalsystem) a Berlini Műszaki Egyetemen (Technische Universität Berlin) készült 1973-ban több diplomaterv eredményeként. Az eltelt idő alatt többször folyton javították a teljesítményét, bővítették a funkcióit, mindezt ugyancsak diplomamunkák és disszertációk keretében. Még ma is folyamatokban karbantartják és bővítik a rendszert.

A FAKYR kísérleti dokumentációs információkereső rendszer, vele az információ szervezésének és keresésének különféle automatikus módszerei és hipotézisek próbálhatók ki. A rendszer afféle módszertani adatbanknak is tekinthető a következő jellemzőkkel:

---

22 Itt hozzárendelő indexelésen azt értjük, hogy az automatikus indexeléssel kiválasztott szavakat szabályozott szótárból választott deszkriptorokkal helyettesítik (a szerk.).

23 A CONDOR program megszüntetéséről Krause 1983-ban ezt írta: „...a programot akkor szakították félbe, amikor már olyan állapotban volt a rendszer, hogy a gyakorlatban is ki lehetett volna próbálni. Az ilyen fejlesztési gyakorlat, amelyben egyre újabb első fokozatú prototípusokat készítenek és aztán hagyják, hogy eltűnjenek a süllyesztőben, teljesen értelmetlen."

– Adatbázis-kezelés: közvetlen és invertált állományok, tezaurusz, tiltott szavak jegyzéke, statisztikai adatok megállapítása.
– Információkeresés: keresés Boole-operátorokkal, rangsorolás 43 asszociációs mérték alapján, keresés életlen halmazok (fuzzy) alapján, klaszterelemzés (dokumentum- és kulcsszóosztályozás).
– Automatikus osztályozási módszerek: gráfelméleti eljárások, single-pass algoritmusok, hierarchikus osztályozások dendrogramjainak automatikus generálása.
– Automatikus tesztelés: a visszahívás, a pontosság, a selejt kiszámítása, visszahívás és pontosság, illetve visszahívás és selejt diagramok megszerkesztése.

A FAKYRból hiányzik a szöveges dokumentumok nyelvészeti ellenőrzési komponense.

A Rank Xerox cég FIRST (**F**lexibel **I**nformation **R**etrieval **S**ystem for **T**ext) információkereső rendszere olyan on-line dokumentációs információkereső rendszer, melyben az adatbázis-kezelő rendszert egyesítették természetes nyelvű keresőkérdések és referátumok automatikus feldolgozásával. A strukturált adatokat (például a bibliográfiai adatokat) és az olyan deszkriptorokat, melyek általánosan érvényesek, külön adatbázisban tárolják. A szöveges adatokat a SMARThoz hasonlóan dolgozzák föl: automatikus nyelvészeti elemzés után a *Dattola* által kidolgozott algoritmus szerint több fokozatban osztályozzák. A dokumentumok hasonlósági keresése után a kapott dokumentumokat e hasonlóság alapján rangsorolják. A FIRST rendszer tehát a szöveges információkat feldolgozó SMART és a CODASYL ajánlásokon alapuló adatbázis-kezelő rendszer integrációja; ennek köszönhetően a bibliográfiai és egyéb dokumentumleíró adatok közvetlenül is hozzáférhetők a kereséskor, és a felhasználó először ilyen adatokkal végezheti el a keresést, ezáltal jelentősen csökkentve az összehasonlítandó dokumentumok számát.

Az adatbázis-kezelő rendszerekkel és a CODASYL ajánlások magyarázatával részletesebben kötetünk automatizálással foglalkozó fejezetének bevezetőjében, az „Automatikus információkereső rendszerek és az adatbázis-kezelő rendszerek” című részben foglalkozunk.

A syracuse-i egyetem SIRE (**S**yracuse **I**nformation **R**etrieval **E**xperiment) rendszerét a konvencionális információkereső rendszerek vonásai jellemzik (invertált fájlok, Boole-algebrai keresés). Kimutatták, hogy az adatszervezés kis változtatásával az ilyen rendszerek is képessé tehetők arra, hogy hasonlósági függvényekkel a dokumentum- és keresőképek között összehasonlító műve-

leteket végezzenek, és ezáltal a talált dokumentumok rangsorolása is megvalósítható. Az invertált állományok hagyományos technikája következtében a a keresés hatékonysága is javul. A SIRE és a hozzá hasonló rendszereket hibrid rendszereknek is nevezik.

A cambridge-i egyetemen működő CUPID (**C**ambridge **U**niversity **P**robabilistic **I**ndependence **D**atamodel) az információkeresés egyszerű valószínűségi modelljén alapszik. A további, kevésbé ismert „kifinomult információkereső rendszerekről" (sophisticated retrieval system) *Salton* számol be 1983-ban megjelent könyvében.[24]

Összegezésként idézünk *Herbert Henrichs* 1983-ban írott tanulmányából: „A késői hatvanas és korai hetvenes években készült kereskedelemben forgalmazott információkereső rendszerek néhány éven belül elavulnak, mert túlhaladja őket az informatikai fejlődés. Szükség lesz olyan rendszerekre, melyek követik őket az automatikus információkeresés területén. Semmivel sem igazolható, hogy még egyszer elölről kezdjék a munkát és újra hatalmas összegeket költsenek az első követő rendszerek programozására."[25] A kereskedelmi forgalomba kerülő rendszerek gyártói ez elől a kérdés elől előbb-utóbb nem térhetnek ki.

[...]

12.2.4. Az automatikus osztályozás és automatikus indexelés közötti viszony

Rövid (például referátumnyi hosszúságú) szövegek esetén az automatikus osztályozás – legalábbis durva közelítéssel – arra használható, hogy hozzárendeljen deszkriptorokat a dokumentumokhoz; hosszú szövegek esetén ez már nem oldható meg egyszerűen. Az automatikus osztályozás tehát semmiképpen sem tekinthető az indexelés alternatívájának, hanem fontos kiegészítőjének, melyet semmilyen indexelési módszerrel nem lehet helyettesíteni. Az automatikus osztályozás nem képzelhető el indexelés nélkül, különösen ami annak eldöntését illeti, hogy milyen szövegszavakat rendeljenek deszkriptorként a dokumentumokhoz, de azért sem, mert az indexelés alapján végezhető el a hozzárendelt (és a nem hozzárendelt) kulcsszavak súlyozása.

---

24 Salton, G., McGill, M. J.: Introduction to modern information retrieval. – New York: McGraw-Hill, 1983.

25 Henrichs,. H.: The growing crisis of traditional information retrieval systems – what is to follow? In: Research and Development in information retrieval. Lexturenotes in computer science 146. [Ed. by] G. Salton and H.-J. Schneider. – Berlin, Heidelberg, New York: Springer, 1983. p. 1–12.

Rövid szövegek esetén egyszerűbb (többnyire morfológiai) nyelvészeti elemzés elegendő a hozzárendelés elvégzéséhez (azaz a jelentéssel bíró kulcsszavak azonosításra és szabványosított információkereső-nyelvi kifejezéssel való helyettesítésükre); hosszabb szövegek esetében ez nem végezhető el megfelelően, mivel hatalmas mennyiségű feldolgozási fölösleggel jár. A CONDORral végzett kísérletek bebizonyították, hogy ha az automatikus információfeldolgozás nem tartalmaz olyan komponenst, melynek segítségével szabványosított indexkifejezéseket (deszkriptorokat) lehet hozzárendelni az automatikusan feldolgozott dokumentumtételekhez, akkor ezt az eredmény nagyon megsínyli.

JURIJ ANATOLJEVIČ ŠREJDER (1928), AVAGY KÍSÉRLET AZ OSZTÁLYOZÁS INTENZIONÁLIS MATEMATIKAI–LOGIKAI ELMÉLETÉNEK MEGFOGALMAZÁSÁRA

A kötet elején bemutatott *Robert A. Fairthorne* és az automatikus osztályozás művelői az absztrakt algebra, illetve a klaszterelemzés eszközeit használva vizsgálták az osztályozás kérdését. Szemléletük szigorúan a fogalmak halmazelméleti, extenzionális értelmezésén alapul. A mai automatizált számítógépes nyelvkezelő eljárásoknak is ez az alapja. Ezek az eljárások legnagyobbrészt csak a szintaxist – a nyelv mondattanát – célozzák meg, azaz azt próbálják algoritmikusan megoldani, hogy a gépi eljárásokban megőrződjenek a mondaton belüli szavak közötti összefüggések. A felmerülő nehézségek miatt mindezek az eljárások évtizedek óta kísérleti szakaszban vannak.

A szavak jelentésével összefüggő – szemantikai – nyelvfeldolgozással még súlyosabb problémák járnak együtt, mivel a nyelv szavainak kettős természete van: a szavak egyrészt vonatkoznak valamire (ez a referenciájuk, a fogalmak terjedelme, extenziója), másrészt jelentésük van és valamilyen értelemben használják őket (ez a jelentésük és értelmük, összefüggésük, a fogalmak tartalma, intenziójuk). A gépi fordítás és automatikus indexelés/osztályozás az elsőnek említett referenciális azonosságon alapul. Ez azt jelenti, hogy ha két deszkriptornak (például a „kutya” és az „eb”) azonos a terjedelme (mindkettő fizikai értelemben az ismert négylábú háziállatra vonatkozik), akkor teljes mértékben azonosak, tehát egymás helyett használhatók. Következésképp a „kutya” deszkriptorral és az „eb” deszkriptorral osztályozott dokumentumok halmazának azonosnak kell lennie. Formálisan – extenzionálisan – ez igaz. De a józan, mindennapi ész szerint egy „Eb ura a fakó” című dokumentum (melyben a Habsburg-ház

elleni magyar függetlenségi küzdelmeket tárgyalják) nem tartozhat bele a találatokba, ha a „kutya” deszkriptorral végzik a keresést. Azaz intenzionálisan a két szó nem azonos, mert nem azonos értelemben használják őket, holott a jelentésük (ezen itt a terjedelmüket, a referenciájukat értjük) egyforma.

Mindez addig nem okoz problémát, amíg az osztályozást intellektuálisan végzik. Amint azonban automatikus indexelést/osztályozást alkalmaznak, az eljárás a példázott esetben „feldobja a talpát”.

A problémát nagyon leegyszerűsítve szemléltettük. Valós körülmények között, a dokumentumok, referátumok, címek szövegeiben az ennél bonyolultabb eseteknek se szeri, se száma. De sem a mai számítógépes nyelvészetben, sem pedig az ennek eredményeit felhasználó, szintaktikai eszközöket használó automatikus indexelésben egyelőre nincsenek reális megoldások erre a szemantikai problémára. A matematikus Jurij A. Šrejder azon kevesek egyike, aki munkásságát lényegében e problémának szentelte és azzal kísérletezett, hogy az automatikus feldolgozás érdekében az osztályozás egzakt elméletének intenzionális alapú megfogalmazásához hozzájáruljon. Szándékosan fogalmaztunk ilyen körmönfontan, mivel a probléma valójában olyan alapvetően érinti az emberi gondolkodást magát is, hogy megoldásával a belátható jövőben aligha számolhatunk. Ezért is van, hogy az osztályozás szakembereinek egy része szkeptikusan tekint az automatikus megoldásokra. (Kétségeiket szemlélteti a *Gerard Saltonnal* kapcsolatban közölt vita részlete, továbbá a *Bar-Hillel* által írt szemelvény.) Annyiban igazuk van, hogy nincs egyedüli üdvözítő eljárás; a megoldás inkább a különféle, adott esetben egymással ellentétes szemléletű eljárások kombinációja lehetne. Erről tanúskodnak azok a vizsgálatok is, melyek kimutatták, hogy a mai könyvtárakban és az online adatbázisokban (továbbra is) együtt alkalmazzák a hagyományos, intellektuális osztályozást/indexelést és a különféle automatikus indexelési eljárásokat.[26]

Šrejdertől magyarul korábban az alábbi, Varga Dénes által fordított tanulmány jelent meg az információ nem fizikai – shannoni értelemben kezelt –, hanem értelmi, szemantikai felfogásáról:

26 Margarete Burkart-Sabsonb, Gernot Wersig: Kombinatorischer Einsatz von Dokumentationssprachen. – Berlin: Progris, 1982. 23 p. (PROGRIS PHS 7/82)

Az információ szemantikai jellemzői

In: A dokumentáció nyelvészeti kérdései I. [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest: OMKDK, 1966. – (A tudományos tájékoztatás elmélete és gyakorlata; 10.) p. 137–153.
Eredeti: On the semantic characteristics of information. In: Information storage and retrieval, 1965, p. 221–133

Az információ Shannon-féle valószínűségelméleti–természettudományi értelmezése alkalmatlan arra, hogy vele a megértés folyamatát modellezzék. Szemantikai jellegének elemzéséhez a tezaurusz fogalma használható föl. Tezauruszon ebben az összefüggésben a világra vonatkozó tudati ismerettár értendő, mely különféle állapotokban lehet. Minden közlemény megváltoztatja a tezaurusz állapotát, azaz annak valamilyen transzformációjával ekvivalens. (Információelméleti megközelítésben a tezaurusz az események és valószínűségeik listájával ekvivalens.) Az információ mennyiségét úgy határozhatjuk meg, mint a tezaurusz változásának mértékét a közlemény következtében. A közlemény hatására keletkező információ mennyisége azonban nemcsak a közleménytől, hanem a tudat tezauruszától is függ.

Nemcsak attól nő, ha a közlemény „tartalmas", hanem attól is, ha a tezaurusz fejlett és strukturált. Más szóval valaki, aki már foglalkozott egy kérdéssel, az ezzel összefüggő közleményekből több információhoz jut, mint az, aki a kérdéssel nem foglalkozott.

A szemantikai információt ez a belső tezauruszon alapuló tulajdonság alapvetően megkülönbözteti az információ klasszikus, információelméleti felfogásától. Ez utóbbi esetében ugyanis a meglévő információ minden esetben csökkenti a beérkező információ mennyiségét, hiszen ha nincs határozatlanság, akkor nincs mit csökkenteni. Azaz a klasszikus információelméletben nem játszik szerepet a közlemény „megértésének" a foka. Ez összhangban van azzal a tapasztalattal, hogy mennél többet tud valaki, annál inkább tisztában van tudásának hiányosságaival, tehát annál nagyobb információmennyiség éri a világból a tezauruszát.

Ha a tezaurusz összetettebb, akkor nagyobb változások játszódhatnak le benne az új szöveg hatására. Az összetettebb tezaurusz intenzívebben változhat, mint az egyszerű.

Az így értelmezett szemantikai információ formális modellje leírható a predikátumhalmazzal, az objektumok halmazával és az események halmazával.

A modellből következik, hogy a szemantikai információ mindig valamilyen közvetítő révén nyerhető, aki vagy ami az információt birtokolja. Ez a közvetítő valójában egy analóg tezaurusz. Az információcsere a két tezaurusz közötti adásvételből áll. Az adott szöveg a vevő szempontjából pragmatika, az adó szempontjából szemantika. (Ezzel magyarázható minden tudományinterpretáció – oktatás – ama jellemzője, hogy nagy súlyt fektetnek benne a szavak helyes használatára.) Két tezaurusz komplexitási fokának összhangban kell lennie ahhoz, hogy megérthessék egymást. Ha az adó „intelligensebb", mint a vevő, az utóbbi nem érti az előzőt. A gépek esetén ilyen helyzet nem állhat elő: nincs olyan gép, amely intelligensebb lenne (komplexebb tezaurusza lenne) az embernél.

Az emberi nyelven megfogalmazott szövegek rendkívül sokrétűek, megértésük összehasonlíthatatlanul nagyobb információs gazdagsággal rendelkező tezauruszokat igényel. Nem ok nélkül hoznak létre például egyre általánosabb programnyelveket, mert az általánosítás fokozásával lehetséges csak közelíteni az ember természetes nyelvéhez.

Šrejder ebben a hatvanas évek elején írt tanulmányában a mai történeti informatika tezauruszfogalmát előlegezte meg. Eszerint a társadalom tagjait különféle tezauruszok kapcsolják össze, és hasonló tezauruszt használó csoportokat – például családot, nemzetet – alkotnak.[27]

Rendszerek és modellek[28]

5. Osztályozási rendszerek[29]

5.1 A természetes osztály és a természetes rendszer

A tárgyak osztályozása a megismerés leghagyományosabb módszere; az osztályozás eredményeként az ismeretek osztályozási táblázatok formájában jelennek meg. A táblázatokban a vizsgált tárgyak célszerűen kiválasztott ismérvek alapján osztályokba (taxonokba) rendeződnek; az ismérvek az osztályozás alapelemeinek tekinthetők.

---

27 A történeti informatikáról lásd: Z. Karvalics László: A történeti informatika a könyvtárról. In: Tudományos és Műszaki Tájékoztatás, 1995, 42. évf, 1. sz., p. 7–16.

28 Sistemü i modeli / Ju. A. Šrejder ; A. A. Šarov. – Moskva : Radio i svaz´, 1982. 152 p. – (Kibernetika)

29 Klassifikacionnüe sistemü. p. 75–90. In: Sistemü i modeli

Ismeretelméleti szempontból az elvi kérdés az, hogy az osztályozás a természetes „káosz" rendezésének eredménye-e, vagy pedig a természet dolgaiban eleve meglévő rend tükröződése? Számos szerző arra a következtetésre jutott, hogy a természet tárgyait a rendszerszerűség jellemzi. Mi az osztályozás módszertanát szeretnénk továbbfejleszteni e következtetés alapján; abból indulunk ki, hogy az osztályozás alapvető feladata a természetben található harmónia tükrözése ***és*** kifejezése logikai eszközökkel. Az osztályozási rendszereket úgy fogjuk fel, mint objektíven létező külső rendszerekre vonatkozó ismeretet. (Külső – „kívülről" felfogott – a rendszer, ha úgy alakítjuk ki, hogy korábban már létező tárgyakat osztályokban *fogunk össze* [azaz a dolgokat egyesítve alkotunk teljességet], szemben a belső – „belülről" felfogott – rendszerrel, melyet úgy alakítunk ki, hogy valamilyen teljességet *összetevőire bontunk* föl.)

> Külső rendszerek például az elemek periódusos rendszere, a biológiai taxonómia rendszerei (pl. a növény- és állatrendszerek), és az automatikus osztályozással előállított rendszerek.
> Belső rendszerek például a hagyományos könyvtári és dokumentációs célú osztályozási rendszerek (ETO, tárgyszójegyzékek, deszkriptorszótárak).

A külső rendszereket nem az egész keretén belül működő „szervek", hanem az általános egész egymáshoz hasonló képviselői alkotják. A belső rendszer részei (szervei) egyediek, a külső rendszer összetevői hasonlóak. A rendszerszerűség problémája ebben az esetben a megfigyelt hasonlóság és az általános lényeg viszonyából adódik. Különösen a biológiában elterjedt az a nézet, hogy az általános lényeg a történelmi vagy származási közösségben nyilvánul meg. E felfogás módszertani gyengéje egyrészt, hogy a rendszerszerűség távolról sem függ össze mindig a közös eredettel (pl. az elemek periódusos rendszerében), másrészt, hogy a közös eredet csak az előzőleg megállapított közös lényegre támaszkodva bizonyítható. Ezért a tárgyak hasonlósága (melyet a hasonlósági vagy tolerancia reláció alapján alkotott csoportjaik tükröznek) és eme tárgyak általános lényege közötti összefüggés megfogalmazása mind időszerűbb feladat. Egy egyszerű módszertani áthidaló megoldás található az 1970-ben írt könyvben, ahol igazoljuk, hogy minden ekvivalencia reláció (azaz minden osztályba sorolás) megadható a „reprezentánsa" relációval.[30] A következő lépést a *Panovával* közösen írt tanulmányunkban[31] fogalmaztuk meg, szembeállítva

30 Šrejder, Ju.: Ravenstvo, shodstvo, porădok. – Moskva: Izd. Nauka, 1970.
Magyarul: Šrejder, Ju.: Egyenlőség, hasonlóság, rendezés. Bevezetés a modern matematika alkalmazásába. – Budapest: Gondolat, 1975.
A vonatkozó rész az idézett mű magyar fordításának 111–114. és 123–126. oldalán található (a szerk.).

31 Lásd a kötetünkben szereplő előző szemelvényt (a szerk.).

szemiotikai szempontból a taxon (az ekvivalencia szerinti csoportosítás) fogalmát az ismérv fogalmával.

A taxon a névvel jelölt tárgyak osztályát (a megfelelő fogalom terjedelmét) képviseli; az ismérv ezzel szemben a névhez fűződő képzetet (a fogalom tartalmát) adja meg. E kettősséget módszertanilag a legtermékenyebben *Mejen* közelítette meg, melyet később közös tanulmányban is kifejtettünk[32]. Eszerint az osztályozás a taxonómia (a tárgyak hasonlóság szerinti csoportosítása) és a meronómia (a tárgyak olyan elkülönítése, amely lehetővé teszi a közöttük fennálló hasonlóság fokának megállapítását) dualitásként értelmezhető.

> Az automatikus osztályozással (mely Šrejder felfogásában taxonómia) hasonló dokumentumok rangsorolt – a hasonlóság fokozatai szerint rendezett – osztályai (klaszterei) alakíthatók ki. A hasonlóság foka itt mennyiségi. Az intellektuálisan szerkesztett tezauruszban a hasonlóság foka helyett az egyes típusaival jellemzik a deszkriptorok közötti összefüggéseket: például a Kutya és a Ragadozó között a generikus, a Kutya és a Falka között a partitív, a Kutya és a Házőrzés között az instrumentális relációval. Az előbbi a taxonómiai, az utóbbi a meronómiai rendszer.

A taxonómia alkotja a tárgyak külső rendszerét, a meronómia pedig belső rendszerüknek tekinthető. Ki fogjuk mutatni, hogy e dualitás alapján az osztályozás olyan általános módszernek tekinthető, mellyel valamely egészet alkotó, azonos nembeliségű tárgyak csoportjaira vonatkozó ismeretek rögzíthetők.

A továbbiakhoz lényeges a külső rendszert alkotó „természetes objektumosztály" fogalma. A természetes osztályra példa az orosz nyelv összes szavának osztálya. Kiderül, hogy nagyon nehéz olyan orosznak festő kváziorosz szót találni, melynek ne lenne közös morfémája más orosz szavakkal. A kitalált szavakat könnyű felismerni nem természetes mivoltuk alapján. Következésképpen értelmetlennek tűnik a kérdés felvetése: „Orosz-e a mesterségesen alkotott szó?" Azonban mégis felvethető és értelmes válasz adható rá.

A természetes osztály másik fontos példája az összes élő dolog osztálya. A fantasztikus regényekben az ismeretlen bolygóra érkező űrhajós többnyire azt dönti először el, hogy az eléje kerülő valami élőlény-e vagy sem? A szerzőknek nem jut eszébe kételkedni e kérdés értelmében. Egyrészt tehát nincs semmiféle konstruktív meghatározás az „élet" fogalmára, az élet diagnosztizálására nem létezik recept, másrészt viszont mégis teljesen értelmetlen a kérdés, hogy valami élő-e vagy sem? Ez azt bizonyítja, hogy az élő szervezetek osztálya termé-

32 Mejen, S. V., Šrejder, Ju. A.: Metodologiceskie aspektü teorii klassifikacii. In: Voprosü Filozofii, 1976. No. 12.

szetes osztály. Megjegyezzük még, hogy az űrhajós példákban semmilyen jelentősége sincs a szervezetek genetikai közösségének. A tudományos–fantasztikus regények szerzői aligha tételezik fel, hogy az élőlények az összes égitesten közös őstől származnak.

A tárgyak természetes osztályai a természetes osztályozási rendszerek segítségével írhatók le. Az osztály természetessége az ontológiai előfeltétele annak, hogy a természetes osztályozási rendszer kialakítható legyen. Az osztályozás természetes rendszerének fogalmát már *Karl von Baer* Szentpéterváron élt német filozófus bevezette.

Az osztályozási rendszer természetes rendszer, ha a rendezési struktúrában elfoglalt helye alapján minden tárgy lényegi tulajdonságai meghatározhatók. Az elemek periódusos rendszerében például az egyes elemek legfontosabb tulajdonságai az elem rendszeren belüli helye alapján kiolvashatók.

E meghatározásból következik, hogy a természetes rendszer nemcsak az ismert tárgyak összességével van összhangban, hanem azokkal az elképzelhető tárgyak sokféleségével is, melyek a rendszer logikájából következő, de a természet szeszélye vagy hiányos ismeretek következtében betöltetlenül maradó „helyekhez" tartoznak. Az ilyen rendszerben ugyanis az összes, logikailag lehetséges tulajdonság kombináció elvileg lehetséges, bár az adott időszakban megfigyelhető tárgyak körében távolról sem mindegyik valósul meg.

A természetes rendszer kialakításában két fontos osztályozási szempont játszik szerepet: a taxonómia és a meronómia.

5.2. Taxonómia

Az osztályozó szempontjából a taxonómia módszer, hogy az objektumokat a hasonlóságuk alapján osztályokba „sorolja". A taxonok szerinti csoportosítás módszere magának a tárgynak az ismeretéből következik. A módszertani szakember számára az a probléma, hogy megalkossa a „taxonomikus szerkezet" fogalmát, meghatározza terjedelmét: a kigondolt taxonómiai struktúra osztályát. A módszertani szakember következő lépése az ilyen struktúrák felépítésének általános logikai leírása. Az első kérdéskörrel foglalkozik ez a fejezet. A második, mint majd meglátjuk, csak a taxonomikus és meronomikus szempontok összevetésével oldható meg.

A taxonok olyan részek, melyekre az osztályozott objektumok osztálya tagolódik. Ehhez először ki kell választani az objektumok bizonyos alaposztályát. Ezt az osztályt „taxonómiai univerzum"-nak nevezzük. Azt mondjuk, hogy ebben az univerzumban adott a taxonómia, ha adott a taxonok, vagyis eme univerzum alosztályainak összessége, melyben megtalálható az egész univerzum (a legnagyobb taxon) is, és a taxonok metszete mindig taxont eredményez.

Megjegyezzük, hogy a taxonómiai univerzum osztály és nem halmaz, mivel nincs pontosan meghatározva, hogy mi lehet az univerzum eleme. A szervezetek osztályozásánál nem világos, mi a szervezet? Mi tekinthető szervezetnek a bambusz esetén? A szár vagy az egész bokor? Valószínűleg a bokor, mivel az összes szár egyetlen tőről fakad. Másfelől, az utódok a partenogenetikai osztódáskor a szülő csaknem pontos másolatai. Különböző szervezeteknek vagy egyetlen szervezetnek kell ezeket tartani? A dokumentumok osztályozásakor szintén meg kell egyezni, hogy mit tartsunk dokumentumnak: a példányt, a kiadást, vagy a művet? Hasonló példák származnak az állatvilágból: több polip telepes, a szifonofarák esetén egységet alkot, nagyobbat, mint az egyes szervezetek egysége, vagyis mindegyik polip a telep egy-egy szerve. Sajátságos példa a gombából és moszatokból álló zuzmó. Ez szervezet vagy sem?

Osztályozáskor gyakran nem az egyes egyedeket érdekes összehasonlítani, hanem a belőlük képzett taxonokat. A pszichológust érdekelheti az egyes személyiségek sajátossága, de a biológust nem a személyiségek érdeklik, hanem a fajok, fajták vagy populációk. Általában kényelmesebb a minimális taxonokat (fajtákat) kijelölni és a taxonómiai univerzum helyett a fajok összességét tekinteni, amely rendszerint pontos halmazt alkot. Ezt a halmazt osztályozási mezőnek fogjuk nevezni.

A „faj" fogalmát bevezetve az osztályozó átalakítja a taxonómiai univerzumot osztályozási mezővé: minimális taxonok halmazává. Ilyenkor a fajok, mint halmazok, általánosítva, metszhetik egymást, vagyis egy objektum tartozhat több fajhoz; vagy előfordulhat, hogy bizonyos objektumok egyáltalán nem tartoznak semmilyen fajhoz: az osztályozáson kívül maradnak. A taxonómia sikeres leírásához csak az a fontos, hogy a fajok összessége halmaz legyen. Ez azt jelenti, hogy az osztályozó meg tudja különböztetni a fajokat; a fajok halmaza legyen jól megszámolható.

A faj maga osztály és nem halmaz, mivel tetszőleges objektum adott fajhoz tartozását távolról sem lehet mindig egyértelműen meghatározni és az ilyen objektum határait nem mindig lehet egyértelműen meghúzni. Abból a föltevésből indulunk ki, hogy meg lehet úgy választani a minimális taxonokat (fajokat), hogy azok már jól elkülönülő objektumok meghatározott összességét (halmazát), azaz terjedelmet alkossanak.

A többi – többségében minimális – taxon most minimális taxonok halmazaként, vagyis fajok halmazaként fogható föl. A fajok vagy maradéktalanul a taxonba tartoznak, vagy egyáltalán nem tartoznak bele.

Tehát, *a taxon nem szervezetek (általánosítva: osztályozott objektumok) halmaza, hanem fajok halmaza.* Ebben az értelemben minden taxon bizonyos minimális taxonból képzett halmaz. Ebből kiderül a minimális taxonok különleges szerepe: az osztályozott objektumok összességének (osztályának) fogalmát (lényegét) képviseli. Egyébként a taxon felfogható a beletartozó minimális taxonok uniójaként is, és ezzel az osztályozott objektumok összességeként is. Az ilyen értelmezés megengedhető, de nem szabad megfeledkezni arról, hogy

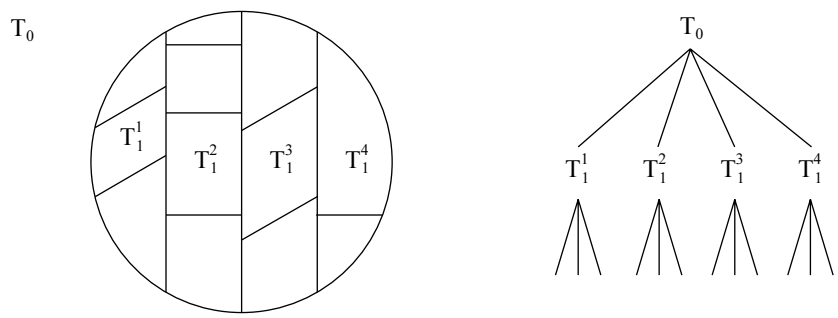
ebben az esetben megváltozik a taxon elfogadott státusza. Például, a biológiai osztályozásban általában a család a fajok halmaza, de nem szabad a családon a megfelelő egyedek halmazát érteni. Kiválasztható az összes fajtából álló minimális taxon, ez lesz a legnagyobb.[33]

Hangsúlyozzuk, hogy a fajok halmaza és a szervezetek halmaza nem egy és ugyanaz.

Vizsgáljuk meg a taxonok közötti legfontosabb relációkat. Az első a taxonok tartalmazási relációja. Ha a T_1 taxon összes faja a T_2 taxonhoz tartozik, akkor azt mondjuk, hogy a T_2 taxon része a T_1 taxon ($T_1 \subset T_2$). A második a metszete reláció: az mondható, hogy a taxonok metszik egymást vagy nem metszik egymást. E relációk a taxonok halmazában bizonyos struktúrát alkotnak. A legegyszerűbb taxonómiai szerkezet a $\subset$ reláció szerinti faszerkezet, ahol a fa gyökere a maximális taxon. Ebben az esetben T_1 és T_2 taxonok metszete csak akkor nem üres, ha az egyik tartalmazza a másikat. Valóban, tegyük fel, hogy $T_1 \cap T_2 \neq \varnothing$, vagyis létezik **t** minimális taxon, amely része mind a T_1 és mind a T_2 taxonnak. Ebben az esetben a fa meghatározásából következik, hogy $T_1 \subset T_2$ vagy $T_2 \subset T_1$. Jelöljük **T**-vel a minimális taxonok halmazát. Akkor az összes taxonok Tax**(T)** halmaza részhalmaza lesz a **T** összes részhalmaz B**(T)** Boole-féle halmazának:

$$\mathrm{Tax(T)} \subset \mathrm{B(T)}.$$

Ha a Tax (**T**) halmazon a tartalmazási reláció faszerkezetet alkot, akkor a taxonómiai szerkezetet hierarchikusnak nevezzük. E szerkezetben minden taxon a fa megadott szintjéhez tartozik. Az 1. a) ábrán a T_0 osztályozási mezőnek T_1^1, T_1^2, T_1^3, T_1^4 első szintű taxonokra osztása és e taxonoknak másodszintű taxonokra osztása látható. A taxonok fastruktúrájának megfelelő részletét az 1. b) ábra illusztrálja.



a) **b)**

1. ábra

33 Rendezett halmaz maximális elemének azt szokás nevezni, melynek nincs további fölérendeltje. A legnagyobb az, amely a többi felett áll. A minimális és legkisebb elem analóg meghatározásából következik, hogy a minimális taxon általában nem a legkisebb.

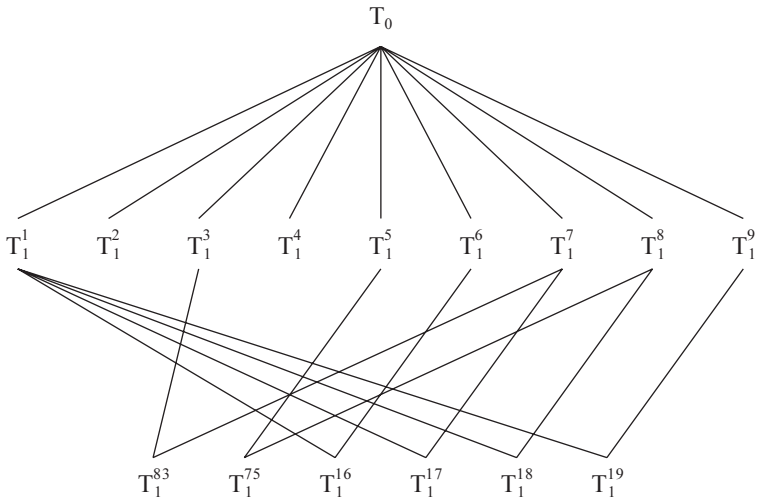
Az élő szervezetek Linné-féle rendszerét például a taxonok hierarchikus szerkezete jellemzi.

A hierarchikus osztályozás fontos példája a könyvek és dokumentumok indexelésére használt Egyetemes Tizedes Osztályozás (ETO). E rendszerben a tudomány egészének jelzete 5, az alosztályok jelzetei 50-től 59-ig terjednek, a speciálisabb tematikai alosztályok hosszabb tizedes jelzeteket kapnak. A felhasználható jelzeteket a szükséges irodalom gyakorlati osztályozására és kereséséhez táblázatokban foglalják össze. A leghosszabb jelzetek megfelelnek a minimális taxonoknak, a rendszerben megkülönböztetett legszűkebb tematikai területnek. A jelzet hossza mutatja azt a szintet, amelyhez a taxon tartozik. A minimális taxonok (ellentétben a biológiai osztályozással) a fa különböző szintjein helyezkedhetnek el. Gyakorlatilag az ETO 10 különböző „fából" áll, azaz a dokumentumok nem egyetlen legátfogóbb osztályba, a tudományba tartoznak, hanem a 10 főosztály közül valamelyikbe. A taxonómiai szerkezet lehet nem hierarchikus is. Az ilyen szerkezet például a fazettás vagy kombinatív taxonómia. Az ilyen osztályozás mindegyik fazettája (szempontja) meghatározza a minimális taxonok halmazának egymást nem metsző első szintű taxonokra való bontását. Az első szintű taxonok páros metszete adja (a lét fazettával meghatározott) második szintű taxonokat. A hármas metszetek adják a harmadik szintű taxonokat és így tovább a legkisebbig, amelyeket az első szintű taxonok n.-metszete határoz meg, ahol n a fazetták száma az adott osztályozásban. A 2. ábra az osztályozási mezőnek két fazetta szerinti felosztását mutatja be. Az elsőnek a T_1^1–T_1^5, a másodiknak a T_1^6–T_1^9 taxonok felelnek meg. E taxonok metszete adja a második szintű taxonokat.

A 3. ábrán a taxonok struktúrája látható a tartalmazási reláció szerint. Ez a struktúra nem fa.

	T_1^1	T_1^2	T_1^3	T_1^4	T_1^5
T_1^6					
T_1^7					T_2^{83}
T_1^8			T_2^{83}		
T_1^9					

2. ábra



3. ábra

A hierarchikus és a fazettás struktúrák bizonyos értelemben ellentétesek. De sok olyan struktúra létezik, amely nem tartozik sem az egyik, sem a másik típushoz. Például lehetséges két olyan hierarchikus struktúra, hogy az első struktúra bármely taxonjának a másik struktúra bármely taxonjával közös metszete új taxont eredményez. Kérdéses, hogy milyen a hierarchikus és fazettás struktúrákat szélsőséges esetként tartalmazó taxonómiai struktúrák lehetséges spektruma. Kiderül, hogy izomorfikus pontossággal az összes ilyen struktúra leírható az idempotens félcsoportok segítségével. Ahány ilyen félcsoport lehetséges, annyi nem izomorf taxonómiai struktúra lehetséges.

Izomorf leképezésről beszélünk, ha az egyik halmaz minden elemének megfelel a másik halmazban egy-egy elem és fordítva. A hasonlóság ennél gyengébb, ha (mint a továbbiakban) homomorf leképzésről van szó. Ebben az esetben az egyik halmaz minden elemének megfelel a másik halmazban egy elem, de fordítva nem: a másik halmaznak lehetnek olyan elemei, melyeknek az első halmazban egynél több elem felel meg. Homomorf leképzésére példa a táj és térképe, vagy a könyv tartalma és a tartalmat kifejező ismérvek (jelzetek vagy deszkriptorok). Minden hierarchikus struktúra homomorf leképzésnek is tekinthető. Izomorf leképzésre példa a térkép és annak mikrofilmes felvétele, vagy a katalóguscédulán szereplő dokumentumleírás és az annak megfelelő dokumentumrekord az adatbázisban.

A taxonómia szempontjából az objektumok annál közelebb állnak egymáshoz, minél kisebb közös taxonra eshetnek szét. Mivel a taxonokat az jellemzi, hogy meghatározott szinthez tartoznak (megjegyezzük, hogy a minimális taxonok nem feltétlenül egy és ugyanazon szintűek), a legkisebb taxon mindkét típusát tartalmazó szint a fajták hasonlóságának mértéke. Ezért taxonómián belül is lehet beszélni a hasonlóságról. De a hasonlóság fogalma nem a taxonómia szintjén merül fel, hanem korábban, mivel a taxonómiai felosztás az objektumok közötti hasonlóság korábbi megállapítása alapján keletkezik. A taxonokra felosztás csak annak a ténynek a rögzítése, hogy az objektumok hasonlóak. Az x és y fajok közötti hasonlóság a hierarchikus rendszerben meghatározható a

$$\rho(x,y) = N–n$$

szabály alapján ahol N a rendszer rangszáma,

n az x és y fajt egyidejűleg tartalmazó legkisebb taxon rangja (feltételezve, hogy az összes faj rangja azonos és egyenlő N-nel).

Könnyű bebizonyítani, hogy a $\rho(x,y)$ a minimális taxonok terében metrikát alkot:

1. $\rho(x,y) \geq 0$
2. $\rho(x,y) = 0$ akkor és csak akkor, ha az x és y faj egybeesik
3. A háromszögszabály: $\rho(x,y)+\rho(y,z) \geq \rho(x,z)$.

Ha a minimális taxonok nem metszik egymást, akkor a $\rho(x,y)$ érték pszeudometrikus az osztályozott objektumok (szervezetek) terében. Ilyenkor csak az első és harmadik metrika-axióma érvényesül.

Még egyszer hangsúlyozni kell két fogalom ellentétét. Az első a taxonómiai univerzum: az osztályozott objektumok összessége, amely általában osztályt alkot, és ezekben nem kell pontosan meghatározni, hogy mi az objektumok azonossága és kétségbe vonható az objektumnak az osztályba tartozása is. A második: a minimális taxonok összessége, vagyis a taxonómiai mező, amely pontos halmazt alkot.

Ez a normális helyzet, de előfordul fordítva is, amikor a taxonómiai univerzum a pontos halmaz és az osztályozott mező a pontatlan. Például legyen az univerzum az épületek halmaza, az osztályozott mező az építészeti stílusok osztályai. Az épületek összessége pontos halmazt alkot, a stílusok összessége pedig eléggé elmosódott osztály. Nem teljesen érthető, hogy mi egy és ugyanazon stílus, nincs módszer annak meghatározására, hogy egy épület milyen stílushoz tartozik. Egy másik példa a jellemek szerint osztályozott emberek halmaza. Itt sincsenek pontos szabályok a jellemek meghatározására, nincsenek kritériumok a jellemek szétválasztására. A biológiában sem olyan pontos

(sőt kifejezetten pontatlan) a fajok összessége. Ez különösen szembetűnik a taxonomikus vizsgálatokban. A fajok száma ilyenkor vagy két-háromszorosára nő, vagy felére-harmadára csökken. Ezért kétségbe vonható, hogy a minimális taxonok kiválasztásának módszere mindig halmazzá alakítja az osztályt.

Ebből következik, hogy kétségbe vonható az a módszer, amely szerint a taxonómiai univerzumot az osztályozott objektumok meglévő (természetes) tulajdonságai alapján osztjuk föl. Ez nem oldható meg a taxonómia keretein belül és más osztályozáselméleti szempont figyelembe vételét igényli.

5.3. Meronómia

Mint láttuk, a taxonómiai struktúra alapján megállapítható az osztályozási mező elemei (a fajok) közötti hasonlóság. A fajok annál közelebb vannak egymáshoz, minél „kisebb" az a közös taxon, amelyhez egyidejűleg tartoznak. A hierarchikus struktúrában ez a közelség a taxon gráfján a fajok közötti távolsággal jellemezhető.

Az a kérdés, hogy ez a közelség az osztályozott objektumok tényleges rokonságát jelenti-e, vagy csak az osztályozó szubjektum önkényéről van szó? A természetes rendszer esetében fel kell tételezni, hogy az osztályozott objektum lényegéből következő, meghatározott taxon-struktúráról van szó. Ez kiderül, ha az osztályozott objektumok belső struktúráit megvizsgáljuk, és a „hasonlóságuk" alapján összevetjük ezeket a struktúrákat. Az objektum szerkezetét alkotó részek szétbontása a meronómia tárgya. Innen következik az osztályozás új szempontja, a meronómia, melynek létezését először *S. V. Mejen* említi.

Amikor az osztályozáskor az objektumok részekre – meronokra – való bontásával foglalkozunk, ezzel a vizsgált objektumok meronómiáját írja le. A meronokra bontás struktúrája alkotja az objektum archetípusát (őstípusát). Így beszélhetünk arról, hogy néhány objektumnak közös archetípusa van. Az ilyen objektumokat az osztályozó nem különbözteti meg az általa létrehozott osztályozásban, ezek egy és ugyanazon fajhoz tartoznak. Az osztályozó jogosult „általánosított archetípus" szerint összehasonlítani az objektumokat, azonosítva a különböző, de valamilyen értelemben hasonló archetípusokat. A természetes rendszerben az osztályozó a közös (általánosított) archetípussal rendelkező objektumok osztályát jelöli ki.

Újra a módszertani szakember feladata, hogy a „meron" és „archetípus" szavakat pontosan meghatározza. A továbbiakban az archetípusokról, meronokról és általánosított archetípusokról részletesebben fogunk beszélni. Egyenlőre megelégszünk a következő megjegyzéssel. Az osztályozott objektumok maguk belső rendszerek; az archetípus fogalma összekapcsolható a belső rendszerrel. A taxonok a hasonló (pl. izomorf) megjelenésű objektumokból (belső rendszerekből) alakulnak ki.

Ebben a megfogalmazásban már ott van a taxonómia és meronómia fontos dualitása. Lényege, hogy a természetes rendszerben minden taxonnak meg kell feleljen bizonyos, a taxon összes reprezentánsával közös fogalom (lényeg), melyet az osztályozás számára az általánosított archetípus képvisel. Ilyenkor a kisebb taxonoknak a taxon elemeire (már nem fajokra, hanem az osztályozott objektumokra) jellemző gazdagabb általános tartalom felel meg. Ezt a gondolatot szokás kifejezni a terjedelme (intenziója) és tartalma (extenziója) terminusokkal.

> A Ló fogalmának a terjedelme kisebb, mint az Állat fogalmáé, viszont a tartalma (ismertetőjegyeinek – meronjainak – száma) nagyobb. Az archetípus e tartalom ismertetőjegyeinek (meronjainak) struktúrája.

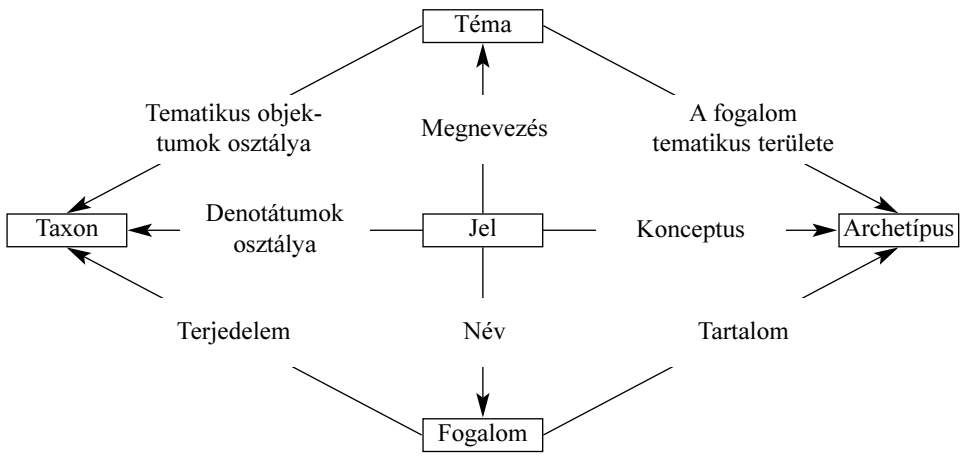
Legyen T taxon, melynek van meghatározott neve, amely jelöli e taxon bármely objektumát. Például a „ló” taxonnév bármely lovat jelölhet. Az egyes konkrét lovak a taxonnév denotátumai. De másként is megfogalmazhatjuk ezt: ha a taxonnal nemcsak a nevet, hanem a fogalmat is összekapcsoljuk, akkor a taxon egyszerűen a fogalom terjedelme. Következésképpen minél kisebb a taxon, annál szűkebb a fogalom terjedelme. Most emlékezzünk, hogy a fogalomnak nemcsak terjedelme, hanem tartalma is van. Analóg módon a névnek nemcsak denotátuma, hanem konceptusa is van. Mi lesz a taxonnév konceptusa és mi a megfelelő fogalom tartalma? Emlékezzünk, hogy a taxont természetes halmazként, azonos tulajdonságú tárgyak összességeként értelmeztük. A taxon az összes olyan és csak olyan objektumból áll, melyeknek az adott taxonra jellemző struktúrájuk van. Ezt a szerkezetet helyénvaló az archetípussal azonosítani. Nevét a szervezet felépítési tervét reprezentáló goethe-i „archetípus” fogalmától vettük át. Pont az archetípus a taxonnév konceptusa és a megfelelő fogalom tartalma. Az archetípus a taxon összes objektumának mintegy felépítési elve vagy terve.

Így a jel (a fogalom-név) megmutatja a taxont alkotó denotátumok (jelöletek) osztályát. E jel értelmét (konceptusát) e taxon archetípusa fejezi ki. A taxon a megfelelő fogalom terjedelmét, az archetípus a tartalmát jellemzi. Ezt a szemiotikai kapcsolatot mutatja a 4. ábra, ahol ezen kívül látható a jel kapcsolata a „témával” is.

Amikor a taxonómiai struktúrát vizsgáljuk, akkor a külső rendszer tanulmányozásának módszeréről van szó. A természetes rendszer taxonjai képviselik az osztályozott objektumok, a belső rendszerek (szervesen hasonló objektumok) lényegi összefüggéseit. A taxonómiai szerkezeten belül kialakuló belső leképzés (tükrözés) homológiái meghatározzák eme leképzések megválasztásának stratégiáját. Ez a választás a homológia formájában játszódik le. A tanulmányozott objektum (belső rendszer) homologizálódik a már ismert,

vele közös taxonba tartozó objektumokkal. Ezek az objektumok a reprezentátorok („kifejező objektumok”) szerepét játsszák, amelyekkel új taxonobjektumok ismerhetők meg. Kiderül, hogy ha a belső rendszert a neki megfelelő külső rendszer adta keretben tanulmányozzuk, akkor a homologikus „tükrözés” segítségével felismerhetővé válik a belső rendszer szerkezete.

Homológiáról akkor van szó, ha két dolog strukturálisan azonos, noha külső jegyeiben különböznek. Például az állatok mellső végtagjai (a madár szárnya, a fóka uszonya, a lovak mellső lába, az ember karja) között homológia áll fenn.



4. ábra

Az archetípus fogalmának bevezetésével nemcsak a meglévő objektumok, hanem az elképzelt objektumok osztályozásáról is beszélhetünk. Például, ha meghatározzuk a páros ujjú patások archetípusát, akkor ennek nemcsak a valós lovak meg az orrszarvúak felelnek meg, de az elképzelt szervezetek, mondjuk az egyszarvúak is.

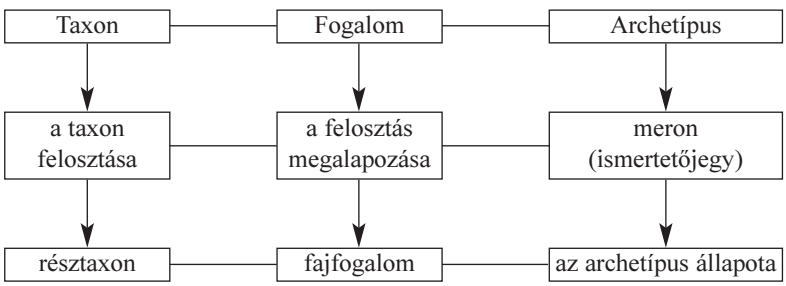
Az objektumok elképzelt sokfélesége mindig jobban „szervezett”, mint a meglévő sokféleség. Ez nagyon fontos módszertani elv, amelynek nagy szerepe van a tudományos leírásokban. Az elképzelt sokféleség általában szimmetrikusabb, szerkezetét könnyebb leírni. A reális objektumok sokféleségére a meghatározott törvényszerűségeken kívül még véletlenszerűségek is rárakódnak. Bizonyos szerkezetek nem valamilyen véletlenszerű okokból realizálódtak. Például, amikor az elemi részecskék osztályozásában a „részecske” szimmetriáit vizsgálják, az osztályozás megszépül. Mindazonáltal antirészecskék nincsenek természetes körülmények között. Bizonyos reális szerkezeteket egyszerűen módszertanilag nehéz felfedezni.

Hasonló a helyzet az elemek periodikus rendszerével. *Mengyelejev* rendszerébe bevezetett néhány kigondolt elemet, s ennek köszönhetően a rendszer szimmetrikus, szép lett. Nem félt üres helyeket hagyni a még fel nem fedezett elemek részére. Az ilyen struktúrák azon alapulnak, hogy nagyon jól ismerjük a létrehozott osztályozás természetes alapjait. Érvényes még egy fontos módszertani elv: *minden, amit a jó elmélet megenged, végül is létezik a természetben.*

Vizsgáljuk meg részletesen az archetípus fogalmát. Az archetípus a jellemző részek bizonyos halmaza. Az archetípus részeinek jelölésére *S. V. Mejen* a „meron" terminust (a görög meros = rész alapján) javasolta. Az archetípus meronokból álló szerkezet és a meronok az archetípushoz úgy viszonyulnak, mint rész az egészhez. Hangsúlyozandó azonban, hogy a meron nem elkülöníthető rész, vagyis az archetípus nem a meronok aggregátuma, hanem egységes szerkezet.

A meronok bizonyos állapottal rendelkezhetnek. Például a növényevők archetípusában van a „patásnak lenni" meron. Ennek a meronnak két állapota lehet a „páros ujjú patás" és a „páratlan ujjú patás". E példán látható, hogy a meron az ismérvnek felel meg, az állapota pedig az ismérv konkrét értékével azonos.

A meron állapotának kiválasztása meghatározza az archetípus állapotát, mely meghatároz valamely résztaxont. (5. ábra).



5. ábra

Általánosságban az archetípust frame formájában írhatjuk le, azaz $K = < M, \{r_i\}, \{P_j\}, \{A_k\}$, ahol

M = a meronok halmaza;

r_i = a meronok közötti relációk az archetípusban,

a P_j konkrét realizációként meghatározza az archetípus meronjainak állapotát,

az A_k axiómák pedig meghatározzák a megengedett állapotok halmazát.

Az általánosított archetípus kialakításakor az osztályozó akarva-akaratlan a taxonok szerkezetét használja. De igaz az ellenkezője is: a taxon helyes

megválasztásához a taxont az olyan objektumokból kell kialakítani, amelyek archetípusai helyesen viszonyulnak a taxon archetípusához. Leszögezzük azt a módszertani nehézséget: az archetípusokat a taxonómia, a taxonokat pedig az archetípusok, vagyis a meronómia határozzák meg. Ezért a taxonómiát és a meronómiát megbonthatatlan kapcsolatban lehet csak vizsgálni.

> Más szóval: ahhoz, hogy valaki például kialakítsa az ETO rendszerét – melynek köztudomásúan teljesen elvont, nem az objektumokon alapuló, tizedes szerkezete van –, elkerülhetetlen, hogy mégis az objektumokból és azok ismeretterületeiből induljon ki, azaz lényegében taxonomikus struktúrából. Ugyanakkor az objektumok bármilyen osztályát (taxonját) is alakítják ki, tudva vagy öntudatlanul figyelembe veszik ezeknek az objektumoknak a belső jellemzőit, az archetípusuk hasonlóságát.

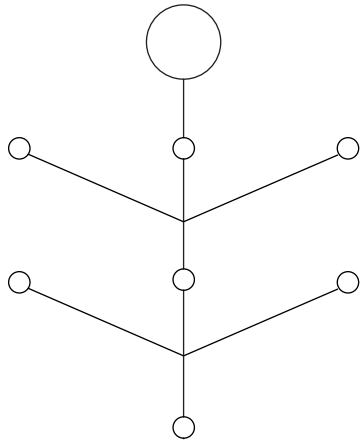
5.4. A dualitás elve

Vizsgáljuk meg a meronómia fogalmát, feltételezve, hogy a taxonómia már adott. Ez a feltevés megengedhető, amíg a meronómiát fogalomként és nem a realitás leírásaként vizsgáljuk, vagyis amíg a módszertani szakember és nem a tárgyanyaggal kapcsolatos osztályozó szempontjából foglalkozunk a témával.

Ha a taxon minden objektumának közös (általánosított) archetípusa van, ez rögtön a taxon objektumai közötti „homológia" fogalmához vezet. Legyen T az α archetípusú taxon, x pedig a T taxon eleme, esetünkben valamilyen élő szervezet. Akkor az archetípus felfedezése a szervezetben azt jelenti, hogy az x szervezet bizonyos részének megfelelnek az archetípus bizonyos meronjai. De az archetípus meronjainak analóg módon megfeleltethetők a T taxon bármely másik y szervezetének részei. Így az archetípuson keresztül az x szervezet egyes részeinek megfeleltethetők az y szervezet egyes részei. Például a négylábúak (tetrapodák) archetípus egyszerűsített vázlata a következő meronokból áll: fej, nyak, jobb mellső végtag, bal hátsó végtag stb. (6. ábra).

Ha tudjuk, hogy a madarak szárnya megfelel az archetípus mellső végtagjának, az embereknél pedig a mellső végtagoknak megfelelnek a karok, akkor levonható a következtetés, hogy a madárszárny homológ az emberi karral. Itt a következő logikai eljárás használatos: meg kell találni a reális objektumban a meronokat és fordítva a szervezet meronjai alapján meg kell találni a másik szervezet e meronnal homológ részét.

A meronok részekre oszthatók. Ezáltal az archetípus részletesebbé válik. Például a négylábúak „mellső végtagja" felkarra, alkarra és kézfejre bomlik. Ilyenkor a szervezetek részeinek részletesebb homológiája állapítható meg: a madárszárny „felkarjának" megfelel az ember felkarja stb.



6. ábra

A homológia tehát az archetípus struktúrájával függ össze. De lehet más logikai úton haladni: feltesszük, hogy az archetípus a homológia alapján épül fel. Először megállapítjuk a homológiát, vagyis bizonyos megfeleltetés alakítható ki az összehasonlítandó szervezetek részei között. Majd meghatározható a meron, mint a taxon összes szervezetére jellemző homológ részek osztálya. Az archetípus meronjai közötti relációk pedig azok a relációk lesznek a szervezet részei között, amelyek megmaradnak a homológia esetén.

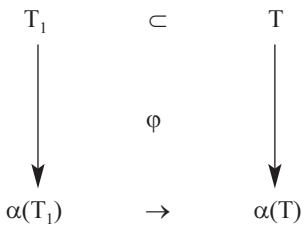
Ilyenkor feltételezzük, hogy van a taxon szervezeteinek részekre osztásához olyan módszer, melynél később megállapítható e részek közötti homológia.

A rendszer részekre bontása azonban általában nem hajtható végre egyértelműen. A rendszernek (ezen belül a szervezetnek) mint részek halmazának különböző megjelenítési formája létezik. A rendszerek felbonthatók úgy is, hogy a taxonokon belül semmilyen homológia nem ismerhető fel. Az ilyen felbontások alapján lehetetlen az archetípus felépítése. A taxon archetípusát „meglátni" nem jelent egyebet, mint hogy megtaláljuk a taxonhoz tartozó objektumok adekvát felbontását. Itt nem-triviálisan összemosódik a „belső" és „külső rendszer" fogalma. Az első a meronómiával, a második pedig a taxonómiával függ össze.

Mindkét út – az archetípustól a homológiához és a homológiától az archetípushoz – egyenrangú. A reális osztályozások általában az archetípus felépítésének iterációs módját használják, mely megszünteti a fent említett módszertani nehézséget. A szervezet adekvát bontásához legalábbis körülbelüli elképzeléssel kell rendelkezni az archetípusról, amelyet közvetlenül vizsgálunk. Majd megállapítjuk a homológiát és pontosítjuk az archetípust. Az archetípus pontosításának alapján már pontosítható a homológiarendszer. Az archetípusról alkotott előzetes elképzelés nélkül ezt lehetetlen lenne megtenni.

Egy egyszerű példa: a kígyónak nincsenek végtagjai. Ugyanakkor helytálló, ha azt mondjuk, hogy a kígyó állat, melynek négy nulla-végtagja van. Itt két dolgot kell megkülönböztetni: a nulla-állapotú meront és a meron hiányát. Például az orsóféregnek nincs végtag meronja. Nincs értelme az orsósféreg végtag meronjának nulla-állapotáról sem beszélni, mivel ezt a meront semmivel sem lehet homologizálni. De a kígyók helyzete más a szervezetek rendszerében: az összes többi döntő ismérve alapján a négylábúak közé tartozik, bár az osztály, melybe tartozik, éppen tőle kapta a „csúszómászók" nevet. Ezért is alakul ki a homológiák természetes szerkezete, mely ez esetben arra kényszerít, hogy a nulla-végtagokról, a meronok nulla-állapotáról beszéljünk.

Most vizsgáljuk meg a következő formális struktúrát: legyen T taxon $\alpha(T)$ archetípusú és tartalmazza a T_1 taxont, mely $\alpha(T_1)$ archetípusú. Az $\alpha(T)$ és $\alpha(T_1)$ archetípusok közötti relációt kifejezi az alábbi diagram:



Valóban, ezek az archetípusok nem lehetnek tetszőlegesek. A T_1 taxon minden objektumának egyidejűleg az archetípusa kell hogy legyen az $\alpha(T)$, mivel ez az objektum az általános taxonhoz is tartozik. Ez azt jelenti, hogy az $\alpha(T_1)$ archetípus „tartalmazza" az $\alpha(T)$ archetípust. Az $\alpha(T_1)$-ben vagy vannak kiegészítő meronok, vagy az $\alpha(T)$ meronjai összetevődnek. Valójában ez azt jelenti, hogy az $\alpha(T_1)$ archetípus homomorf módon tükröződik az $\alpha(T)$ archetípusban. Az $\alpha(T_1)$-ben lehetnek kiegészítő meronok, melyek nincsenek kapcsolatban a közös archetípus meronjaival. Például az emlősök tejmirigyei semminek sem felelnek meg a gerincesek közös archetípusában. Így, ahogy a T_1 taxon T taxonba tartozik, úgy az $\alpha(T_1)$ archetípus homomorf módon tükröződik az $\alpha(T)$ archetípusban.

A harmadik fejezetben már szó volt egy nagyon fontos fogalomról – kategóriáról –, az objektumok osztályáról, és az egyik objektumot másikba leképező tükrözések (morfizmusok) halmazairól. Itt teljesülnek bizonyos axiómák. A tartalmazása műveletű taxonok halmaza kategóriaként fogható fel. A taxonkategóriák objektumai maguk a taxonok, a morfizmusok pedig a taxonok tartalmazási leképezései. Az archetípus-kategóriák objektumai az archetípusok. Létezik tehát az archetípusok kategóriája és a taxonok kategóriája, a taxonoknak az archetípust megfeleltető leképezése pedig funktor, azaz az eredeti szerkezetet érintetlenül hagyó leképzés (könyvünk korábbi, 3. fejezetében mutattuk ezt ki). *Az osztályozás tehát nem más, mint a taxonok kategóriájából*

az archetípus kategóriájába vezető funktor. A különböző funktoroknak különböző osztályozások felelnek meg.

Ha a T_1 taxon a T taxonba tartozik, akkor a T_1-nek megfelelő fogalom terjedelme szűkebb a T-nek megfelelő fogalom terjedelménél. De az első fogalom tartalma nagyobb, ha neki bonyolultabb archetípus felel meg. Minél nagyobb a fogalom terjedelme, annál kisebb a tartalma és fordítva. Ez az elv jól ismert még Arisztotelész idejéből, de jelen esetben ez egyben az osztályozáselmélet dualitáselve is.

A meronómia bevezetése lehetővé teszi az objektumok hasonlóságának felértékelését az archetípusokon keresztül, mégpedig minél teljesebben tükröződik a t típusú archetípus szerkezete a t' típusú archetípus szerkezetében, annál közelebb állnak ezek a típusok egymáshoz. A dualitás elve biztosítja a hasonlóság-értékelés két módszerének – a taxonómiának és a meronómiának – az összehangolását.

Az osztályozó tevékenység tehát felbontható taxonómiára és meronómiára. Mivel kell kezdeni: a taxonómiával vagy a meronómiával? A gyakorlatban az osztályozáselmélet iterációsan épül fel. Soha nincs előre kész taxonómia vagy meronómia. Először elkészül egy „nyers" taxonómia, erre épül a „nyers" meronómia, majd erre épül a taxonómia. Ha a taxonómia nem nagyon elégít ki minket, akkor módosítjuk a meronómiát stb. Munkahipotézisünk volt például, hogy a cet hal (azaz a Hal fajtája a Cet). De az osztályozás megújítása során arra a következtetésre jutunk, hogy a cet nem hal, hanem emlős. Ez a következtetés a tudomány eredménye. Egy másik példa: próbáljuk meg az emlősök archetípusát egy ismérv alapján meghatározni, mely szerint „utódjaikat tejjel etetik". De hiszen a méhek a lárváikat szintén „tejjel" etetik. A méheket nem sorolhatjuk az emlősök közé. Következésképpen korrigálni kell az emlősök feltételezett (hipotetikus) archetípusát. Hangsúlyozandó, hogy a taxonómia és a meronómia viszonya duális és nem kiegészítő. Minél jobban ismeri az osztályozó a taxonok reális szerkezetét, annál pontosabban határozhatja meg az archetípust és valósíthatja meg a homológiát. És fordítva az archetípusok pontosabb meghatározása elősegíti a taxonomikus szerkezet pontosabb meghatározását. A kiegészítés az ellenkezőjét jelentené, amikor az osztályozásnak csak egyik aspektusát lehet jól leírni.

Az osztályozás tehát nemcsak az objektumok hasonlóságának megismerési módszere, hanem a hasonlóságot a természetük egységével szembeállító módszer is. A belső rendszerek tanulmányozásakor az alapkérdés: „Milyen részekből áll az egész és mi az adott résznek a más részekhez fűződő viszonya által kifejezett szerepe?". A külső rendszerek tanulmányozásakor arra várunk választ, hogy „miként csoportosulnak a hasonló objektumok és mi határozza meg hasonlóságukat?". De ha beszélni lehet az ilyen objektumokról, máris felmerül még egy kérdés: „Hány ilyen van?". Majd ez után a metakérdés „Létezik-e megalapozott törvényszerűség arra vonatkozóan, hogy hány van?". A

megfigyelés szintjén az ilyen törvényszerűségek egész sor olyan szakterületen fölvetődtek, melyek a biológiai és a szociális rendszerekkel függnek össze: a városok megoszlása lakosságuk száma szerint, a születések megoszlása típusonként vagy regionálisan, a lakosság megoszlása jövedelem szerint, az ökológiai rendszerben a fajok megoszlása létszám vagy biotömeg szerint, a szavak megoszlása a szövegben előfordulásuk gyakorisága szerint, a folyóiratok megoszlása az adott témában publikált cikkek mennyisége szerint stb.

Meglepő, hogy e törvényszerűségek nagyon hasonlóak. Tartalmilag egy és ugyanazon mennyiségi törvényszerűség formájában jelenik meg a nyelvészetben (Zipf-törvény), a földrajzban (ugyancsak Zipf-törvényként, ez ugyanaz a Zipf, mint a nyelvészetben), a biológiában (Wilks-törvény), a szociológiában (Pareto-törvény), az informatikában (Bradford-törvény) stb. És ezek a törvényszerűségek mind az úgynevezett rangsorolt eloszlások jellemzőit határozzák meg. Ezzel túllépnek a konkrét tudományos eredmény keretein és általános rendszerelvvé válnak, aminek módszertani jelentősége van. Ez az elv adja a rendszer „természetességének" tartalmi és műveleti kritériumát és „megsúgja" leírásának célszerű útjait.

A kettősség elve az osztályozáselméletben[34]

1. Az osztályozás helye a tudományos kutatásban

Az osztályozásnak az egyes tudományokban más és más a szerepe. Van, ahol – például a könyvtárakban, dokumentációs intézményekben – az osztályozás a kutatások előzetes előkészítését szolgálja. Más területeken tudományos vizsgálatának végleges eredménye, célja az osztályozás. A matematikában például először a tárgyak vagy objektumok bizonyos osztályainak axiómáit határozzák meg, ezt követi ezeknek a tárgyaknak a tartalmi osztályozása, mint „végleges" eredmény, amelyhez hozzátartozik a tárgyak vizsgált osztályának a pontos leírása és lényege is. Ilyen eredmény például a mátrixelméletben a Jordan-féle normálforma alkalmazásának elve (melyben megadják a lineáris leképezés összes invariánsait); vagy az egyszerű Lee-csoportok osztályozása stb.

Minden ilyenféle osztályozás elve a következő: először felépül a formális axiomatikus elmélet. Ezután tisztázódik, hogy az elmélet modelljei „konstruktív módon" leírhatók valamilyen természetes hasonlósági leképezés (morfizmus) alapján – vagyis struktúratípusok állapíthatók meg. Az axiomatikus meghatározástól a megengedhető lehetőségekig vezető út a matematika egyes szakterületeinek története egyben.

---

34 Princip dvojstvennosti v teorii klassifikacii / Nadežda Semenova Panova, Jurij Anatoljevič Šrejder. In: Naučno-Techniceskaa Informaciă, Ser. 2. 1975, No. 10, p. 3–10.

Mindez valójában nem megy végbe zökkenőmentesen; egyes szakterületeken különböző axiomatikus megállapításokat variálnak, és helyesbítik azokat attól függően, hogy mennyire eredményes a megfelelő modellek osztályozása. A fizikában jobban látható az empirikus extenzionális (induktív) és a természetes (deduktív, intenzionális) osztályozás ellentéte.

Az induktív osztályozással az elért eredményeket hozhatjuk alkalmas formátumra; a természetes (deduktív) osztályozással pedig az (osztályozott) tárgyak lényeges tulajdonságai ismerhetők föl. Például: az ismert elemi részecskék feloszthatók töltésük, tömegük, vagy különböző kölcsönhatásokban való részvételük jellege szerint (ebből áll az empirikus osztályozásuk). Másrészt a részecskéket deduktív módon sikerül osztályba sorolni valamilyen szimmetriatulajdonság alapján, vagy elméletet dolgoznak ki a részecskéknél kisebb kvarkok szerkezetéről, s ezáltal az összes létező és lehetséges részecske leírható, mint a feltételezett szubrészecskék, a kvarkok által alkotott struktúra.

Az empirikus (extenzionális) osztályozások, ellentétben a természetes (intenzionális) osztályozással, a tudomány fejlődésének kezdeti szakaszában jelennek meg, amikor is magát a kutatás tárgyát határozzák meg pontosabban. A természetes (intenzionális) osztályozással az adott tudományban új felismerésekhez lehet jutni, általa tovább fejlődik a tudomány.

A dokumentációban és katalogizálásban alkalmazott mai osztályozások ez ideig csak az információtömegek gyakorlati csoportosítását képviselik. Nem találhatók meg bennük az objektív, mélyben rejlő állandók. Az empirikus (induktív) osztályozásból az átmenet a természeteshez (a deduktívhoz) elválaszthatatlan azoknak a deduktív elméleteknek a fejlődésétől, amelyek a tárgyak axiomatikus leírását összekötik e tárgyak minden lehetséges modelljének osztályaival. Az is nagyon fontos, hogy a tárgyak osztályainak leírásakor felhasználhatjuk a logika intenzionális kategóriáit.

Ezzel kapcsolatban érdemes hangsúlyozni: az utóbbi években az a tendencia érvényesül, hogy a tiszta extenzionális tudományos leírást fokozatosan megpróbálják kiegészíteni az intenzionális leírással.

2. Taxonómia és meronómia

Korábbi munkánkban az osztályokat extenzionálisan írtuk le, mint az osztályozási mező sajátos részhalmazait. Ezeket a részhalmazokat neveztük taxonoknak; köztük a szokásos halmazelméleti viszonyok állapíthatók meg (üres, és nem üres metszetek stb.)[35]

35 Az extenzionális fogalmát abban az értelemben használjuk, ahogy azt R. Karnap vezette be.

A hierarchikus osztályozás extenzionálisan megfelel annak az esetnek, amikor a taxonok sokasága A tartalmazásra nézve fát alkot; a taxonok sokasága – akárcsak egy tetszőlegesen lemetszett fa egyik fele – az osztályozási mező egyik felét képviseli. A kombinatív, fazettás osztályozás ezzel ellentétben extenzionálisan annak az esetnek felel meg, amikor a kettéosztott osztályozási mezőt alkotó csoportokban (fazettákban) mindegyik taxon valamely fazettából álló taxonok metszetéből adódik.

Intenzionálisan az osztályokat úgy írhatjuk le, hogy az osztályozási mezőt olyan univerzumig bővítjük, amely minden gondolt objektumból áll és a taxonok szerkezetével izomorf szerkezetet nem az anyagi dolgok, hanem az osztályozási ismérvek alkotják.

R. Karnap szerint a predikátum (az állítás) intenzionálisan nem más, mint az összes képzetes objektumok osztálya, melyre a predikátum igaz értéket vehet föl. De akkor a gondolat önmagában nem létezik, hanem csupán a gondolt megnevezések osztályában tárul fel. Ebből a tézisből kiindulva korábban áttekintettük az ismérvek szerkezetét, azok jelentését, mint amelyek megfelelnek, összhangban vannak az osztályozási univerzum hozzájuk tartozó taxonjainak szerkezetével. Kiindulásuk alapján az ismérvek önmagukban, a taxonoktól függetlenül nem léteznek. Így a következő rejtett ellentmondásba ütköztünk: egyrészt kijelentettük, hogy az extenzionális és intenzionális osztályozás egymás ellentétei, vagyis elválasztottuk egymástól a taxonokat és az azokat meghatározó ismérveket; másrészt bebizonyítottuk, hogy a taxonok és ismérveik szerkezete izomorf egymással (azaz nem választhatók el).

Ebben a cikkben azt próbáljuk megmutatni, hogy a valódi osztályozáselméletnek (és minden osztályozási rendszernek) két független komponense van: a taxonómia (a faj–nem kapcsolatban álló taxonok szerkezete) és meronómia (az egymással asszociatív kapcsolatban álló ismérvek szerkezete).

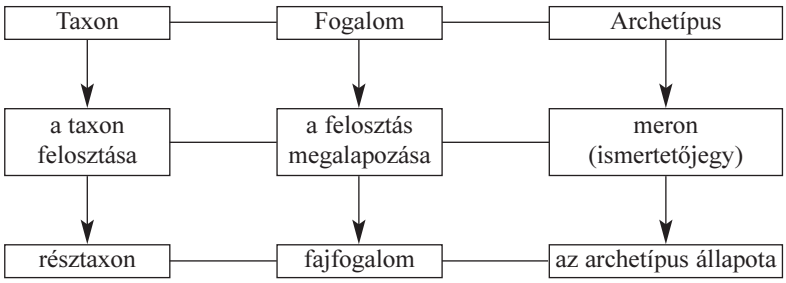
A továbbiakban számunkra nagyon fontos pontosítani, hogy mi a fogalom. Vizsgálódásunk tárgya a nem elvont fogalmak, más szóval a nem valódi osztályok (azaz olyan fogalmak, amelyek konkrét tárgyat, jelenséget vagy körülményt reprezentálnak) és nem azok a valódi osztályok, elvont fogalmak, amelyek a felsoroltak ismérveit, tulajdonságait képviselik. Másképpen fogalmazva: az olyan típusú fogalmakat vizsgáljuk, mint a ló, az eladás, a vörös fogalma, nem pedig a „lóság", „árusítás", „vörösség". Megkülönböztetjük a fogalomneveket (fogalmakat jelölő szavakat) maguktól a fogalmaktól.

A fogalom terjedelme azoknak az objektumoknak az osztálya, amelyeket az adott fogalom reprezentál, képvisel, azaz e fogalom nevének denotátumai által alkotott osztály. (Itt már eltávolodunk *R. Karnap* nézetétől, aki szerint a képzetes denotátumok osztálya az általános vagy fogalomnév intenzionálisa, és inkább a fogalom tartalmához, mintsem a terjedelméhez kapcsolódik.)

A fogalom tartalmát vagy (ahogyan a szemiotikában elfogadott) a fogalomnév (az általános név) konceptusát összekapcsoljuk valamilyen struktúrával (az archetípussal), amely felfedezhető a hasonló taxonok minden objektumában.

Megjegyezzük, hogy a fogalomnév egyúttal a taxon megnevezése. A név ebben az esetben többes számban is szerepelhet: a „ló" a fogalom, a „ló" taxonját a „lovak" nevezi meg. Az archetípus névalakját a megfelelő elvont fogalom képviseli (példánkban a „lóság").

A leírt fogalmak kapcsolatait az 1. ábrán láthatjuk.



1. ábra

Az archetípuson a részek (morfológia) és külső funkcionális kapcsolatok (ökológia) ama szerkezetét értjük, amely az adott fogalom taxonjába tartozó mindegyik objektumra jellemző. Ezeket a részeket *Sz. V. Mejent* követve meronnak hívjuk. A fogalom eme értelmezése alapján az osztályozási rendszer a fogalmak rendszere, mely egyrészt meghatározza a hozzá tartozó (kapcsolódó) taxonok megfelelő szerkezetét, másrészt az archetípusok szerkezetét. A taxonok saját szerkezete alkotja az osztályozás taxonómiai összetevőjét, a taxonómiát. A taxonómia úgy is jellemezhető, mint az osztályozás extenzionális megközelítése. Az osztályozási fogalmaknak megfelelő archetípusok rendszere, éppúgy, mint az osztályozott objektumokban levő archetípusok megállapítási, elemzési módszerei a meronómiai összetevőjét, a meronómiát alkotja. Az osztályozás intenzionális megközelítését a meronómia képviseli.

Az osztályozás problémáinak tisztán taxonómiai megközelítése azért problematikus, mert ebben a megközelítésben is az ismérvek közötti intenzionális viszony alapján elemzünk, miközben az intenzionális viszony természetéről semmit sem tudunk. Noha a közös osztályozási ismérv alapján a taxont részekre kell tudni osztani, ezt az elvet sehogy sem sikerül bizonyítani. Ha azonban úgy járunk el, hogy a fogalomhoz kapcsoljuk a konkrét objektumok – extenzionális – halmazát, amely elemek e fogalom „megtestesítőivé" válhatnak, akkor az, ami az intenzionális részt képviseli, általában magával a fogalommal azonosítható. Más szóval: maga a fogalom azonos az eszmével (platóni értelem-

ben), amely a hasonló taxonokat alkotó objektumokban „testesül meg". Úgy is mondhatjuk – eszme helyett –, hogy a fogalom tartalma (a fogalomnév képzete) valamilyen absztrakt struktúra.[36]

A fogalom eme analitikus értelmezésben a tartalom és terjedelem szintézise. A fogalom terjedelme a fogalomnévvel jelölt taxon. Beszélhetünk valós fogalomterjedelmekről – a való világ létező objektumainak halmazáról – és a képzeletbeli (eszmei) fogalomterjedelmekről – az adott fogalomnévvel jelzett képzelt objektumok halmazáról. A fogalom tartalmát struktúraként fogjuk fel, és ezt a struktúrát nevezzük archetípusnak. Az osztályozási univerzum minden olyan objektuma, amely az adott struktúrával rendelkezik, ugyanabba a taxonba tartozik. Az osztályozott objektumokat ezentúl a meronómiában nem egyszerűen mint egészet alkotó egységeket szemléljük, hanem mint bizonyos szerkezettel bíró objektumokat (archetípusokat).

A későbbiekben pontos meghatározást is adunk, addig pedig fogadjuk el, hogy az archetípus meronokból áll. Az archetípus – a részek és a közöttük fennálló kapcsolatok szerkezete – az adott taxon minden objektumára jellemző. A közös archetípusok megállapításának jellegzetes szakterülete a biológiai morfológia: itt az élő formák kutatása során kezdtek intenzívebben foglalkozni a közös archetípusokkal. Világítsuk meg az archetípus fogalmát egy szándékosan egyszerű példán, az asztal fogalmán („bútortárgy" jelentésében). Bármely asztal esetében három meron választható ki: a munkafelület (tető, deszka), az alapzat (lábak, talapzat) és az asztal rendeltetése. Az első két meron egymással a „tartja" viszonyban van. Az első és harmadik a „rendeltetése" viszonyban. A szokásos ebédlőasztal esetén az alapzat maga is további meronokra oszlik (az egyes lábak). Néhány (kinyitható) ebédlőasztal esetén a munkafelület két meronra oszlik: a fő és a kiegészítő (csukott állapotban rejtett) részre. Az íróasztaloknál az aljzat meronjai tűnnek ki, a hivatali íróasztaloknál kiegészítő meronként szerepelnek az asztal lapján található rekeszek. Ezen a példán világosan látható a fontos elv: a kisebb taxonnak bonyolultabb archetípusok felelnek meg.

Mindegyik meron lehet ilyen vagy olyan állapotban (például az asztal alapzata állhat lábakból vagy talapzatból). A taxonnak tehát meghatározott archetípus (struktúra) felel meg, amely a taxon minden objektumára vonatkozóan azonos, és csakis azokra nézve azonos. Ugyanezért beszélhetünk a taxon elemei közötti homogenitásról is: találhatók bennük egyforma archetípusok, az adott taxon különböző objektumaiban felfedezhető meronok pedig kölcsönösen egyértelműen megfeleltethetők egymással. Ily módon természetesen felvetődik a meronok hasonlósága is, ezek a közös archetípus esetén egy és ugyanazon meronnak felelnek meg. A hasonlóság annak következménye, hogy taxonoknak közös archetípusa van. Lehetséges egy másik szempont is: előbb létrejön a két

36 Ezáltal eltávolodunk R. Karnap álláspontjától, de közeledünk G. Frege nézetéhez, ami szerint a predikátum értelme az információ, melyet a lehetséges denotátumokban hordoznak.

elem közötti hasonlóság mint a szerkezetek valamilyen egymásra tükröződése. Azután pedig ezeknek a tükröződéseknek eredménye a taxon archetípusaként jelenik meg.

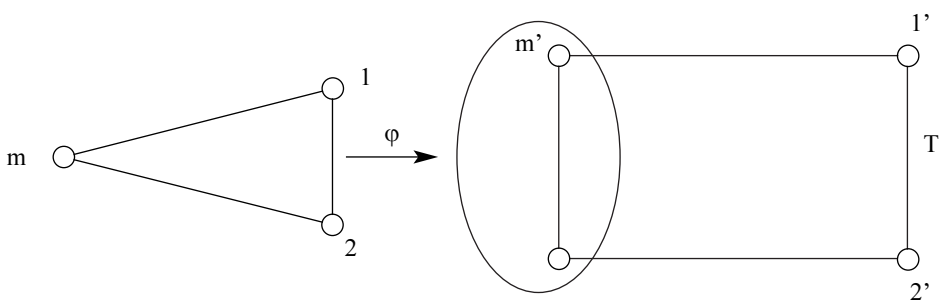
Csak akkor keletkezik meron, ha felismerhetővé válik a taxonok objektumainak minden párja között a hasonlóság. Például: a pata fogalmát csak akkor ragadhatjuk meg, ha megtanultuk minden patásállatnál egymással összevetni a láb szaruképződményeit. Sőt, csak miután felismertük a fejlett gerincesek végtagjai közötti hasonlóságot, beszélhetünk a kígyóknál elcsökevényesedett végtagokról. A férgeknél vagy a tengeri csillagnál viszont a láb nem vetődik fel végtagfogalomként. Hangsúlyozzuk: az archetípusban hiányzó meron (a féreg végtagjai) és a zérus állapotú meron (a kígyók végtagjai) között alapvető, elvi különbség van.

Az archetípus szerkezetével, és az archetípust alkotó (morfológiai és ökológiai) részekkel – a meronokkal – foglalkozó szakterületet a továbbiakban meronómiának nevezzük, és az osztályozás intenzionális formájával kötjük össze. Hasonló gondolatmenetet követve alkotta meg *S. Lesznyevszkij* lengyel logikus a merológia fogalmát, melyen részekre tagolt objektumok formális elméletét értette. Számunkra a meronómia a taxonómia kiegészítése. Más szóval az osztályozás intenzionális leírása.

A meronómiában lehetséges egy realista megközelítés (amely az objektumok hasonlóságát megszabó közös archetípus megállapításán alapszik) és egy nominalista megközelítés (amely szerint előbb a részhasonlóságokat határozzák meg, és ennek alapján az absztrakció segítségével jutunk el az archetípushoz.)

Mindkét megközelítés esetében beszélhetünk az adott archetípushoz tartozó meronok szerkezetéről. Mindegyik meron különböző állapotokban lehet (különböző modalitással bírhat). Minden meron összevethető a taxon archetípusában az ebben a taxonban működő, az adott meronnal azonos nevű ismérvvel, az ismérv jelentése pedig megfelel majd a meron állapotainak.

Valamely <u>m</u> meron T archetípusbeli állapota értelmezhető úgy, mint valamilyen másik T' archetípus leképzése a T-re, ahol is több m' meron egyetlen m meronba képződik le (2. ábra).



2. ábra

Az említett archetípusok megengedhető leképezései az m meron állapotainak felelnek meg.

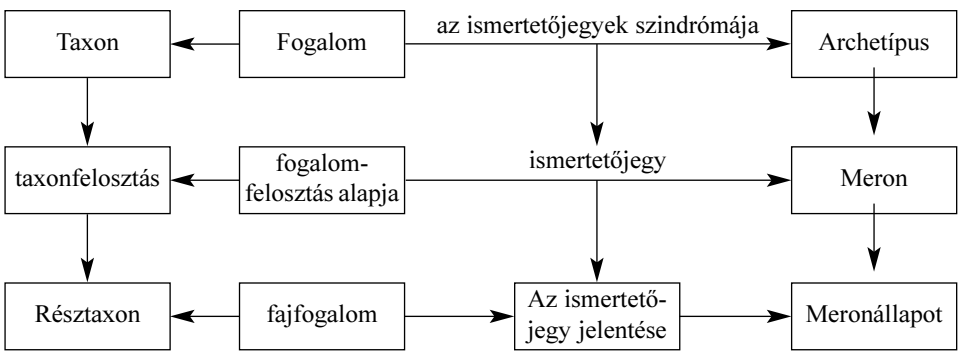
Például a növényevő emlősöknél egymáshoz hasonló a „pata” meron megléte alapján olyan taxonok választhatók ki, mint „párosujjúak” és „páratlan ujjúak”, a „pata” meron két lehetséges állapotának megfelelően.

A meronómia mindegyik archetípusának megfelel a taxonómia egyik taxonja. Az adott archetípusnak megfelelő taxonba kerülhet az osztályozási univerzum minden olyan objektuma, amelyben fellelhető az illető archetípus.

Ezáltal az adott archetípus összekapcsolódik az adott archetípusra jellemző objektum fogalmával, s ez lesz a fogalom tartalma. Például: az elefánt fogalmának archetípusa meghatározott felépítésű élő szervezet, mely többek között elefántormánnyal rendelkezik.

Most megpróbáljuk vázolni a taxonómia és a meronómia közötti kölcsönös viszonyt.

A 3. ábrán tüntettük fel a kapcsolatot a taxonómia és a meronómia kategóriái között.



3. ábra

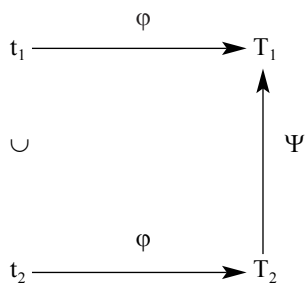
Az osztályozási univerzumban a taxon képviseli a fogalom terjedelmét (halmazát, pontosabban osztályát). Más szóval a taxon az adott fogalommal reprezentált objektumok osztálya. E taxon neve általában megegyezik a fogalommegnevezés többes számú alakjával. A névelőt használó nyelvekben a fogalom neve (a taxon neve) névelő nélkül szerepelhet, a taxon pedig mindazokból az objektumokból tevődik össze, amelyekre ugyanazzal a névvel mutathatunk rá a határozott névelőt használva. A fogalom tartalma pedig a taxon minden objektumára jellemző archetípus.

Az archetípuson nem csupán a taxon objektumainak a belső szerkezetét értjük. Az archetípus az objektumokkal társítható szerkezet, amely nemcsak azok sajátos felépítését, hanem a külső kapcsolatait is leírja (úgy, ahogy a biológiai osztályozások számára nemcsak a morfológia fontos, de az ökológia is).

Térjünk át most a 3. ábra második sorára. A fogalom generikus felosztását meghatározott szempont (genus species) alapján végezhetjük el. Ehhez olyan ismérvet kell választani, amely lehetővé teszi a fajták meghatározását. Megjegyezzük, hogy az ismérv neve egybeeshet a fogalom nevével (a szín egyszerre fogalom és ismérv) és különbözhet (a ló: fogalom, az állat szőrének a színe: ismérv; az asztal: fogalom, rendeltetése: ismérv). Fontos, hogy az ismérv értelmezési tartománya megegyezik a taxonéval vagy annál tágabb.

Általában véve az ilyen ismérvekből sok lehet, de a 3. ábrán mi csak az egy ismérvvel leírható esetet vázoltuk. A fogalmak felosztásának a taxonómiában a taxonok felosztása felel meg. Ha az ismérv megkülönböztethető, akkor ez a taxon kettéválását eredményezi. Ellenkező esetben átfedő alosztások keletkeznek.

A meronómiában az ésszerű ismérvnek egy meron felel meg. Ez azt is jelenti, hogy az ismérv alapján a taxon felosztható. Végül a harmadik sorban a taxon felosztásáról áttérünk a meghatározott felosztáshoz, a fogalom esetében az ismérv konkrét jelentésének megfelelő fogalomfajtához, a merontól pedig a konkrét helyzethez, ami az archetípus állapotának felel meg. Az alábbi szerkezetben bemutatjuk a taxonómiában és meronómiában az átmeneteket. A taxont t és az archetípust T jelöli, a Ψ pedig az archetípusok megfelelési szabálya.



Ez a struktúra a matematika nyelvén azt jelenti, hogy az osztályozáselmélet az a funktor, amely leképezi az archetípusok kategóriájába a megfelelő taxonkategóriát.

A klasszikus logika felől nézve az ismert kettősséget, a fordított viszonyt tapasztalhatjuk a fogalom terjedelme és tartalma között. Teljesen nyilvánvaló, hogy a taxon a fogalom terjedelme. Majdnem annyira világos, mint az, hogy az archetípus a fogalom tartalmának a kifejtése. A fenti funktorban az archetípusok jobb oldalon, a taxonok bal oldalon látható leképzései mutatják, hogy a fogalom tartalmának a növelése (a gazdagabb archetípusba való átmenet) a fogalom terjedelmének csökkenésével jár. A legszegényebb archetípus a teljes osztályozási univerzumnak felel meg. A leggazdagabb a minimális (a legkisebb) a taxonnak.

Ezzel sikerült reprezentálni a fogalom terjedelme és tartalma közötti klaszszikus, fordított megfelelés pontos értelmét.

A vizsgált ábra az osztályozási rendszerek alapszerkezetét szemlélteti. Az egy ismérv alapján végzett felosztás után visszatérhetünk az eredeti taxonhoz és egy másik ismérv szerint újra elvégezhetjük a felosztást; ennek megfelelően egy másik meront választunk ki. Néhány ilyen ismétlésből kialakul az eredeti taxonhoz tartozó kombinatív (fazettás) szerkezet. Megfordítva a folyamatot, a felosztásokra alkalmazva a taxonok és az ismérveik hierarchiája alakul ki.

Amikor a fogalom tartalmát az archetípus alapján határozzuk meg, akkor ez hasonló ahhoz, amikor a szó jelentését a fogalmak közötti asszociatív szerkezet révén határozzuk meg. A kettő azonban mégse ugyanaz. Például az élő sejt felépítése lehet majdnem olyan bonyolult, mint az élő szervezeté, hiszen a sejt az organizmus genetikus típusát tartalmazza, és a sejt elemei nagyon bonyolult felépítés létrehozására alkalmasak. Ezért a fogalom tartalmának archetípussal való meghatározását sohasem kapacsolhatjuk össze azzal, hogy a fogalmat egyszerűbb fogalommá redukáljuk. Ez teljesen különböző probléma.

Az ideális (intenzionális, lényegi) osztályozási rendszer archetípus-struktúrákkal dolgozik. Például az elemek periódusos rendszere minden elemhez az elektronhéj meghatározott struktúráját rendeli, s a különböző izotópok (amelyek mint kémiai elemek nem különböztethetők meg) izomorf elektronhéjjal rendelkeznek, amelyek az atommag felépítésében és összetételében különböznek. Azonban az objektumok gyakorlati osztályozásakor általában egész sor olyan összevethető (diagnosztikai) ismertetőjelet használunk, amelyek nem találhatók meg közvetlenül az archetípusban, hanem korrelálnak vele. Így fedezzük föl a vas ama tulajdonságát, hogy vonzza a mágnes, a kismacskát a macskákhoz sorolják, minthogy macskától született stb.

Ezért az ideális osztályozási rendszer megléte nem foglalja magában a „szűk” osztályozás szükségességét, azoknak a külső ismertetőjelek meghatározását, amelyek lehetővé teszik, hogy az objektumot ehhez vagy ahhoz a taxonhoz soroljuk.

A szervezetek és élőlények természetes rendszere csak mint etalon, mint ideális modell játszik szerepet.

4. Osztályozási rendszer és tezaurusz az informatikában

A dokumentum az informatikában az osztályozás természetes tárgya. A dokumentációs célú osztályozási rendszerek segítségével a felhasználó kiválaszthatja a releváns dokumentumok taxonjait. A taxont, amelybe a dokumentum kerül, a dokumentum jelzete határozza meg. Ily módon a dokumentum jelzete a taxon neve. Például az ETO-ban az az ETO-jelzet az osztályozott taxon neve, amely minden e jelzettel bíró dokumentumból áll.

Fel kell ismernünk egy lényegi különbséget. A dokumentum jelzete vagy a deszkriptor neve nem egy tudomány területén belül megvizsgált fogalom

nevével azonos, azaz nem vizsgálati tárgyat vagy tudománykategóriát jelöl. Magának a tudománynak a megnevezése sem jelent semmiféle tudományos fogalmat. Sőt, amikor a tudományról vagy olyan ismeretterületekről van szó, amelyeken belül az objektumok konkrét osztályát vizsgálják (élő szervezeteket, vegyi kapcsolódásokat, fémötvözeteket, gépkocsikat stb.), akkor e terület elnevezése vagy egyszerűen nem tartalmazza a vizsgálat tárgyát (biológia, kémia stb.) vagy a következők szerint képződik: ötvözetek elmélete, gépjárműismeretek stb. Néha ugyanaz a szó jelenthet tudományos fogalmat és az osztályozási rendszer keretében a fogalom vizsgálatával foglalkozó szakterületet. Például: programozás, lézer, repülőgép. De minden ilyen témamegnevezés elválaszthatatlan az „adott téma dokumentumai" fogalmától és a dokumentumok taxonját nevezi meg.

Ugyanakkor a téma megnevezését úgy is felfoghatjuk, mint nem a tudományok, hanem a dokumentumosztályozás körébe tartozó fogalom megnevezését. E fogalom terjedelmét nem a tudomány objektumai, hanem a témába vágó dokumentumok alkotják. A fogalom tartalmát pedig a dokumentumokban található gondolati szerkezet képviseli, melynek alapján az osztályozó az adott dokumentumot az adott jelzethez vagy deszkriptorhoz sorolta.

Mivel ezért a dokumentumok körében az ilyen osztályozó fogalom tartalma kevésbé pontosan határozható meg, mint a tudományos értelemben használt fogalom tartalma, a gyakorlati osztályozáskor bizonytalanság keletkezik. Igen gyakran véleménykülönbségek és határozatlanság tapasztalható egy és ugyanazon dokumentum indexelésekor. Más szóval: a dokumentációs vagy osztályozó fogalmak archetípusának hiánya a fogalmak önkényes értelmezéséhez vezet. A dokumentumok taxonjainak kétértelműségét a megfelelő meronómia kidolgozatlansága okozza.

Az előzőekben kifejtettük, hogy a tárgykör taxonokra való felosztása csak akkor lehetséges, ha ezeknek a taxonoknak elég pontos tartalmú fogalmakat feleltetnek meg, azaz kidolgozták a meronómiájukat.

Az informatikában alkalmazott osztályozási rendszerek a dokumentumok, nem pedig a tudomány taxonómiáját határozzák meg. A tudomány számára pedig meghatározza a meronómiát. Valójában a „fizika" tudománya a tudományos kutatások nagy területét alkotja. Ennek a területnek része az „optika", amely maga is tudományág. De az „optika" nem fajfogalma a „fizikának". Azonban a dokumentumok osztályozásában létezik a „fizikával kapcsolatos dokumentumok" és az „optikával kapcsolatos dokumentumok" fogalma. Ezért közöttük leszármazásbeli viszony áll fenn, amely a megfelelő dokumentumtaxonok között a tartalmazása reláció áll fenn.

Az osztályozási jelzetek között csak kétféle viszony létezik: a leszármazási (eredeteredménye) és szinonímia. Ez ténylegesen a következési relációt (implikációt, tartalmazást) és a megfelelő fogalmak ekvivalenciáját jelenti (predikátumok). Az első a következőt jelöli: minden olyan x dokumentum, amelynek jel-

zete I(x), rendelkezik az I_1 (x) jelzettel is. A második reláció pedig azt jelenti, ahogy két jelzet egyazon képzetes dokumentumosztályt jelöl.

Határozzanak meg minden egyes I(x) jelzetet a $\{D_1, D_2, ... D_n\}$, minden I_1 jelzetet pedig a $\{D_1, D_2, ...D_m\}$ deszkriptorok. A legegyszerűbb esetben a következtetési reláció fennállása kettőjük között azt jelöli, hogy minden D_i-hez létezik valamely más régebbi D_j deszkriptor. Jóval bonyolultabb, grammatikával és tezaurusszal rendelkező deszkriptornyelvek esetén a jelzetek között fennálló következési relációk a deszkriptorok jóval finomabb tulajdonságait jelölik.

Figyelemre méltó, hogy a jelzetek között fennálló egyszerű következés és szinonímia relációk a deszkriptorok között fennálló meglehetősen bonyolult tartalmi viszonylatok alapján adódnak. A legegyszerűbb deszkriptornyelvekben például ez a deszkriptorok – korábbi – leszármazási relációján keresztül nyilvánul meg, amely reláció sokféle különböző tartalmi viszonyt egyesít. Éppen ezek a tartalmi relációk tükröződnek vissza az adott tudományterület tezauruszában. Ez már azt sugallja, hogy a tezaurusz a meronómiával kapcsolatos. Logikusnak látszik a következő föltevés: a dokumentációs taxonnak megfelelő fogalom mint „a megfelelő tudományág X témájához tartozó dokumentum" fogalmazódik meg. E fogalom archetípusa olyan tematikus tezauruszként interpretálható, amelyben csak azok a szemantikai összefüggések őrződnek meg, amelyek e téma keretein belül közösek. Ilyen értelmezésben a témának megfelelő dokumentációs fogalom archetípusának struktúrája nem egyéb, mint e témakörben a tezaurusz szemantikai relációinak rendszere. Más szóval a tezaurusz elemei között fennálló relációkat a meronómia határozza meg. A téma szűkítése (az átmenet a dokumentációs fajfogalomhoz) speciális szemantikai reláció és lexikai egységek formájában jelenik meg a tezauruszban.

A dokumentumokból álló taxon szűkítésekor keletkezik a megfelelő archetípushoz (a megfelelő tezauruszhoz) való átmenet, amelynek eredményeképp jóval szűkebb tudományos témakörbe kerülünk. A tezauruszban arra kell törekedni, hogy a szűkebb és tágabb tárgyak közötti relációkat a lehető leggondosabban kidolgozzák.

A meronómia ilyen, dokumentum-taxonómiával összefüggő értelmezésében meg kell különböztetnünk egyrészt az osztályozási rendszert, másrészt a tezauruszt. A tezaurusz itt mint az adott téma dokumentumaiban foglalt szemantikai információkészlet interpretálható, azaz mint az ismeretek struktúrájának leírása. Ilyenfajta értelmezés esetén eltérünk a tezaurusz hagyományos informatikai fogalmától, mint a kutatás közvetlen eszközétől, s áttérünk arra, hogy a tezauruszt az adott témakörbe tartozó ismeretek rendszereként értelmezzük.

Az az alapvető probléma, amely az informatika különböző konkrét ágaiban felbukkan, az ismeretek struktúrájának olyan leírásából áll, amelyben megközelítőleg jól „kiértékelődnek" a megnevezések jelentései és a dokumentumok-

ból álló taxonok tartalma. Az informatikát éppen csak annyira érdekelte a tudomány és technika konkrét területének tartalma, hogy felépíthesse a dokumentumok taxonómiájának megalapozásához szükséges meronómiát.

Az a tény, hogy a dokumentumok taxonómiája a meronómiával függ össze, a meronómia pedig nem más, mint a tudományterület alapfogalmainak struktúrája, megköveteli – a tudományra vonatkozó metainformáció leírása szempontjából –, hogy elég sok fontos információval rendelkezzünk. A taxonómia metainformációt nyújt, azaz arról tájékoztat, mi az információtartalma a taxonoknak. Másrészt, a meronómia az archetípusok (a tezauruszok) az adott tudományterületre vonatkozó konkrét adataiból áll. Magától értetődik, hogy a taxonómiáról és a meronómiáról az informatikában kialakított álláspont egyenlőre heurisztikus jellegű, s hogy azt konkrét tartalommal tölthessük meg, nagy mennyiségű tényanyagot kell feldolgozni. A tudományelméletben nem a dokumentumok osztályozását vizsgálják, hanem magukat a tudományokét. Ahhoz, hogy magukat a tudományokat taxonok szerint felosztható ssuk, mindenekelőtt össze kell gyűjteni a tudományokat (ismeretterületeket). Hiszen az olyan osztályozás, amelyben egy taxonba kerülnek az egyes tudományok és azok részei (matematika és algebra, fizika és optika stb.), se nem esztétikusak, se nem felelnek meg a világ tényleges szerkezetének.

De tegyük föl, sikerült összeállítanunk a tudományok „kiegyenlített", jól rendezett jegyzékét. A nevezetes osztályozásokban arra törekednek, nem indokolva azt speciálisan, hogy ilyen „kiegyenlített" jegyzékeket vizsgáljanak. Akkor nyilvánvaló a következő. Először is, a tudomány osztályozása nem más, mint ami közvetlenül szükséges az informatika számára. Ez utóbbi számára nem a taxonómia kell, hanem a tudomány meronómiája, amely a dokumentumok taxonómiájának felel meg. Ez a tény jól magyarázza, hogy a tudomány osztályozása a gyakorlatban miért nem használható az informatika számára, hogy cseppet sem becsülhető le a tudományok filozófiai–ismeretelméleti osztályozása. Másodsorban felmerül az az ismeretelméleti probléma, hogy mi a tudományok taxonómiájának megfelelő meronómia.

Feltételezhető, hogy a megfelelő meronómia archetípusok olyan rendszere, amelyek már nem a tudomány struktúráját írják le, hanem az általa vizsgált valóság kategoriális struktúráját.

A tudományok ama felosztása, mely az anyag és az energia szervezettségi szintjein alapszik, megfelel ennek a szempontnak.

Azon kívül megállapíthatjuk, hogy a valóság alapvető archetípusainak leírása nagyon fontos feladat, és az ismeretterületek (tudományok) taxonómiája függ a világról és a természet szerkezetéről alkotott felfogásunktól. Komolyan foglalkozni ezzel a problémával e helyen nem kívántunk. A feladatunk sokkal szerényebb, fölhívni a figyelmet a meronómia és a taxonómia kettőssége elvének heurisztikus értékére, amely az osztályozás elméletének minden jelentős problémáját áthatja.

MEGKÖZELÍTÉS A NYELVÉSZET SZEMPONTJÁBÓL

A tartalmi feltárás egyik alapvető problémája, hogy miként őrizhetők meg az információkeresés számára a szavak szövegen belüli összefüggései, más szóval hogyan lehet mondattanilag (szintaktikailag) is feldolgozni a szövegeket. Az intellektuális feldolgozás terén *William J. Perry*, *Jason L. Farradane*, *Jean M. Perreault*, *Derek Austin* és mások a hetvenes évekig szintaktikai relációk (szerep-, illetve kapcsolatjelölők) kidolgozásával próbálták megoldani a problémát (a felsorolt szerzők szemelvényei az első kötetben találhatók).

Az intellektuális tartalmi feldolgozás két módjának – az osztályozásnak és indexelésnek – egységes nyelvészeti tárgyalására *William J. Hutchins* vállalkozott 1975-ben megjelent alapvető monográfiájában (részletei ugyancsak az első kötetben olvashatók).

Az automatikus indexeléskor a probléma még nagyobb súllyal jelentkezik. A mai üzemszerűen működő automatizált információkereső rendszerek túlnyomó többségében a szintaktikai összefüggéseket nem elemzik, csak kvantitatív (statisztikai jellemzőkön alapuló) nyelvészeti elemzésre kerül sor (a rendszerek szintaktikailag érzéketlenek).

Például a „francia autók exportja Magyarországra” szövegben a szintaktikailag érzéketlen indexelőprogramok legegyszerűbb fajtái csak a „francia”, az „exportja”, az „autók” és a „Magyarországra” szavakat invertálják. Az igényesebbek ragokat választanak le, szótöveket állapítanak meg. Ezt nevezik konflálásnak, eredménye az előbbi példa esetén a „francia”, „autó”, „export”, „Magyarország” szavak (indexkifejezések). A még igényesebbek esetleg még szabályozott szótárt (tezauruszt) is használnak, és ebben az esetben előfordulhat, hogy az invertált fájlba a „Franciaország”, „jármű”, „külkereskedelem” és „Magyarország” kerül. (A kiválasztott szövegszavakat összehasonlítják a tezaurusz kifejezéseivel és minden nemdeszkriptornak minősülő szót az előírt deszkriptorral váltanak föl. Ez a hozzárendelő indexelés.) Néhány, a gyakorlatban is felhasznált rendszer

arra is képes, hogy a kifejezések előfordulási gyakorisága alapján súlyozott relációkat állapítson meg, melyek a kiválogatott szövegszavakat a deszkriptorokkal összekapcsolják.
A szintaktikai feldolgozást is végző, ún. „parserekkel" működő indexelőrendszerek immár évtizedek óta kísérleti stádiumban vannak. Ennek oka, hogy az egyszerűbb automatikus eljárásokat már a késői hatvanas, korai hetvenes években elkészítették, amikor a kutatás még nem állt elő a fejlettebb – szintaktikai feldolgozásra, vagy akár csak súlyozásra, valószínűségi – indexelésre alkalmas algoritmusokkal, és ezek az egyszerűbb rendszerek azóta is úgy ahogy kielégítik a nagy tömegű információkeresés igényeit. A szoftvergyártó cégek pedig ennek láttán egyelőre nem sokat áldoznak a továbbfejlesztésre.
A számítógépes nyelvészet – a gépi fordítás és a nyelvészeti eszközökkel végzett információkeresés – az ötvenes és hatvanas években kezdett kialakulni. Az első időszakot derűlátás jellemezte, melyet aztán a felmerült és egyelőre tökéletesen nem megoldhatónak bizonyult szemantikai problémák miatt kiábrándultság követett. Klasszikus megfogalmazója ennek a kettősségének az ausztriai születésű, izraeli *Jehoshua Bar-Hillel.*
A hatvanas évek eredményeiről a matematikai nyelvész és informatikus *Varga Dénesnek* köszönhetően számos magyar fordítás született, melyek önálló kötetekben jelentek meg.[1] *Bar-Hillel* szemelvénye előtt ezeket ismertetjük.

P. L. GARVIN

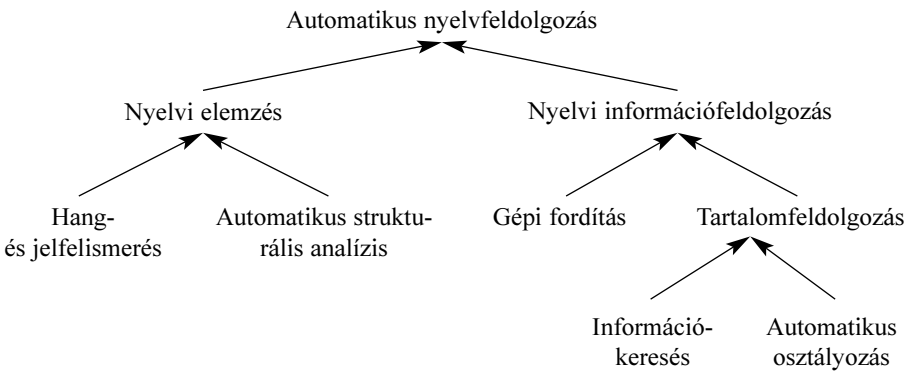
A nyelvi adatfeldolgozás a nyelvész szemszögéből

In: A dokumentáció nyelvészeti kérdései I. [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest: OMKDK, 1966. – (A tudományos tájékoztatás elmélete és gyakorlata; 10.) p. 110–127. Eredeti: A linguist's view of language-data processing. In: Natural Language and the Computer, McGraw-Hill, 1963, p. 109–127.

Az automatikus nyelvi adatfeldolgozáson a számítógépeknek a természetes nyelvekre való alkalmazását értjük. Ennek egyik célja a nyelvi elemzés, a másik az információk kezelése (gépi fordítás, in-

1 Varga Dénes (szerk.): A dokumentáció nyelvészeti kérdései I. [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest: OMKDK, 1966. – (A tudományos tájékoztatás elmélete és gyakorlata; 10.) 156 p.
Varga Dénes (szerk.): Dokumentáció és nyelvészet. Az értelmi összefüggések formalizálása. Szemelvénygyűjtemény. [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest : OMKDK, 1979. 174 p.

formációkeresés, automatikus kivonatkészítés), más szóval a tartalmi feldolgozás.
Az elemzés legformálisabb részét a hang- és jelfelismerés képviseli. A következő fokozat az automatikus strukturális analízis, melynek beviteli anyagát a szövegek alkotják, a kimenetét pedig az e szövegek által képviselt természetes nyelv rendszerének nyelvészeti leírása. A nyelvi információfeldolgozás egyik fő területe a gépi fordítás, a másik a tartalomfeldolgozás. Az utóbbi abban különbözik a gépi fordítástól, hogy a tartalmat értékelik is valamilyen releváns feltétel alapján. A tartalmi feldolgozásnak egyik része az információkeresés, a másik pedig az automatikus osztályozás. Összefoglalva címkézett, irányított gráffal (az élek a faj–nem relációt jelölik):



(A gráf részletesebb formában az automatikus indexeléssel és osztályozással foglalkozó bevezetőben is tanulmányozható.)

L. E. PŠENIČNAĂ–E. F. SKOROHOĎKO

Információkeresés értelmi kódok alapján

In: A dokumentáció nyelvészeti kérdései I. [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest: OMKDK, 1966. – (A tudományos tájékoztatás elmélete és gyakorlata; 10.) p. 68–72.
Eredeti: Informacionnüj poisk po smüslovüm kodam. In: Naučno–tehničeskaă Informaciă, 1964, No. 6, p. 25–26.

A szovjet kibernetikai intézetben 1960–68 között dolgozták ki a *Perry* és *Kent* szemantikai kódjaira emlékeztető, de annál algoritmikusan lényegesen fejlettebb „tárgy (X)–reláció (R)” (RX) rendsze-

rét és kezelőprogramját, a Bit automatikus információkereső rendszert. A rendszerben predikátumlogikai és gráfelméleti eszközöket használnak. Az RX kódokat a referátumok automatikus indexelésére és a referátumszövegekben végzett információkeresésre tervezték. A szintaktikai relációkat ebben a rendkívül fejlett, de mindvégig kísérleti stádiumban maradt rendszerben fák, a paradigmatikus (értelmi) összefüggéseket pedig gráfok képezik le. Az elemzést a következő példán mutatjuk be (az orosz nyelvi szerkezetet a szemléltetés miatt nem lehetett teljesen megszüntetni):

A rét olyan földterület, melyen fű nő, és folyók árterében található

$X_1 = R_1X_2 \quad R_2 \quad X_3 \quad X_4 \quad R_3$

Az erdő olyan földterület, melyet fák borítanak

$X_5 = R_1X_2 \quad X_6 \quad R_2$

A sztyeppe olyan földterület, melyen fű nő,

$X_7 = R_1X_2 \quad R_2 \quad X_3$

száraz klímazónában található

$R_3 \quad X_8$

kevés nedvességgel rendelkezik, nem borítják fák

$X_9 \quad R_4 \quad -R_2 \quad X_6$

A kódok szerkesztéséhez a fenti formájú szövegek szükségesek. A meghatározások a szavak jelentését adják meg. A szemantikai kódok ezzel szemben a megnevezett dolgok összefüggéseit fejezik ki. Az alábbi három megszerkesztett kód például:

Rét $X_1 = R_1\ X_2\ R_2\ X_3\ R_3\ X_4$

Erdő $X_5 = R_1\ X_2\ R_2\ X_6$

Sztyeppe $X_7 = R_1\ X_2\ R_2\ X_3\ R_3\ X_8\ R_4\ X_9\ -R_2\ X_6$

A kódok alapján már egyértelmű, ami pusztán a Rét, az Erdő és a Sztyeppe megnevezésekből még egyáltalán nem következik, hogy mindhárom esetében olyan növényzettel borított földterületről van szó (R_1X_2), melyek a fekvésükben, illetve a növényzet fajtájában különböznek (X_3, X_4, X_6, X_8). Az eljárásban tehát a szerepjelölők szerepét játszó szemantikai összetevőket állapítják meg.

URIEL WEINREICH

Értelmi összefüggések[2] felhasználása természetes nyelvek mondatstruktúrájának interpretálására

In: Dokumentáció és nyelvészet : Az értelmi összefüggések formalizálása : Szemelvénygyűjtemény / [szerk.] Varga Dénes ; [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest : OMKDK, 1979. 174 p.
Eredeti: Explorations in semantic theory. In: Current Trends in Linguistics 3., The Hague : Mouton, 1966, p. 395–479.

Az amerikai lexikológus Weinreich a Chomsky-féle generatív nyelvelméletet kiegészítő szemantikaelmélet kidolgozásával foglalkozott. Szemantikai jegyek olyan rendszerét próbálta meg kidolgozni, melynek alapján a mondat jelentése algoritmikusan levezethető részeinek ismert jelentéséből. A ma már klasszikusnak számító tanulmányt még a hatvanas évek optimizmusa hatja át.

A. K. ŽOLKOVSKIJ–I. A. MEĹČUK

Értelmi összefüggések felhasználása természetes nyelvek mondatainak szintézisére

In: Dokumentáció és nyelvészet : Az értelmi összefüggések formalizálása : Szemelvénygyűjtemény / [szerk.] Varga Dénes ; [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest : OMKDK, 1979. 174 p.

Az Oroszországban kialakult Rozencvejg-féle szemantikai iskola képviselői ebben a tanulmányukban egy minőségileg magasrendű gépi fordítórendszer kidolgozását fogalmazták meg. A szerzők a lexikai függvényekben vélték megtalálni azt a hatékony eszközt, amellyel az egyedi mondatok sokfélesége visszavezethető mélyebben rejlő

2 Az angol szakirodalomban inkább az „analitikus relációk” kifejezés fordul elő, olykor „báziskapcsolatok”. A nyelvészetben „paradigmatikus összefüggésekről” beszélnek. A 60-as évek orosz szakirodalmában az „értelmi összefüggések” kifejezést is használták. Minden esetben szövegfüggetlen, a fogalmak, ill. a szavak között eleve meglévő összefüggésekről van szó.

logikai–szemantikai struktúrára, s formális szabályok segítségével létrehozhatók ugyanannak a gondolatnak különböző kifejezési formái. A két tanulmány magyarra fordítása olyan időszakra esett, amikor a matematikai nyelvészeti s ezzel a gépi információkeresési kutatások – átmenetileg rendkívül föllendültek Magyarországon.

JEHOSHUA BAR-HILLEL (1915), AVAGY A GÉPESÍTÉS SZEMANTIKAI PROBLÉMÁI

Bar-Hillel 1915-ben született, osztrák származású izraeli filozófus, matematikus és nyelvész. Már egyetemi tanulmányai alatt megismerkedett *Carnap* logikai és szemantikai munkáival, és később a nemzetközi hírű amerikai műszaki főiskola, a Massachusetts Institute of Technology munkatársaként *Carnappal* együtt a jelközvetítés és a szemantikai információ elméletén dolgozott. 1953-tól a jeruzsálemi egyetem professzora. Tanulmányaiban egyre szkeptikusabbá válik régi vesszőparipáival, a gépi fordítással és az automatizált információkereséssel szemben. 1963-ban, amikor a szemelvényünket tartalmazó kötetét kiadták, már nem hitt abban, hogy a számítógép valóban jó minőségű fordításokat készíthet, sem abban, hogy az információáradat olyan fenyegető, amilyennek beállítják. Erre vonatkozó érveit sorakoztatja fel az itt közölt, nagy visszhangot kiváltott tanulmányában. Az írás a számítógépes információfeldolgozás hőskorát követő állapotokat tükrözi, amikor szükségszerűen váltotta fel a kezdeti lelkesedést a kétkedés. Bar-Hillel tanulmányára azóta is sokan hivatkoznak, hiszen olyan szakember kétségeinek adott hangot, aki már a gépi információfeldolgozás születésénél is ott bábáskodott, e téren jelentős eredményeket ért el, mégis mindent megkérdőjelezett.
Kétségeivel Bar-Hillel nem állt egyedül. A generatív nyelvészet egyik megalapítója, *Noam Chomsky* kezdettől fogva azt az álláspontot képviselte, hogy a nyelvészetnek nem lehet feladata az algoritmikus relevancia vizsgálat; más szóval az automatikus nyelvi elemzés nem logikus célkitűzés.

Az irodalomkeresés gépesítésének elméleti aspektusai

In: A dokumentáció nyelvészeti kérdései I. [közr. az] Országos Műszaki Könyvtár és Dokumentációs Központ. – Budapest: OMKDK, 1966. – (A tudományos tájékoztatás elmélete és gyakorlata; 10.) p. 128–136.
Eredeti: Theoretical aspects of the mechanization of literature searching. In: Language and Information, Jerusalem, 1964, p. 356-364.

A számítógépeket nem arra kellene használni, hogy teljesen automatizált információkereséseket végezzenek velük, mert ez a valódi kutatók számára soha sem hozhat kielégítő, megfelelő részletességű és ténylegesen releváns eredményt, hanem arra, hogy a felhasználó úgy használhassa ezeket a gépeket, mint valamilyen szerszámot, mellyel az információk tömkelegében saját belátása szerint navigálhat. Bar-Hillel ebben a tanulmányában több mind harminc évvel a világméretű hálózati on-line hozzáférés, az internet korszaka előtt megfogalmazta a gépi információkeresés leghatékonyabb módját: az interaktív információkeresés technikáját. „A számítógépek információkeresésre való hatékony alkalmazását nagymértékben akadályozta a terméketlen elméletieskedés, az az értelmetlen törekvés, hogy nagytekintélyű tudományos, főleg matematikai diszciplínákat juttassanak szóhoz, akadályozta továbbá a kutatást a szabadjára engedett fantázia és annak az intellektuális erőfeszítésnek a lebecsülése, amely egy jó összefoglalás, egy jó index elkészítéséhez, vagy a témakörök közelségi fokának megítéléséhez kell. Ez a lebecsülés jórészt a szemantikai nézeteknek azzal a nagyfokú nyersességével magyarázható, ami az információkeresés számos szakemberét jellemzi. Mert mi az eredménye, ha kifinomultabb fogalmakat alkalmazunk az intellektuális feldolgozás eredményeként? Ez részint arra vezet, hogy negatívabban kell értékelnünk az irodalomkeresés automatizálása terén megfigyelhető sok kísérletet."

Válságban az információkeresés?[3]

Amikor az információkeresés, a fordítás vagy más szöveges adatfeldolgozás gépesítésének kérdése felvetődik a nyilvánosság előtt – népszerű cikkben, kutatási jelentésben, kongresszusi bizottságban vagy tudományos összejövetelen, az Egyesült Államokban, a Szovjetunióban vagy bármelyik nagyobb országban – érvelésünket mindig azokkal a statisztikai adatokkal kezdjük, amelyek szerint elképesztő mértékben megnőtt az utóbbi évtizedekben a tudományos, műszaki és egyéb kiadványok száma. Szaktekintélyekre hivatkozunk, akik megjósolták, hogy a növekedési ütem az elkövetkező években változatlan marad, sőt, felgyorsul. Majd – különösen amerikai hallgatóság előtt – elkerülhetetlenül *Vannevar Bush* találó mondásainak egyikét idézzük, esetleg éppen azt, amelyik arra figyelmeztet, hogy „a tudomány ugyanúgy be-

3 Is information retrieval approaching a crisis? In: Bar-Hillel, J.: Language and information. Selected essays on their theory and application. Reading : Addison Wesley ; Jerusalem : The Jerusalem Academic Press, cop. 1964. p. 365–372.

lefulladhat saját eredményeibe, ahogy a túlszaporodott baktériumtenyészet saját burjánzásába".

Ilyen alkalmakkor gyakran ítélik válságosnak a helyzetet, mondván: az információrobbanás előbb-utóbb a tudományos kommunikáció teljes csődjéhez vezet. Általános az a vélemény, hogy a tudósok egyszerűen képtelenek lesznek megbirkózni az információáradattal, ha az információgyűjtést és -keresést a megszokott módon folytatják. Csak a szokások radikális megváltoztatása mentheti meg a helyzetet. A végkövetkeztetés többnyire az, hogy a referátum- és kivonatkészítést, az indexelést, a fordítást, az irodalomjegyzékek összeállítását és a szöveges adatfeldolgozást teljesen gépesíteni kell. Meggyőződésem, hogy ez az érvelés majdnem teljesen hamis. Hajlottam volna rá, hogy ne is foglalkozzam vele, mert csak a szerzett jogok és érdekek fejeződnek ki benne; de sok tekintélyes tudós is a bólogatók, sőt a kezdeményezők táborába tartozik. Valóban kevés olyan tudós van, aki alkalmanként ne panaszkodnék az információkeresés és az irodalomkutatás tökéletlenségeire vagy a fordítások megbízhatatlanságára. Mindig készek annak bizonygatására, hogy a dolgok ezeken a területeken elég rosszul mennek. Bár e panaszok megalapozatlansága egy röpke elemzés és elmélkedés után nyilvánvalóvá válna, a kérdés fontossága miatt messzebb szándékozom menni, mint amennyire a probléma valódi érdeme szerint kellene. Úgy érzem, magyarázatot kell találni arra, hogy miért emlegeti olyan sok tudós a panaszokat, és miért hisz olyan mélyen a tudományos kommunikáció területén fenyegető válságban.

Bár mondanivalómat csak a természettudományos információra korlátozom, az alábbi ellenérvek többsége nemcsak a természettudományok szöveges adatainak feldolgozására és értékelésére érvényesek, hanem azokéra is, amelyek például a hírszerző szolgálatokat érdeklik.

Az „információáradat" tételével szemben érvelésem lényegében roppant egyszerű: az érv a fenti formában nem helytálló. Lássuk ezt részletesebben. Az idézett számadatokat általában az éves növekedési ütem, illetve a kétszereződési idő formájában adják meg. Ezek a formulák, és természetesen a többi is, matematikailag megegyeznek, és csupán a velük kiváltható pszichológiai sokkhatás tekintetében különböznek, ha egyáltalán különböznek.

A becslések 5–8%-os növekedéssel, illetve 10–15 éves kétszereződési idővel számolnak. Nem vonom kétségbe a számok helyességét, még az értelmüket sem, mert ez mit sem változtatna kritikám megalapozottságán. Mit bizonyíthat a következő állítás: „Ha a tudományos irodalom termelése napjainkban 2000 oldal percenként, akkor valamivel 1975 előtt percenként 4000 oldal lesz". Önmagában éppenséggel semmit! Az, hogy valami mértani sorozat szerint (vagy, hogy tekintélyesebben hangozzék „exponenciálisan") növekszik, lehet rossz, de lehet jó vagy semleges is. Mit szólnánk például a következőkhöz: „Ha a személyi jövedelem egy évvel ezelőtt 4000 dollár volt, akkor valamivel 1975 előtt 8000 dollár lesz". És mi a véleményünk erről: „Ha az évi

autótermelés jelenleg 6 000 000, akkor valamivel 1975 előtt 12 000 000 lesz". A másodjára említett fejlődés problémákat okozhat, ha más dolgok, például az útépítés, nem fejlődnek ennek megfelelően. De a megnövekedett útépítés maga is problémát okozhat, hacsak... és így tovább. Tulajdonképpen bármi problémává válhat, még a stagnálás is. Nem kétséges, hogy az autógyártás jelenlegi ütemének folytatódása hamarosan válságot idézhet elő (vagy némelyek szerint a már meglévő válságot súlyosbíthatja), ha az utakat nem tartják megfelelően karban. (Az olvasó bocsánatát kérem érvelésem némiképp gyermeteg volta miatt, de úgy gondolom, hogy lényegét – akár gyermeteg módszerekkel is – jól a fejükbe kell vésnem.)

Ez rejlik tehát az „információáradattal" érvelők gondolkodásának hátterében. Miért hat ez mégis olyan sok emberre? Néhányan valószínűleg így gondolkodnak: a tudósnak, aki 1961-ben úgy vélte, hogy naponta átlag 20 oldalt kell elolvasnia, hogy szakterületének kutatásaival lépést tartson, 1973-ban ugyanezért naponta majd 40 oldalt kell elolvasnia, feltéve, hogy szakterületén a publikációk száma eme időszakban duplájára növekszik. Joggal feltételezhetjük, hogy az egy oldalhoz szükséges olvasási idő ez alatt a 12 év alatt nem fog változni jelentősen. Így a tudós a következő dilemmával találja majd magát szemben: vagy az olvasásra fordított időt növeli a duplájára és ezért (ha a teljes kutatásra fordított idő változatlan) kevesebb időt fordíthat alkotó kutatásra, vagy kutatómunkájának minősége fog romlani, mivel a kívánatosnál kevésbé lesz tájékozott arról, hogy szakterületén mi folyik. Sőt, mindkét esetben, de különösen az elsőben, több időt kell az elolvasandók kiválasztására fordítania, hiszen kétszer akkora anyagot kell majd átböngésznie, hogy meghatározza, mit is kell alaposabban tanulmányoznia. Bárhogy döntsön is, nemcsak saját kreatív tevékenysége szenved csorbát – amit esetleg nem is bán majd túlzottan, mivel sok olyan tudós van, aki boldog lenne, ha idejéből többet fordíthatna olvasásra – de egyúttal annak a társadalomnak az együttes erőfeszítései is, amelybe a tudós tartozik; ennek pedig nyilvánvaló a káros hatása.

Midőn ezt az érvelést fogalmaztam, vigyáznom kellett, nehogy megadjam magam látszólagos erejének és vonzásának. Mert természetesen teljesen hibás. Vitathatatlan tény (amit személyes tapasztalataim és megfigyeléseim is alátámasztanak, és amiről az olvasó is meggyőződhet), hogy a tudósok 1961-ben nem töltöttek átlagosan több időt olvasással, mint 12 évvel ezelőtt, bár időközben a tudományos kiadványok száma valóban közel a kétszeresére emelkedett. Másrészt nem hiszem, hogy bárki is komolyan gondolja, hogy eközben romlott a tudományos kutatás minősége. Ebből következik, hogy a dilemmának biztos van valami megoldása. De mi az? Mindenki tudja: a specializáció. 1961-ben egy tudós kutatásai általában fele akkora területet érintenek, mint 1949-ben, és ez 1973-ra ismét megfeleződik. Egyszerű megoldás! Vagy mégsem? Ha a helyzet

valóban ilyen egyszerű lenne, hogyan magyarázhatjuk meg a tudósok aggodalmait és panaszait a tudományos információ helyzetével kapcsolatban? Különböző magyarázatok léteznek. Önmagában egyik sem kielégítő, de együttesen elegendőnek kell lenniük. Először is a specializáció szinte illetlen szónak számít sokak számára, beleértve e tudósokat is. Nem szándékozom most boncolgatni ennek az érdekes szociológiai jelenségnek az eredetét. Mindenesetre sok tudós érzelmileg meg van győződve arról, hogy a specializáció rossz és még azt is kikéri magának, hogy szükséges rosszként emlegessék. Bár minden hosszabb tapasztalattal rendelkező tudós tisztában van azzal, hogy csak úgy maradhat egy terület szakértője, ha érdeklődési körét többé-kevésbé folyamatosan szűkíti; mégis, olyan régen ki van téve már az úgynevezett humanisták propagandájának (ami talán leginkább abban az ostoba mondásban testesül meg, hogy „egyre többet tudni egyre kevesebbről"), hogy tudat alatt elfogadta ennek a propagandának az igazságát. Mindig kész ezért a fejlődésért a társadalmat hibáztatni és odaadó figyelemmel hallgatja azokat, akik erre a képzelt betegségre gyógyírt kínálgatnak.

Másodszor: ezt a propagandahatást csak fokozza az az igazán természetes érzelmi reakció, amit a folytonos specializálódás tényleges nyomása vált ki: „Mi az ördög, csak nincs valami bajom? Ha ma képtelen vagyok az irodalommal lépést tartani azon a területen, amelyen tíz éve még könnyedén kiigazodtam, az egész biztosan azoknak a fickóknak a hibája, akiknek az a dolga, hogy lefordítsák, referálják, kivonatolják, tömörítsék, indexeljék, összefoglalják, áttekintsék és elbírálják az irodalmat, és akik rosszul végzik munkájukat, mivel valószínűleg még mindig középkori vagy tizenkilencedik századi módszereiket használják, ahelyett, hogy ráébrednének a huszadik század technológiájára. Nem hallottak ezek még a komputerekről?" Gyakran hallani manapság a bűntudatról, amely akkor támad fel a tudósokban, amikor irodájukba lépve szembe találják magukat a jóakaratú barátok és kollégák által oly készségesen küldözgetett különlenyomatok és az előfizetett újságok legfrissebb számainak a hegyével, amelyeket valahogy sosem sikerül olyan alaposan átolvasniuk, ahogy szeretnék. Végül persze csak találnak valami megoldást, mondjuk egy időre lemondják az egyik folyóirat előfizetését, vagy új asszisztenseket vesznek fel, ha meg tudják szerezni rá a szükséges pénzt és rájuk testálják eredeti érdeklődési körük egy részét. De ezt mindig vonakodva teszik és éppen ezért késve, ami önmagában is feszültségekhez vezet.

Harmadszor: van objektív oka is annak, hogy az Egyesült Államokban és néhány más országban (amelyek közé már a Szovjetunió is tartozik) a tudósok úgy érzik, hogy valami egyre rosszabb lesz az információkeresés területén. Talán ugyanez az ok készteti az amerikai kormányt és a hadügyi szerveket arra, hogy különös figyelmet szenteljenek az információs problémáknak, s hogy ha-

talmas pénzösszegeket és sok időt fordítsanak automatizált megoldásukhoz vezető kutatásokra. Az ok a következő: bár a publikációk száma világszerte a tudósok számával arányosan emelkedik és az amerikai publikációk száma is az amerikai tudósok számával arányosan emelkedik, ám a tudományos publikációk száma általában nem ugyanolyan arányban növekszik, mint az amerikai tudósok száma. Éppen ellenkezőleg. Közismert, hogy a tudományos publikációk száma világszerte gyorsabban növekszik, mint az Egyesült Államokban, mert a tudósok száma az Egyesült Államokon kívül az elmúlt években gyorsabb ütemben növekedett mint magában az USA-ban. Az okok ismertek, e helyen nem kell őket külön részleteznünk. Mivel a nem-amerikai publikációkat természetesen nehezebb beszerezni és többnyire elolvasni is, hiszen nem feltétlenül angolul íródtak, az amerikai tudósnak több erőfeszítésébe kerül a munkájához szükséges publikációknak a megszerzése és elolvasása. Nagyon könnyen megeshet, hogy ma több időt kell erre fordítania, és ezért kevesebb ideje jut más tevékenységekre, mint 12 évvel ezelőtt.

Másrészt persze a nem-amerikai tudósok helyzete arányosan könnyebb, de ez a dolgok mai állása mellett valószínűleg nem vigasztalja meg az amerikaiakat.

A legutóbbi időkig sok amerikai természettudós és műszaki szakember nagyon jól elboldogult, ha az Egyesült Államokból származó publikációkat elolvasta és még többen közülük, ha az angol nyelven írt publikációkat elolvasták. Friss becslések szerint az összes természettudományos publikáció közel hatvan százaléka angol nyelven íródik. Ez a hányad jóval nagyobb, mint az angol és amerikai tudósok aránya a világ tudóstársadalmában. Az a tény, hogy az amerikai tudósnak többet kell idejéből az őt érintő publikációk megszerzésére fordítania, valóban, bizonyos értelemben lecsökkenti kreatív tevékenységét és ennek következménye egészében véve az lehet, hogy az Egyesült Államok fölénye a természettudományokban csökken, mielőtt egy új egyensúly létrejönne.

A fennálló politikai helyzetben ez a folyamat, amelyet más körülmények között a fejlődő országok tudományos és technológiai haladásának jeleként üdvözölhettünk volna és amely az életszínvonal általános növekedésével járna együtt szerte a világon, nyilvánvalóan potenciális veszélyt jelent az Egyesült Államok számára, éppen ezért tökéletesen érthető, hogy az Egyesült Államok kormánya és katonai szervezetei készen állnak arra, hogy erőteljes ellenintézkedéseket tegyenek. Éppen ennyire érthetően először azzal próbálkoznak, hogy a pénzügyi erőforrásokban és számítógépes technológiában meglévő fölényükkel ellensúlyozzák az ilyen irányú változásokat. Nem hangsúlyozhatjuk azonban eléggé, hogy a vágyak és a lehetőségek felcserélése több kárt okozhat mint hasznot. Az a hiedelem, hogy a megoldást egyszerűen azért fogják megtalálni, mert sürgető szükség van rá, könnyen a pénzügyi források és

még inkább az értékes kutatási idő helytelen, sikerrel alig kecsegtető befektetéséhez vezethet.

Más helyen már megpróbáltam bizonyítani, hogy az információkezelés sok mozzanata, például a fordítás, a referálás vagy az indexelés nem automatizálható teljesen. A részleges automatizálás ugyanakkor elméletileg keresztülvihető és valószínűleg gazdaságilag is kifizetődő. A szükséges vizsgálatokat megfelelő felsőoktatási intézményekben, a kipróbált tudományos módszerekkel kell elvégezni. A probléma néhány apró részlet kivételével még bizonyára nem érett meg arra, hogy átkerüljön az ipari kutatóintézetekbe, és még kevésbé arra, hogy a tényleges fejlesztő munkához ösztönzést nyerjenek belőle.

Nem igaz, hagy összedől a világ vagy belefullad az információáradatba, ha nem gépesítjük az információfeldolgozást, vagy legalábbis nem végzünk rajta valamiféle „generáljavítást". Természetesen nem kétséges, hogy az információfeldolgozás, mint bármely más emberi tevékenység, javítható. Nagyon könnyű számtalan olyan vonatkozást kimutatni, amelyet jobbítani lehetne és könnyű azt is megmondani, hogy az ilyen helyi javításokat hogyan kellene elvégezni. Azt már nem ilyen egyszerű bebizonyítani, hogy ezek a helyi javítások nem vezetnek majd az egész romlásához. Köztudott, hogy a részletek optimalizálása nem vezet szükségszerűen az egész optimalizált működéséhez. A helyzet még bonyolultabbá válik, ha – mint esetünkben is – politikai megfontolások is közrejátszanak, és ezzel olyan kiszámíthatatlan tényezők lépnek be a folyamatba, amelyek igencsak megnehezítik az egyes javaslatok racionális értékelését. Nem kétséges például, hogy a szovjet Össz-szövetségi Tudományos és Műszaki Információs Intézethez – a VINITI-hez – hasonló szervezet létrehozása nagymértékben javítaná a tájékoztatás jelenlegi helyzetét az Egyesült Államokban, sőt valószínűleg szélesebb körben is javuláshoz vezethetne. De nagyon nehéz megbirkózni az ellenérvekkel, amelyek szerint az ilyen központosítás veszélyezteti az amerikai életfelfogást.

Bár véleményem szerint világméretekben nincs válságban az információügy, nem ez a helyzet bizonyos területein. Ezeken valamilyen ok miatt kedvezőtlen feltételek uralkodnak, amelyeknek halmozott hatása lehet. Gondolok itt például a szabadalmak vagy az angolszász világban a jog területére.

A kutatás általános gyakorlatában a régebbi publikációk használatának mértéke korukkal nagyjából fordítottan arányos, így sok kutatási témában a tíz évnél régebbi irodalmat alig forgatják. Egyáltalán nem ez a helyzet a szabadalmak vagy a jogi irodalom esetében, különösen azokban az országokban, amelyeknek joggyakorlata a precedenseken alapul. Ezért valóban előfordulhat, hogy nemcsak a szabadalmakra vonatkozó kérdések száma növekszik állandóan – feltehetően ugyanolyan arányban, mint az egyéb tudományos vagy műszaki tevékenység, ami az egészséges és természetes fejlődés jele – de nő az egyes keresésekre fordított idő mennyisége is az említett kumulatív hatásnak megfelelően. Amíg tudományon kívüli okokból a szabadalmi törvények változatlanok maradnak, ennek

következménye valóban az lesz, hogy a szabadalmi irodákban az alkalmazottak számának gyorsabb ütemben kell növekednie, mint a feltalálók számának. Idegesítően hangzik, de nem szabad megfeledkezni arról, hogy ez nem a tudományos információs tevékenység, hanem bizonyos társadalmi konvenciók kóros kinövése, amelyeket talán újból felül kellene vizsgálni.

A helyzet hasonló a jogi irodalom terén is. Nemcsak a jogi ügyek száma növekedik állandóan (ne foglalkozzunk azzal, hogy ez a fejlődés örvendetes vagy sajnálatos), hanem az ügyek előkészítésére fordított idő is, megint csak a fent említett kumulatív hatásnak köszönhetően. Valóban lehetséges, hogy jogi eljárások, különösen az angolszász országokban, befulladnak saját világukba, ha a jelen eljárási renden nem változtatnak. De ismétlem, a probléma forrása politikai és nem tudományos jellegű.

A kumulatív hatás sok könyvtárban is észlelhető. A mai könyvtári gyakorlat ritkán számol a dokumentumok kora és használata között a tudományos kutatásban érvényesülő – már említett – fordított aránnyal. Egy régi könyvnek ugyanannyi helyre van szüksége, mint egy újnak, és ugyanolyan hozzáférhetőséget is biztosítanak számára. E gyakorlat miatt a könyvtári berendezések vásárlásának aránya mind az egyetemek, mind az ipari létesítmények költségvetésében növekszik. Az így kialakult problémák nem elhanyagolhatók, de messze nem olyan súlyosak, mint a szabadalom és a jog területén felmerülők, mivel a szükséges módosítások az intézményi politika radikális megváltoztatása nélkül is végrehajthatók. Ezeket a változtatásokat lassan bevezetik a különböző intézményekben és a helyzet belátható időn belül elfogadható egyensúlyi állapotba kerül, anélkül, hogy „válságról" vagy „csődről" kellene beszélni.

Összefoglalva: a tudományos és műszaki kiadványok számának mértani sorozat szerinti növekedése nem okoz különösebb problémákat, nem idéz elő semmiféle fenyegető helyzetet és nem követel mentőakciókat. Ez a növekedés mértékeiben megfelel a tudományos és műszaki kutatók számbeli gyarapodásának, és a valóságban semmi más, mint ennek közvetlen eredménye. A küszöbön álló válság érzését – ami még tényleg illetékes tudósok között is széles körben elterjedt – a hagyományos érdekeknek megfelelő propaganda váltja ki, amely nyitott fülekre talál a tudósoknál, akik pszichológiailag tökéletesen érthetően berzenkednek a beszűkítő specializációtól. Ám a válságérzet leginkább annak köszönhető, hogy helyi szinten a tudományos és műszaki élet vezető országaiban kisebb a tudományos publikációk növekedési üteme, mint az olyan országokban, amelyek egészen mostanáig hátul kullogtak. Ez a helyzet arra készteti a már magasan fejlett országok tudósait, hogy idejük egyre nagyobb hányadát fordítsák a más országokból származó, idegen nyelvű irodalom áttekintésére. Ezért az információkeresés úgynevezett problémája – amennyiben egyáltalán problémának tekinthető – szociológiai, pszichológiai és legnagyobbrészt poli-

tikai jellegű és éppen ezért ezeken a szinteken kell szembeszállni vele. Az egyedüli területek, ahol az információkeresés feltételei a valóságban is romlanak, azok, amelyeken a kumulatív hatás jelentkezik, tehát a jog, a szabadalmak és a könyvtári állománygyarapítás területe. Ám ezeken a területeken a természetes orvosság inkább az intézményi politikában és a gyakorlat megváltoztatásában és nem a keresési műveletek forradalmi átalakításában rejlik. Az a hozzáállás, mely az információtárolás és keresés fenyegető válságának feloldását egyedül a gépesítéstől reméli, objektíve nem indokolt és szubjektíve is veszélyes, mivel az elfogulatlan tudományos elemzést vágyálmokkal akarja helyettesíteni és így az értékes kutatási időt utópisztikus spekulációkra pazarolja. Az információtárolás és -keresés egyes mozzanatainak részleges gépesítése elméletileg kivitelezhető, de gazdasági kivitelezhetőségét csak az e kérdésben anyagilag nem érdekelt tudományos intézmények által folytatott kiterjedt kutatásokkal és kísérletekkel lehet megállapítani.

HEURISZTIKUS ÉS LÉLEKTANI MEGKÖZELÍTÉS, AVAGY A SZUBJEKTÍV TÉNYEZŐK MEGRAGADÁSÁNAK KÍSÉRLETE

Az információkeresés automatizált eljárásai a keresés teljesen formalizált végletét képviselik. Ezzel szemben áll az a tény, hogy a keresés teljesen, illetve bizonyos intellektuális része soha sem formalizálható és ezért nem is gépesíthető. Az információkeresésnek ez a része mindig ún. hiperkomplex, az emberi intuíción alapuló eljárás marad, melynek vannak egyrészt stratégiai–taktikai, másrészt pedig lélektani összetevői is.

A keresési stratégia módszereire jelentősen hatottak a cranfieldi vizsgálatokkal kezdődő hatékonysági kísérletek. Az első stratégiai összefoglalásokat *Frederick W. Lancaster* írta. Azóta minden valamirevaló on-line kézikönyvben külön fejezetben foglalkoznak vele (a cranfieldi vizsgálatokkal kötetünk korábbi részében foglalkoztunk, a keresési stratégiáról pedig *Stephen P. Harter* könyvéből mutatunk be tömörített részletet). A taktikai–heurisztikus megközelítések még korábbra, a század harmincas éveire vezethetői vissza. Az ilyen jellegű kutatások legjelentősebb képviselője *Marcia John Bates,* aki a hetvenes években publikálta kísérleti eredményeit, melyet azóta szinte minden on-line kézikönyvben idéznek. Az információkeresés lélektani–kognitív összetevőinek elszórt vizsgálataira az orosz *N. D. Kravčenko* és a dán *Peter Ingwersen* tanulmányai a példák.

A lélektani tényezőkkel az osztályozás más területén is számoltak. Az első kötetben bemutatott J*ason L. Farradane* az ötvenes években a szintaktikai relációit kognitív lélektani alapon dolgozta ki.

STEPHEN P. HARTER (1921)

On-line információkeresés. Fogalmak, elvek és technikák[1]

7. fejezet. Keresési stratégia és heurisztika[2]

Gyorskeresés, keresőfogalom-alkotás, leválogatási és hólabda-stratégiák

A keresési stratégiát különféle értelemben használják az információkeresésben. Néhány szerző már a parancsnyelvek használatát is a stratégia körébe sorolja, noha ezt helyesebb a keresőkérdés megformálásának tekinteni. A stratégia a keresésen belül meghatározott szempont érvényesülését jelenti, annak eldöntését, hogy mi a keresés lebonyolításának legjobb útja. A taktika e cél közvetlen megvalósítására alkalmazott konkrét módszereket képviseli. A stratégia tehát az átfogó keresési terv, a koncepcionális (fogalmi) szint, a taktika pedig a konkrét megoldás, a „fogások" összessége. A taktika helyett a szakirodalomban használják a heurisztika fogalmát is: ez a feltalálás, a rájövés módszerének tudománya, a rávezető, kitaláltató módszereinek összessége, mely ötleteken, tapasztalatokon és intuíción alapul.

A problémamegoldás heurisztikus eljárásai rendkívül fontosak az on-line keresésben. A legfontosabb fogalmait *Charles Meadow,* főleg pedig *Fredrick W. Lancaster* dolgozták ki és vezették be a gyakorlatba.

A legfontosabb stratégiai fogalmak a következők:

Egyszerű gyorskeresés (briefsearch, quick and dirty search)

Az egyszerű gyorskereséssel egyszeri, Boole-operátorokkal végzett rövidre szabott keresést valósítanak meg annak érdekében, hogy előzetes képet kapjanak az adatbázis rekordjainak a kérdéssel kapcsolatos jellemzőiről. A felhasználó által közvetlenül megadott kifejezésekkel vagy azok csonkolt változataival hajtják végre; az utóbbi esetre akkor kerülhet sor, ha ismerünk legalább egy témába vágó szerzőt, címet, kiadót stb. A kapott találati tételekből megállapítható számos, az adatbázisban a vonatkozó témával összefüggésben használt deszkriptor, mellyel a részletes keresés folytatható.

---

1 On-line information retrieval : Concepts, principles, and techniques / Stephen P. Harter. 3th ed. – San Diego [etc.] : Academic Press, 1990. – (Library and Information Science Series. A Series of Monographs and Textbooks.)

2 Search strategies and heuristics, p. 170–175. In: On-line information retrieval.

Keresőfogalmak alkotása (építőkocka-technika, building blocks)

A keresőfogalmak alkotásán (építőkocka-technikán) alapuló stratégia az on-line keresés legátfogóbb megvalósulása. Egyes lépcsőit az alábbi táblázatban mutatjuk be:

1. Interjúkészítés a kérdezővel.

2. A keresőcél megfogalmazása. Mi fontosabb: a nagy visszahívás, a nagy pontosság, vagy a visszahívás és a pontosság arányos megvalósítása?

3. Az adatbázis és a keresőrendszer kiválasztása.

4. **A legfontosabb keresőkifejezések vagy összetevők és a közöttük fennálló logikai ÉS, VAGY, NEM kapcsolatok megállapítása.**

5. **a) Az egyes keresőfogalmakat jelölő keresőkifejezések megállapítása: szavak, szöveges kifejezések, szótöredékek, deszkriptorok, azonosítók, kódok, nem szemantikai jellegű bibliográfiai ismérvek.**
   b) A mező meghatározása, melyben keresni kell.

6. **Minden önálló keresőfogalom (építőkocka, fazetta) részére az építőkockán belül meg kell szerkeszteni az ismérvláncok halmazát. Az egyes építőkockákba összevont keresőkifejezések között a VAGY operátort használjuk.**

7. **Az egyes építőkockákat ÉS, NEM (esetleg VAGY) operátorokkal összekapcsoljuk.**

8. Alternatív változatok elkészítése.

9. A rendszer parancsnyelvén megfogalmazzuk a kezdő kérdést.

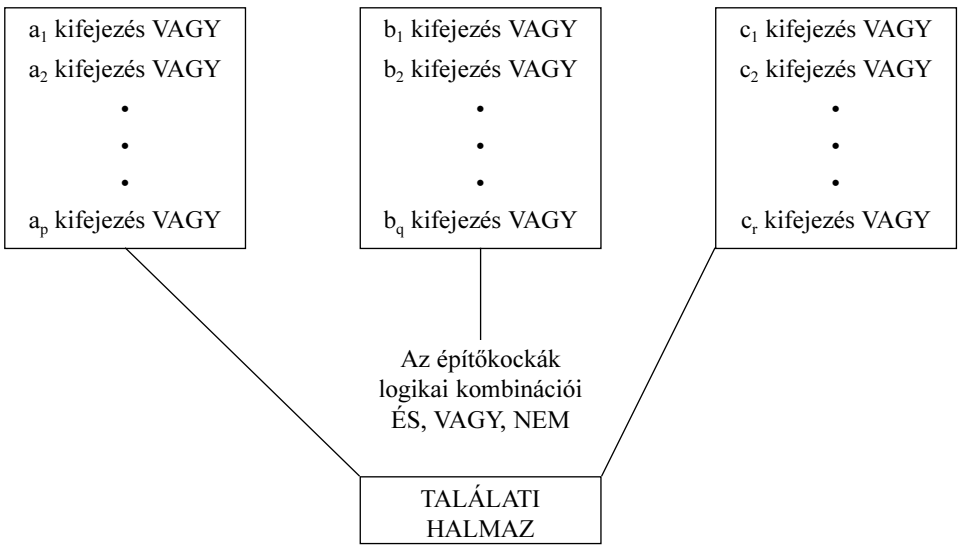
10. Bevisszük a rendszerbe a kérdést.

11. Értékeljük a közbenső eredményeket.

12. Iteratív folyamatok: a rendszer interaktív lehetőségeit kihasználva heurisztikus (taktikai) módszereket, fogásokat alkalmazunk a keresési eredmény teljessé tételére.

A lényege az, hogy az első lépésben megállapítjuk a keresőkifejezések valamilyen láncát. Ezt követően minden egyes keresőkifejezés esetén végiggondoljuk, milyen más kifejezés jöhet még az adott kifejezéssel kapcsolatban szóba, és ezeket egy-egy építőkockába vonjuk össze oly módon, hogy az egy építőkockába tartozó kifejezéseket vagylagosan összekapcsoljuk. Az egyes építőkockák egy-egy keresőfogalmat reprezentálnak. A keresőfogalmat az építőkockába vagylagosan összevont összes kifejezés együttesen nevezi meg.

Az eljárás központi részét alkotó 4., 5., 6. és 7. műveleteket grafikusan is ábrázoltuk:



A legfontosabb, hogy minden jelentősebb keresőkifejezést, illetve összetevőt megragadjunk. Az építőkockát (fazettát) a kereső szempontjából ekvivalens ismérvek alkotják. A keresőkérdésben több építőkocka is lehet, de egy kérdésen belül lehetőleg ne használjunk háromnál vagy négynél többet.

A következő lépésben szelektálunk az ismérvek között. Az egy építőkockába (fazettába) kerülő keresőszavak a kereső szempontjából ekvivalensek. Egy építőkockán belül az uniójukat kell képezni a VAGY operátorral.

Az építőkockán belül lehetnek szinonimák, kváziszinonimák, nagyon közelálló kifejezések, átfogóbb kifejezések stb. A lényeg, hogy ezek mindegyike egymáshoz képest vagylagosan kerüljön be a keresésbe.

Ezt követően az egyes építőkockákat ÉS, VAGY, NEM operátorokkal kapcsoljuk össze.

Például: A kérdés a technológia, a szabályozott piaci struktúrák és a hírközlés közötti összefüggésekre vonatkozik. Sem nagy pontosságra, sem nagy visszahívásra nincs szükség.

Az egyes ismérvblokkokat az alábbi táblázatban foglaltuk össze:

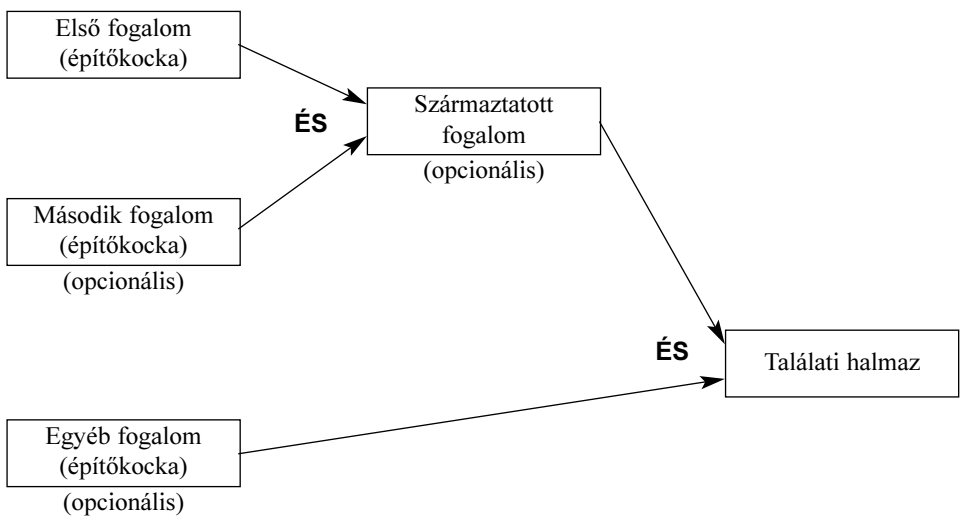
1. építőkocka	2. építőkocka	3. építőkocka
SZABÁLYOZÁS	TECHNOLÓGIA	HÍRKÖZLÉS
irányítás	technika	telekommunikáció
piaci struktúra	innováció	műhold
iparirányítás	technológiaváltás	rádió-műsorszórás
monopóliumok		távbeszélő
		telefon
		távíró
		televízió
		TV

Egymásutáni leválogatás (successive fractions)

A teljesség növelésének heurisztikus módszere, hogy valamelyik kész építőkockát (keresőfogalmat reprezentáló kifejezések csoportját) töröljük és a maradékkal keresünk. Ez persze a leggazdaságtalanabb eljárás, hiszen gondosan megszerkesztett építőkockáról mondunk le.

Ha feltehető, hogy az összes fontos keresőfogalom (építőkocka) felhasználásával túl kevés a találat (vagy egyáltalán nincs találat), vagy egyes építőkockák csak bizonytalanul reprezentálnak valamilyen keresőfogalmat, akkor módosítani kell a keresőfogalmakon alapuló stratégiát. A keresés többnyire nagy visszahívást (sok találatot) eredményező kérdésformával kezdődik, és a következő kérdésformák fokozatosan csökkentik a találati halmazt a már kezelhető, értelmes méretre.

Az eljárást az alábbi ábrán foglaltuk össze:



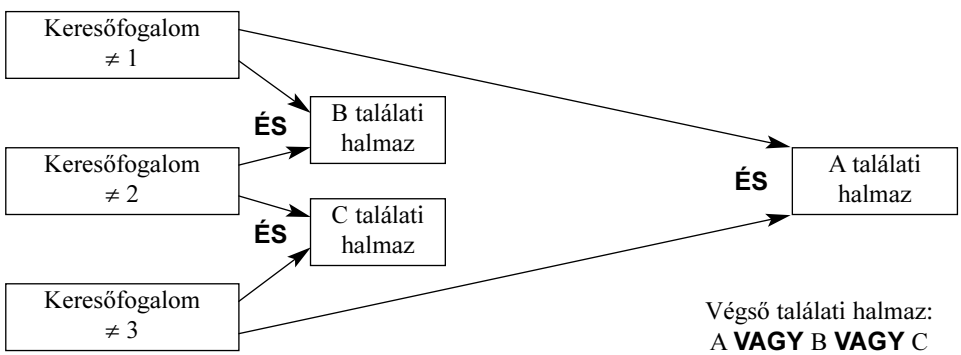
Az eljárás alapelve, hogy a származtatott keresőfogalmakat (származtatott építőkockákat) egyesével kell megszerkeszteni, lépésről lépésre és mindig szükség szerint, nem pedig az összeset egyszerre. Minden lépésben az előző eredmény segítségével szerkesztik meg a következő új származtatott keresőfogalmat. A keresés akkor ér véget, amikor a teljesség és a pontosság kívánt mértékét sikerült elérni, előnyös esetben még azelőtt, hogy az összes szóba jöhető építőkockát meg kellett volna szerkeszteni.

Kétféle leválogatási stratégiát ismerünk: az „először a legspecifikusabb fogalom" és az „először a legkevesebb találatot eredményező fogalom" leválasztását. Első esetben a legspecifikusabb fogalommal kezdjük a keresést, a második esetben pedig azzal, amely önmagában a legkisebb számú találatot eredményezte.

Az egymás utáni leválogatás tehát úgy kezdődik, hogy első lépésben létrehozunk egy kezdő találati halmazt, és annak méretét szükség szerint, lépésről lépésre csökkentjük. A legfontosabb különbség a korábban leírt stratégiákhoz képest, hogy a kezdő találati halmaz gyakran nem is jellemezhető valamilyen fogalommal, azaz nincs feltétlenül szemantikai tartalma (nincs jelentése), hanem például dokumentumtípus, nyelv, vagy a kiadás éve. Mindezek a stratégiák olyan találati halmazzal kezdődnek, melyre a nagy visszahívás (teljesség) a jellemző, és ennek mértékét csökkentik a további lépésekben az ÉS operátorral összekapcsolt fogalmakkal, vagy más módon. Különösen olyankor célszerű a lépésenkénti leválogatás stratégiájával élni, amikor a keresés tárgya, témája nem körvonalazható pontosan, vagy túl általános, vagy egyéb okból hasznos de nem lényeges korlátozások engedhetők meg a kérdésben.

Páronkénti leválogatás (pairwise facetets)

Ha mindegyik keresőfogalom (építőkocka) nagyjából azonos mértékben specifikus, azaz egyformán fontos, akkor páronként képezhetjük metszetüket. A találati eredményeket vagy külön-külön (a páronkénti keresés eredményeként), vagy az egész keresési folyamat uniójaként nyomtathatjuk ki. A stratégiát az alábbi ábra mutatja be.



Többszörös egyszerű gyorskeresés (multiple briefsearch)

Ezt a stratégiát a páronkénti leválogatás esetéhez hasonlóan akkor célszerű követni, ha az összes specifikus keresőfogalom metszetének eredménye várhatóan a nullához közelít. Alapelve, hogy több adatbázisban kell egyszerű, nagy visszahívással járó gyorskeresést végezni. Ezáltal nagyon különböző jellemzőket kapunk, ami a keresés témáját illeti, melyet a további lépésekben hasznosíthatunk. Ugyanazt a tárgykört az egyes adatbázisokban ugyanis másmás módon osztályozzák (indexelik) és dolgozzák föl formai szempontból. Ezáltal sok olyan jellemzőt megállapíthatunk, melyeket egyetlen adatbázisból körülményesebb kielemezni.

Hólabdakeresés (pearl growing)

Ezt a stratégiát olyankor célszerű használni, ha nagyon kevés találatot várunk. Ellentétben az eddigi stratégiákkal, itt nem a nagy visszahívással kezdünk, hanem éppen ellenkezőleg. A leggyakoribb, hogy a felhasználó megadja az egyik, általa ismert és a tárgyba vágó dokumentum adatát, melyet kikeresünk. Ebből megállapítható, milyen ismérvek (deszkriptorok, tárgyszavak, jelezetek) reprezentálják a tartalmát – tehát a keresett tárgy fogalmát. Most már ezekkel végezhető a keresés, hogy még több hasonló adathoz jussunk stb.

Gyakori, hogy a hólabdakeresést hivatkozási indexben végzik.

Az alábbi táblázatban foglaltuk össze az egyes stratégiákat (az ábécé A, B, C nagybetűi, illetve a nagy- és kisbetű kombinációk – mint Aa, Ab stb. – a keresőkifejezéseket jelölik, ha nincsenek találati halmazként megnevezve):

Egyszerű gyorskeresés

A ÉS B = Találati halmaz

Keresőfogalmak alkotása

Aa VAGY Ab VAGY Ac	= 1. halmaz (nagy halmaz)
Ba VAGY Bb VAGY Bc	= 2. halmaz (nagy halmaz)
Ca VAGY Cb VAGY Cc	= 3. halmaz (nagy halmaz)

1. halmaz ÉS 2. halmaz ÉS 3. halmaz = Találati halmaz

Egymás utáni leválogatás

A ÉS B	= 1. halmaz (nagy halmaz)
A ÉS B ÉS C	= 2. halmaz (az 1. halmaz származtatott része)

A ÉS B ÉS C NEM D = Találati halmaz

Páronkénti leválogatás

1. FOGALOM ÉS 2. FOGALOM = B találati halmaz
2. FOGALOM ÉS 3. FOGALOM = C találati halmaz
3. FOGALOM ÉS 1. FOGALOM = A találati halmaz

A találati halmaz VAGY B találati halmaz VAGY C találati halmaz = Végső halmaz

Többszörös egyszerű gyorskeresés

1. adatbázisban: A ÉS B	= 1. találati halmaz
2. adatbázisban: A ÉS B	= 2. találati halmaz
3. adatbázisban: A ÉS B	= 3. találati halmaz

Hólabdakeresés

Formai adat keresőszóként	= 1. találati halmaz (1 vagy néhány találat/formai adat)
1. találati halmaz formai adatai	= 2. találati halmaz (több találat/formai adat)
2. találati halmaz formai adatai	= 3. találati halmaz (sok találat/formai adat)

A további kereséshez a tételeket jellemző tartalmi ismérvek alapján állapítjuk meg a tartalmi keresőkifejezéseket.

MARCIA JOHN BATES (1942)

Marcia J. Bates a washingtoni egyetem könyvtáros iskolájának (School of Librarianship, University of Washington, Seattle) tanára. A hetvenes években az információkeresés taktikáinak és kognitív lélektani összefüggéseiknek a vizsgálatával foglalkozott és módszeresen feltérképezte az intuitív alapon végzett információkereső műveleteket.

Az információkeresést megkönnyítő taktikák[3]

Bevezető

Az információkeresés automatizálásában elért eredmények ellenére semmi nem ér még fel a tapasztalt ember ez irányú tudásával, képességeivel. Ugyanakkor keveset tudunk ezekről az ismeretekről, és azt sem tudjuk, mi is voltaképpen a különbség e téren a tapasztalt szakember és a kezdő tudása között.

A keresést megkönnyítő taktikák bibliográfiai és faktografikus keresésekre, a legkülönfélébb keresés-típusokra és témakörökre, manuális és on-line rendszerekre egyaránt alkalmazhatók. A taktikák komplexebb, többlépcsős keresésekhez készültek, illetve azokra a szituációkra, amelyekben nem tudni fejből a forrást. Néhány taktika közismert, de még ezek közül sem volt korábban mindegyiknek neve, és vannak köztük kevésbé ismertek is. Egy-egy taktika világos leírása és elnevezése már önmagában is segítség: tudatosabbá teszi alkalmazását, és gyorsabban „ugrik be" a kereső agyába.

A keresési taktika fogalma

Az információkeresés közben legalább négyféle magatartásmodellnek van létjogosultsága: a keresést idealizálónak, reprezentálónak, oktatónak és a keresést megkönnyítőnek–egyszerűsítőnek.

Az itt bemutatandó modell elsődlegesen a negyedik kategóriába tartozik, másodsorban oktatási célú. A gyakorlat fogja eldönteni, hogy a benne leírt taktikák melyik minőségükben mennyire lesznek megfelelőek.

A taktikák egész halmazát tekintve vannak közöttük átfedések és bizonyos hierarchikus viszonyok is. Hosszú távon lehet célunk kis számú nagyon hatékony taktika. Rövid távon azonban az a cél, hogy összeszedjük a lehető legtöbb taktikát, amely potenciálisan hasznos lehet. Ezután tesztelhetjük őket, és kiválaszthatjuk a legjobbakat. Ha idő előtt keressük a majdani kevés kiválasztottat, elveszíthetünk néhány értékeset.

Amikor a keresést megkönnyítő–egyszerűsítő modellről van szó, pszichológiai okokból is jobb, ha a taktikák nagyobb választéka áll rendelkezésünkre, akár átfedésekkel is. Ezeknek a taktikáknak a használata a kreatív problémamegoldás egyik formája. Ilyen esetekben az agy nem mindig logikus, szabályos sémák szerint működik. Sok különböző irányból közelíthet meg egy-egy problémát, egyszer ilyen, másszor pedig olyan taktikát alkalmazhat ugyanabban a

3 Információkeresési taktikák. In: Tudományos és Műszaki Tájékoztatás 28 (1981) 3. p. 122–125. Eredeti: Information search tactics. In Journal of the American Society for Information Science, 1979, Vol. 30, No. 4, p. 205–214.

problematikus esetben. Más szóval: egy ilyen modellel szemben egyenesen követelmény lehet a bizonyos fokú redundancia.

A stratégia átfogó tervezéssel foglalkozik, a taktika rövid távú célokkal és manőverekkel. Ha ezeket a kifejezéseket adaptáljuk az információkeresés területére, elfogadhatjuk az alábbi általános és egyszerű definíciókat:

Keresési taktika egy lépés a keresés folytatására.

Keresési stratégia egy egész keresés terve; illetve mint kutatási terület, a keresési stratégiák és taktikák felállításával és alkalmazásával kapcsolatos elméletnek, elveknek és gyakorlatnak a tanulmányozása.

A modellünkben javasolt taktikák valószínűleg javítják a keresések hatékonyságát. A valószínűleg szó azonban fontos, mivel ezek a taktikák heurisztikusak, s ezért nem szükségszerű, hogy segítsenek. Továbbá, valamely taktika jó lehet az egyik szituációban, de egy másikban nem.

A hatékonyságot közvetlenül javító taktikákat keresőtaktikáknak, a közvetve javító taktikákat pedig képzelettaktikáknak nevezzük.

A keresőtaktikáknak négy fajtája van:

1. Felügyelő taktikák – céljuk, hogy a keresést a megfelelő vágányon tartsák, és a hatékonyságra figyeljenek.
2. Fájlszerkezet-taktikák – hogyan találja meg a kereső az utat a széles értelemben vett információhordozók között, pontosabban az információ szerkezetében, a kívánt forrásig vagy a forrásban lévő információig.
3. A kérés megfogalmazásával kapcsolatos taktikák – a keresőkép tervezésében, módosításában segítenek: nem korlátozódnak a számítógépes keresésekre.
4. Szóhasználati taktikák – a kérés megfogalmazása közben a kifejezések kiválasztásának módosításában segítenek.

A fájl (a rendezett állomány) fogalmát a hagyományosnál szélesebben értelmezzük: beletartozik az információs egyedek bármely rendezett halmaza. A könyv tipikusan ilyen információs egyed: például egy könyvtár törzsállománya egy jelzettartományon belül rendezve szintén fájl. A fájlstruktúra jelentése értelemszerűen szintén tágul. A lényeg az általánosabb szemléletmód. Például minden indexelő- és osztályozórendszerhez tartozik valamilyen struktúra, ám most nem a struktúra milyensége, hanem meglétének puszta ténye a fontos. A taktikák, amelyek azzal foglalkoznak, hogyan jutunk tovább a fájl szerkezetében, függetlenek az információ adott szervezési módjától.

Ezeknek a taktikáknak a használata a kreatív problémamegoldás egyik formája. Ilyenkor az agy nem mindig működik logikusan, szabályos sémák szerint. Sok különböző irányból közelíthet meg egy-egy kérdést, egyszer ilyen, másszor olyan taktikával. Ezért a taktikák alább következő választékának redundanciája eleve nem tekinthető hibának: kifejezetten jó, ha a taktikák között átfedések vannak.

Keresőtaktikák

M Felügyelőtaktikák

M1 ELLENŐRZÉS

Az eredeti kérdés összehasonlítása a keresés pillanatnyi tárgyával: azonosak-e még.

M2 MÉRLEGELÉS

A keresés egy vagy több pontján felbecsüljük, mennyibe kerülnek vagy mit eredményeznek a következő lépések. Megfontolható, hogy másféle megközelítés nem volna-e eredményesebb.

M3 KAPTAFA

A gyakori típuskérdések megszokott keresési sémákhoz vezetnek. Itt arról van szó, hogy a kérdés tipikussága és az előhúzandó kaptafa tudatosuljon a keresőben, a megoldási sémát azonban át kell gondolni, és ha nem maximálisan hatékony vagy esetleg már idejétmúlt, akkor módosítani kell.

M4 JAVÍTÁS

A keresés tárgyának helyessége tényszerűen és helyesírásilag egyaránt fontos. A hibát elkövetheti eleve az olvasó, azután a kereső amikor emlékezetből keres, nem lévén az írás a kezében. Így fordul elő, hogy neurológia helyett neuralgia lesz a keresés tárgya, ami pedig igencsak eltérő fogalom.

M5 JEGYZETELÉS

Maradjon nyoma, hogy milyen utakat járt végig a kereső, melyeket nem próbált, és melyeket szakított meg.

F Fájlszerkezet-taktikák

F1 „BIBLIZÉS”

Ez az elnevezés neologizmus, a bibliográfia rövidítéséből származik. Kész bibliográfia keresését jelenti, mielőtt nekiállnánk, hogy készítsünk egy újat. Általánosabban fogalmazva, annak ellenőrzéséről van szó, hogy a keresést nem végezték-e már el, az eredmény nem elérhető-e valamilyen használható formában.

F2 PROBLÉMALEBONTÁS

Komplex kérések lebontása részproblémákra. Az egyes részeken külön, egymás után lehet dolgozni, és a részmegoldásokat össze lehet kötni az egész probléma megoldásává.

F3 ÁTTEKINTÉS

Keresés közben a döntések előtt választható lehetőségek áttekintése. Helytelen idő előtt leragadni egyetlen forrás vagy megközelítési mód mellett. Az áttekintés révén az ember ellenállhat az ilyen kísértésnek. Például az elsőnek éppen eszünkbe jutó mutató böngészése helyett gondolatban vegyük sorra a téma nagyobb mutatóit, és válasszuk ki közülük az adott kéréshez leginkább illőt, azután pedig ne lapozzunk rögtön egy tárgyszóhoz, hanem nézzük át a tezauruszt, keressük meg benne a legjobb terminus(oka)t.

F4 KIZÁRÁS

Alapvetően fontos taktika. Ha többféleképpen kereshetünk, válasszuk azt a lehetőséget, amely azonnal a lehető legnagyobb területet zárja ki a további keresésből. Azaz, ha a Smith és Brzustowicz szerzőpáros könyvét keressük, az utóbbi név alapján lényegesen hamarabb jutunk eredményre. Ebben a felfogásban végzik a keresést a kézi koordináltindexeléskor is.

F5 KITERJESZTÉS

Az információforrásokra természetesen olyan összefüggésben szoktunk gondolni, amilyen jellegű használatra szántuk őket. Mégis, szinte minden forrás eredményesen használható nem tervezett célokra is. Ehhez persze nemcsak a rutinszerű használati módokat, hanem a forrás egész információtartalmát ismerni kell. Ha például sikertelen volt a nyomozás egy mérnök címe után, a keresőnek eszébe juthat, hogy a szabadalmakon a feltaláló neve mellett rendszerint a munkáltató is szerepel. Amennyiben az illető mérnöknek van valami szabadalma, a címe is meglesz a szabadalmi nyilvántartásban.

F6 KÖRÜLÁLLVÁNYOZÁS

Amikor az épület elkészült, az állványokat lebontják, de nélkülük az épület nem épült volna fel. Az információkeresés néha ugyancsak ilyen körülállványozásra kényszerül. Olyan információelemekkel kell dolgoznia, amelyeknek közvetlenül ugyan semmi közük sincs a válaszhoz, de lehet, hogy végül mégiscsak hozzásegítenek. Ha például egy jelentéktelen költőről nem sikerül információt találni, a kereső utánanézhet, kik voltak a kortársak, és náluk kereshet tovább, hátha valahol említik az illetőt.

F7 HASOGATÁS

A bináris keresés alkalmazása, ami formálisan hatékonyabb a soros vagy a véletlenszerű keresésnél, de mivel emberekről van szó, pazarlás lenne mindig mereven ragaszkodni a felezéshez. Aki az Ajax Corporation telefonszámait keresi, semmiképpen sem kezdi a telefonkönyv közepét nézegetni. Különösen a nagy, ismeretlen fájlok esetén mégis jó, ha valahol a tudatunk előterében van a bináris keresés elve.

S A kérdés megfogalmazásával kapcsolatos taktikák

S1 SPECIFIKUSSÁG

Minden osztályozási rendszerben és információkereső nyelvben megkövetelik, hogy amennyire csak lehet, a dokumentumokhoz tartozó leírások a tartalomnak megfelelő mértékben specifikusak legyenek. A jelek kivételesen átfogóbb fogalmak alatt is szerepelhetnek, de specifikus szerepeltetésük jóformán mindig kötelező. Ezért a keresést is a specifikus terminusokkal érdemes kezdeni.

S2 RÉSZLETES KÉRDÉSMEGFOGALMAZÁS

Ez a taktika a kérés első megfogalmazásába beviszi annak összes vagy legtöbb elemét, vagy a már kész keresőképet bővíti egy vagy több keresési elemmel. Minél bővebb, azaz minél több elemet kapcsolnak össze ÉS relációval a keresőképben, annál szűkebbek a kérdések, tehát annál kevesebb találat várható eredményképpen.

S3 SZŰKSZAVÚ KÉRDÉS MEGFOGALMAZÁS

Ez a taktika a kérés első megfogalmazásában minimalizálja az elemek számát, vagy a már kész keresőképből vesz el egy vagy több elemet. Minél kevesebb elemet kapcsolnak össze ÉS relációval a keresőképben, annál tágabb a kérdés, tehát annál több találat várható eredményképpen.

S4 HASONLÓSÁGOK MEGENGEDÉSE

A keresőkép szélesítése, bővítése szinonimák vagy egyéb rokon értelmű kifejezések bevonása által; voltaképpen a VAGY kapcsolat alkalmazásáról van szó.

S5 PONTOSÍTÁS

Az előző taktika ellentéte, a keresőkép lehető legpontosabb megfogalmazása a rokon értelmű kifejezések számának minimalizálása vagy legalábbis csökkentése és a legtalálóbb kifejezések megtartása által.

S6 KIZÁRÁS

A kérés megfogalmazásával kizárjuk a válaszból azokat a tételeket, amelyek – önmaguk vagy mutatóik – bizonyos kifejezés(eke)t tartalmaznak, azon az áron is, hogy releváns dokumentumokat veszítünk szem elől. Ez a taktika voltaképpen a DE NEM logikai művelet megfelelője. Azért kizárás a neve, hogy a NEM fent említett kényes oldalára felhívja a figyelmet: A nemkívánatos kifejezést tartalmazó dokumentum kizárása kívánatos információ elvesztésével járhat.

T Szóhasználati taktikák

T1 FELJEBB LÉPÉS (Hierarchiaszint emelés)

Feljebb lépés a hierarchiában az általánosabb, fölérendelt kifejezéshez. A keresőt segítheti a tezaurusz, de lehet, hogy saját ismereteire támaszkodva kell ezt a kifejezést megállapítania.

T2 LEJJEBB LÉPÉS (Hierarchiaszint csökkentés)

Lejjebb lépés a hierarchiában egy specifikusabb, alárendelt fogalomhoz.

T3 ASSZOCIÁCIÓ (Bővítés rokonsági kapcsolatok bevonásával)

Koordinált kifejezés keresése – „oldalirányú" lépés a hierarchiaszinten belül.

T4 SZOMSZÉDKERESÉS

Ez a taktika a szomszédos kifejezések között keres továbbiakat, akár a betűrend, akár a tartalmi hasonlatosság, akár valami más szomszédság alapján. A szomszédkeresés a forrásválasztásra is kiterjeszthető, például amikor megvizsgáljuk a referensz polcain egymás mellé került rokon forrásokat.

T5 NYOMOZÁS

A már fellelt információ megvizsgálása olyan újabb kifejezésekért, amelyek továbbvihetik a keresést. Egyik mindennapi formája az on-line keresés-kor kapott találatok deszkriptor jegyzékének áttekintése. A másik változat: végignézzük a tárgyszavakat, amelyek az adott dokumentum katalóguskártyáján az éppen aktuális besorolási elemen kívül szerepelnek. Ezeket a tárgyszavakat „nyomoknak" hívják, innen származik a taktika elnevezése.

T6 VARIÁCIÓK

A kifejezések módosítása, helyettesítése.

T7 AFFIXUMOK VARIÁLÁSA

Próbálkozás különféle prefixumokkal, suffixumokkal és infixumokkal. Az ún. csonkoló rutinokkal egyidejűleg több ilyen művelet is elvégezhető.

T8 SZÓRENDVÁLTOZTATÁS

Minden olyan rendszerben, ahol egy kifejezés több szóból is állhat, a szórend befolyásolhatja a keresés sikerét. A többszavas terminusok esetén minden lehetséges – vagy legalábbis többféle – értelmes szórend kipróbálását érdemes elvégezni.

T9 ELLENTÉT SZERINTI KERESÉS

A kívánt információt leíró kifejezés logikai ellentétével keresünk, például az „együttműködés" sikertelensége esetén a „verseny" terminussal próbálkozunk.

T10 HELYESÍRÁSI VÁLTOZATOK SZERINTI KERESÉS

A javítás (M4) egyebek között a jó helyesírás megőrzésével is foglalkozik. A jelen taktikánál nem a helyességen, hanem a hatékonyságon van a hangsúly. Főként az on-line rendszerekben igen tarka helyesírással találkozunk, s a jó eredmény érdekében gondolni kell a lehetséges változatokra. A taktikára a manuális rendszerekben is szükség van, gondoljunk például csak az angol–amerikai helyesírási különbségekre.

T11 EGYBEÍRÁS – KÜLÖNÍRÁS

A lehetséges változatok figyelembevétele döntő fontosságú lehet. A probléma a kézi rendszerekben is súlyos. A besorolási szabályok két alapvető változata, a szavankénti és a betűnkénti besorolás a szóköz kezelésében tér el egymástól. Mindkét változatot használják. Ha a kereső csak az egyikre gondol miközben a másik változatot alkalmazó állományban keres, értékes információt veszíthet.

A keresési taktikák és a keresési stratégiával kapcsolatos kutatások

1. *Bizonyos taktikák csoportokat képeznek.* Például a lehetséges reakciók olyan szituációkban, ahol a keresés túl sok vagy kevés találatot eredményez: a lejjebblépés (T2), a részletes kérdésspecifikáció (B2), a pontosítás (S5) és a kizárás (S6), illetve a feljebb lépés (T1), az asszociáció (T3), a szűkszavú kérdésmegfogalmazás (S3), a hasonlóságok megengedése (S4), a szomszédkeresés (T4), a nyomozás (T5) és a variációk (T6).

Ha megkülönböztetjük a keresések tipikus szakaszait és megkeressük a nekik megfelelő taktikákat, kialakíthatunk stratégiai modelleket is. Ha a kereső tudja, hogy egy adott szakaszban a taktikák melyik kis csoportjával számíthat leginkább sikerre, akkor abban a szakaszban csak erre a néhányra kell koncentrálnia.

2. A tényleges keresésen kívül *a tájékoztatási folyamat más elemeire is kidolgozható taktikák.* Ilyen területek a holtpontra jutott kérések kimozdítása, a referensz interjú, ennek részeként a konzultáció a felhasználóval a keresés előtt, alatt és után, a kérés kezdeti elemzésével kapcsolatos taktikák (például a számításba jövő források rendszere aszerint, mennyire valószínű, hogy ténylegesen segítenek, a felhasználótól jövő visszajelzéssel, a visszacsatolással kapcsolatos taktikák, és végül azok, amelyek az eredmény relevanciájának értékelésében segíthetnek). Ezek az eljárások már átvezetnek a képzelettaktikákhoz. A taktikák egyes csoportjainak összeállítása után végcélként felmerül egy nagy, átfogó taktika-készlet. Lehetővé válna általa, hogy egységes szemlélettel tekintsük át az egész referensz folyamatot, és magját képezhetné egy referensz információkeresési tananyagnak.

3. A könyvtári–informatikai gazdaságossági elemzések általában terjedelmes tanulmányokon és matematikai modelleken alapulnak. A mérlegelés (M2) azzal foglalkozik, amit az emberek fejben, néhány másodperc alatt kiszámíthatnak. Egyszerű szabályokra van szükség, megalkotásuk azonban bonyolult munkát kíván. Míg a rendszerkutatók a gazdaságossági elemzések jól fejlett tudományát hívják segítségül, addig az információkeresők számára nincs ilyen tudomány, márpedig *olyan keresési–döntési szabályok kellenek, amelyek minimalizálják a szellemi erőfeszítést.*

4. A keresési stratégia egyik alapvető kérdése, *mikor kell megállni, hogyan állapítható meg, elég információ gyűlt-e össze,* illetőleg mikor kell döntenünk a sikertelen keresés feladása mellett.
   Az áttekintésről (F3) feltehető, hogy a keresés hatékonyságát minőségileg és mennyiségileg egyaránt emeli, és valahol határnak kell lennie, egy ponton túl egyre kisebb a haszon, csakhogy ezt a pontot még meg kell találni.
   A mérlegelésnek (M2) is vannak határai. Például, miután a forrásokat relevanciájuk valószínűsége szerint sorba rendeztük, hol az az optimális pont, ahol az egyik forrásban való keresést abba kell hagyni, és át kell térni a következő forrásra? Alighanem jóval hamarabb érünk ehhez a ponthoz, mint kimerítenénk az adott forrásban rejlő valamennyi lehetőséget.

Képzelettaktikák[4]

A keresőtaktikák fizikailag létező, konkrét eszközök alkalmazására vonatkoznak; céljuk a keresés hatékonyságának közvetlen növelése.

A képzelettaktikák kizárólag értelmi műveletek; céljuk, hogy a keresés holtpontjain átsegítsék a keresőt. A problémamegoldás közben gyakran elakad a gondolkodás, ilyenkor a kérdést új szemszögből, más módszerrel célszerű megközelíteni. Arra már ritkábban gondolunk, hogy az új irányba terelő ötleteknek gyakran az először alkalmazott, többnyire megszokott elképzelések állják útját. A képzelettaktikák új ötletek, megoldási módszerek, elképzelések keletkezését segítik elő, fellazítva a szellemi megkötöttségeket, olykor arra a felismerésre építve, hogy az információk szellemi feldolgozása és az adatok fizikai elhelyezése egymástól elválaszthatatlanok.

Az alábbiakban röviden közöljük az ismertté vált képzelettaktikákat. Ha azt akarjuk, hogy a taktikák eleven, késztető szerepet játsszanak bennünk, felszólító módba tett, tömör és kifejező *igék* formájában célszerű őket nyilvántartani.

E taktikák a képzelet fejlesztését, a bevált sémák feloldását és a képzeletben, illetve fizikailag tárolt információk egyeztetését segítik elő.

A taktikák készlete

I1 GONDOLKODJ

A gondolkodás annyira magától értetődik, hogy meglehetősen pazarlóan bánunk vele. A kereséskor is sok időt és energiát fordítunk rá – sikertelenül. Úgy is mondhatjuk, hogy „gondolkozgatunk" ahelyett, hogy „gondolkodnánk". A programozásban a GONDOLKODJ a jelszó szerepét játssza: azt jelenti, hogy „jusson eszünkbe valami helyes", azaz vegyük észre, hogy gondolkodni kell. Idézzünk elő és tartsunk fenn olyan szellemi állapotot, amelyben valóban gondolkodunk.

I2 ÖMLESSZ (brainstorm)

A kritikus értékelést teljesen félretéve engedjünk utat szabad ötleteinknek, anélkül, hogy nyomban szabatosan meg is fogalmaznánk és értelmeznénk őket.

I3 MEDITÁLJ

Minden felidéző, kereső folyamat tartalmaz befelé forduló, „képszerű" szakaszokat, melyek nehezen ragadhatók meg fogalmilag. E szakaszok akti-

4 Idea tactics. In: Journal of the American Society for Information Science. 1979, Vol. 30, No. 9. p. 280–289.

vizálása érdekében nem annyira konkrét gondolatok szükségesek, mint inkább szellemi állapot, melyben az intuitív és a racionális gondolkodás párhuzamosan működik és egymásra talál. Ezt az állapotot fejezi ki a meditáció.

I4 KONZULTÁLJ

Ötletszerző, tájékozódó beszélgetés a kérdésről valaki mással.

I5 MENTSD, AMI MÉG MENTHETŐ

Ellenőrizzük a már eredménytelennek bizonyult megközelítések még ki nem próbált változatait, mielőtt végleg más irányba indulnánk, nehogy a fürdővízzel együtt a gyereket is kiöntsük.

I6 BÖNGÉSSZ

Böngésszünk a segédletek és a szemünkbe ötlő egyéb eszközök között, hátha valamelyik alkalmas forrásnak bizonyul arra, hogy a keresés új kiindulópontja legyen.

I7 OCSÚDJ (tettenérés)

Vegyük észre, ha zsákutcába kerültünk, és jusson eszünkbe, hogy a megközelítés módosítása haszonnal járhat.

I8 SÖPÖRJ LE MINDENT

Bizonyos fajtájú kérdések gyakrabban fordulnak elő, és megválaszolásuk is gyakrabban sikerül. Az ezekből szerzett tapasztalatok akarva-akaratlanul is a keresés általánosított formájának beidegzéséhez vezetnek. Ezzel az általánosított modellel viszonylag hamar és kevés fejtöréssel célhoz érhetünk. Előfordul azonban, hogy éppen ez okozza adott esetben a keresés sikertelenségét. Fordítsunk tehát hátat az eddigieknek, és söpörjük le a színről a megszokott keresési modellt (vagy legalábbis felejtsük el egy időre). A kérdésben rejlő egyedi problémákhoz talán a keresés teljesen más modellje révén jutunk el, mely éppen azért ismeretlen, mivel az adott probléma eddig csak ritkán vetődött fel, s megoldása nem járt modellalkotó általánosításokkal.

I9 NYISS

Tágítsuk ki a keresési tartományt. Lehet, hogy a kérdéshez választott vagy feltételezett szakterület, tudományág túl szűk, a benne felhalmozódó ismeretek nem elegendőek a válasz megfogalmazásához.

Q5 ALKUDOZZ (a felhasználó befolyásolása)

Vizsgáljuk meg az adott kérdés vonatkoztatási rendszerét, azt a keretet, amelyben a FELHASZNÁLÓ a problémáját megközelítette. Fogalmazzuk meg azt, hogy szükség esetén tisztázhassuk vele, mennyiben változtatható az ő kiindulópontja anélkül, hogy az alkudozás eredményeként túlzott torzulások keletkeznének. Az igen-nem válaszok eredményeként az n-edik lépésben megszülethet a felhasználó számára még elfogadható, a kereső számára pedig megoldható kérdésmegfogalmazás.

I10 TÁROLJ

Bonyolult kereséskor számtalan forrást nézünk át, míg meg nem találjuk azt, amely a leginkább megfelel a kérdésnek. A közbenső információk is elárulhatnak valamit a leginkább megfelelő információ természetéről, esetleg azt is jelezhetik, hogy az eredeti kérdést zagyván fogalmaztuk meg, s ezért újra kell fogalmazni. Azaz jó, ha emlékezetünkben számon tartunk minden nyomravezető elemet, melyről feltételezhető, hogy megváltoztathatja a kérdés természetéről vagy a legmegfelelőbbnek tartott információról alkotott egyik vagy másik elképzelésünket.

I11 HÖKKENTS

Előfordulhat, hogy hiába alkalmaztuk a fent felsorolt taktikákat. Ilyenkor talán a leleményesség és a szellemesség segíthet: találjunk ki valamilyen lehetetlen, eszeveszett vagy legalábbis drámaian új megoldást a problémára! A hökkentés „oldalazó" gondolkodás. Segítségével mintegy rendeződnek a meglévő információk, felismerhetők a hagyományoktól eltérő modellek és teljesen átfogalmazható a kérdés. A bevált sémák meghatározzák a gondolkodás folyamatát, a gondolati „félrelépéssel" vagy „oldalazással" mintegy rápillantunk arra a modellre, mellyel azonosultunk, s ezért megköti a képzeletünket.

I12 VÁLTS

Változtassunk meg valamit – bármit – a keresési viselkedésben; próbálkozzunk más forrásokkal, más kifejezésekkel, más témakörrel stb. A váltással egyben a megszokott keresési modell is automatikusan törlődik. Az új viselkedés termékeny gondolatokat sugall.

I13 SZŰKÍTS

Vizsgáljuk meg a kérdést közelebbről oly módon, hogy (1) a kérdés egészéről a kérdés egyik részletére irányítjuk a figyelmünket, vagy oly módon, hogy (2) a kérdés általánosabb értelmezésére térünk át (akár alkalmazhatjuk mindkettőt).

I14 BŐVÍTS

Vizsgáljuk meg a kérdést távolabbról oly módon, hogy (1) a kérdés egyik részletéről az egészre irányítjuk a figyelmünket, vagy, hogy (2) a kérdés speciálisabb értelmezéséről egy általánosabb értelmezésre térünk át (akár alkalmazhatjuk mindkettőt).

I15 UGORJ

Változtassunk a kérdés megközelítésén oly módon, hogy (1) az összetett, több részből álló kérdés egyik részlete helyett egy másik részlet keresésébe fogunk, vagy oly módon, hogy (2) a kérdést más szemszögből vizsgáljuk meg, amely se nem tágabb, se nem szűkebb, mint a korábbi megközelítés, hanem egyszerűen más.

I16 ÁLLJ

Függesszük fel a keresést és foglalkozzunk valami mással.

N. D. KRAVČENKO

Az indexelés pszicholingvisztikai problémái[5]

A tudományos kommunikáció rendszerén belül négy alapvető összetevőt különböztetünk meg. Nevezetesen: az információk alkotóit, az információkat, az információátvitel csatornáit és a címzetteket (az információk felhasználóit). Annak, hogy az információk az információátvitel csatornáin keresztül gyakran *töredékesen* és *zajok* által zavarva jutnak el a címzettekhez, illetve, hogy a címzettek sok esetben elégtelenül és torzítva használják fel a hozzájuk érkezett információkat, négy fő oka – ún. akadálya – van:

1. ***a nem tudás,*** azaz ha a felhasználó nem tudja vagy nincs tudatában annak, hogy a számára szükséges információ valahol már megszületett és rögzítették;
2. ***a nyelven belüli akadály,*** amikor az információ ismeretbeli és vele együtt nyelvezeti szintje nem egyezik a közvetítő vagy a felhasználó ismeretbeli–nyelvészeti felkészültségével;
3. ***a nyelvek közötti akadály,*** amikor a közvetítő vagy a felhasználó nem ismeri (nem ismeri jól) az információ rögzítésénél használt idegen nyelvet vagy szaknyelvet;

5 Az indexelés pszicholingvisztikai problémái. In: Tudományos és Műszaki Tájékoztatás. 1977, 24. köt, 6. sz., p. 134. Eredeti: Psiholingvističeskij podhod k probleme indeksirovaniă. In: Naučnüje i Tehničeskie Biblioteki SSSR, 1976, Tom. No. 10, p. 14–20.

4. ***a kapcsolathiány,*** amikor nem lehet hozzáférni a szükséges információhoz.

Látható, hogy a négy akadály közül az első három pszicholingvisztikai természetű, csak a negyedik technikai. Ebből következik, hogy a tudományos kommunikáció rendszerét a pszicholingvisztikai szempontok érvényesítése nélkül nem lehet érdemben leírni.

Most, hogy az információátvitel csatornáinak szélesítésében és elmélyítésében mind nagyobb szerepet szánunk a számítógépnek, hangsúlyoznunk kell, hogy e csatornák meghatározó eleme továbbra is az az ember, akinek információközvetítő munkáját – az osztályozást és az indexelést – éppen úgy befolyásolják a pszicholingvisztikai tényezők, mint az információk alkotóiét és felhasználóiét. Vagyis az információközvetítés éppen úgy alkotó munka, mint a másik kettő.

Az indexelés – az információátvitel leglényegesebb mozzanata – nem más, mint valamely természetes nyelven írt dokumentum tartalmának lefordítása valamely formalizált információkereső nyelvre, s ezáltal megbízható keresőkép készítése a dokumentum tartalmáról. E keresőkép segítségével válik a keresés algoritmizált eljárássá.

Függetlenül attól, hogy az indexelés valamilyen könyvtári osztályozó rendszer vagy tezaurusz igénybevételével folyik, valamint feltételezve, hogy az indexelők a lehető legpontosabban betartják az általuk alkalmazott rendszer tezaurusz szabályait, két indexelő ugyanarról a dokumentumtartalomról eltérő keresőképet alkothat. (Más kérdés, hogy a könyvtári osztályozó rendszereknek és a tezauruszoknak egyaránt megvannak a maguk előnyei és hátrányai, s ezért az előbbiek az átfogóbb könyvtári keresésre, az utóbbiak specializált ismeretek sok szempontú keresésére alkalmasak.)

A dokumentum, az indexelő és a keresőkép közötti háromszögű kapcsolatrendszer eddig meglehetősen elhanyagolt kutatási terület volt. Itt a következő, egyelőre végleges válasz nélkül maradó kérdések vethetők fel: formalizálható-e a dokumentum tartalmi értelmezésének a folyamata, formalizálható-e a természetes nyelv bármiféle keresőnyelvre történő fordítása? Lehetséges-e az olvasók valamennyi kérdését előre látni, illetve pontosan leírni, az indexelés melyik módját lehet formalizálni, és melyiket nem lehet? Létezik-e elvileg olyan indexelési mód, amely algoritmizálható?

Annak ellenére, hogy a fenti kérdésekre nem tudunk választ adni, a számítógép lényeges segítséget nyújthat az információátvitelben, ha eleve tudjuk: a számítógép nem válthatja fel az embert, viszont megnövelheti a különféle szellemi tevékenységek hatékonyságát. Ezzel kapcsolatban érdemes emlékeztetni *Jesse Shera* következő intelmére: „A könyvtári–információs gyakorlatban még egyetlen jelenség sem váltott ki olyan széles közérdeklődést, mint a gépesített információkeresés, de egy sem volt improduktívabb nála, minthogy a figyelem homlokterébe a gépek kerültek, nem pedig az emberi, a logikai, a nyelvészeti stb. aspektusok."

PETER INGWERSEN

A dán Peter Ingwersen a könyvtárhasználók és a dokumentumok közötti interakciói, a használó és a könyvtáros megbeszélései és a könyvtáros keresési műveletei alapján az információkeresés pszichológiai szempontokat figyelembe vevő modelljét fogalmazta meg.
A tanulmány alapját képező vizsgálatokat a dán közművelődési könyvtárakban végezték 1976 és 1980 között.

Keresési eljárások a könyvtárban. Kognitív szempontú elemzés

A könyvtárosok szívesen fognak bele a keresésbe anélkül, hogy átgondolnák az előttük álló problémát. Ha a felhasználó nincs jelen, a könyvtáros információkereső tevékenységét három, a motívumokkal és a keresési rutinokkal és lehetőségekkel kapcsolatos elvárásokhoz tapadó attitűd határozza meg: fontos szerepet játszanak a fogalmai ismeretek, a korábbi keresési tapasztalat és a munkaterület. Az attitűdök kihatnak a rutinok használatával kapcsolatos célokra, valamint a keresőfogalmak használatára.

[...]

3.1 A kognitív szempont

Ezt a megközelítési módot általában ismeretelméleti keretbe ágyazva tárgyalják, annak a központi gondolatnak a jegyében, hogy bármiféle információfeldolgozás, legyen az érzéki vagy szimbolikus, ama kategóriák és fogalmak közvetítésével játszódik le, amelyek az információfeldolgozó „készülék" számára a világ modelljét jelentik, függetlenül attól, hogy ez a „készülék" emberi lény vagy gép. A világ eme modelljét az egyéni és a társadalmi (kollektív) tapasztalatok, az oktatás, a képzés stb. határozzák meg. Az ismeretstruktúrákat – a kategóriák és fogalmak rendszerét – úgy is elképzelhetjük, mint az iskolák tantermeiben függő térképek együttesét. A különböző térképek vonatkozhatnak ugyanarra a területre, más-más uralkodó szempont szerint rendezve. A vita során – a fogalmi utat követve – más és más térképek kibontása válhat szükségessé. Az is előfordulhat, hogy a térképek teljesen átalakulnak a tudati műveletek vagy a beszélgetés következtében. Ebből a megközelítésből az egyéni ismeretstruktúrák olyan változatossága keletkezik, amilyenekre a tanulással összefüggő újabb vizsgálatok is következtetnek. Az információkeresés feladata ily módon az, hogy

a meglévő igény kielégítése érdekben a szerzők, a rendszertervezők és az osztályozók kognitív struktúráit összhangba hozza az információs szakember és a felhasználó kognitív struktúráival. A gyakran paradigma-elméletekben megfogalmazott kollektív kognitív mechanizmusok is befolyásolják az osztályozási és indexelőrendszereket, és ezért hatnak a szakirodalomban és az információs igényekben a témák és fogalmak kapcsolataira is.

[...]

3.4 Az információkeresés kognitív modellje

Modellünkben a hangsúly az információs folyamaton van. A világról alkotott kép minden emberben a különböző ismeretstruktúrák konglomerátumából áll. A modell három alapvető individuum világképével számol, amelyeket bizonyos fokig össze kell hangolni ahhoz, hogy az információkereső helyzetben sikeresek legyenek: a felhasználónak, az emberi közvetítőnek (interfésznek) és a dokumentum (a mű) alkotójának a világkép. A kiválasztó, a rendszertervező és az osztályozó/indexelő világképei – melyek ugyancsak belejátszanak az információkereső rendszer képébe – a dokumentum reprezentációhoz, és ezen keresztül az alkotónak a világképéhez tartoznak.

A modell jobb hasábja szemlélteti a felhasználói igény megformálásának a folyamatát. Ha a használó fogalmi ismeretstruktúráit meghatározott területen hiányosnak ítéli (ezt a helyzetet szokták „rendellenesnek" nevezni), az ismeretstruktúrák rendellenes tudásállapottá alakulnak. Ez egyben a felhasználó ama állapota, melyre az információ igény a jellemző. A rendellenes tudásállapottól az emberi közvetítőnek föltett kérdésig vezető folyamatra számtalan transzformáció jellemző, melyek nyomán lépésről lépésre (transzformációról transzformációra) alakul ki a megfogalmazott felhasználói kérdés.

A könyvtáros helyzetére a szakmai problémamegoldó feladat a jellemző. A modellben erre három ismeretstruktúra jellemző:

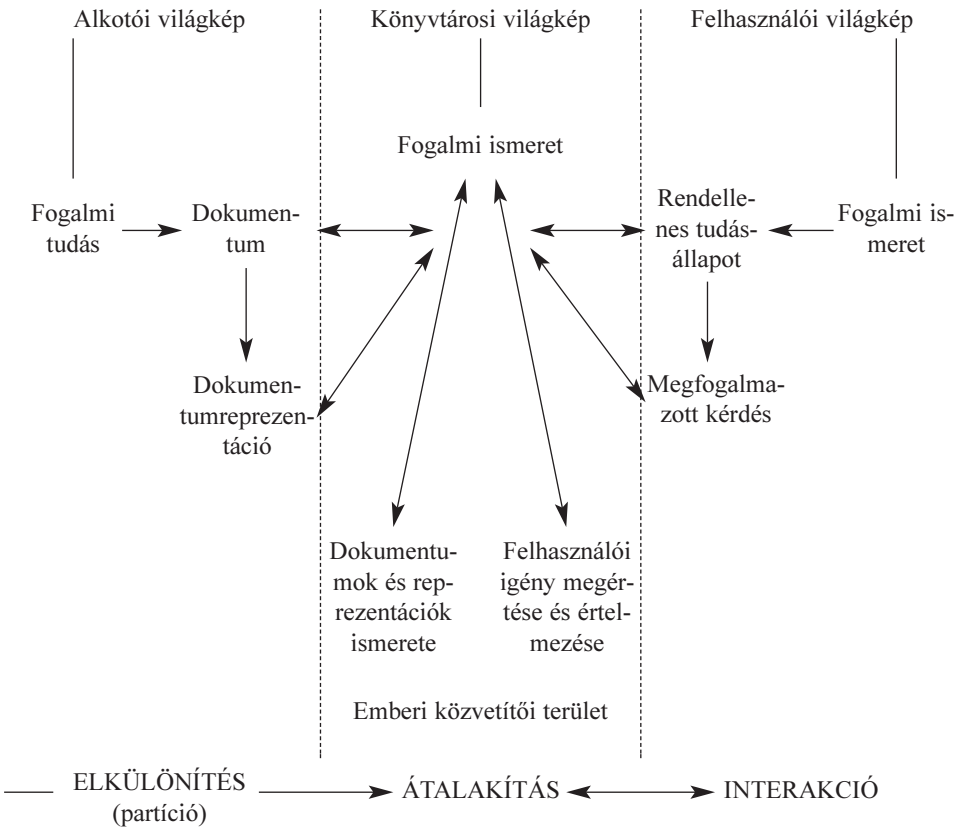
- a könyvtári–információs munkára vonatkozó információs struktúrák (dokumentumok, szurrogátumok, keresési rutinok);
- a fogalmi ismeretekre vonatkozó struktúrák;
- a használó verbálisan megformált információs igényének és problémahelyzetének a tudati felfogása, átélése, melynek nyomán benne valamilyen kép, képzet alakul ki.

Az interakcióhelyzetbe kerülő könyvtáros fogalmi ismeretstruktúrái aktivizálódnak. Ezek a struktúrák a téma vonatkozásában többé-kevésbé megalapozottnak tekinthetők. A kérdés megfogalmazása alapján, és egyéb lelki és interperszonális kapcsolatok segítségével a közvetítő képet alkot a felhasználói

igényről, esetleg még arról a helyzetről is, melyben igénylése következtében a felhasználó van. A megbeszélés során ez a kiinduló kép finomodik, és ebben a könyvtáros fogalmi ismeretei, a dokumentumokra vonatkozó tudása, valamint a felhasználó rendellenes tudásállapotára vonatkozó ismeretei is érintve lesznek.

A közvetítő könyvtáros dokumentumokra, a dokumentum reprezentációkra és a keresésre vonatkozó ismeretei dinamikusan kapcsolódnak a fogalmi ismereteihez, és aktivizálódnak, amikor a közvetítő az igényről alkotott képét megpróbálja összehasonlítani a dokumentumok, pontosabban a dokumentumreprezentációk rendszerével.

Az ábra bal oldali hasábján szerepel a szövegek, a dokumentumok, a művek alkotója, aki közölni akarja saját értékeit, szándékait, céljait. A potenciális felhasználókra vonatkozó ismereteit pedig a dokumentum megjelölésével (címadással stb.) fejezi ki; azaz olyan megjelöléseket választ, melyről feltételezi, hogy azok megfelelnek a felhasználó tudásállapotának.



A könyvtári kommunikációs rendszer kognitív modellje

A modellből kiolvashatók azok a lehetséges következmények és bonyodalmak, amelyekkel számolni kell, ha az emberi közvetítőt „intelligens" információkereső rendszerrel – például automatikus információkereső rendszerrel – akarják helyettesíteni. Számolni kell a fogalmi ismeretek, a felhasználói igények állapota, a szakmai hozzáértés és a keresési eljárások közötti rendkívül bonyolult interakciókkal is, amikor a folyamatokat algoritmizálják.

Az információfeldolgozás és keresés elméleti és gyakorlati szakemberei előtt új szerepek állnak: ha ismeretalapú (tudásbázisú) információkereső rendszereket akarnak létrehozni, vizsgálniok kell a tudati információfeldolgozást is, hogy „intelligens" on-line keresést támogató szakemberként cselekedhessenek és tervezhessenek.

[...]

A felhasználót a nyitott – heurisztikus eszközöket használó – keresés és a szimmetrikus megbeszélési helyzet jellemzi. Ez utóbbi azt jelenti, hogy – együtt a dokumentumokkal és a segédletekkel – a közvetítő könyvtáros számára a felhasználó az információforrás szerepét játssza, és javítja a közvetítő képét a keresési igényről. A közvetítőt az motiválja ebben, hogy igyekszik az igény megértéséhez sok releváns információhoz jutni, mielőtt a keresési probléma tényleges megoldásába egyáltalán belekezdene.

Ugyanakkor feltételezhető aszimmetrikus és félkötött megbeszélési helyzet is. Ebben a felhasználónak passzívabb a szerepe, a közvetítő könyvtáros erősebben támaszkodik a felhasználó által konkrétan megadott keresőfogalmakra, ezeken nem akar különösebben módosítani, nem kérdez vissza, és így kísérli meg közvetlenül megtalálni a releváns dokumentumot.

A teljesen kötött, algoritmikus, minden visszakérdezés nélkül végzett keresés normális munkaköri helyzetben, amikor nem kell túlhajszoltsággal vagy negatív lélektani beállítódással számolni, a kísérleti eredmények szerint ritkán fordul elő. Ez talán annak is tulajdonítható, hogy a felhasználóknak nem lebecsülhető a hatása a közvetítő könyvtárosokra. Amikor a keresés azonnal, aszimmetrikus és kötött kiindulás alapján nem vezet eredményre, a könyvtárosok utat engednek az ismétlődő magatartásnak a keresési eljárásban.

Az „intelligens" on-line információkereső segéderők számára a nyitott keresési módok alkalmazása a leghatékonyabb, mert ebben összekapcsolódnak a keresés elején alkalmazott heurisztikus vonások a későbbi formálisabb megoldásokkal, melyekben az információkereső rendszerben meglévő, beépített algoritmusokat használják ki.

Az információátvitel döntő mozzanata az esetleg rosszul megfogalmazott kérdés mögött meghúzódó információigény pontos megértése és ezt követően átalakítása az információkereső rendszer struktúráinak megfelelően. Ehhez a használónak segítségre és ismeretekre van szüksége.

P. J. VIGIL

Az on-line keresés pszichológiája

In: Tudományos és Műszaki Tájékoztatás, 1984, 31. évf, 11. sz., p. 324–327.
Eredeti: The psychology of on-line searching. In: Journal of the American Society for Information Science, 1983, Vol. 34, No. 4, p. 281–287.

A felhasználó és az on-line katalógus közötti interakció alacsony színvonalú, a felhasználók túl egyszerű stratégiákat használnak. Szükség volna zártciklusú relevancia klaszteráló algoritmusra, melyben a NEM operátort használják. A logikai műveletek közül ugyanis a NEM művelet fogható föl a legnehezebben. Ez pszichológiai tény, a tagadó mondatok több gondolkodási időt igényelnek, mint az állító mondatok. Ezért a negációt a felhasználók a redundáns hivatkozások kiküszöbölésére sem használják. A VAGY műveletet önmagában használva a zaj növekszik.

AZ ON-LINE INFORMÁCIÓKERESÉS ELTERJEDÉSE ÉS A KÉZIKÖNYVEK

Ameddig az adatbázisokhoz kizárólag az adatbázist kezelő nagyszámítógépeken, illetve a hozzájuk kapcsolódó távközlési rendszereken keresztül lehetett hozzáférni, inkább csak a távoli on-line keresőszolgáltatások rendszeres használata volt a jellemző. (Helyi használatra csak a nagyszámítógép monitorjait lehetett igénybe venni, ami nem kedvezett az információkeresés ad hoc természetének.) Ahány adatbázis, annyiféle kereséstechnikai jellemzőre kellett számítani, amiből következett, hogy elsősorban a professzionális felhasználó tudott élni a lehetőségekkel. Az on-line hozzáférésen és a távoli használaton ebben az időszakban jóformán ugyanazt értették.

A nyolcvanas évek közepétől a személyi számítógépek és a helyi hálózatok megjelenésével ugrásszerű változás állt be: a helyi hálózatba kapcsolt személyi számítógépek és terminálok használója bármikor hozzáférhetett a központi gépen kezelt helyi – tehát „saját" – adatbázishoz. Ennek köszönhetően jelentek meg többek között az on-line nyilvános hozzáférésű katalógusok (angol rövidítéssel: OPAC; On-line Public Acces Catalogue). Mindez azzal járt, hogy jelentősen megnőtt azok száma, akik az on-line szolgáltatásokat igénybe vehették; a rendszeres helyi felhasználók pedig legalábbis a saját adatbázis-kezelő rendszerűk professzionális on-line használatát a saját érdekükben igyekeztek elsajátítani. A nyolcvanas évek végétől egyre-másra jelentek meg az ehhez szükséges kézikönyvek. Többségüket az áttekinthető felépítés, az egyszerű nyelvezet, a nagyfokú gyakorlatiasság jellemezte. Minden más kérdést, így a dokumentációs nyelvet, a tezauruszhasználatot, a csereformátumot és a rekordszerkezetet a keresésnek alávetve tárgyaltak bennük.

A kilencvenes évek elején még ennél is jelentősebb változás körvonalai rajzolódtak ki. Az internet megjelenésével ugyanis a távoli on-line keresőszolgáltatások is hozzáférhetővé váltak a személyi számítógépek használói számára, azaz világméretekben tömegessé vált a hálózatok használata.

A lehetőségek eme rendkívüli „demokratizálódásával" párhuzamosan leegyszerűsödött és nem utolsó sorban olcsó – és ami a puszta hálózati használatot illeti, ingyenes – lett a távoli hozzáférés. Az egyes adatbázisok specifikus, nem mindenki által rögtön megismerhető, elágazó, lassú (ennél fogva drága) parancs- vagy menüvezérelt keresőrendszerét az internet gyorsabb, ablaktechnikát alkalmazó grafikus felülete váltotta föl. A World Wide Web eredetileg elsősorban a laikus felhasználók számára készült. Ha arról van szó, hogy a konkrét dokumentációs adatbázisokban (különösen azokban, melyek szolgáltatásai pénzbe kerülnek) kell keresni, a professzionális felhasználó továbbra is az adott adatbázis konkrét kereső eszközeit kénytelen megismerni és gyakorlottan használni, ha azt akarja, hogy keresése hatékony és főleg kifizetődő legyen.
Mindez még inkább növelte az igényt a könnyen érhető, ugyanakkor minden fontos keresési információt tartalmazó szakkönyvek iránt. Bennük, ami a tartalom szerinti - jelentéssel bíró szavak alapján végzett - feltárást és keresést illeti, szinte kizárólag a szabad tárgyszavakkal, a szabad szövegen belüli kereséssel és a deszkriptoros nyelvekkel (a tezaurusz használatával) foglalkoznak, de szigorúan csak annyira, amennyire ez a keresési technika megismeréséhez szükséges.

A legfontosabb on-line szakkönyveket az alábbi táblázatban foglaltuk össze:

Az on-line információkereséssel foglalkozó művek

1980	Busha, Ch., Harter, S. P.: Research method in librarianship
1981	Penichel, C. H., Harter, S. P.: Survey of on-line searching
	Markey, K., Atherton, P.: On-line training and practice. Manual for ERIC data base searchers. (2. kiad.)
1986	Harter, S. P.: On-line information retrieval. Concepts, principles, and techniques (2. kiad.: 1990)
1989	Lee Pao, M.: Concepts of information retrieval
1990	Hartley, R. J. [et al.]: On-line searching. Principles and practice
1992	Carande, R.: Automation in library reference sercvice. A handbook
1994	Hewitt, J. A. [et al.]: Advances in library automation and networking
1997	Reading in information retrieval. Ed. by Sparck Jones, K., Willet, P.

Az alábbiakban néhány ismertebb kézikönyvből mutatunk be részleteket, melyek a rendezéssel, a dokumentációs nyelv használatával összefüggésben főként ökonomizmusukkal tűnnek ki. Szerzőik nem elsősorban tudományos kutatók elméleti értelemben (mint *Michael Keen*, aki már a het-

venes években részt vett a második cranfieldi jelentés és a SMART rendszer készítésében, vagy *Stephen P. Harter*), hanem többnyire olyan könyvtárosok, akik az információkeresés tanárai az egyetemeken (mint *Charles Meadow*, *Puline Cochrane*, *Dick Harley*, *Miranda Lee Pao*). Könyveikben alig van szó elméletről, mindent alárendeltek a gyakorlati felhasználásnak. Az osztályozás az on-line használatban technikai kérdéssé vált.

CHARLES MEADOW–PAULINE COCHRANE (ATHERTON) (1929)

Charles Meadow a philadelphiai Drexel University, Pauline Cochrane pedig a New York-i Syracuse University professzora. *Robert Freemannal* együtt vett részt az ETO on-line használatával foglalkozó nemzetközi fejlesztési programban. Könyvüket az oktatás céljára írták, az alapokból kiindulva a legmagasabb szintig tárgyalták benne az on-line információkeresést.

Meadow-tól magyarul korábban megjelent:

Vissza a jövőbe: az adatbázisipar kronológiája

In: Tudományos és Műszaki Tájékoztatás, 1989, 36. évf., 6. sz., p. 266–271.
Eredeti: Back to the future: making and interpreting the database industry timeline. On-line database industry timeline. In: Database, 1988, Vol. 11., No. 5, p. 14–31.

Az adatbázisiparnak is kialakult már a történelme. Az első WORM (write-once, roead-many, egyszer írható, sokszor olvasható) adathordozók a mezopotámiai agyagtáblák voltak. Az on-line adatbázisok eredete a 17. századig vezethető vissza, amikor megjelentek az első tudományos folyóiratok, ezt követően jöttek létre a témakör szerinti könyvtári katalógusok, indexek és referáló kiadványok (pl. az Index medicus 1879-ben). A számítógépek történeti fejlődésének leírása után kronologikus táblázatban ismerteti az információtudomány fejlődését.

Az on-line keresés alapjai[1]

Keresés szövegben[2]

1. A kezdetek

Amikor az 1960-as évek elején az első információkereső rendszerek kialakultak, a fájlok (a számítógépben tárolt adatállományok) rekordjai (a hivatkozási tételek) általában nem tartalmaztak tartalmi kivonatot (pl. referátumot, tömörítvényt stb.). Drága volt a kivonatok konvertálása (újbóli előállítása) géppel olvasható formába, és nagyon drága volt a gépi tárolásuk is. Mivel a felhasználók mindezt nem tartották gazdaságosnak, ezért nem is igényelték a kivonatokat. Az on-line keresőrendszerekben manapság használt alapvető parancsok léte azon az elgondoláson alapszik, hogy bizonyos specifikus kereshető mezőkből ún. szótárfájlokat készítenek, melyek tartalmazzák a keresőkifejezésekkel összehasonlítható szavakat vagy rövid kifejezéseket. A gépi memória olcsóbbodásával megjelentek a gépi tárolású rekordokban a tartalmi kivonatok is, és a bennük előforduló szavak[3] bekerültek a szótárba (lásd az 1. ábrát).

Ma ez a hagyományos eljárás. A fejlődés még le sem zárult, amikor már azt is igényelték, hogy egyedi szavak szerint egy-egy címen, sőt többszavas deszkriptoron belül is kereshessenek. Így aztán az is lehetővé vált, hogy egy-egy mezőn, tehát deszkriptormezőn, címeken, kivonatokon belül is keresni lehessen, a szótárak tartalma tehát bővült, túllépett az indexelők által kijelölt, illetve a címbeli szavak révén adódó deszkriptorok körén. Egy *mezőn* belül többféle módon lehet keresni, amint azt az alábbiakban majd leírjuk. Előbb azonban azok számára, akik nem ismerik a fogalmakat, megvizsgáljuk a dokumentumok indexeléséhez használatos, *szabályozott* és *szabályozatlan* szókincs (és szótárak) közötti különbséget.

2. A szabályozott és szabályozatlan szókincs

Ha visszamegyünk a legkorábbi számítógépes információkeresési kísérletekhez, azt látjuk, hogy a rekordok rövid mezőkből állnak, amint erről már volt is szó. A mezők adatelemeket tartalmaznak, például a *címet,* a *szerzőt,* a *megjelenés idejét* és a *deszkriptorokat.* A szabályozott szótárak alapján kijelölt

1 Basic of on-line searching / Charles Meadow, Pauline Cochrane. – New York: Wiley and Sons, 1981. XIV, 245 p. – (A Wiley Interscience Publication) (Information Science Series)

2 Text searching. In: Basic of on-line searching, p. 94–104.

3 A szótárak készítésekor majdnem mindig elhagyjuk a névelőket, elöljárókat stb. Ha az ellenkezőjére nem utalunk, akkor ez azt jelenti, hogy ezeket a szavakat eleve kihagyjuk a kérdés tárgyalásából.

deszkriptorok szellemi munka eredményei; a célja az, hogy a dokumentum tartalmát előre meghatározott szó-, kifejezés- vagy kódkészlet segítségével leírják (feltárják). A deszkriptorok lehetnek hagyományos osztályozási jelzetek (pl. az ETO jelzetek), tárgyszavak vagy kulcsszavak. A kulcsszó rendszerint egyetlen szó vagy rövid kifejezés, amely a mutatóban más kulcsszavakkal kombinálva érthetőbben reprezentálja a tartalmat, mint a hierarchikus osztályozási jelzet vagy a hagyományos tárgyszó[4]. A kulcsszavak kombinációival az indexelő könnyen szerkeszthet elemi nyelvi építőkockákból egyedi és jellemző dokumentumleírásokat; így előre nem látott, váratlanul felbukkanó tartalmak is könnyen leírhatóak.

A kulcsszavak használatának egyik módja, hogy előre elkészül az elfogadható kulcsszavak jegyzéke. Noha a jegyzék módosítható, az indexelőnek vagy a keresőnek első lépésben az adott jegyzékben kell keresnie a legtalálóbb kifejezést. Az ilyen szótárnak szabályozó–ellenőrző szerepe van. Idővel a szókincset többek között hierarchikusan is strukturálhatják, s kialakulnak a tezauruszok. A legfontosabb relációk a nem–faj (az általános és a specifikus kifejezés közötti kapcsolat) és a rokonság reláció. A szinonimákat a „lásd" (inverzüket a „helyett") utalás jelzi.

Egyes adatbázisokban az indexelők *szabad kulcsszavakat* is használhatnak, azaz a használat pillanatában kiválasztott szavakat és kifejezéseket. A szabad kulcsszavak kiválasztása mindig csak ahhoz igazodik, hogy az adott dokumentum leírásához mire van szükség, és nem ahhoz, hogy mi van a tezaurusban. A szabad kulcsszavak formájukban nem különböznek a szabályozott (ellenőrzött) kulcsszavaktól, de a használatukra vonatkozóan nincs irányadó forrás, és más mezőbe kerülnek.

Egyes rendszerekben különösen a hely- (földrajzi név) és márkaneveket kezelik szabad kulcsszóként. A szabad kulcsszavak számára fenntartott külön mező megkönnyítette az indexelést, különösen az új, fejlődésben levő témákkal foglalkozó dokumentumok esetén. Sajnos a múltban a keresőnek kevés fogódzója volt ahhoz, hogy az indexelő a szabad kulcsszavakkal hogyan írja le a témát.

A kulcsszavakon kívül manapság bizonyos mezők minden szavát is invertált fájlba teszik, ahol közvetlenül hozzáférhetők: ezek a mezők a rekordban az ún. lekérdezhető mezők. A szótárfájl vagy invertált fájl a számítógép által kiválogatott kifejezések állományának tekinthető, bár tartalmazhatja az indexelők által kijelölt kifejezéseket is. Az 1. ábrán invertált fájl megjelenése, a mutató látható.

4 A szerző itt a történetileg kialakult Cutter-féle, hierarchikus al-, fő- és melléktárgyszavakból álló tárgyszavakra gondol (a szerk.).

Hivatkozási fájl	A invertált fájl Ismérv	Tételek száma	B invertált fájl Ismérv	Tételek száma
Tételazonosító: 1	SZ = BAKER	3	SZ = BAKER	3
Cím: A referátum készítés	SZ = MILLER	2	SZ = MILLER	2
elmélete	SZ = SMITH	1	SZ = SMITH	1
Szerző: Smith	DE = REFERÁTUM	1	REFERÁTUM/DE	1
Deszkripotorok:	DE = DOKUMENTÁCIÓ	1	REFERÁTUM/CI	1
Referátum	DE = INFORMÁCIÓKERESÉS	1	AUTOMATIZÁLÁS/DE	2
Dokumentáció				3
Információkeresés		3	DE = DOKUMENTÁCIÓ	1
Tételazonosító: 2	DE = KÖNYVTÁRI AUTOMATIZÁLÁS	2	DE = INFORMÁCIÓKERESÉS	1
Cím: Keresés		3		3
Szerző: Miller	DE = ON-LINE KERESÉS	2	DE = KÖNYVTÁRI AUTOMATIZÁLÁS	2
Deszkriptorok:				3
On-line keresés			DE = ON-LINE KERESÉS	2
Könyvtári automatizálás			DOKUMENTÁCIÓ/DE	1
			INFORMÁCIÓ/DE	1
Tételazonosító: 3				3
Cím: Az információkeresés elmélete			INFORMÁCIÓ/CI	3
			KÖNYVTÁR/DE	2
Szerző: Baker				3
Deszkriptorok:			ON-LINE/DE	2
Információkeresés			INFORMÁCIÓKERESÉS/DE	1
Könyvtári automatizálás				3
			INFORMÁCIÓKERESÉS/CI	3
			KERESÉS/DE	2
			KERESÉS/CI	2
			ELMÉLET/CI	1
				3
			CI = REFERÁTUMKÉSZÍTÉS ELMÉLETE	1
			CI = KERESÉS	2
			CI = INFORMÁCIÓKERESÉS ELMÉLETE	3

1. ábra. Szótár- vagy invertált fájl. Az A invertált fájl csak az indexelő által meghatározott szerzőket (SZ) és deszkriptorokat (DE) tartalmazza. Ebben az elrendezésben ezek a kereshető kifejezések. A B invertált fájl címeket (CI) is tartalmaz, valamint önálló szavakat, amelyek minden címben vagy deszkriptorban szerepeltek; kivételek a leggyakoribb általános vagy funkcionális szavak (pl. Az). A B invertált fájlban az önálló szavakhoz kapcsolt utótagok – pl. /DE – mutatják, milyen mezőből származik az illető szó. Tehát keresni lehet teljes kifejezést – pl. információkeresés elmélete –, vagy egyetlen szót, pl. elmélet. A teljes mezőtartalmakat előtagok jelölik (pl. DE). Az ábra célja, hogy megközelítő képet adjon, de nem pontos változata egyetlen rendszernek sem.

Jól felismerhető a különbség, milyen az invertált fájl akkor, ha csak az indexelő választotta szavakat tartalmazza, és milyen akkor, ha minden lekérdezhető mező minden szavát tartalmazza. Az invertált fájlok a lekérdezhető mezők minden szavát tartalmazzák, kivéve a leggyakoribb általános szavakat, amelyekről az a vélemény, hogy nincs információtartalmuk. Ilyen szavak például a kötőszók, névmások, névelők, névutók, de a túl általános jelentésű szavak is (pl. Valami, Dolog, Jelenség, Különbség, Időnként). Ezeket a szavakat jegyzékbe foglalják, amelyet negatív szótárnak, irreleváns szótárnak (stoplistának) neveznek. Ezt használják arra, hogy kizárják a benne szereplő szavakat a feldolgozásból. Így a dokumentum (vagy a cím, vagy a tartalmi kivonat) szövegében szereplő összes szónak közel 50 százaléka kizárható anélkül, hogy lényeges információveszteség keletkeznék. Az invertált fájl valójában nem más, mint az összes lekérdezhető mezőszöveg szavainak a keresés céljára rendezett változata.

Az Országos Széchenyi Könyvtár NEKTÁR adatbázisában használt negatív szótár látható a 2. ábrán. Mivel a nemzeti könyvtárban különféle nyelvű dokumentumokat dolgoztak fel, a negatív szótár nemcsak a magyar nyelvű tiltott szavakat tartalmazza.

a	ko	mellett	v
...	kod	melletti	vaan
alatt	kohtaan	mellől	vagy
alatti	konečno	mert	vai
ale	körül	merthogy	vaikka
alebo	következésképpen	miedzy	vajha
aleer	közé	miatt	vajon
alhoewel	között	míg	valószínű
ali	közötti	mindamellett	van
alig	közül	mint	vanaf
aligha	krome	minthogy	varten
alighanem	kroz	mit	...
all’	...	mitsamt	with
alla	m	mivel	wskutek
...	ma	mivelhogy	wśród
ellen	malgré	múltán	y
ellenben	már	múlva	ynnä
ellenére	másrészt	myös	z
elől	medu	...	za
előtt	még	u	zamiast
en	mégis	über	zatem
ener	meglehetősen	überm	zato
enimvero	mégpedig	übers	zbog
ennélfogva	mégse	...	ze
...	mellé		...

2. ábra. A Magyar Nemzeti Bibliográfia tiltott szavainak jegyzéke

Összefoglalva: a rekordokhoz tartozhatnak az indexelők által kijelölt, szabályozott és szabályozatlan kifejezések, valamint olyan kifejezések, amelyeket a számítógép „válogat ki", mivel szerepeltek valamelyik lekérdezhető mezőben. A kifejezések két halmaza rendszerint összekeveredve található az invertált fájlban. Az indexeléskor használt többszavas deszkriptorok, mint az Információkeresés elmélete invertált fájlban megjelennek egyszer összetett kifejezésként és az egyes összetett kifejezésként az egyes összetevő szavaknak megfelelő helyen. Példánkban három bejegyzés jelenik meg: az Információkeresés, az Elmélet és az Információkeresés elmélete.

3. A szövegben végzett keresés haszna

Bármely könyvtári gyűjteményben végzett keresésnek – legyen az kézi, vagy gépi keresés – az a hagyományos módja, hogy egy vagy több osztályozási jelzetet, tárgyszót vagy deszkriptort használunk. A keresőnek meg kell próbálnia kikövetkeztetni, milyen kifejezéseket használt a katalogizáló vagy az indexelő az adott témaleírására. Ez az egyik oka annak, hogy általában szükség van a hivatásos könyvtáros segítségére, akinek nagyobb gyakorlata van abban hogy egyeztesse a felhasználó igényeit és a katalogizáló vagy indexelő döntéseit, mint a téma szakértőjének.

A számítógépek tovább bővítik a keresés módjait. Segítségükkel bármelyik szóval lekérdezhető a bibliográfiai rekord vagy a rekord legtöbb mezője. Ebből az is következik, hogy a keresőnek nem kell törnie a fejét a deszkriptorokon, ehelyett egyszerűen azokat a szavakat használja, amelyeket ismer, vagy amelyekről reméli, hogy ismerik, és használta a szerző vagy a referáló a téma leírására. Ha a keresőnek és a szerzőnek ugyanaz a szakterülete, ez sokkal egyszerűbb lehet, mint az indexelők segítségének igénybevétele. Ez a lehetőség különösen két esetben jár haszonnal. Amikor új szó születik a szaknyelvben vagy a szakzsargonban, időbe kerül, amíg a szó a szabályozott szótárba kerül. Így a gyorsan fejlődő szakterületen dolgozó kereső (például aki az elemi részecskék fizikájával foglalkozik), azt tapasztalhatja, hogy legújabb szakszavait a bibliográfusok nem ismerik. Az egyedi, szabályozatlan szavakkal végzett keresés második hasznát csak akkor tapasztalja a kereső, amikor csak nehezen talál valamit a témában, illetve amikor finomra kell „hangolnia" az egyébként túlságosan sok találati tételt eredményező keresést. Ilyenkor kiegészítheti a szabályozott szókincset néhány szabályozatlan kifejezéssel, amelyekkel a téma finomabban leírható, mint egyébként. A szabadságnak ezen a fokán (hogy ti. bárhol a szövegben kereshetők szavak) alakul ki az a keresési mód, amelyet szabad, szövegen belüli keresésnek (free text searching) hívnak.

A keresők igényei azonban soha nincsenek teljesen kielégítve. Ha lehet is lekérdezni a *Toxicitás* és a *Vizsgálat* szavakkal, nincs biztosíték arra, hogy

ha egyidejűleg[5] szerepelnek valamelyik hivatkozásban, akkor ott *Toxicitás vizsgálat* a tárgyalt téma. A biztonság kedvéért a keresőnek szüksége van még tárgyszóra vagy osztályozási jelzetre, vagy arra, hogy az összetevő szavak helyett magával a teljes kifejezéssel végezhessék a lekérdezést. Ezt a lekérdezési módot nevezik szövegen belüli vagy szövegkeresésnek (text searching). Ebben tehát nemcsak azt lehet megszabni, hogy milyen szavak szerint keressenek, hanem azt is, hogy az illető szavak milyen sorrendben vagy egymáshoz milyen közel forduljanak elő a szövegben.

A szövegben végzett kereséseknek („szövegkereséseknek") az alábbi fajtáiról van szó:

Szabad szövegen belüli keresés:

Megkötések nélkül szövegben végzett keresés. Ilyenkor nem szabhatják meg, hogy a kifejezés milyen távolságban legyen valamilyen másik kifejezéstől, nem csonkolnak, és nem határoznak meg más keresőszó tulajdonságot sem.

Kötött szövegen belüli keresés:

Szövegben különféle feltételek, megkötések megadásával végzett keresés. Általában a cím vagy a referátum szövegében végzett kereséskor alkalmazzák.

Teljes szövegben végzett keresés (lásd az utolsó fejezetet):

A dokumentum teljes, rövidítetlen szövegében végezhető, különféle feltételek, megkötések megadásával végzett keresés.

4. Keresés szótöredékek szerint

Reméljük, most már érti az olvasó, hogy amikor az egyes rendszerekben a keresésre a *select,* a *find*, a *search* parancsot[6] adunk valamelyik szóra, akkor arra számítunk, hogy a szót egy deszkriptoron, címen, kivonaton, vagy valamilyen más mezőn belül találjuk meg. Nem kívánjuk tehát egyben azt, hogy egyedül a szó maga alkossa tartalmát. Ez azt jelenti, hogy választhatunk: a deszkriptorokkal és teljes címekkel is kereshetünk, vagy kereshetünk a bennük előforduló egyedi szavakkal. Tehát még akkor is, ha csak szabályozott szókincsre támaszkodunk, kereshetünk többszavas deszkriptor részét alkotó szóval, vagy ugyanennek a szónak bármilyen előfordulásaival, ha az egy kivonaton belül szerepel.

---

5 Az angol nyelvben a Toxicity testing szakkifejezést mindig különírják. Magyar dokumentum esetén ilyenkor számolni kell azzal, hogy egyaránt előfordulhat a különírt és az egybeírt változat.

6 A SELECT parancsot általában a DIALOG rendszerben, a FIND parancsnevet általában a CCL nyelvű rendszerekben, a SEARCH parancsnevet pedig általában a STAIRS rendszerekben használják keresőparancsként.

Mihelyst eltávolodunk attól a követelménytől, hogy szabályozott kifejezéseket használjunk, belekerülünk a szóalak- és helyesírási változatok dzsungelébe. Ha például a számítógépek témája iránt érdeklődünk és történetesen az ERIC tezauruszból választjuk ki a keresőszót, a tezauruszból kiderül, hogy ennek a szónak a szabályozott írásmódja a többes számú alak: *Számítógépek.* Ez tehát a szabályozott deszkriptor, és csak ezt az alakot szabad használni, ha szabályozott szótárra támaszkodva végezzük a keresést. Tegyük fel, hogy ha a Számítógépek alakot használtuk, feltűnően kevés találatunk van, noha gyanítjuk, hogy több van. Az egyik lehetőség az, hogy keressük a *Számítógépek* szót vagy bármely kváziszinonimáját (ilyen pl. az *Analóg számítógépek)* egy másik deszkriptor vagy egy cím, illetve kivonat részeként. Ha a szó deszkriptorként fordul elő, természetesen azt jelenti, hogy az indexelő szerint a számítógépek dokumentum lényeges témája.

Az a tény, hogy a szó a kivonatban jelen van, nem jelenti szükségképpen azt is, hogy reprezentálja az is az adott dokumentum tartalmát. Döntenünk kell, hogy kérjük-e a szónak eme járulékos előfordulásait, kockáztatva a hamis találatokat, olyan tételek visszanyerését, amelyek formálisan kielégítik a kérésünket, de ténylegesen a nem kívánt témáról szólnak. Például a referátum szólhat számítógépek felhasználásáról, adószámítási eljárásról és valamelyik mondatában előfordulhat az *Adójogi könyvtár* kifejezés is. Ezt a dokumentumot megkaphatnánk, amikor számítógépeket keresünk (a *Számítógépek* szóval vagy változataival). De ha megkapnánk, a találat hamis lenne, mert nem a számítógépek könyvtári alkalmazásáról szól. Amikor a cím vagy a referátum szerzője a számítógépekre utal, nem mindig a *Számítógépek* szót használja. Helyette időnként használja a *Számítógép, Számítástechnika* vagy *Számítógépes* szavakat, amelyek lényegében mind ugyanazt jelentik a jelzett téma szempontjából. Az információkereső rendszerek nem várják el tőlünk, hogy valamennyi lehetséges szinonimát megadjuk az OR (‚vagy' műveletjelölő) szócskával összekapcsolva; elég ha szócsonkot adunk meg (töredékszót, vagy a „tövet") és speciális szimbólummal jelezzük, hogy bármiféle karaktert elfogadunk a szóréssze kezdődő szó további részeként. A szóban forgó esetben a következőket (a parancsokat a DIALOG, az ORBIT és a BRS információkereső rendszerek előírásai szerint fogalmazzuk meg):

select számít?	(DIALOG)
számít:	(ORBIT)
számít$	(BRS)

Ebben az esetben a három szimbólum, a **?**, **:** és **$** mind ugyanazt jelenti. Mindegyik azt „specifikálja", hogy a kereső elfogad bármilyen szót, amelyik a töredékszóval kezdődik és tetszés szerinti (esetleg tehát nulla) számú további betűt tartalmaz. Más szóval, ezekkel a parancsokkal mind kereshető a „számít" szó fontosabb elsorolt összes változata.

Az eljárás alkalmazásának vannak határai. Például a *szám* is része a *számítógépnek,* s ezért a select *szám**?*** legális parancs, de nem különösképpen értelmes, mert nagyon sok szó kezdődik ezzel a szórésszel (számháború, számító, számelmélet, számolás, számrendszer, számvitel stb.). Minden keresőrendszerben léteznek eszközök a szótöredékkel végzett lekérdezés finomítására. Ha a *seprű* szót változatai szerint keressük, csonkításszimbólumot – ún. „maszkot" – használhatunk a szó végén a birtokos alak *(seprű**nek**)* a kiiktatására és hasonló szimbólumot használhatunk a szó belsejében különféle tőváltó formák együttes kiiktatására (a seprű szó tevékenységet jelentő alakja ui. seper). Az ORBIT rendszerben azt mondanánk: *sep**#**r* – annak jelölésére, hogy elfogadunk bármilyen – de csak egyetlen – betűt a **#** jelölt pozícióban és akárhány betűt az után. Az utolsó **?**, ha egy szóközzel elválasztják a „szótőtől", azt jelzi, hogy a szóközt megelőző ***?***-eket úgy kell kezelni, hogy mindegyik egyetlen karakternek felel meg. A BRS-ben a *sepr**$4*** kifejezés azt jelenti, hogy a *sepr* után még legfeljebb négy karakter következhet.

5. Keresés szóláncok szerint

Az egyedi szavakkal még a csonkítás különféle formáival együtt sem kérdezhetünk le olyan pontosan, mint amilyenre időnként szükség van. Használtuk korábban a *Toxicitás vizsgálat* kifejezést. A toxicitás vizsgálat speciális típusa az *in vitro* toxicitás vizsgálat. Annak ismeretében, hogy az *In vitro toxicitás vizsgálat* kifejezést szinte biztosan használják a témával foglalkozó dokumentumok címében vagy kivonatában, a kutató joggal kívánhatja, hogy az egész kifejezés szerint keressen. Így ugyanis kiküszöbölhető a szavak más lehetséges kombinációiból keletkező „zaj". Hasonló a helyzet a *Számítógépes grafika* kifejezés esetén; *Számítógép* és *Grafika* szavak együttes előfordulása ugyanabban a címben vagy referátumban nem szükségképpen a számítógépes grafikával foglalkozó dokumentációra utal. Minden információkereső rendszerben megoldható, hogy azokat a tételeket keressük, melyekben az ilyen kifejezések – mondjuk a *Napenergiát hasznosító eszköz* – egy-egy deszkriptor-, cím- vagy referátummezőn belül előfordulnak. A megoldások azonban már különböznek.

Az ORBIT megengedi a kereséshez használt kifejezés közvetlen specifikálását (meghatározását és a rendszer előírásai szerinti leírását), amennyiben már létezik találathalmazunk. Ha a *Napenergiát hasznosító eszköz* kifejezést akarjuk, egyszerűen ezt mondjuk. A parancs azonban más, a neve „string" (lánc). (A lánc betűkből, szóközökből és esetleg egyéb szimbólumokból álló sorozat.) Ez a láncparancs legalább egy találathalmazt képező parancs után kell hogy következzék, azaz *csak egy előzetesen létrehozott találathalmazon működik.* (Más szóval előbb „durvább" eljárással – deszkriptorral, szótöredékkel stb. – ki kell válogatni a valószínűbb tételeket.) Bizonyos értelemben a *string* korlátozó parancs, csak előzetesen kiválogatott halmaz csökkentésére használatos. A példában találato-

kat gyűjthetünk tágabb értelmű kifejezéssel (ilyen pl. az *Energia* vagy a *Számítógép,* és utána lekérdezhetünk a speciálisabb jelentésű lánc szerint).

A DIALOG és a BRS hasonlítanak egymáshoz: a kereső specifikálja a számára értékes szavakat, és azt, hogy ezeknek a szavaknak milyen összekapcsolódását kívánja. Például a DIALOG-ban a

select Toxicitás ***(w)*** *Vizsgálat*

parancs azt mondja a rendszernek, hogy azokat a tételeket keresse meg, amelyekben a *Toxicitás* szó a *Vizsgálat* szóval együtt (ez esetben mellette) fordul elő a rekord valamelyik mezőjében. A *w* elé együtthatót (numerikus prefixumot) tehetünk és a szavak között hosszabb távolságot specifikálhatunk. A

select Toxicitás ***(2w)*** *Vizsgálat*

parancs megtalálná azokat a tételeket, amelyekben a *Toxicitás vizsgálat* kifejezés is és a *Toxicitás és környezetvédelmi vizsgálat* kifejezés is előfordul az *és* figyelmen kívül hagyásával.[7]

Ennek a jelölésnek a neve *infixum*. A DIALOG-ban egy **f** infixum [pl. *Toxicitás* ***(f)*** *Vizsgálat*] azt jelenti, hogy a kereső két szó szerint keres, legyenek azok bárhol egyazon mezőn belül. Egy ***c*** infixum viszont azt jelenti, hogy a két szó bárhol előfordulhat egy tételen belül. A select *Toxicitás* ***(c)*** *Vizsgálat* parancs ugyanazt eredményezi, mint a

select Toxicitás
select Vizsgálat
combine 1 and 2 sorozat, vagy a
select Toxicitás Vizsgálat parancs.

A BRS az infixumokkal egyenértékű különleges szavakat használ, ezek az ***adj*** (az adjacent [szomszédos] megfelelője), ***with*** (-val/-vel) és ***same*** pedig ugyanabban a bekezdésben vagy mezőben. Egy lehetséges BRS parancs

Napenergiát ***(adj)*** *Hasznosító* ***(adj)*** *Eszköz*

minden olyan rekordot kér, amely tartalmazza a *Napenergiát hasznosító eszköz* kifejezést. Ha ***with*** kapcsolat volna, a rendszer megkeresne minden olyan rekordot, amelyben ez a három szó előfordul ugyanabban a mondatban, a ***same*** esetén pedig azokat a rekordokat keresné meg a rendszer, amelyekben ez a három szó ugyanabban a mezőben vagy bekezdésben szerepel. A BRS jelölési rendszere egészen bonyolult lehet. Például a *Napenergiát hasznosító Eszköz* ***(with)*** *Lak****$4*** kifejezés jelentheti a *Napenergiát hasznosító eszköz* kifejezést egy mondaton belül, a *Lakás* vagy *Lakóház* szóval vagy a szó egyéb változataival.[8]

---

7 Az angol példa: Information retrieval as information storage and retrieval.
8 Az angol példa: Solar thermal (adj) Energy (with) Residen$4.

A csonkításszimbólumokat keresőparancsokban lehet használni, tehát mondjuk:

woman: athlet:
vagy
select wom?n? (lw) athlet?
vagy
(woman$ women$ adj athlet$3)

és mindegyik esetben a női atlétika (*women's athletics*), nők az atlétikában (*women in athletics*), atlétanő (*woman athlete*) kifejezésekkel játszódik le a keresés. Végezetül néhány szót az általánosan használt szavakról. Mindegyik rendszernek van negatív szótára: ebben azok a szavak szerepelnek, amelyeknek nincs érdemben témameghatározó szerepük. Ezek a szavak figyelmen kívül hagyhatók a szövegen belüli keresésben. Például a szavak egymásmellettiségének megállapításakor a *women in athletics* (nők az atlétikában) kifejezés *in* (-ban/-ben) szócskáját az ORBIT nem veszi figyelembe.

A 3. ábrán láthatók a különféle szövegen belüli kereséshez használható parancsok és az általuk kapható eredmények példái.

Bibliográfiai tétel
Cím: On-line bibliográfiai keresés

Abstract: This work discusses the automation of reference searching through various systems for storage and retrieval of document surrogates using on-line, mechanized search systems, such as DIALOG, BRS, ORBIT, or MEDLINE. These are described, as are some research projects involving automatic indexig and abstracting of documents by computer. Published in a special library binding.

BRS keresőkifejezések	**Eredmény**
	Megkapjuk a tételt
1. *literature **adj** search**$3***	A két szó együtt, a megadott sorrendben jelenjék meg a címben.
2. *dialog **with** medline*	Mindkét szó ugyanabban a mondatban szerepeljen.
3. *searchng **with** on-line*	Mindkét szó ugyanabban a mondatban szerepeljen mindegy lehet.
4. *library **same** automation*	Mindkét szó a referátumban forduljon elő, de nem kell, hogy ugyanabban a mondatban legyenek.
	Nem kapjuk meg a tételt
5. *on-line **adj** searching*	A két szó nem szomszédos egymással.
6. *library **with** automation*	A két szó nem szerepel ugyanabban a mondatban.
7. *literature **with** automation*	A két szó nem szerepel ugyanabban a bekezdésben.

DIALOG keresőkifejezések	**Eredmény**
	Megkapjuk a tételt
1. *on-line* ***(2w)*** *search*	A két szó vagy a címben vagy a referátumban együtt szerepel.
2. *on-line* ***(f)*** *searching*	Mindkét szó a címben vagy a referátumban (de nem feltétlenül együtt) szerepel.
3. *library* ***(f)*** *automation*	Mindkét szó előfordul a referátumban.
4. *literature* ***(c)*** *search*	A *literature* a címben, a *search* a referátumban fordul elő, de ugyanabban a tételben.
	Nem kapjuk meg a tételt
5. *on-line* ***(w)*** *searching*	A két szó nem szomszédos egymással.
6. *literature* ***(w)*** *search*	A két szó nem szomszédos egymással.
7. *literature* ***(f)*** *search*	A két szó nem található ugyanabban a közös mezőben.

ORBIT keresőkifejezések	**Eredmény**
	Megkapjuk a tételt
1. *search:*	Megtalálható a címben vagy a referátumban.
2. *search system#*	A két szóból álló kifejezés megtalálható a referátumban.
3. *abstracting documents*	Az *abstracting of docment* megtalálható a referátumban; az *of* tiltott szó, nem számít.
4. *abstract****:*** *document****:***	A helyzet ugyanaz, mint az előbb.
	Nem kapjuk meg a tételt
5. *search system*	Csak a *search systems* fordul elő.
6. *abstract # document #*	Mivel a # jel csak egy karaktert engedélyez, nincs találat.
7. *orbit dialog brs*	A kifejezés ugyan létezik, de nem a megadott sorrendben.

3. ábra. Szövegben végzett keresés a BRS, a DIALOG és az ORBIT rendszerben. A bal oldali oszlopban a parancsok láthatók, a jobboldali oszlopban pedig a parancsok eredménye olvasható.

6. Mikor keressünk a szövegben?

Néhányszor rámutattunk már, mikor érdemes a szövegben keresni. Ismételjük át ezeket a pontokat, azután pillantsunk távolabb, ki tudja, hány évet, és nézzük meg, hogy néhány technikai újdonság hogyan befolyásolja az irodalomkeresést, a technikai fejleményeket.

A dokumentumok témájának leírására ma jóformán minden referáló és indexelőszolgáltatásban használnak valamiféle szabályozott nyelvet: osztályozási jelzeteket, tárgyszavakat, deszkriptorokat vagy ezeknek valamilyen kombinációját. El kell dönteni, hogy milyen tárgyszavakat, deszkriptorokat stb. használjunk. Valahányszor feltételezhető, hogy a kereső jól ki tudja vá-

lasztani a megfelelő szabályozott keresőkifejezéseket, akkor a kifejezések szerinti lekérdezés a leghatékonyabb eljárás.

Ha finomításokra van szükség, például hasonló témák megkülönböztetésének olyan fokára, amely kívül van az szabályozott nyelv lehetőségein, akkor a szövegben érdemes keresni. Ez érvényes akkor is, ha a kereső bármilyen oknál fogva nem tudja kiválasztani a megfelelő szabályozott kifejezéseket, például olyan időszakban, amikor új szavak áramlanak a tudomány szókincsébe. Tulajdonképpen rendkívül hasznos eljárás, ha szabad szövegen belüli kereséssel összeszedünk sok jó hivatkozási tételt, azután lekérdezzük őket a legalkalmasabb deszkriptorokkal, majd ezek felhasználásával újrafogalmazzuk a keresést további releváns tételek érdekében.

A szövegben végzett keresés viszonylag obskurus anyag keresésekor is értékes, például akkor, amikor a dokumentum éppen csak érint egy témát, és a kereső szerint az indexelők valószínűleg figyelmen kívül hagyták a fogalmat. Nyilvánvaló a szövegben végzett keresés haszna, ha a kifejezés bármilyen oknál fogva nem szerepel a szabályozott nyelvben.

A könyvtárak története azt bizonyítja, hogy a tartalom, a tárgykörök szerint felépített és megszervezett információtárolás valamilyen formája (hagyományosan a katalógus) nélkül lehetetlen megtalálni az adott témáról szóló művet, kivéve, ha valaki emlékezetből tudja azt, hol található, ami a nagy könyvtárak esetén valószínűtlen dolog. A modern számítógépek segítségével eljön a nap, amikor kényelmes is és gazdaságos is lesz, hogy a dokumentumok egész szövegét gépi adathordozókra rögzítsék és teljesen elhagyjanak mindenféle osztályozást vagy indexelést.

Helyettük a szövegben végzett keresés fejlett módszerére támaszkodnak majd a keresők. *Salton* többéves kísérletei a Cornell Egyetemen[9] arra utalnak, hogy az ilyenfajta keresésnek semmivel sem kell kevésbé eredményesnek lennie, mint a szabályozott nyelvet használó indexelésnek.

7. A legújabb: A teljes – eredeti – szövegben végezhető on-line keresés

1980-ban a BRS új kísérletet ismertetett. Hivatkozások és referátumok helyett folyóiratcikkek teljes szövegét vették fel on-line adatállományaikba:

> *„Az adatbázist magáncélból, kísérleti jelleggel állították fel a Journal of Medicial Chemistry (Orvoskémiai Folyóirat) (1976–78) mintegy 1000 cikkének teljes szövegéből. Az adatállomány címeket, kivonatokat és a cikkek teljes szövegét tartalmazta, s ezek mind közvetlenül kereshetők voltak. A vállalkozás célja az volt, hogy a primer folyóirat-*

9 Lásd kötetünkben *Gerard Salton* szemelvényét (a szerk.).

fájl létesítésének műszaki feltételeit megállapítva értékelje a fájl használat az erre a célra alakult felhasználói csoporttal.

A teljes szövegben végezhető on-line keresés egyik, legigényesebb fajtáját »környezet feltáró« (in-context) keresésnek nevezik. Amikor a keresést a primer szövegeket tartalmazó adatbázisban végzik, a rendszer nemcsak a szokásos eredményeket közli, megjelölve a felhasznált kifejezést tartalmazó cikkek számát, hanem a környezeti adatokat is megjeleníti, megjelölve a kifejezés pontos helyét a cikken belül (bekezdés, mondat, a szó/szavak sorszáma/sorszámai). A kereső ezután böngészhet a szövegben, kérheti csak azokat a bekezdéseket, amelyek tartalmazták a keresett kifejezést.

Mivel az egész – egyébként részekre osztott – cikk kereshető, a keresők összehasonlítják a cikkek törzsében szereplő módszereket, eredményeket, tényeket. Hasonlóképpen, a keresők könnyen megtalálhatják a cikkek leglényegesebb pontjait és finom részleteit, mert a szöveg végigböngészhető és a kulcsbekezdések kinyomtathatók például az első és utolsó bekezdés, amelyek rendszerint a célokat és a végkövetkeztetéseket tartalmazzák."

Példa teljes szövegben végzett keresésre:

Parancs és válasz	**Megjegyzések**
1_: melting adj point$1	E kifejezés alapján keresésre használhatók a *melting point* (olvadáspont) vagy *melting points* (olvadáspontok) szókapcsolatok.
RESULT 86	86 dokumentum elégíti ki a kérést.
2_:...print 1 oc/doc = 1	A felhasználó kéri a környezetfeltáró (in-context) kereséseredményét.
OC PARAGRAPH SENTENCE WORD TX(8) 3 7	Ez a pontos hely: bekezdés (a 8.), mondat (a 3.) és a szó (a 7.) sorszáma.
*tx(8) 1	A felhasználó kéri megtekintésre a 8. bekezdés szövegét.

TX PARAAGRAPH 98 OF 20. EXPERIMENTAL SECTION. MELTING POINTS WERE TAKEN ON A THOMAS–HOOVER CAPILLARY MELTING POINT APPARATUS AND ARE CORRECTED. WHERE ANALYSES ARE INDICATED ONLY BY THE SYMBOLS OF THE ELEMENTS, ANALYTICAL RESULTS WERE OBTAINED FOR THESE ELEMENTS WERE WITHIN ±0.4% FOR ALL COMPOUNDS AND WERE FOUND TO BE CONSISTENTA WITH THE ASSIGNMENT STRUCTURES.

RICHARD J. HARTLEY

Az alábbi szemelvényünk olyan bevezető tankönyvnek íródott, mely az on-line szakirodalomban gyakori közös munka eredménye. Mindent az alapoktól kezdve ismertetnek benne, teljesen tájékozatlan olvasót feltételezve. Az egyes témakörökben (adatbázisok szerkezete, adatcsere-formátumok, parancsnyelvek, tezaurusz használat) nem mélyednek különösebben el, mindig az on-line kereső praktikus információigényeit tartják szem előtt. Ha valaki részletesebb információkat igényel, a speciálisabb szakkönyvekhez kell fordulnia.

Richard J. Hartley a wales-i egyetem információtudományi és könyvtári tanszékének (Department of Information and Library Studies, the University College of Wales) professzora, jelenleg főleg a nyilvános on-line katalógusokban végzett tartalmi keresés gyakorlati kérdéseivel foglalkozik. Korábban 12 éven át akadémiai és nyilvános könyvtárakban, majd a brit nemzeti könyvtárban dolgozott.

A vele egy tanszéken dolgozó szerzőtársa, Michael Keen 1966-ban *Cyrill Cleverdonnal* és *Jack Millsszel* együtt részt vett a második cranfieldi kísérletekben, később pedig *Gerard Saltonnal* dolgozott együtt a SMART automatizált információkereső rendszer fejlesztésén. Számos on-line tanfolyam szervezése fűződik a nevéhez, tanulmányaiban elsősorban az információkeresés stratégiájával foglalkozik.

Andy Large a montreali könyvtári és információtudományi iskola (Graduate School of Library and Informations Studies at McGill University in Montreal) igazgatója, az egyik legismertebb on-line keresési kézikönyv szerkesztője, kutatási programok és on-line konferenciák szervezője, az Education for Information című folyóirat szerkesztője.

Lucy Tedd szabadfoglalkozású on-line szakértő, egyetemi előadó, a Program című szaklap szerkesztője.

On-line keresés. Elvek és gyakorlat[10]

On-line keresés lokális adatbázisokban[11]

Bevezetés

Az on-line keresést e könyvben eddig többnyire távoli, főleg kereskedelmi szolgáltatók nyilvánosan hozzáférhető adatbázisainak lekérdezésével ösz-

10 On-line searching : Principles and practice / R. J. Harley, E. M. Keen, J. A. Large and L. A. Tedd. – London : Bowker-Saur, 1990. 408 p.

11 On-line Searching of Locally Stored Databases. p. 256–281. In: On-line searching...

szefüggésben tárgyaltuk. Ebben a fejezetben a helyi számítógépes rendszerekben működő adatbázisokban végezhető keresésre fordítjuk figyelmünket. A helyi vagy lokális adatbázisok lehetnek nyilvánosak vagy házon belüliek, tárolhatók mágneses vagy optikai lemezen.

Bibliográfiai adatok tárolására és keresésére először *Luhn* alkalmazott számítógépeket 1957-ben az IBM-nél. Eleinte ez abban merült ki, hogy a dokumentumok címében szereplő kulcsszavakból a számítógép segítségével mutatótételeket készítettek, és ezeket a kulcsszavak betűrendjében kinyomtatták. A mutató KWIC (Keyword in Context, Kulcsszavak szövegkörnyezetben) indexként váltak ismertté. Az 1. ábrán négy dokumentumcím KWIC-indexe látható.

Walks in west Wales	1
Hiking up hills in Wales	2
Hill-walking: some useful hints	3
Wales: a guide to walking in the hills	4

GUIDE	
Wales: a guide to walking in the hills	4
HIKING	
Hiking up hills in Wales	2
HILLS	
Hiking up hills in Wales	2
Wales: a guide to walking in the hills	4
HILL-WALKING	
Hill-walking: some useful hints	3
HINTS	
Hill-walking: some useful hints	3
WALES	
Hiking up hills in Wales	2
Wales: a guide to walking in the hills	4
Walks in west Wales	1
WALKING	
Wales: a guide to walking in the hills	4
WALKS	
Walks in west Wales	1
WEST	
Walks in west Wales	1

1. ábra. Egyszerűsített KWIC-típusú index

A hatvanas évek közepére néhány nagy szervezet már a korábban különféle kártyarendszereken (például peremlyukkártyán, nyolcvanoszlopos fénylyukkártyán) kezelt információ tárolására és keresésére is felhasználta helyben lévő nagyszámítógépét. Ezeket a bejelölt kártyákat elektromechanikus eszkö-

zökkel lehetett használni. A számítógépes rendszereket eredetileg nyomtatott termékek, gyarapodási jegyzékek, KWIC indexek, SDI szolgáltatások előállítására, néha napján a teljes géppel olvasható rekordkészlet retrospektív lekérdezésére stb. használták. A hatvanas évek végén kimondottan ilyen célokat szolgáló szoftver volt például az ICI mezőgazdasági részlege által kifejlesztett ASSASSIN (Agricultural System for the Storage and Subsequent Selection of Information rövidítése). Az ASSASSIN-t kezdetben a nyilvános adatbázisok (például az INSPEC, a Chemical Abstracts Condensates) házi feldolgozására használták.

A hetvenes évek közepén már számos információs részleg és szakkönyvtár kezdte használni a távoli kereskedelmi keresőszolgáltatásokat nyilvános információ on-line lekérdezésére, de még mindig szükségük volt külön rendszerekre, amelyekkel a helyi vagy bizalmas információkat – a cég jelentéseit, a laboratóriumi vizsgálati eredményeket, piaci felméréseket stb. – tárolták és keresték. *Tedd* 1979-ben számolt be arról, hogy 12 európai információs részleg használ (helyben lévő vagy távoli) számítógépeket nyilvános vagy zárt bibliográfiai adatgyűjtemények lekérdezésére. A távoli keresőszolgáltatások közül néhány lehetővé teszi a használói számára, hogy magán jellegű adatokat tároljanak, amelyekben a megfelelő parancsnyelv és távközlési rendszer segítségével kereshetnek. A nyolcvanas években azonban a helyi számítógépes források gyors fejlődése nyomán egyre inkább elterjedt a helyi adatbázisok mikroszámítógépes használata. A számítógépesítés költségeinek csökkenése sok szervezet számára elérhetővé tette az ilyen információ házon belüli feldolgozását biztosító szoftver használatát. Ezt a szoftvert gyakran információkezelő szoftvernek nevezik, amely *Kazlauskas* 1987-ben írt meghatározása szerint

> *„... olyan számítógép-program, amely elősegíti változó hosszúságú szöveges rekordok létrehozását, kezelését és karbantartását, illetve ezekből további termékek előállítását. Ezek a rekordok jellemzően neveket, címeket, dátumokat, a kiadással kapcsolatos információkat, forrást, a lelőhelyet, az indexkifejezéseket, a referátumokat, a tartalomra vonatkozó szöveges és számszerű információkat, azonosító számokat és hasonló adatokat tartalmaznak. Ezeket az adatokat könyvek, dokumentumok, jelentések, audiovizuális és mágneses információhordozók, cikkek és különlenyomatok, emlékiratok és levelezések, szerződések, tájékoztatók, periratok és más jogi anyagok, és számos tájékoztató és forrástájékoztató dokumentum leírásai és teljes szövegei tartalmazzák."*

Kazlauskas becslései szerint a hetvenes évek végén Észak-Amerikában mintegy húsz kereskedelmi forgalomban megszerezhető programcsomag állt rendelkezésre információkezelő adatbázisok létrehozására; 1984-re mintegy

100 ilyen csomag volt, az 1987-es jegyzékébe pedig már több mint 200 csomagot vett fel. *Kazlauskas* a következőképpen osztja fel az információkezelő szoftvereket:

1. Fájlkezelő szoftverek.
2. Általános adatbázis-kezelő szoftverek.
3. Speciális adatbázis-kezelő szoftverek.
4. Könyvtári/irattári/információs központok egyedi adatbázis-kezelő rendszerei.
5. Könyvtári/irattári/információs központok integrált rendszerei.
6. Szöveges információkereső rendszerek (szövegben kereső rendszerek).

A szöveges információkereső rendszer a távoli on-line adatbázisokban végzett kereséshez hasonló funkciókat biztosít; ebben a fejezetben elsősorban ezeknek a programoknak a helyi keresésben való felhasználásával foglalkozunk.

Bizonyos szervezetek CD-ROM-on szereznek be helyben kereshető adatbázisokat. Ezekben az esetekben keresőprogram (általában szövegkereső szoftver) is járul az adatokhoz.

Ez a fejezet a nyilvánosan hozzáférhető CD-ROM-adatbázisok, illetve a saját szerkesztésű adatbázisok on-line keresésével kapcsolatos kérdésekkel foglalkozik.

Helyi rekordok és adatbázis szerkezet

CD-ROM-adatbázisok rekordjai

A jelenleg beszerezhető CD-ROM-adatbázisok közül nem egy távoli on-line szolgáltatásokon át is elérhető. Időnként vannak azonban eltérések, ha mégoly csekélyek is, a rekordszerkezetben, illetve a keresési lehetőségekben.

A 2. ábra például azt mutatja milyen a LISA (Library and Information Science Abstracts) rekord a Silver Platter által előállított CD-ROM-adatbázison, illetve a Dialogon. Természetesen a rendszer kézikönyvéből lehet megtudni, hogyan kell az adott rendszerben bizonyos adatbázisban keresni. Az adatbázis előállítója által kiadott kézikönyv általában olyan hasznos összevetéseket is tartalmaz, amelyek befolyásolhatják, hogy a kereső melyik rendszer mellett dönt.

A 2. ábrán látható, hogy a CD-ROM-változatban a szavak közé többször is kötőjeleket iktattak be, hogy lehetővé tegyék az összetett kifejezésekre való keresést, amilyen például az On-line–Computer–Library–Center, Bulletin–des–Bibliotheques–de–France vagy az Information–storage–and–retrieval.

(a) ***Silver Platter CD-ROM***

Silver Platter v 1.4 LISA (1/69–9/88)

TI: The **OCLC–DBMIST** agreement
TO: L´accord **OCLC–DBMISZ**
AU: **Darrobers, – Martine**;
D–B–M–I–S–T–(Direction–des–bibliothèques, –des–musees–et–de –l´information–secientifique–ete–technique),–France; Online–Computer–Library–Center–(**OCLC**)
SO: Bulletin–des–Bibliothèques–de–France, 30 (6) 1985, 537–538. 3 refs
PY: 1985
LA French
AB At the end of 1985, the French Directorate of Libraries, Museums, and Scientific and Technical Information (DGMIST) signed an agreement with the OS On-line Computer Library Center (OCLC) to cooperate in provision of cataloguing services and aresearch project. OCLC´s international data base already inculdes 700.000 French notices, and provides content summaries and locations as well as bibliographic descriptions. This step qwill allow France to develop its own national catalogues, and although the move may be opposed on the grounds that it means abandoning French standards for American, the move to adopting international practices is essential now that databases are internationally accessible on-line.
FH: On-line Cooperation, France, Direction des bibliothèques des musees et de l´information scientifique et technique and On-line Computer Library Center
DE: France-; Technical–processes–and–services; Information–storage–and–retrieval; Information–retrieval; Cataloguing–; Computerised–cataloguing; On–line–cataloguing; Cooperation–
CC: TogsNccD44 Togs
DA: 1987
AN: 87–1485

(b) ***Dialog***

179715 87–1485 Library and Information Science Abstracts (LISA)
The **OCLC–DBMIST** agreement
L´accord **OCLC–DBMISZ**
: **Darrobers, – Martine**;
Bulletin–des–Bibliothèques–de–France
SOURCE: 30 (6) 1985, 537–538. 3 refs
LANGUAGES: French
At the end of 1985, the French Directorate of Libraries, Museums, and Scientific and Technical Information (DGMIST) signed an agreement with the OS On-line Computer Library Center (OCLC) to cooperate in provision of cataloguing services and aresearch project. OCLC´s international data base already inculdes 700.000 French notices, and provides content summaries and locations as well as bibliographic descriptions. This step qwill allow France to develop its own national catalogues, and although the move may be opposed on the grounds that it means abandoning French standards for American, the move to adopting international practices is essential now that databases are internationally accessible on-line.
NOTE: D B M I S T (Direction des bibliothèques, des musees et de l´information secientifique et technique), France; On-line Computer Library Center (**OCLC**)
DESCRIPTORS: France; Technical processes and services; Information storage and retrieval; Information retrieval; Cataloguing; Computerised cataloguing; On line cataloguing; Cooperation
SECTION HEADINGS: CATALOGUING
SECTION HEADING CODES: TogsNccD44 Togs

2. ábra. A LISA referálószolgáltatásnak a Silver Platter (a) által előállított CD-ROM és Dialog (b) rekordja

A CD-ROM-változat a szerző mezőben tartalmazza a szerző munkahelyét is (jelen esetben D–B–M–I–S–T), amit a Dialog rekordokban nem találhatunk meg. Egy másik kisebb eltérés, hogy a CD-ROM-változatban szerepel explicit dátum is (DA), amely a Dialógban (a CD-ROM-adatbázisban is) rejtett adatként jelenik meg a tétel ún. gyarapodási azonosítójában. A jelentősebb indexelési döntéseket kivétel nélkül az adatbázis-készítők hozzák meg (a LISA esetében a British Library Association), ám a keresési kódok, a kereshető mezők, a nyomtatási formátumok stb. meghatározása már az adatbázist letöltő szervezet hatáskörébe tartozik, legyen az CD-ROM-gyártó (mint a Silver Platter), vagy adatbázis-szolgáltató (például a Dialog).

Képernyő-oldal (VDU)	Képernyő-mező azonosítója	Szinonima-fájl mező-azonosítója	Mezőnév	Mező-hosszúság×mező száma	Szinonima	Mutató-típus	Mező-végjel
1. oldal		1	Elérési azonosító	7×1		F	
	06	6	Felhasználható mező 2	1×6	acc		
	02	2	Felhasználható mező 1	1×6			
	03	3	Biztonsági kód	1×8	sec	F	
	04	4	Átfogó tárgyszókód	4×1	abllh	F	
	05	5	Fájlazonosító	6×1	fil		
	07	7	Felhasználható mező 3	1×1			
	08	8	Felhasználható mező 4	1×1			
	09	9	Felhasználható mező 5	56×2			
	10	10	Raktári jelzet	63×2	loc		
	11	11	ETO-jelzet	63×2	ETO	F	
	12	12	Dátum	4×1	dat	F	
	13	13	Szerző	223×2	aut	A	
	14	14	Cím	383×3	tit	T	
2. oldal	02	15	Kapcsolat	222×2	ref	T	
	03	16	Kiadó	222×1	pub	T	
3. oldal	02	17	Terjedelem	298×4	col	T	
	03	18	Szerző típusa	1×1	auc	F	
	04	19	ISBN	20×1	isb	T	
	05	20	A szöveg nyelve	58×2	lat	A	
	06	21	Dokumentumtípus	27×2	doc	A	
	07	22	Felhasználható mező 6	20×2			
	08	23	IDRI-tanulmány azonosító	5×1	pro	F	
	09	24	Átfogó tárgyszó	58×3	buh	F	
	10	25	Gyors hozzáférés	58×2	acb	F	
4. oldal	02	26	Köröztetés	50×2	lon	A	
	03	27	Visszaérkezési dátum	50×1	rdt	F	
	04	28	Helyi tárgyszavak	210×1	shd	A	
	05	29	Felhasználható mező 7	1×1			
	06	30	Felhasználható mező 8	1×1			
	07	31	Felhasználható mező 9	1×1			
5. oldal		32	Deszkriptorok		des		

A = automatikus; F = teljes mezőtartalom (full field); M = kézi, eseti, opcionális (manual); T = jelölt (Tagged)

3. ábra. A CAIRS információkereső rendszer rekordjának mezői (az eredetileg angol mezőnevek magyar fordításban szerepelnek)

Helyi – házon belüli – adatbázisok rekordjai

Az on-line kereshető lokális adatbázisok létrehozásának sarkalatos pontja a rekordok – a mezők számának, lehetséges hosszúságának, e mezők indexelési módjának stb. – tervezése. A 3. ábrán látható a CAIRS terminológiája szerinti képernyő-meghatározási táblaként ismert rekordszerkezet. Ebben mind a 32 mezőre különböző paramétereket határoznak meg, például a mező nevét, a mező indexelésének módját.

A CAIRS csomag számos lehetőséget nyújt keresőkifejezések generálására. Ezek közül az alábbiakat használják:

A (automatikus)	–	a tiltott szavak jegyzékén szereplők kivételével az összes szót felhasználják.
F (teljes mező)	–	a mező teljes tartalmát keresőkifejezésként viszik be az indexbe.
M (kézi)	–	a keresőkifejezéseket kézzel jelölik ki.
T (jelölt)	–	a szavakat kijelölik (a < > jelek közé) és így kerülnek keresőkifejezésként a mutatóba.

Néhány helyi adatbázis sokkal egyszerűbb, mint ahogy azt bemutattuk. A 4. ábrán például feltételezett idegenforgalmi információs központ sétányokról készített rekordja látható, amely a sétaút jellegéről, hosszáról stb. tartalmaz részleteket.

NÉV:	BROBRYN
HOSSZ:	5
TEREP:	Mocsaras
TÉRKÉP:	OS 135
KOR:	8+
LÁTNIVALÓK:	Vízesés; Bánya; Ragadozó madarak
ÚTMUTATÁS:	Induljon a National Trust parkolójától (SN) és haladjon a jelzett úton...

4. ábra. Egyszerű rekord

Teskey szerint gyakran nem mérik fel a rekordszerkezet tervezésének kihatását a keresőrendszer teljesítményére. Nemegyszer a rekordon végrehajtott apró változtatások is jelentősen befolyásolják a tételek keresési idejét.

A rekordszerkezet szabványai

Még az azonos típusú adatbázisok (bibliográfiai, numerikus, szöveges) rekordjai is jelentős elétéréseket mutatnak, ennek ellenére megkísérelték ezt

az összevisszaságot országos, sőt, nemzetközi szabványok keretei közé szorítani. A szabványosított rekordszerkezetek alkalmazása megkönnyíti az új adatbázis megismerését, és az ilyen szabványokra elengedhetetlenül szükség van, ha valamikor könnyedén szeretnénk adatokat áttölteni egyik adatbázisból a másikba.

A szabványosítás különösen fontos a bibliográfiai rekordok esetében. Például sok országban állítanak elő géppel olvasható nemzeti bibliográfiákat, amelyek a határok között megjelent könyveket veszik számba. Az így kapott rekordokat azután hatalmas egyetemes bibliográfiába lehetne összeolvasztani, amely a világ teljes dokumentumterméséről számot adna. A középkor óta kergetjük ezt az álmot. A bibliográfiai rekordok ilyen cseréje jóval egyszerűbb lenne, ha valamennyi ország azonos szerkezetű rekordokat generálna – azonos mezők követnék egymást azonos sorrendben. E cél érdekében a hatvanas évek végén megállapodás jött létre Nagy-Britannia és az Egyesült Államok között. Ezt a számos mezőt és almezőt tartalmazó adatcsere-formátum szabványt nevezik MARC-nak (MAchine Readable Cataloguing; Géppel olvasható katalogizálás).

Bár a MARC formátumot jelenleg jó néhány ország átvette, az egyes országok aprólékos kívánalmainak megfelelő nemzeti változatok miatt sajnos még nem működik igazi nemzetközi szabványként. Ennek folytán a közelmúltban másik szabványt dolgoztak ki, az UNIMARC-ot, amelyet nemzetközi szabványnak szántak. Néhány ország elhatározta, hogy nemzeti bibliográfiai rendszere számára átveszi. Az UNIMARC egyben olyan közös adatcsereformátum, amely lehetővé teszi, hogy a különböző MARC formátumokat az UNIMARC-on keresztül bármely másik változatra konvertálják. Ez azt jelentené, hogy bármely MARC formátumhoz csak UNIMARC-ra konvertáló programra lenne szükség. Az UNIMARC-ról azután bármely másik MARC formátumra átkonvertálhatóvá válna.

Nemzetközi szabványt (ISO 2709) is kidolgoztak a bibliográfiai rekordok cseréjére. Ezt a szabványt széles körben elfogadták és használják, különösen a MARC rekordok esetében. A MARC formátumot azonban elsősorban könyvek géppel olvasható rekordjainak cseréjére szánták, és nem igazán megfelelő a folyóiratcikkek részletes leírására. E probléma leküzdésére hozták létre az Unesco és az International Council of Scientific Unions Abstracting Board (ICSU–AB), az International Federation of Library Associations and Institutions (IFLA) és az International Standardization Organization (ISO) összefogásával a Közös Kommunikációs Formátumot (Common Communication Format; CCF). Az Unesco éveken át foglalkozott a helyi adatbázisok fejlesztésével olyan fejlődő országokban, amelyek számára a távoli kereskedelmi on-line szolgáltatások a nagy költségek, a távközlés hiányosságai vagy hasonló okok miatt nem elérhetők. A CCF olyan adatcsere-formátum, amelyet az információs világnak ezek a

fejlődő országokbeli szervezetei használnak, amelyeknek elsősorban csak a bibliográfiai rekordok cseréjére futja.

Az adatcsere-formátumokról kötetünkben részletesen *Vajda Erik* és *Mirna Wilmer* szemelvényeiben olvashatunk.

A helyi adatbázisok szerkezete

Az e fejezet elsődleges témáját adó szöveges információkereső rendszer a kereshető kifejezésekből alkotott invertált fájlokon alapul. *Ashford* felsorolja az ilyen kezelőrendszerek jellemzőit:

1. Az információ szövegét egyetlen rugalmas rekordban tárolják, amely mező- és rekordazonosítókat tartalmaz a mezők és a rekordok egyértelmű meghatározására.
2. A tárolt információhoz olyan invertált fájlon keresztül lehet hozzáférni, amely valamennyi, a szövegben előforduló szignifikáns kifejezést mutatónévként tartalmazza.
3. A kereső parancsnyelv segítségével érintkezik a rendszerrel, amely lehetővé teszi a Boole-operátorok alkalmazását, a rekordok megjelenítését, a szavak előfordulási gyakoriságának követését az invertált fájlban stb.
4. A szoftver biztosítja az invertált fájl karbantartását rekordok bevitelekor, módosításakor vagy törlésekor.

Helyi adatbázisok építésére széles körben használt szöveges információkereső programcsomagok az ASSASSIN, a CAIRS, a POLYDOC, a STATUS, illetve ezek mikroszámítógépes változatai.

Más adatbázis-szerkezetre épülő információkezelő szoftvereket ismertet *Kazlauskas*. A fájl- vagy adatkezelő programot egyetlen, máshoz nem kapcsolódó fájl létrehozására és kezelésére lehet felhasználni. Ezeknél összetettebb szerkezetűek az adatbázis-kezelő rendszerek (ABKR; angolul Data Base Managament System; DBMS), de összetettségük meglehetősen különböző lehet. Maga az adatbázis-kezelő rendszer fogalma különböző szakemberek számára mást és mást jelenthet. Az adatbázis-kezelő rendszereket eredetileg a hatvanas években fejlesztették ki a nagyszámítógépekkel foglalkozók, amikor egyetlen – integrált – adatbázist akartak építeni (mondjuk a gyár termékeinek nyilvántartására), amelyet a különböző részlegek (például a kereskedelmi, a marketing, a termelési, a kutatási és fejlesztési) saját igényeiknek megfelelően tudtak felhasználni. Az adatbázis-kezelő programok ezért mind nagyszámítógépeken, mind miniszámítógépes környezetben (például az ADABAS, a FOCUS,

az IDMS vagy a TOTAL) bonyolultak, nem pedig olyan szoftverek, amelyet végső felhasználó vagy újsütetű kereső közvetlenül információkeresésre tudna felhasználni. Ezért gyakran csak keretnek, nyers alaprendszernek tekintik azok, akik alkalmazási programokat írnak. A hierarchikus és a hálós adatbázis-kezelő rendszerekhez képest a legfejlettebb a relációs ABKR, amelyben az információt táblázatok képviselik; az oszlopok megfelelnek a hagyományos rekordok mezőinek, a sorok pedig maguk a rekordok.

Az adatbázis-kezelő rendszerekkel részletesebben kötetünknek az automatikus indexeléssel és osztályozással tárgyaló részében, az „Automatikus információkereső rendszerek és az adatbázis-kezelő rendszerek" című fejezetében foglalkozunk.

Egyre több adatbázis-kezelő rendszer készül mikroszámítógépekre is. Léteznek olyan programrendszerek is, mint például az INFOText, melyben a szabványos adatbázis-kezelő rendszert információkereső résszel egészítik ki. A szöveges információkereső rendszerek (szövegben kereső rendszerek) és dokumentációs célú adatbázis-kezelők világa meglehetősen szerteágazó, és a kereskedelmi forgalomban kapható rendszerekről eligazodni csak nagyon részletes szakértelem alapján lehet.

Keresés helyi adatbázisokban

A helyi adatbázisok kereső eszközei gyakran hasonlóak a kereskedelmi jellegű távoli on-line keresőszolgáltatások által használtakéhoz. A szöveges információkereső programcsomagok között vannak olyanok, amelyeket közvetlen kapcsolat fűz a távoli szolgáltatások által használt szoftverhez. A MicroQuestel például, amelyet a Télésystémes–QUESTEL állított elő, hasonló keresőutasításokat tartalmaz, mint amelyek a távoli on-line szolgáltatásokban használatosak, és tartalmaz a dokumentumok nagyszámítógépbe való áttöltéséhez szükséges utasításokat is, amelyet a Questel keresőszolgáltatás alkalmazhat. Ehhez hasonló a BRS/Search, amelynek az utasításai a BRS keresőszolgálatéval csengenek össze. A távoli on-line keresőszolgáltatások közül több betört a CD-ROM-piacra. A Dialog például különböző CD-ROM-adatbázisokat forgalmaz (ezek közül látható néhány a következő táblázaton), amelyek a távoli szolgáltatáséhoz hasonló (ezért a használók számára ismerős) keresési lehetőségeket biztosítanak.

A CD-ROM-lemezen kívül rendelkezésre áll a Dialog On Disc Discovery Preview című, a keresést oktató segédlet. A szöveges információkereső szoftverek előállítói közül némelyek olyan programcsomagokat alakítottak ki, amelyek egyaránt alkalmasak a CD-ROM-adatbázisok, és a helyi, saját készítésű adatbázisok keresésére. A Harwell Computer Power forgalmazza például a

STATUS-t és a Micro-STATUS-t, és a holland Samson és Philips cégekkel együttműködve gyártanak olyan CD-ROM-adatbázisokat, amelyek a STATUS utasításaival használhatók. A különböző szöveges információkereső rendszerek keresési lehetőségeit Kimberley (1989) adattára ismerteti.

Az adatbázis neve	Szakterület
AGRIBUSINESS USA	Mezőgazdasági ipar
CANADIAN BUSINESS AND CURRENT AFFAIRS	Kanadai országos és tartományi cég-, termék, ipari és pénzügyi információ
ERIC	Oktatás és nevelés
MEDLINE	Orvosbiológiai irodalom
MEDLINE CLINICAL	Klinikai orvostudomány
CONNECTION NTIS	Az Egyesült Államok kormánya által finanszírozott kutatás és fejlesztés
STANDARD AND POOR'S CORPORATIONS	Köz- és magáncégek adatai

A parancsnyelv

A szövegben végzett keresésre alkalmas programcsomagok többsége parancsvezérlésű, bár néha menü képernyők segítségével jelenítik meg a használható utasításokat. A Micro-CAIRS például menü képernyők segítségével biztosít közvetlen hozzáférést az adatbázisok készítéséhez és módosításához, a rekordok beviteléhez és módosításához, a kereséshez, az output formátumok megtervezéséhez, a mutatók generálásához stb. használható funkciókhoz. A TINman ugyanakkor nem ad felismerhető keresési utasításokat: a használók az információ rendezett listájában böngészhetnek, majd egyetlen billentyű lenyomásával juthatnak el az érdekes tételhez.

A Silver Platter CD-ROM-adatbázisokon a következő utasítások használhatók:

HELP	segítség a rendszer funkcióival kapcsolatban.
FIND	keresőkifjezések (szavak vagy összetett kifejezések) bevitele.
GUIDE	tájékoztatás a használt adatbázisról.
SHOW	a talált rekordok vagy egy részük megjelenítése.
INDEX	a keresőkifjezések invertált fájljának megtekintése.
PRINT	a talált rekordok kinyomtatása.
RESTART	adott keresési ülés befejezése.
XCHANGE	váltás másik Silver Platter CD-ROM-lemezre.
PREVIOUS	az előző rekord megjelenítése.
NEXT	a következő rekord megjelenítése.

A Boole-operátorok használata

A kereső programcsomagok többségében mód van a Boole-operátorok segítségével végzett keresésre. A CAIRS, BRS/Search, STATUS és INMAGIC szoftverek esetében a keresőkérdések zárójelezhetők, így a kérdés egyetlen sorban összefoglalható, például:

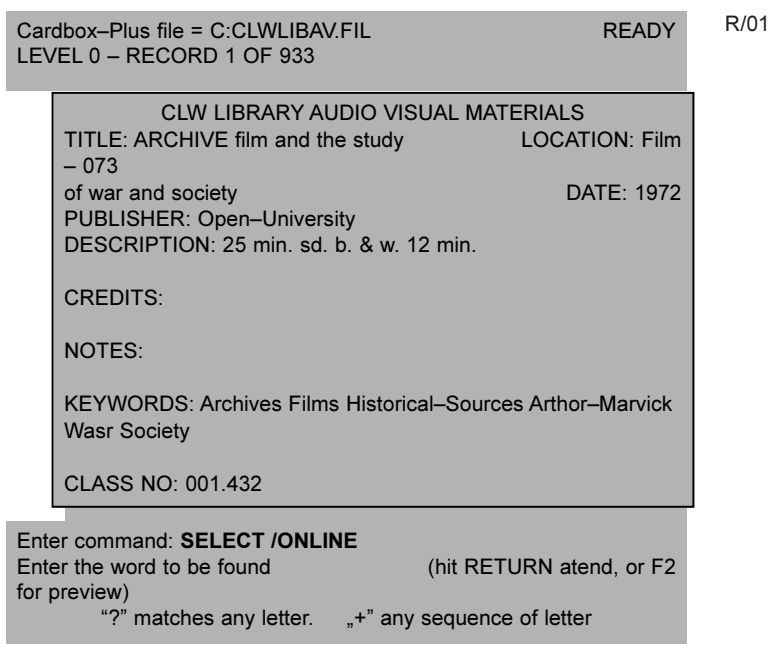
(COMPUTERS VAGY MICROCOMPUTERS) ÉS (SOFTWARE VAGY PACKAGE) ÉS RETRIEVAL

Más esetekben a Boole-logikai keresés körülményesebben hajtható végre.

Az 1. keresési példa a College of Librarianship Wales (CLW) könyvtárának audiovizuális dokumentumai között, a CARDBOX–PLUS szoftvercsomaggal végzett keresést mutatja be. Az adatbázis 933 rekordja közül az első egy filmé *(Archive film and the study of war and society)*. A használónak kell bevinnie az utasítást.

SELECT/ONLINE

Ennek az eredményeként 37 rekordot választott ki a rendszer, amelyek tartalmazzák az „online" szót. Ezek közül egy jelenik meg a képernyőn, az OCLC-nek a stratégiai tervezési stratégiájáról szóló videofilm.



R/01

Cardbox–Plus file = C:CLWLIBAV.FIL READY
LEVEL 1 – RECORD 1 OF 37

CLW LIBRARY AUDIO VISUAL MATERIALS
TITLE: OCLS´S straategic planing LOCATION: VHS/C – 623
challenges. DATE: 1985
PLACE: Dublin, Ohio
PUBLISHER: Online–Computer–Library–Center Inc.
DESCRIPTION: 1 videocassette (CHS) (88 min.); sd, col. NTSC standard

CREDITS: By Rowland–Brown

NOTES: An OCLC Video Communications Program N. B. NTSC standard: must be played on multistandard player in academic block.

KEYWORDS: OCLC Cataloguing Online Housekeeping Automation USA

CLASS NO: 021.650973

Enter command: **INCLUDE /DIALOG**
Enter the word to be found (hit REWTURN atend, or F2 for preview)
"?" matches any letter. „+" any sequence of letter

R/01

Cardbox–Plus file = C:CLWLIBAV.FIL READY
LEVEL 0 – RECORD 2 OF 40

CLW LIBRARY AUDIO VISUAL MATERIALS
TITLE: The DIALOG of information LOCATION: VHS/C – 362
retrieval DATE: 1981
PLACE: Paolo Alto
PUBLISHER: Dialog–Marketing–Department
DESCRIPTION: 1 videocassette (VHS) (15 min.); sd., col.

CREDITS:

NOTES:

KEYWORDS: DIALINDEX DIALORDER Databases USA Online Dialog Computers

CLASS NO: 024.04

Enter command: **INCLUDE KE/DATABASE+**
Enter the word to be found (hit RETURN atend, or F2 for preview)
"?" matches any letter. „+" any sequence of letter

```
Cardbox–Plus file = C:CLWLIBAV.FIL                    READY
LEVEL 0 – RECORD 1 OF 49

          BCLW LIBRARY AUDIO VISUAL MATERIALS
TITLE: DATABASES                                  LOCATION:
Tape/S – 073
                                                  DATE: 1985
PLACE: London
PUBLISHER: Prismatron
DESCRIPTION: 59 slides: col. + 1sound cassette (22 min.);
1  7/=8  ips, mono
CREDITS:

NOTES: Computer avareness series

KEYWORDS: Databases–Management–Systems Cataloguing
Books StructureDMS  DBMS

CLASS NO: 001.6442

Enter command: QUIT
      (now hit RETURN)
```

1. keresési példa. *CARDBOX–PLUS* keresés

A következő utasítás:

INCLUDE/DIALOG

a már meglévő találati halmazhoz további rekordokat csatol, amelyekben szerepel a DIALOG kifejezés, így újabb 40 rekordból álló halmaz képződik. A halmaz második rekordja a *The DIALOG of Information Retrieval* jelenít-hető meg a jobbra mutató nyíl segítségével, amely az első rekordról a más-odikra való átlépést biztosítja. A következő utasítás:

INCLUDE KE/DATABASE+

azoknak a rekordoknak a felvételét idézi elő, amelyek a kulcsszó mezőben a „database" szótövet tartalmazzák. Az eredmény 49 rekord, amelyek közül az első egy hangosított diaprogram az adatbázisokról. Bár nem használtunk Boole-operátorokat, ez a keresés megfelel a SELECT (ONLINE OR DIALOG OR KE/DATABASE+) „mondattal" reprezentált keresésnek. A SELECT utasítás a CARDBOX–PLUS-ban úgy is működik, mint az AND (ÉS), a NOT (NEM) operátor pedig az EXCLUDE utasítással helyettesíthető. A QUIT utasítás a CARDBOX–PLUS kereső programból való kilépésre szolgál.

Keresés meghatározott mezőkben

Számos programcsomag biztosítja a lehetőséget arra, hogy a keresést meghatározott mezőkre szűkítsék (mint az 1. keresési példában a kulcsszó – KE – mezőre).

Keresés a keresőszavak közelsége alapján

Némely programrendszerben, így az ASSASSIN, a BASIS, a VAIRS, az INMAGIC, a MINISIS és a POLYDOC szomszédos vagy egymáshoz adott számú szótávolságú kifejezéseket kereshetünk. Az e célra használható szótávolsági operátorok az INQUIRE rendszerben például a következők:

ADJ	szomszédos kifejezések keresése
SEN –	ugyanabban a mondatban előforduló kifejezések keresése,
WITHIN±N words	meghatározott távolságra lévő szavak keresése.

Ez a lehetőség különösen hasznos teljes szöveges adatbázisokban való kereséskor.

Szócsonkolás és alakváltozatok bevonása a keresésbe

Sok programrendszerben lehetséges a jobboldali csonkolás. A baloldali csonkolás nem olyan elterjedt, de sokszor szükség van rá kémiai nevek esetében, például

SEARCH ?SULPH?

ha a METABISULPHATE-ra, a SULPHUR-ra, a SULPHUROUS-ra stb. keresünk. A SULPHUR-hoz hasonló kifejezések helyesírási változatai is gondot okozhatnak. Ha a helyi adatbázisba különböző forrásokból visznek be vagy töltenek le rekordokat, szükség van olyan keresési eszközökre, amelyek a különböző szóalakok szerint végzett keresést biztosítják. Néhány programrendszer *karaktercsonkolással* vagy *helyettesítő karakterrel* oldja ezt meg, például

SEARCH SUL*UR

ami mind a SULFUR, mind a SULPHUR szerinti keresésre alkalmas. Az egyik csomag, a Southdata Ltd Superfile-ja fonetikus keresési eljárással szűri ki a

helyesírási változatokat. Így a THOMSON-ra való keresés kihozza a Thompson, Tommson, Tomson, Tomasson stb. írásmódú neveket is. A Superfile-t sok hollandiai könyvtár használja a holland Jóléti, Egészségügyi és Kulturális Minisztérium által készített WORM lemezeken található nagy adatbázisokban végzett keresésre.

Keresés numerikus intervallumok alapján

A szöveges információkereső rendszerekben is szükség lehet numerikus keresésre (például időpontok vagy árak szerint). Az intervallum alapján végzett keresés – a GE (nagyobb mint), EQ (egyenlő), LT (kisebb mint) operátorok segítségével – sok információkereső rendszerben lehetséges. A FIND PRICE LT 60 utasítás például az olyan rekordok keresésére alkalmas, amelyeknek az ár mezőjében 60-nál kisebb szám szerepel. Egyes rendszerekben (a Micro–CAIRS és az InMAGIC–Micro) az is lehetséges, hogy a számszerű értékekkel egyszerű számításokat végezzenek.

A tezaurusz bevonása a keresésbe

Egyes információkereső rendszerekben tezauruszok készítése és karbantartása is lehetséges; a tezauruszok fölhasználhatók a kereséshez. Egyes esetekben a kurrens keresőkifjezésekből áll (időnként go [„gyerünk”] listaként is emlegetik), más rendszerekben teljesebb tezaurusz készülhet, amely a keresőkifejezések közötti relációkat (fölérendelt, alárendelt, rokonsági, szinonima) is feltünteti. A tezaurusz használata javíthatja a keresés eredményességét. A tezaurusz a beviteli stádiumban is alkalmazható, amikor a tételekhez deszkriptorokat vagy kulcsszavakat rendelnek. *Pasqual* (1986) a STATUS használatáról írva a Western Australia Department of Agriculture-ban bemutatja, hogyan használható a STATUS tezaurusza arra, hogy a búza (wheat) betegségéhez az adatbázisban előforduló összes specifikusabb (alárendelt) kifejezést előhívja:

Q Wheat disease?

Ha nincs tezaurusz, a keresőtől függ, hogy sikerül-e a megfelelő fölé- és alárendelt, illetve rokon vagy szinonim kifejezéseket megtalálnia.

A mutató megjelenítése és böngészése

A szöveges információkereső rendszerekben többnyire megjeleníthetők a betűrendes mutató vagy invertált fájl bizonyos részei, amelyben láthatók a

keresőkifejezések, és a találatok száma is, amelyekben a kifejezéseik ismérvként előfordulnak. A 2. keresési példán látható az ‚ERYTHROMYCIN' kifejezés környezete a Consumer Drug Information on Disk (CDID) mutatójában. Az American Society of Hospital Pharmacists (ASHP) által készített CDID olyan adatbázis, amely hajlékony lemezen elérhető program segítségével IBM vagy azzal kompatibilis személyi számítógépeken helyi keresésekre használható.

Az adatbázis célja, hogy az egészségügyi dolgozók és a nagyközönség kereshesse a gyakran felírt gyógyszerekre vonatkozó információkat. Az ASHP nemrégiben a CDID-t a MedTeach-csel váltotta fel, amely menürendszerű szoftverjével a *Medication Teaching Manua*l-on alapul. A 2. keresési példa a CDID keresésével lehívott rekord részletét mutatja.

A helyben kialakított adatbázisokat gyakran használják különböző nyomtatott jegyzékek, mutatók vagy cédulakészletek előállítására és on-line keresésre. Tulajdonképpen a számítógéphasználat legfőbb ösztönzője a nyomtatott termékek előállításának támogatása volt, és az on-line keresés az adatbázisokban csak hab volt a torán. A használók meghatározhatják mely mezőket akarják kinyomtatni és milyen formában. Bizonyos csomagokban a tárolt szöveget és a kulcsszavakat is be lehet építeni. Datta példaként a CAIRS-zel előállított a nyomtatott gyarapodási jegyzéket (kövérrel és aláhúzva), cédulákat és jegyzékeket hozza fel. *Green* ismerteti egy általános R.DBMS, a Paradox és a WordStar szövegszerkesztő-csomag használatát a szelektív információ szétsugárzására, újdonságértesítő jegyzék összeállítására és a retrospektív keresésre használt házi adatbázis építésére (egy műszaki kutatóközpont könyvtárában). Egyes csomagok, például az ASSASSIN-PC lehetővé teszik a kulcsszavas mutatók nyomtatását is.

A szöveges információkereső rendszerek kezelői a szakkönyvtárakban vagy információs részlegekben már hozzászokhattak ahhoz, hogy felhívják használóik figyelmét a számukra releváns, újonnan kiadott vagy beszerzett dokumentumokra. Ezt a funkciót szelektív információ-szétsugárzás vagy SDI néven ismerik. Az SDI szolgáltatáshoz szükséges eszközök a távoli keresőszolgáltatásoknál, de néha a helyi szöveges információkereső rendszerekben is elérhetők. Az alapvető követelmény ama keresési profilok tárolásának lehetősége, amelyeket – a megfelelő idő intervallum kiválasztásával – össze lehet vetni az adatbázisba újonnan bevitt vagy átvett rekordok állományával.

A helyi adatbázisok keresésével kapcsolatban az egyik előre lépés, hogy egyre nagyobb mértékben használják a kereső- és a számítógépes rendszer közötti interakciót támogató eszközöket.

keresőszó: ERYTHROMICIN

Erypar
EryPed
Erythntyl Tetramnitrate
Erythrocin Stearate
Erythromycin
Erythromycin Base Filmtab
Erythromycin Estolate
Erythromycin Ethylsoccinate
Erythromycin Stearate
Eserine Sulfate

// 1./ 5 oldal ///////

and = Moves | F1 = Maim Main menu | F9 = End Grouip | ENTER = Choose | F2 = Summary | F) = Item
Please select a term
Alt–C = Colors	F3 = Page Back	F10 = Help
Alt–A = First	F4 = Page Ahead	O = Group items
Alt–Z = Last	F7 = Group Items	Type a Response

CONSUMER DRUG INFORMATION DISK
– page 1

Ł 1 of 2 ....

MONOGRAPH TITLE: Erythromycims (eh rith roe mye`sins)

GENERIC NAME: Erythromycin Ethylsuccimnate/ Erythromycin Stearate/ Erythromycin/ Erythromycin Estolate

DRUG CLASSIFICATION: Erythromycins

MEDICAL CODITION: Infections–General

ROUTES AND DOSAGES: Oral Capsules/ Oral Tablets/ Oral Liquid, Slution, Syrup, etc.

REGISTRY NUMBER: 41342–53–4/643–22–1 / 114–07–8 / 33521–62–8

PRODUCT INFORMATION. . . . .

E–Mycin/ ERYC/ Ery–Tab/ Erythromycin Base Filmtab/ Illotycin/ PCE Robimycin/ RP–Mycin/ Ilosone/ EES/ E–Mycin E/ EryPed/ Pediamycin/= Wyamycin/ Bristamycin/ Eramycijn/= Erypar/ Erythrocin Stearate/ Ethryl/ Pfizer–E/ BK–Erythromycin/ Wyamycin S/ Pediazole

USES . . . . .

The erythromycins are available in a number of chemical foms, including erythromicin (base), estolate, ethylsuccimate and stearate. All of these fomrs of erythromycin share the same uses, side effects and precautions except erythromycin estolate/, which produces liver problems more frequently than other erythromycins (see Undesired Effects and Precautions).

The erythromiycons are systemic antibiotics udes to treat a wide variety of enfections, including throat, ear skin infections, pneumonia and diphteria. They are considered good drugs to treat or prevent „trep" infections in people wwho may be sensitive or ellergic topenicillins.

The erythromycins are the preferred drugs toeliminate diphtehria–causing bacteria from people who show no signs of the diesease but are infecting others. There is some evidence that erythromycins are effective against Legionnaires` disease.

2. keresési példa *CONSUMER DRUG INFORMATION* adatbázis

Sok mikroszámítógép, közöttük az IBM PC-k és az ezekkel kompatibilis gépek külön funkcióbillentyűkkel rendelkeznek, amelyeket meghatározott utasítások vagy utasításcsoportok tárolására előre beprogramoztak. Ezek használatával a kereső mentesül az utasítások beírásától. A 3. keresési példa a Silver-Platter LISA CD–ROM adatbázisának nyitó képernyőjét mutatja. Ezen röviden ismertetik az egyes funkcióbillentyűk használatát. A rendszer FIND utasítás alapján kezdi a keresést, így az első keresőkifjezés, a HOSPITAL–PATIENT–LIBRARIES (a LISA deszkriptor jegyzékéből származó kifejezés) az input. A következő kifejezés csonkolt: CHILD*, majd a két kifejezést az AND operátorral kapcsolják össze. A 23 talált tétel közül az első az F4 (SHOW) funkció billentyű megnyomásával jeleníthető meg. A kifejezések pedig, amelyek alapján megtalálták a rekordot, kiemelten jelenek meg a képernyőn.

Másik lehetőség, amely szintén a keresést támogatja, a helyőrmozgató billentyűk (←, ↑, →, ↓) használata. Ebben az esetben is a lehető legkevesebb billentyű megnyomása a cél.

```
Silver Platter v 1.4                                    LISA (1/69 – 9/88)

TITLE SCREEN                                                          1
OF 1

                              The LISA Database
                          1969. September 1988

The LISA Database contains summaries of the world`s literature in
librarianship, information science and related disciplines.

To search LISA:  type your searchrequest, the press RETURN

To learn abaout the system:                          press F1 (HELP)

To learn about the LISA database:                    press F3 (GUIDE)

FIND:
Type a search request, then press RETURN; or press F1 for HELP

Silver Platter v 1.4                                    LISA (1/69 – 9/88)

No.          Request                                            Records

Ł1:          HOSPITAL–PATIENT–LIBRARIES                            124
Ł2:          CHILD*                                               4516
Ł3:          Ł1 and –Ł2                                             23
```

SHOW fields: ALL Records: ALL
Press RETURN tostart with the first record; or press F1 for HELP.

Silver Platter v 1.4 LISA (1/69 – 9/88)

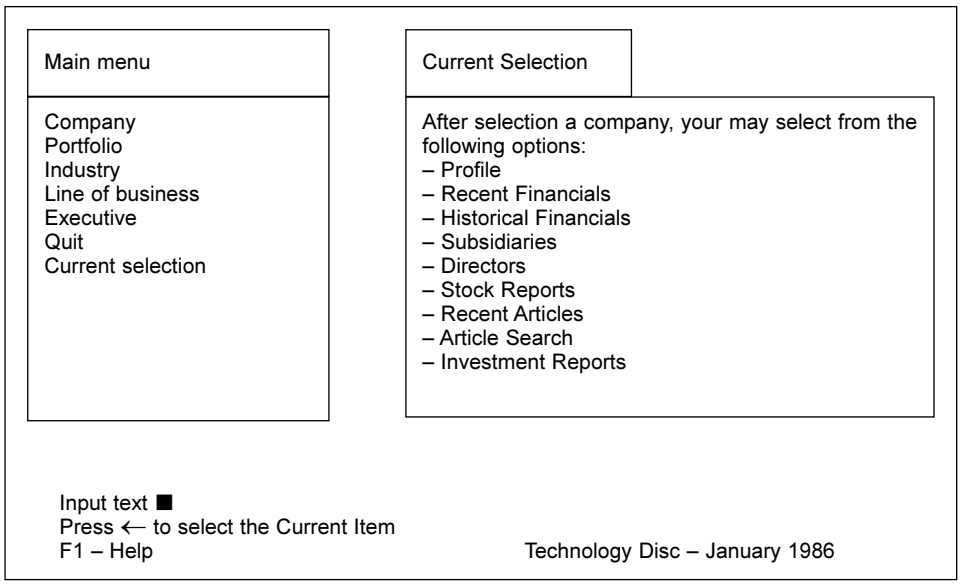
TI: Hospital outreach programme at the Montreal Children`s Library
AU: Walsh, –Molly; Montreal–(Quebec–Province)–Children`s Léibrary
SO: Bulleti–ABQ/QLA–Bulletin, 30 (1) Jan–Apr 88, 21–22
PY: 1988
LA: English
AB: Describes the hospital outreach programme provided by the Montreal **Children**`s Library wich serves several departments of the Montreal **Children**`s Hospital, Shirner`s Hospital,l and schedules visits for Papillomn day care groups from athe Quebec Society for Disabled **Children**.
FH: Hospital patient libraries. **Children**`s libraries.
DE: Canada–; Public–libraries; **Children**–; Welfare–services; Hospital–libraries; Handicapped–; Institutional–libraries; Isolated–; **Hospital–patient–libraries**
CC: HuEfo& Hu
DA: 1988
AN: 88–3854
SHOW fields: ALL Records: ALL
Press CTRL F2 taoselect terms from record for searching, PgDn for more;
F10 Next Record; F2–Find– Fl–Help; ESC–Command Menu

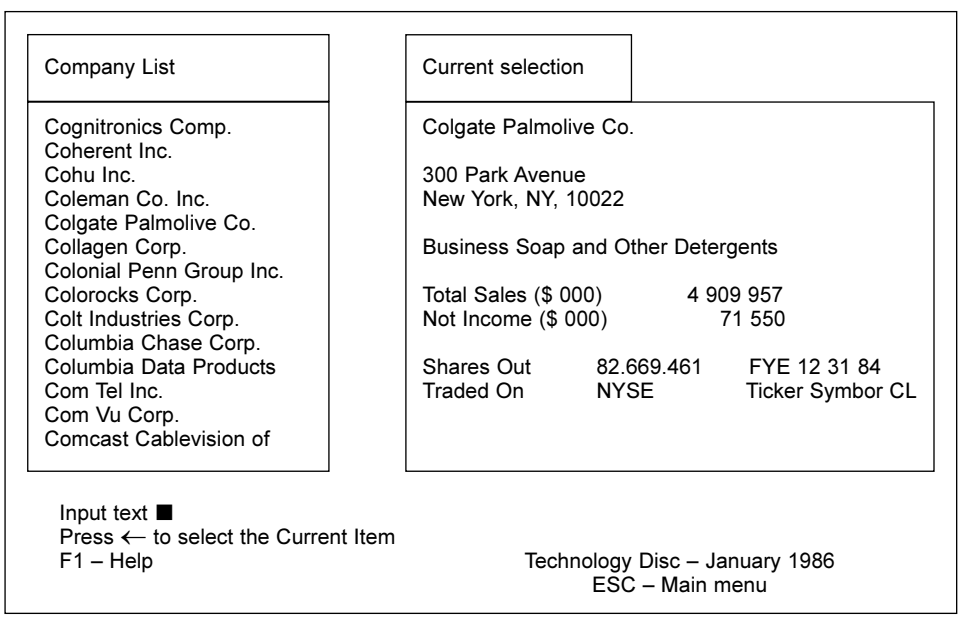
3. keresési példa. Keresés a LISA referáló szolgáltatás CD–ROM adatbázisában

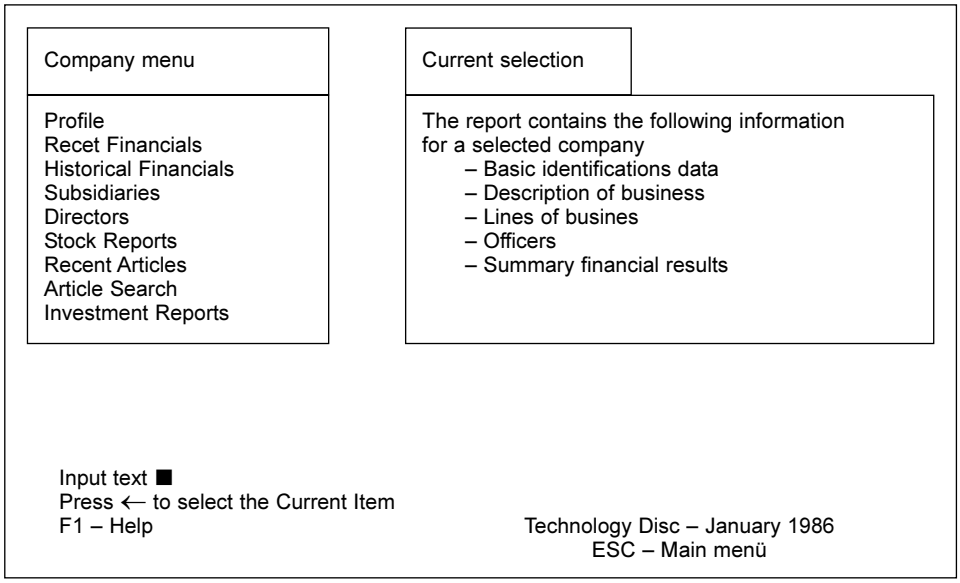
A Datext Corporate Information CD–ROM-adatbázis a Dartmouth Collegeban (New Hampshire, USA) egyesíti a több mint 10 000 amerikai vállalatra vonatkozó, különböző adatbázisokból (például a Predicasts PROMT, DISCLOSURE II, INVESTEXT, ABI/INFORM, MEDIA GENERAL'S, MARKET FILE és WHO'S WHO IN FINANCE AND INDUSTRY) származó bibliográfiai, szöveges és numerikus adatokat. A 4. keresési példán a Colgate Palmolive Company keresése során megjelent képernyőkből láthatunk néhányat. A cég nevét kell beírni, és ennek nyomán megjelenik a vállalatok nevét tartalmazó betűrendes mutatójából a Colgate környezete. A nyilakkal lehet azután a „profil"-nak megfelelő helyre menni, és a kívánt rekordot megjeleníteni.

Másik, a helyőr (kurzor) mozgatására alkalmas eszköz az *egér.* Ez a neve annak a kicsiny doboznak, amely hosszú vezeték segítségével kapcsolódik a munkaállomáshoz, és megfelelő felszínen szabadon mozgatható; a helyőr ennek megfelelően mozog a képernyőn. Ennek megfelelően használható az egér arra, hogy a képernyőn valamit kiválasszunk az egéren lévő gomb megnyomásával. Az egér szöveget vagy adatokat tartalmazó ablakok *megnyitására* és azok manipulálására is alkalmas. Ez a keresési interfész egészen másként működik,

mint a megszokott soros keresés, amely a távoli keresőszolgáltatásokat jellemzi. További fejlemény a nyelv helyett inkább a képekre építő *ikonok* használata. A fájlokat egyszerűen a fájlnév kiválasztásával, az egér billentyűjével adott utasítással („ráklikkeléssel"), a fáljnév félrehúzásával, és a megfelelő ikonon az egérbillentyű újabb lenyomással törölhető. Ez a keresőkörnyezet a WIMP (Windows, Icons, Mice and Pointers – Ablakok, Ikonok, Egér, Mutatók).







4. keresési példa. Keresés a DATEX testületinformációs-szolgáltatás CD–ROM-adatbázisában

Az *ablakok* használata (az egér nélkül) magától értetődő, ha a Bowker BOOKS IN PRINT PLUS CD-ROM-adatbázisában keresünk (5. keresési példa). A műveletek a képernyő tetején láthatók, amelyek közül jelenleg a Search folyik. A használó a ←, → helyőrmozgató billentyűk segítségével térhet át más műveletre. A megosztott képernyő mutatja, hogy milyen keresőkódok használhatók, és itt van felület a keresőkép bevitelére is. A kórházban fekvő gyerekekkel kapcsolatos könyvek kulcsszavas keresése az alábbiakban látható:

KW = gyerek$
KW = kórház$
CS = 1 és CS = 2

Az F10 billentyű való a rövid megjelenítési formátum előhívására. Ezt megnyomva, megnyílik a rövidített bibliográfiai leírásokat tartalmazó ablak. A ↓ billentyűvel mozgathatjuk a kurzort a nyolcadik tételre, majd az F10 ismételt megnyomásával megjelenik Charlotte Krall és Judith Jim könyve, *Fat Dog's First Visit: A Child's View of the Hospital.*

A végső felhasználók, közelebbről a medikusok egyre gyakrabban végzik maguk a keresést a CD–ROM-on, és lassan megjelennek az első elemzések a keresési technikájukról.

Search Browse Format Action Options Databases Books in Print Plus

au	=	Author
bn	=	ISBN
kw	=	Keyword
lc	=	LCCN
pu	=	Publisher
su	=	Subject
ch	=	Children´s Subject
tc	=	Title Code
ti	=	Title
se	=	Series Title
tk	=	4.4 Author Title
cs	=	Combine Set
ac	=	AudieceGrades
il	=	Illustration
la	=	Language
pr	=	Price
py	=	Publication Year

1. kw	=	csild$	14805
2. kw	=	hospital$	1443
3. kw	=	1 and cs = 2	100

F1 → Help ESC → Menu Bar
Enter new SearchStatement & press ENTER F10 → Brief Citation Serch Completed

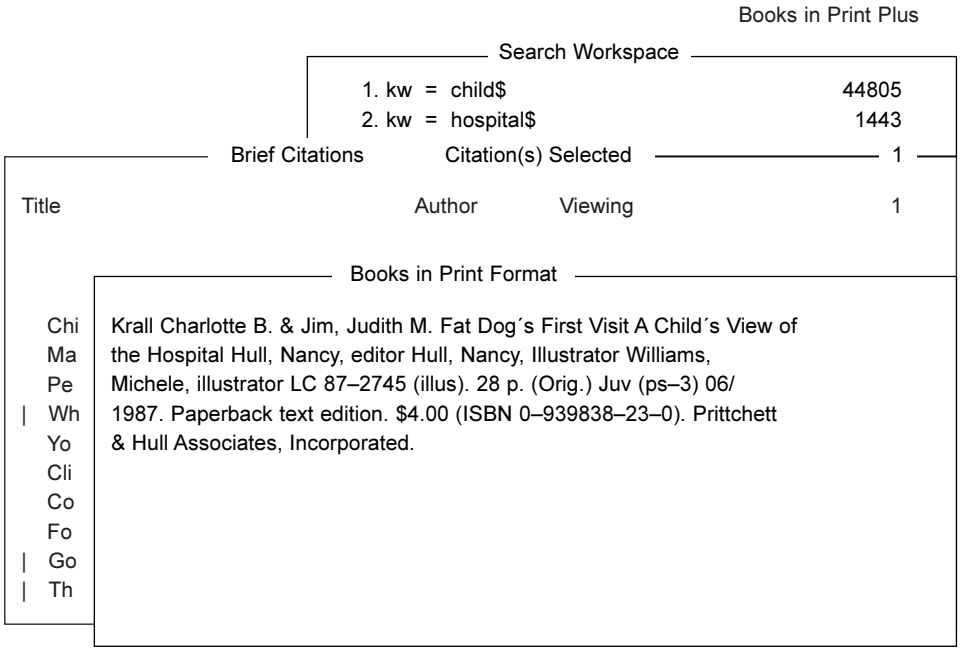
Books in Print Plus

Search Workspace

1. kw	=	child$	44805
2. kw	=	hospital$	1443

Brief Citations

Title	Author	Price	Date	ISBN
Children´s Hospitals in t	Rothman, David	$40.00	1988	082241765834
Manual of Pediatric Thera	Children´s Hosp	$24.50	08/1988	0324076886
Pediatric Hospitalization	Knafi, Kathleen		1988	0673397327
| What Teenagers Want to Kn	Boston Children	$16.95	05/1988	03106250635
Your Hospital Stay lt 1	Rosenstock, Jud	$4.95	11/1988	09622127024
Clinical Pastoral Care fo	Hesch, John B.	$9.95	05/1987	08091287123
Coping with a Hospital St	Carter, Sharon	$12.95	10/1987	0823906825
Fast Dog´s First Visit: A	Krall, Charlott	$5.00	06/1987	0939838230
For Your Hospital Visit	Gregg-Schroeder		10/1987	0835815700
| Gong to the Hospital	Civardi, Anne	$.95	1987	0746000731
| The New Child Healt Ency	Boston Children	$19.35	11/1987	0352295979



5. keresési példa. Keresés a Bowker cég *BOOKS IN PRINT PLUS*TM CD–ROM-adatbázisában

Az Erskine Medical Library (Edinburgh University) értékelte a végső felhasználók kereséseit a Medline on CD–ROM-on. Eszerint:

1. A keresők inkább megismétlik a kifejezéseket, minthogy az OR (VAGY) operátort használnák, például

   CEREBRAL PALSY AND CHILD ABUSE
   CEREBRAL PALSY AND SOCIAL WORK
   CEREBRAL PALSY AND FOSTER HOME
   CEREBRAL PALSZ AND CHILD PRE–SCHOOL

   ahelyett, hogy így írnák:

   (CHILD ABUSE OR SOCIAL WORK OR FOSTER HOME OR CHILD PRE–SCHOOL) AND CEREBRAL PALSY

   amivel időt takaríthatnának meg, és megkímélnék magukat a rekordok többszörözésétől.

2. A keresők releváns szinonimákat hagynak ki a keresésből. Pusztán azt adják meg, hogy

TUMOR

ahelyett, hogy

TUMOUR OR TUMOR OR NEOPLASM.

3. A keresők helytelenül használják a csonkolást:

ETHICS? AND HANDICAPPED

ahelyett, hogy

ETHIC? AND HANDICAPPED.

Az ilyen vizsgálatokból az derül ki, hogy a végső felhasználók általában elégedettek a keresésük eredményével, de nem árt őket figyelmeztetni: ha naprakészebb és átfogóbb keresést akarnak végezni, akkor érdemes a közvetítő segítségét igénybe venniük.

STEPHEN P. HARTER (1921)

Stephen P. Harter a bloomingtoni Indiana Egyetem könyvtáros és információtudományi iskolájának professzora. A hetvenes években az automatikus indexeléssel foglalkozott és kidolgozta a valószínűségi indexelés egyik modelljét. Később figyelme az on-line információkeresés és a keresési stratégiák felé fordult. Néhány módszertani kézikönyv után 1986-ban adta ki az on-line információkeresés egyik legsikeresebb tankönyvét, mely 1990-ben a harmadik kiadását érte meg, ami azért figyelemre méltó, mert ezen a területen a változás és vele a könyvek avulása még rendkívül gyors.

On-line információkeresés. Fogalmak, elvek és technikák[12]

1.3 Adatbázisok[13]

Több útmutató (afféle kalauz) is készült azokról az adatbázisokról és keresőszolgáltatásokról, amelyekben naprakész információhoz lehet jutni a nagyközönség számára elérhető és on-line keresésre alkalmas adatbázisok tartalmáról és jellemzőiről. Ilyen például a *Directory of On-line Information Resources,* a *Datapro Directory of On-line Services, Computer-Readable Data Bases: A*

12 On-line information retrieval : Concepts, principles, and techniques / Stephen P. Harter. – San Diego [etc.] : Academic Press, 1990. 320 p.

13 1.3 Databases. In: On-line information retrieval... p. 5–10.

Directory and Data Sourcebook és a *Directory of On-line Databases.* Figyelemre méltó az adatbázisok számának emelkedése. *Martha Williams* számol be róla, hogy míg 1965-ben húsznál kevesebb információkeresésre alkalmas adatbázis állt a nyilvánosság rendelkezésére, 1975-ben ez a szám már háromszáz fölé emelkedett. Kilenc évvel később a *Directory of On-line Databases* nem kevesebb mint 2453 adatbázist tartalmazott, amelyeket 1189 adatbázis-szolgáltató állított elő, és amelyeket 362 on-line keresőszolgáltatáson keresztül lehetett elérni. Az útmutatóba (*Directory*-ba) való bekerülés kritériuma az on-line elérhetőség, a *nyilvánosság* és az volt, hogy on-line keresőszolgáltatáshoz vagy hálózathoz távközlési vonalon keresztül kapcsolódhasson.

Az adatbázisokat a Cuadra Associates által kialakított taxonómia alapján csoportosítjuk. Ezeket foglalja össze az 1. táblázat. A *referensz-adatbázisok,* beleértve a *bibliográfiai* és a *forrástájékoztató* adatbázisokat, meglévő adat-, információ- vagy ismeretforrások reprezentációit vagy szurrogátumait (dokumentumleírásokat stb.) tartalmazzák. A keresőt másik, teljesebb információt tartalmazó forráshoz kalauzolják. A bibliográfiai adatbázisokban a rekordok az emberiség grafikus vagy nyomtatott emlékeinek – folyóiratcikkeknek, kutatási jelentéseknek, szabadalmaknak, könyveknek stb. – intellektuális tartalmára és fizikai jellemzőire vonatkozó keresési belépési pontokat, kulcsokat (ismérveket) tartalmaznak. A forrástájékoztató adatbázisok a nem nyomtatott forrásokhoz – személyekhez, szervezetekhez, kutatási jelentésekhez, nem nyomtatott információhordozókhoz stb. – vezetnek el.

REFERENSZ-ADATBÁZISOK – elsődleges források helyettesítői és azokra való hivatkozások. Tartalmazhat referátumokat és összefoglalókat. Alosztályai:

BIBLIOGRÁFIAI ADATBÁZISOK – az elsődleges források: kiadott vagy nem kiadott (nyomtatott) dokumentumok. Tartalmazhat referátumokat. Példák:

BIOSIS PREVIEWS. Előállító: BioSciences Information Service. Több mint hárommillió hivatkozás az élettudományban kiadott művekre. 1969–

MEDLINE. Előállító: U.S. National Library of Medicine. Több mint négymillió hivatkozást tartalmaz tágan értelmezett orvosbiológiai művekre. 1964–

SOCIAL SCISEARCH. Előállító: Institute for Scientific Informations. Több mint egymillió hivatkozást tartalmaz 1500 társadalomtudományi folyóirat cikkeire, és a természet-, fizikai és orvosbiológiai tudományok köréből válogatott cikkekre. 1972–

ERIC. Előállító: National Institute of Education. Több mint 500 000 hivatkozást tartalmaz az oktatási és nevelés folyóirataira és efemer irodalmára. 1966–

CA SEARCH. Előállító: Chemical Abstracts Service. Több mint ötmillió hivatkozás a kémia és alkalmazásai irodalmára. 1967–

FORRÁSTÁJÉKOZTATÓ ADATBÁZISOK – az elsődleges források személyek, szervezetek, folyó kutatások, audiovizuális dokumentumok stb. Példák:

ELECTRONIC YELLOW PAGES. Előállító: Market Data Retrieval, Inc. Ezek az adatbázisok 4800 amerikai telefonkönyv szakmai részének a sárga oldalaknak géppel olvasható változatai. Csak kurrens adatokat tartalmaz.

TRADE OPPORTUNITIES. Előállító: U.S. Department of Commerce. Kb. 70 000 rekordot tartalmaz az Egyesült Államokban előállított árukra és szolgáltatásokra vonatkozó nemzetközi vásárlási ajánlatokkal.1977–

DEVELOP. Előállító: Control Data Corp. Több, mint 15000, a fejlődő országok rendelkezésére álló termék, szolgáltatás és segély leírását tartalmazza. 1979–

FORRÁSADATBÁZISOK – elsődleges adat- vagy információforrások, amelyek az eredeti forrás teljes szövegét tartalmazzák. Tartalmazza a kimondottan elektronikus terjesztésre szánt dokumentumokat is. Alosztályai:

NUMERIKUS ADATBÁZISOK – eredeti numerikus vagy statisztikai adatok, például pénzügyi, választási adatok, kutatási eredmények stb. Gyakran idősoros formában. Például:

DEFENSE DATA BANK. Előállító: Data Resources, In. 5000 idősor a U.S. Department of Defense kiadásaival és költségeivel kapcsolatban.

U.S. AGRICULTURE. Előállító: Chase Economcetrics/Interactive Data. 1100 heti, havi, negyedéves és éves idősort tartalmaz az U.S. Department of Agriculture által bizonyosított adatokból. Tartalmazza a farmerek által kapott és fizetett árakat, a farmok bevételeit, a fogyasztási és nagykereskedelmi árakat, a takarmánygabona árakat stb. 1950–

PTS TIME SERIES. ELőállító: Predicasts, Inc. Az adatbázis lefedi a termelést, a fogyasztást, az árakat, a mezőgazdasági, bányászati, termelési, szolgáltatási használati statisztikákat és az általános gazdasági és demográfiai adatokat az Egyesült Államokban és nemzetközi szinten. Az évek változóak, de a legkorábbi adat 1957-ből származik.

SZÖVEGES-NUMERIKUS ADATBÁZISOK – a mezők szöveges és számszerű adatokat felváltva tartalmaznak. Címtári és kézikönyvi adatok. Példák:

CHEMSEARCH. Előállító: DIALOG Information Retrieval Service, a Chemical Abstracts Service adatainak felhasználásával. Valamennyi kémiai anyagra kiterjed, amelyre a *Chemical Abstracts* utolsó hat számában utaltak.

DISCLOSURE II. Előállító: Disclosure, Inc. Közel 9000 köztulajdonban lévő vállaltnak a U.S. Securities and Exchange Commission által őrzött jelentéséből származó információ. Kurrens.

U.S. EXPORTS. Előállító: DIALOG Information Retrieval Service, a U.S. Bureau of the Census által nyújtott adatok felhasználásával. Statisztikai

idősorok az Egyesült Államokból más országokba irányuló exportforgalomról. 1978–

TELJES SZÖVEGES ADATBÁZISOK – eredeti szöveges dokumentumok, elsődleges források, mint pl. enciklopédiák, bírósági határozatok, újság vagy folyóiratcikkek. Példák:

ACADEMIC AMERICAN ENCYCLOPEDIA: Előállító: Grolier Electronic Publishing, Inc. 32 000 általános enciklopédia szócikk középiskolai és főiskolai diákok, illetve érdeklődő felnőttek számára. 1980–

HARVARD BUSINESS REVIEW. Előállító: John Wiley & Sons, Inc. a Harvard Business Review-val kötött szerződés alapján. Valamennyi, a *Harward Business Review*-ban 1980 óta megjelent cikket tartalmazza.

CHRONOLOG NEWSLETTER. Előállító: DIALOG Information Retrieval Service. A DIALOG havi újságának on-line változata. 1981–

1. táblázat. A géppel olvasható adatbázisok felosztása. Az osztályozás forrása: Cuadra Associates *Directory of On-line Databases* 5 (1) (1983 ősz)[14]

A referensz-adatbázisokban megkísérlik az elsődleges források lényegi jellemzőinek kiemelését a katalogizálás, az osztályozási rendszerek és az indexelés könyvtári és információs munkában kialakult fogalmi és elemző eszközeinek és technikáinak felhasználásával. Eme erőfeszítések nyomán rövidebb és szabatosabb szurrogátumok készülnek az eredeti művekről. Sajnos a fogalmi elemzés során nem kerülhető el az információveszteség, és az eredeti értelem óhatatlanul csorbul. Olyan probléma ez, amelyre a keresés során nagyon oda kell figyelni és számba kell venni.

A *forrásadatbázisok* elsődleges adat- vagy információforrások, amelyek a kimondottan elektronikus terjesztésre szánt információ teljes vagy komplett szövegét tartalmazzák. Nincs kivonatkészítés, a forrás adatbázisban nincs információveszteség. A kereső a teljes rekordhoz hozzáfér, legyen az enciklopédia vagy újságcikk, bizonyos vegyületek jellemzői vagy sajátos demográfiai adatok Indiana állam lakosságáról 1940-től napjainkig. A forrás adatbázisokat három alosztályra lehet bontani: *numerikus, szöveges–numerikus* és *teljes szöveges* adatbázisok. Az 1. táblázat minden típust tartalmaz, és példákkal is illusztrálják őket. Viszonylag új keletű fejlemény, hogy ma már a referensz adatbázisoknál jóval több nyilvános forrás adatbázis van. A 2. táblázatban láthatók a *Directory of On-line Databases*-ből vett randomminta vizsgálatának eredményei. A min-

14 Az adatbázisok és szolgáltatóközpontok választékával magyar nyelven részletesen foglalkozik Roboz Péter: Adatbázisok és szolgáltatóközpontok kiválasztása on-line információkereséskor. In: Tudományos és Műszaki Tájékoztatás, 1989, 36. évf., 1. sz., p. 3–13.

tában szereplő adatbázisok 55%-a forrás, és mindössze 34% volt referensz-adatbázis. A szoftver felvétele a forrásadatbázisok közé az útmutató további kiadásaiban még tovább bővíti ezt az amúgy is terjedelmes osztályt.

Az adatbázis típusa	Százalékos aránya
Forrásadatbázis	
teljes szöveges	16.0%
numerikus	29.8
szöveges/numerikus	9.0
szoftver	0.50
(összesen)	55.3%
Referensz-adatbázisok	
forrástájékoztató	11.2%
bibliográfiai	23.4
(összesen)	34.6%
Kevert típusok (többféle)	10.1%

2. táblázat. A különböző típusú adatbázisok arányai*

* A százalékokat Cuadra Associates, Inc. *Directory of On-line Databases* 6 (No.1.) (Fall 1984) adattárából véletlenszerűen kiválasztott 188 adatbázisból álló minta alapján számoltuk ki.

Egyes adatbázisoknak megvannak a nyomtatott megfelelői, amelyeket sok közművelődési, felsőoktatási vagy szakkönyvtárban, illetve információs központban megtalálunk. Tulajdonképpen sok adatbázis a korszerű és egyre inkább gépesített nyomdatechnikára való áttérés melléktermékeként vált géppel olvashatóvá. Így a nyomtatott indexelő és referáló folyóiratok, például a *Psychological Abstracts* vagy a *Current Index to Journals in Education* történetileg csak azért váltak géppel olvasható formában hozzáférhetővé, mert számítógépet használtak a szedéshez.

Az on-line információkeresés súlyának növekedésével számíthatunk rá, hogy az adatbázisok géppel olvasható formái saját jogukon is egyre értékesebbé válnak, és még inkább igaz ez a kizárólag elektronikus formában készülő adatbázisokra. Jelenleg az a helyzet, hogy gyakran értékes tapasztalatokat szerezhetünk az on-line adatbázisokról, ha keresés előtt megvizsgáljuk nyomtatott változatukat.

A 3. táblázatban szerepel több géppel és nyomtatásban egyaránt olvasható adatbázis.

Adatbázis	Nyomtatott változat
AGRICOLA	Bibliography of Agriculture
BIOSIS (PREVIEWS)	Biological Abstracts
	Biological Abstracts/RRM
BOOKS IN PRINT	Books in Print
CA SEARCH	Chemical Abstracts
CIS Index	Index to the Publications of the United States Congress
COMPENDEX	Engineering Index
ERIC	Current Index to Journals in Education
	Resources in Education
FEDERAL REGISTER	Federal Register
HISTORICAL ABSTRACTS	Historical Abstracts
INSPEC	Physics Abstracts
	Electrical and Electronic Abstracts
	Computer and Control Abstracts
MEDLINE	Abridged Index Medicus
PSYCINFO	Psychological Abstracts
SCISEARCH	Science Citation Index

3. táblázat. Néhány adatbázis és nyomtatott megfelelőik*

* A géppel olvasható adatbázis nem felelhet meg teljesen nyomtatott változatának. Ha adott adatbázist egy bizonyos terjesztőnél kívánunk használni, az adatbázis részletes leírását kell a terjesztő dokumentációjában tanulmányozni ahhoz, hogy az adatbázis szervezésében, tartalmában vagy elérésében mutatkozó különbségeket megismerjük.

Ez a rövid bevezető az on-line keresésre rendelkezésre álló adatbázisok világába elégséges hátteret nyújt a továbbiakhoz. A legfontosabb jellemzőket, a keresési problémákat és az adatbázisok főbb típusainak értékelési kritériumait a könyv más fejezetei taglalják.

1.4 Adatbázis-szolgáltatók

Az on-line hozzáférhető adatbázisok szolgáltatói vagy terjesztői olyan szervezetek, amelyek a felhasználók vagy az információs szakemberek részére on-line hozzáférést biztosítanak a géppel olvasható formában elérhető adatbázisokhoz. Az adatbázis-szolgáltató beszerzi vagy egyes esetekben maga építi azokat az adatbázisokat, amelyekről úgy véli, hogy használói igénylik. Olyan szoftverrendszereket készít és fejleszt, amelyek lehetővé teszik a keresést ezekben az adatbázisokban. Nagyszámítógépeket lízingel vagy vásárol, azokat – a kiegészítő berendezésekkel együtt – karbantartja. Dokumentációt készít, amely bemutatja, hogyan lehet e rendszer segítségével a leghatékonyabban és gazdaságosabban használni az adatbázisokat.

A legtöbb terjesztő kereskedelmi szervezet. Valamennyi díjat számít fel a nyújtott szolgáltatásért. Rendszerint a díj nagyobb részét az adott keresésre fordított idő függvényében határozzák meg, amelyet a *kapcsolati idővel* jellemeznek, vagyis azzal az idővel, amelyben a kereső terminálja közvetlen kapcsolatban állt a gazda-számítógéppel. Az ERIC adatbázis például a DIALOG szolgálatnál 25 dollár/kapcsolati óra alapdíjért használható.

Adatbázis-szolgáltató	A szolgáltatások és használóik jellemzése
Bibliographic Retrieval Services (BRS)	Nyolcvannál több, főként forrástájékoztató és bibliográfiai adatbázis. Elsődleges felhasználói a felsőoktatásban dolgozók és információs szakemberek; végfelhasználók.
Compuserve Information Corporation	200-nál több adatbázis elérése, népszerű referensz- és forrásinformációk, pl. hirdetés, tanácsadás, újságírás, hírek, üzleti információ, sport, időjárás, bevásárlás, utazási tanácsadás és számos egyéb fogyasztói és vezetői szolgáltatás. Végfelhasználók.
Control Data Corp./Business Information Services	Húsznál több adatbázis az üzleti élet és a gazdaság terén. Szakkönyvtárosok és más információs szakemberek.
DIALOG Information Services, Inc.	Több mint 170, főként forrástájékoztató és bibliográfiai adatbázis. Felsőoktatási és szakkönyvtárosok és más információs szakemberek; végfelhasználók.
Dow Jones & Co., Inc.	15-nél több referensz- és forrásadatbázis, amelyek hírek, piaci ajánlatok, pénzügyi statisztikák faktografikus adatait tartalmazzák. Szakkönyvtárosok és más információs szakemberek; végfelhasználók.
Mead Data Central	Harmincnál több referensz- és forrásadatbázis a jogi kutatással kapcsolatban. Jogi könyvtárosok; végfelhasználók.
National Library of Medicine (NLM)	Több mint húsz, orvosi kutatással és gyakorlattal foglalkozó adatbázis. Orvosi egyetemek, kórházak és más orvosi létesítmények orvosi könyvtárosai; végfelhasználók.
The Source	Több mint negyven adatbázis referensz- és forrásinformációval, mint pl. hirdetések, tanácsadás, újságírás, hírek, sport, időjárás és más szolgáltatások; végfelhasználók.

System Development Corporation (SDC Information Service)	Hatvannál több, elsősorban forrástájékoztató és bibliográfiai adatbázis. Felsőoktatási és szakkönyvtárak és más keresőszakemberek; végfelhasználók.
I.P. Sharp Associates, Ltd.	170-nél több üzleti és gazdasági forrásadatbázis. Szakkönyvtárosok és más információs szakemberek.

4. táblázat. Néhány jelentősebb adatbázis-szolgáltató*

* Az adatok a következő adattárból származnak: Cuadra Associates, In. *Directory of On-line Databases* 6 (No. 1) (1984 ősz).

Különösen azoknak a felsőoktatásban vagy kereskedelmi környezetben dolgozóknak a figyelmébe ajánlom ezeket az eltéréseket, akik ahhoz szoktak hozzá, hogy a számítógép-használatért a tényleges használati idő és a nagyszámítógép központi egységének használata (CPU idő) után fizetnek. A kapcsolati idő alapján történő számlázás bünteti az időveszteséget – vagy egyszerűen a gondolkodást –, ha az on-line történik. Ez pedig meghatározza, hogyan kell hatékony on-line információkeresést végezni. Azt sugallja, hogy amenyire lehetséges, a kereső még a számítógéppel való kapcsolatfelvétel előtt a lehető legteljesebben tervezze meg a munkáját. Az is fontos, hogy az on-line kereső gyorsan és határozottan reagáljon a rendszer visszajelzéseire, és nagy valószínűséggel közelítsen a kívánt eredményhez.

Bár már az 50-es és 60-as években bemutatták és kísérletileg tesztelték az on-line rendszerek alkalmazását az információkeresésben, a megfelelő technológiai háttér, vagyis az időosztásos számítógépek, a távoli terminálok és a távközlési kapacitás költségei és hozzáférhetősége csak a hetvenes évek elején tették lehetővé a kiterjedt információkeresést. Az egyik legkorábbi rendszer a Lockheed DIALOG, a System Development Corporation ORBIT és a National Library of Medicine MEDLINE rendszere. A 4. táblázatban látható néhány jelentős keresőszolgáltatás, adatbázisaik és használóik jellemzésével együtt. A *Directory of On-line Databases* 1984 őszén 362 on-line szolgáltatót sorolt fel, ám ezek többsége viszonylag szűk körben működik, és nem használják őket annyian, mint a 4. táblázaton szereplőket.

Az adatbázis-szolgáltatók és az adatbázis-előállítók megkülönböztetéséről kell még szót ejtenünk. Bár az adatbázis-előállító időnként keresőszolgáltatást is biztosít (mint a Lexis és Westlaw, a National Library of Medicine, az OCLC Inc.), a szolgáltatók többsége legfeljebb egy-két adatbázist épít azok közül, amelyekhez hozzáférést nyújt. Gyakoribb, hogy gazdasági tényezőktől vezetve, a keresőszolgáltatásokkal kereskedők megvásárolják azokat az adatbázisokat, amelyek, reményeik szerint, hasznot hoznak a számukra. Egyes adatbázisok, mint az ERIC, a Medline, az Academic American Encyclopedia

és a BIOSIS, több keresőszolgáltatáson át is elérhetők. Más adatbázisoknak egyetlen szolgáltatóval van kapcsolatuk. Ismét mások, például az Institute for Scientific Information vagy a National Library of Medicine által készített adatbázisok, az adatbázis-készítőn, és más adatbázis-szolgáltatókon (pl. a BRS és az SDC) keresztül is elérhetők.

[...]

2.5 Ellenőrzött szótárak[15]

Több érv is szól amellett, hogy az információ reprezentálására mesterséges nyelvet használjanak. Mint e fejezetben már említettük, a természetes nyelv sok gondot okoz az információkeresésben, ilyen például a szinonímia, a homográfiából származó szemantikai bizonytalanságok, a „lágy" tudományokat jellemző szemantikai pontatlanságok, a kontextuális félreértéseknek betudható téves találatok vagy a generikus keresések végrehajtásának buktatói. Az ellenőrzött szótárak használata, az információ reprezentálása mesterséges nyelvekkel jelentős részben kiküszöböli ezeket a problémákat. Az ellenőrzött szótáraknak is megvannak a maguk problémái. Nem csodaszerek. A következő oldalakon az ellenőrzött szótárral végzett indexelés jellemzőit, előnyeit és hátrányait vizsgáljuk.

2.3 ábra. Információleíró (dokumentációs) nyelvek, a természetes nyelvtől való eltávolodás rendjében

A 2.3 ábra egy a természetes nyelvtől való távolságot mutató kontinuum mentén elhelyezve sorolja fel az információleíró (dokumentációs) nyelvek főbb osztályait. A kontinuum bal felén szerepelnek a természetes nyelven alapuló információreprezentációk, közöttük a cikkek teljes szövege, a referátumok és a cí-

15 2.5 Controlled vocabulries. In: On-line information retrieval... p. 41–51.

mek. A negyedik osztály, amelyet *azonosítónak* – pontosabban másodlagos azonosítóknak – hívunk, azokra a kulcsszavakra utal, amelyeket az indexelő az eredeti szövegből emel ki. Az azonosítókkal később foglalkozunk.

A szótárellenőrzés első és legegyszerűbb formája a *deszkriptorok* használata, amelyeket a *tezaurusz* tartalmaz. A tezaurusz a deszkriptorok ellenőrzött szótára, amelynek szavai általában dinamikusan növekvő dokumentumgyűjteményből származnak, a szótár elemei között pedig bizonyos értelmi–szemantikai (paradigmatikus) kapcsolatokat határoztak meg. A tezaurusz funkciója, hogy felsorolja és meghatározza a szótár érvényes (megengedett) és érvénytelen (tiltott) elemeit, és megmutassa, milyen kapcsolat van az érvényes elemek között. A korábban a természetes nyelvvel összefüggésben megfigyelt problémák közül sokat megold a tezaurusz struktúrája – a homográfokat, a szinonimákat, a generikus keresést, de még a hamis egymás mellé rendelést is. A magyar szakmai telefonkönyv tezauruszának részletét láthatjuk az alábbiakban.

A tezaurusz szótárának elemei között az alábbi alapvető kapcsolatok fordulnak elő:

BT	broader term	–	fölérendelt fogalom általában	F
NT	narrower term	–	alárendelt fogalom általában	A
RT	related term	–	rokonsági kapcsolat	X
USE	use	–	lásd	L
USE&	use in combination	–	lásd ÉS-kapcsolatban	L&
USEV	use alternative	–	lásd VAGY-kapcsolatban	LV
UF	used for	–	helyett	H
UF&	used for in combination	–	helyettesít ÉS-kapcsolatban	H&
UFV	used for in alternative	–	helyettesít VAGY-kapcsolatban	HV
SN	scope note	–	megjegyzés	M:

ÁCSOK ÉS ÁLLVÁNYOZÓK
X Építészek és építőmesterek

Adatbankok
L Adatbázisok

Adatbázis-kezelés
L Szoftverek

ADATBÁZISOK
H Adatbankok
HV Dokumentációs szolgáltatás
T Informatika
Számítástechnika
R Információszolgáltatás
X Adatfeldolgozás
Irattárrendezés
Könyvtárak
Multimédia (K)
Szoftverek

ADATFELDOLGOZÁS
T Számítástechnika
P Adatrögzítés (K)
X Adatbázisok
Adathordozó kártyák
Díjbeszedés
Szoftverek

ADATHORDOZÓ KÁRTYÁK
H Mágneskártyák
T Számítástechnika
X Adatfeldolgozás
Vonalkódtechnika

ADATRÖGZÍTÉS (K)
T Adatfeldolgozás

ADÓSSÁGBEHAJTÁS
T Pénzügy

ADÓÜGYEK, ADÓTANÁCSADÁS
F Pénzügyi tanácsadás
Szakértők
T Pénzügy
X Könyvviteli szolgáltatások

ÁFÉSZ-ek
L Fogyasztási és Értékesítési Szövetkezetek

AJÁNDÉKTÁRGYAK
HV Bizsu
F Vegyesiparcikkek
X Díszek, dísztárgyak
Trafikok

Ajtók
LV Garázskapuk
Kapuk (K)
Nyílászárók

Ajtómentés
M: *Telefonkönyvben nem szerepel*
L Kulcsok, kulcsszolgálat

Albérlet-közvetítő irodák
L Ingatlanforgalmazás és -közvetítés

Állomások
LV Közlekedési vállalatok
Személyszállítás
Vasúti közlekedés
Vízi közlekedés
Díjbeszedés

Az L, L& és LV, illetve a H, H& és HV az érvényes szótári elemek felé, illetve azoktól elmutat. Ha a tezauruszszerkesztők az „adatbázist" választották deszkriptornak (vagyis érvényes lexikai egységnek, amelyet az „adatbank" és a többi kváziszinonima helyett használni kell), akkor az e döntést tükröző utaló a következő:

ADATBÁZISOK
H Adatbankok

A tezaurusz tartalmaz egy inverz (fordított) utalást az adatbanknál, amely a deszkriptorhoz vezeti a felhasználót. Így küszöbölik ki a tezauruszokban a természetes nyelvi szinonímiát. Ha a természetes nyelv több olyan kifejezést tartalmaz, amely ugyanazt, vagy közel ugyanazt a fogalmat jelöli, a tezaurusz egyetlen érvényes kifejezésre szűkíti a választás lehetőségét. Lehet persze, hogy önkényes a deszkriptor kiválasztása a szóba jöhető kifejezések közül.

Az L, L& és LV, illetve H, H& és HV utalások iránymutatásai különösen sokat segítenek az információkereséssel foglalkozó szakembernek, akinek (legalábbis elméletben) nem kell többé adott fogalom valamennyi megnevezési lehetőségét számba vennie, ha az adatbázist, amelyben keres, tezaurusz segítségével indexelik. A deszkriptor használható a keresésre, ahogyan a fogalom valamennyi megnevezésének indexelésére is azt használták az adatbázisban.

A *megjegyzéssel* lehet támogatni a felhasználót a lexikai egység használatáról, definíciójáról. A homonimákat az ún. hátravetett értelmezőkkel különböztetik meg egymástól (pl. ÁR [SZERSZÁM], ÁR [GAZDASÁG]).

RELEVANCIA (OKTATÁS)
M: Az iskolákban tanított ismeretek alkalmazhatósága a tanulók és a társadalom igényeinek és érdekeinek megfelelően

[...]

RELEVANCIA (INFORMÁCIÓKERESÉS)
M: A keresőkérdésre visszahívott összes dokumentum és a talált – releváns – dokumentumok aránya

A RELEVANCIA deszkriptornak tehát két jelentése van, és hátravetett értelmezőik alapján egyértelmű zárójeles minősítők alapján egyértelmű, hogy melyikről van szó.

Az F és az A a generikus fölérendelt (nemfogalmat jelölő) és alárendelt (fajfogalmat jelölő) kifejezésekre utal, és hierarchikus reláció. A T és a P a partitív fölérendeltet (az egészet) és alárendeltet (a részt) jelentő kifejezéseket kapcsolja össze. Az R és az E a különböző oksági, eszközrendeltetési, eredete–eredménye, tevékenység–tevékenység tárgya stb. összefüggéseket jelöli.

A rokonsági reláció (X) arra utal, hogy két fogalom között *valamilyen* kapcsolat van. Ez azonban sem nem hierarchikus (F, A), sem szinonimareláció (L, H). A rokonsági kapcsolatok ezeknél sokkal árnyaltabbak, és az adott kifejezés alatt rokonsági kapcsolattal jelölt kifejezések általában heterogén együttest alkotnak. Ezek a kifejezések semmiképpen sem tekinthetők szinonimáknak. Csak azt jelölik ezzel a relációval, hogy az adott rendszerben, gyűjtőkörben valamilyen érdekes összefüggés áll fenn a két fogalom között, amelynek ismerete segítheti az indexelőt vagy a keresőt az adott információs probléma reprezentálásában, de lehet teljesen érdektelen is.

Végül megmutatjuk hogyan kezeli egy tezaurusz a hibás egymás mellé rendeléseket. Ha egy kifejezést – amilyen például az „információelmélet” – gyakran használnak, akkor az ilyen összetett formában is deszkriptor lehet, s ez célszerűbb, mintha két külön szóként szerepelne. Így a két fogalom, az információ és az elmélet, úgymond *prekoordináltan* kerül a szótárba. A prekoordinált rendszer olyan ellenőrzött szótár, amelyben a két vagy több egyszerű fogalom metszetét képviselő specifikus, összetett tárgykörök maguk is az indexelőnyelv elemei. Az ERIC-ben például az információelmélet fogalom iránt érdeklődő kereső használhatja az „adatfeldolgozás” deszkriptort, az „adat” és a „feldolgozás” kifejezések helyett. Ezzel kiküszöböli a *posztkoordináció* – vagyis a természetes nyelvi kifejezések keresés közben történő összekapcsolása – okozta hibás találatok problémáját.

Az előbbi példáknak megfelelő tezauruszcikk a következő:

KUTYA
- F Ragadozó
- A Agár
- Farkaskutya
- Puli
- Uszkár
- T Falka
- P Tépőfog
- R Házőrzés
- E Ebtenyésztés
- X Ugatás
- Veszettség

A többi másodlagos *azonosító* olyan természetes nyelvi szó vagy kifejezés (kulcsszó, szabad tárgyszó), amelyek további elérési pontok az információforráshoz; olyan elérési pontok, amelyek máshogy nem adódnának, mert a tezauruszban lexikai egységként nem szerepelnek (többnyire azért, mert túlságosan speciális, vagy bizonytalan jelentésű kifejezések. Az ilyen szabadon megadható kifejezések mélyítik az indexelést. A kulcsszavak és egyéb szabad tárgyszavakat tehát a tezaurusz nem tartalmazza, lehetnek akár tulajdonnevek is.

A szakkifejezéseket képviselő szabad tárgyszavak például az új fogalmak megnevezései, az olyan ötletek, amelyek éppen kísérleti stádiumban vannak. De az is lehetséges, hogy egyszerűen csak túl specifikus vagy a szakirodalomban ritkán használt kifejezésekről van szó, amelyeket a szerkesztők nem tartottak érdemesnek arra, hogy a tezaurusz deszkriptorai legyenek. Így az olyan kifejezések, mint a „valószínűségi indexelés", és a „sorban állás elmélet" nem jelennek meg egy könyvtári tezauruszban, annak ellenére, hogy mindhármat használják a könyvtár- és információtudományban. Az indexelők azonban hozzárendelhetik a „sorban állás elmélet" kulcsszót ahhoz a cikkhez, amelyben ezzel az elmélettel foglalkoznak.

A tulajdonnevekkel az olyan információk írhatók le, mint a földrajzi helyek, a kísérletek neve vagy a jogszabályok. A keresésben betöltött szerepük nyilvánvaló.

A szabályozott (hagyományos, *Cutter* nyomán kialakult) tárgyszójegyzékek a tezauruszokhoz kapcsolódnak. Ezekre talán a hagyományosabb amerikai könyvtárakban használt tárgyszójegyzékek a legjobb példák: a *Sears List of Subject Headings*, amelyet a legtöbb közművelődési könyvtár átvett, és a *Library of Congress Subject Headings*, amelyet az egyetemi könyvtárak többsége használ. Ezekben a relációkat kevésbé részletesen szokták kidolgozni. Az L helyett többnyire a Lásd szerepel, és az inverz relációját („helyett") általában nem tüntetik föl.

A tezauruszok és a tárgyszavak között fontos filozófiai különbség is van. A tárgyszójegyzékek általában kísérletek az univerzum (vagy legalább egy részhalmaza) szerkezetének teljes tükrözésére, és nem kötődnek egy adott dokumentumgyűjteményhez. A tezauruszokat ezzel szemben általában meghatározott szakterület meglévő, élő és növekvő könyvgyűjteményeiből, cikkeiből stb. származtatják. E tezauruszok szótárát az e gyűjteményben felmerülő szinonímia, szemantikai pontatlanság felszámolására hozzák létre. A tezaurusz-építés pragmatikus vetülete nem jelenik meg a tárgyszójegyzékek vagy osztályozási rendszerek készítőinek filozofikus munkájában.

A dokumentumok reprezentálására szolgáló mesterséges nyelvek közül az utolsó az osztályozási rendszer. Ezek közül a legismertebb az ETO és az Egyesült Államokban Dewey Tizedes Osztályozása (a TO). Az Egyesült Államokban elsősorban könyvek osztályozására használják a TO-t, amely az emberi tudás teljességének reprezentálására törekszik egyetlen hatalmas hierarchiában. Minden osztályt alosztályokra bontottak, azokat ismét al-alosztályokra. A tizedes rendszer absztrakt eszköz volt a fogalmak hierarchikus kapcsolatainak reprezentálásához. Nincs azonban jó módszer a rendszerbe logikusan nem illeszkedő új ismeretek beépítésére. *Charles Meadow* részletesen taglalja a „mesterséges földi műholdak" példáját; ez a TO-ban a 629.138.82 jelzetet kapja. Ez az egyetlen olyan hely, amelyen a kifejezés logikusan megjelenhet a TO-ban, bár egyértelmű, hogy nem ide tartozik:

600	Technológia (alkalmazott tudomány)
620	Mérnöki tudományok
629	Egyéb mérnöki ágazatok
629.13	Aeronautika
629.138	Légi járművek használata
629.138.8	Űrrepülés
629.138.82	Mesterséges földi műholdak

A mesterséges földi műholdak és az űrrepülés egyaránt a „légi járművek használata" alá kerülnek, holott egyáltalán nem levegőben repülő járművek. Ráadásul a műholdak nem űrrepülés, mert eszközről és nem tevékenységről van szó, mégis alárendelődik a tevékenységnek.

E rövid bevezetés csak a felszínt súrolta, de akit érdekel a téma, érdemes mélyebbre ásnia.

Sokat foglalkoztunk e fejezetben a probléma kiválasztással és a dokumentációs nyelvekkel kapcsolatos kérdésekkel. Kulcskérdés ez, mivel az on-line keresőnek az adatbázisban használt reprezentációs nyelvvel foglalkozva általában számos nehézséggel kell megküzdenie. A legtöbb adatbázisban lehetőség van a természetes nyelvű keresésre a címben, a referátumban, a szabad tárgyszavak között, de még a teljes szövegben is. Az adatbázistól függően deszkriptorok, tárgyszavak és osztályozási jelzetek is keresőszavak lehetnek.

A kereső a bőség zavarával küzd. Melyik lehetőség, vagy a lehetőségek milyen kombinációja a legelőnyösebb adott keresési probléma megoldására? Mik az előnyei és a hátrányai a különböző megközelítési módoknak? Deszkriptorok? Osztályozási jelzetek? Bár lehetetlen határozott, minden keresési problémára és adatbázisra érvényes válaszokat adni, a lényeg az, hogy az eredményes információkeresés összetett folyamat, melyben több irányból célszerű megfogalmazni a keresés tárgyát: deszkriptorokkal, szabad tárgyszavakkal, osztályozási jelzetekkel stb. Mindezek együttes használata vezet csak optimális eredményhez.

MIRANDA LEE PAO

Lee Pao az információkeresés egyik klasszikusának, *William Goffmannak* a tanítványa. A clevelandi Case Western Reserve University-n *Tefko Saracevic* munkatársaként tartott előadásai alapján írta könyvét. Az elsősorban praktikus felhasználói igényeket kielégítő on-line kézi- és tankönyvekhez képest Pao könyvét az elméletibb megközelítés jellemzi, az amerikai egyetemi könyvtárosképzésben ezért is számít ma „az" egyetemi tankönyvnek. A szerző szűkebb szakterülete az automatikus indexelés.

Az on-line információkeresés fogalma[16]

12. Az intelligens információkeresés felé[17]

Az információkeresés kutatásában látványos eredmények születtek. A könyvtártudományi és informatikai kutatók úttörő munkát végeztek a dokumentumok és kérdések tartalmának reprezentálása terén. Különösen a szabályozott nyelvekkel, természetes nyelvekkel, automatikus indexeléssel és tartalmi kivonatkészítéssel megvalósított dokumentumreprezentáció vizsgálata fejlődött sokat. Olyan alternatív keresési stratégiákkal végeztek sok éven át kiterjedt és mélyreható kísérleteket és szűrtek le tapasztalatokat, amelyekben figyelembe vették a bizonytalanság és a részleges megfelelés kérdését, és nem ragaszkodnak szigorúan ahhoz, hogy az eredményt determinisztikusan, szigorúan a relevancia szempontjából értékeljék. E megközelítések elméleti alapjai a valószínűségi elveken nyugszanak. Az információkeresés terén folytatott vizs-

16 Concepts of information retrieval / Miranda Lee Pao. – Englewood: Libraries Unlimited, 1989. – 285 p.

17 Toward intelligent information retrieval. In: Concepts of information retrieval, p. 243–249.

gálatok másik fontos hozadéka a kifejezések közötti relációk, a dokumentum és kifejezés, illetve a dokumentum és dokumentumrelációk és a szerzői relációk vizsgálata, amelynek legfontosabb eredményeit a szöveges dokumentumok elemzésével érték el. A könyvtári informatikában is hosszú múltra tekint vissza a rendszerek és a rendszerelemek értékelése és kipróbálása.

A számítástechnika, a távközlés, a lézertechnika működő rendszereiben és más új műszaki megoldásokban elért hihetetlen eredmények forradalmasították az információkeresést. Az információkereséssel foglalkozó kutatók élen jártak a számítógépre alapozott információkereső rendszerek tervezésében és fejlesztésében. Egyszerre csak nem az volt már a kérdés, hogy a számítógép elérhető-e, hanem az, hogyan alkalmazható hatékonyan a keresőfolyamatban. Manapság az információkeresés összekapcsolódik az on-line kereséssel. A magán adatbázisok névtelen tömegén túl, minden elképzelhető tárgykörnek van távolsági kereskedelmi forgalomban elérhető adatbázisa. A keresések száma is ennek megfelelően alakul.

A számítógéppel támogatott keresési lehetőségek nyomán kialakult derűlátás dacára a tárgyi/tartalmi keresés még mindig a kulcsszavas és deszkriptoros indexelésen és keresésen alapul, a keresőkifejezéseket Boole-operátorok kapcsolják össze, bármilyen adatbázisról legyen is szó. Vagyis a dokumentumok tartalmi reprezentálása és az alapvető fájlszerkezet a manuális rendszerek óta nem változott. Elbátortalanít az a felismerés, hogy még a mai fejlett technológiával is csak olyan kis keresési teljesítmény érhető el mint régen, csak gyorsabban. A működő keresőrendszerek belső szerkezetének gyengeségei mellett egyre nagyobb a bizonytalanság abban a tekintetben is, hogyan lehet a legjobban előhívni a keresőkérdést a használóból. Nem tudjuk, hogyan fejeződnek ki az információs igények, és ez hátráltatja az olyannyira kívánatos fejlődést a kérések és a releváns dokumentumok megfeleltetésében. A képet tovább árnyalja, hogy a keresési szakemberek, de a tényleges információhasználók keresési eredményei is meglehetősen eltérőek. Más szóval, a keresés megbízhatósága egyenlőre csak ködös álom.

Szerencsére a legfrissebb közlemények új kutatási irányokra utalnak az átlagos keresés e makacs gondjainak megoldására. Elsősorban azokra a próbálkozásokra utalok, amelyek új technikákat kívánnak felhasználni az „intelligens" információs rendszerek tervezésére és kifejlesztésére. Egyre nagyobb az érdeklődés az emberi, tudati információfeldolgozás tanulmányozása iránt. Vizsgálják a problémamegoldásban, az információ felfogásában és felhasználásában, az információ értékelésében, a tanulásban, illetve az ismeretek alkalmazásában betöltött szerepét. Ha ismerősen hangzanak ezek a témák a mesterséges intelligencia (MI) kutatásából, ez nem véletlen. A gondolkodás és a tanulás alapvető problémái szorosan összefüggenek azzal, hogyan fogalmazzák meg a kérdéseket az emberek, hogyan jellemezhetők az információs problémák a természetes nyelven, és hogyan emelkednek ki bizonyos stratégiák az

információkeresés folyamatából. Ezek a MI munkásainak, a kognitív tudományok művelőinek, a kérdés–válasz rendszerek tervezőinek és az információs szakembereknek a központi kérdései. Miután a MI kutatók számos alkalmazási területen ígéretes eredményeket értek el, jelenleg nagy reményeket fűznek hasonló technikák alkalmazásához az információkeresési technikák területén. Több szakértői rendszer prototípusa készült el, amelyek különböző keresési feladatokat hajtanak végre. Ezek a rendszerek az általuk lefedett tárgykör terjedelmét, illetve a funkciók bonyolultságát tekintve széles skálán mozognak. Ha a MI kutatás eredményeit az információkereső folyamatokra alkalmazni tudják, a mai életmód jelentős, felmérhetetlen gazdasági, társadalmi, sőt politikai hatással járó változásával számolhatunk.

Mesterséges intelligencia

A MI új, összetett szakterület. Még a szakértők között sincs egyetértés abban, mi a MI, és milyen témák tartoznak hozzá. A MI kutatás tárgya azonban világos. Azoknak a gondolkodási folyamatoknak a feltárására irányul, amelyek az intelligens magatartásmódokhoz kapcsolódnak, és azokat a számítógépes programozási megoldásokat keresik, amelyekkel az egyébként emberi intelligenciát igénylő funkciók elvégezhetők. *Roger C. Schank*, az egyik élenjáró MI kutató, a laikusok számára is élvezetes olvasmányt jelentő könyvében 1984-ben megállapította, hogy a mesterséges intelligencia kutatása két megközelítési módot alkalmaz, a termék vezérelte és az elmélet vezérelte megközelítést. A robotika és a szakértői rendszerek a MI termék vezérelte ágából nőttek ki. Ezeket a termékeket bizonyos szükségletek kielégítésére tervezték: például arra, hogy rokkantaknak robotok segítsenek egy sor házkörüli feladat megoldásában, a szakértői rendszerek pedig esetleg segítik az egyéneket annak megállapításában, hogy jogosultak-e bankhitel felvételére. A szakértői rendszerek tervezői megpróbálják megragadni a szakértői ismereteket, és a következtetéseket egy sor számítógépes programban a ha/akkor szabályok halmazára igyekeznek redukálni. Például orvosi diagnosztikai programban addig kell e szabályok láncolatán előre haladni, amíg a feltételezett differenciáldiagnózist el nem érik. Robbanás észlelhető a szakértői rendszerek használatában az orvostudomány, a geológia, a műszaki tudományok, a termelés, a pénzügyi szolgáltatások, a gépészeti diagnosztika területén és számos más szakmában is. Ezek hihetetlen technológiai előrelépéseket jeleznek.

Másrészt az elmélet vezette MI kutatás középpontjában az áll, hogyan jelenik meg a tudás az emberi elmében, hogyan dolgozzák fel az emberek a gondolatokat, mit tudnak, hogyan használják ismereteiket, és hogyan tanulják meg azt, amit tudnak. A mégoly kevéssé bonyolult szakértői rendszer megszerkesztéséhez a rendszertervezőnek részletesen ismernie kell, mi megy végbe a rend-

szer emberi megfelelőjében. Vagyis tisztában kell lennie azzal, hogyan gondolkodnak és érvelnek az emberek, ha intelligensen akarnak „funkcionálni". Nyilvánvaló, hogy a MI az elme ősrégi filozófiai problémáit feszegeti. A tanuláselmélet megalkotásában például az elmélet vezette kutató számítógépes modelleket szerkeszt e mentális folyamatok reprezentálására. Az elmélet helyességét azután számítógépes programok tesztelik. Mivel a számítógépes programok csak az adott utasításokat képesek végrehajtani, az elméletnek aprólékosnak, és az emberi feladat teljességére kiterjedőnek kell lennie. A számítógép értékes feladatot lát el a MI kutatásban, mivel nem hagy teret a feltételezéseknek. Ennek eredményeként a számítógépet rábírják a nyelv megértésére, a tapasztalatokból való tanulásra, szimbólumok értelmezésére és az input alapján a fontos fogalmak közötti kapcsolatteremtésre. Mindez megvilágítja, hogyan végzik el az emberek ezeket a feladatokat. Az emberi információfeldolgozás jobb megismerése a MI kutatókat is segíti az értelem tanulmányozásában. Bármely ismerettöredék tudományos előrehaladást jelez. A számítógép csak akkor képes intelligens lény módjára cselekedni, ha érti a nyelvet. Ezért az egyes emberben lezajló természetes nyelvi „feldolgozás" a MI és az információkeresés fontos kutatási területe. A MI szorosan kapcsolódik a kognitív tudományokhoz, a pszichológiához, a nyelvészethez és a filozófiához.

Szakértői rendszerek

A szakértői rendszerekről igen sok szó esett, ezért sokan a mesterséges intelligenciát intelligens számítógéprendszerek felépítésével kapcsolják össze. A szakértői rendszer olyan összetett program, amely az emberi szakértők problémamegoldó folyamatát utánozza. Olyan ismeretalapú rendszer, amely két lényeges összevetőből áll: (1) a szakterület megfelelő ismeretanyagából, amely kiterjed az alkalmazási szabályokra is, és (2) olyan következtető „gépből", amely a problémamegoldó mechanizmust biztosítja. Az ismeretanyagot egy vagy több emberi szakértő ismereteinek összefoglalásával nyerik, és ezt egy számítógépben jelenítik meg. A képzett programozó, az ún. „knowledge engineer", az „ismerettechnikai mérnök" dolgozik a szakértőkkel, és a szakértő ismereteit és érvelési, illetve problémamegoldási módját a számítógép által használható formára fordítja. Ez a szakterület a „knowledge engineering", az „ismerettechnika". Azt tapasztalták, hogy a szakértők – például az on-line keresőszakemberek – pontosan körülhatárolható eljárások nélkül oldják meg problémáikat. Ehelyett *heurisztikus* vagy *tapasztalati* eljárásokat követnek. Ez a heurisztikus megközelítés a tevékenység megbízható vezetője, még akkor is, ha nem lehet tudományosan igazolni, hogy beválik. A szakértői rendszer révén általában a problémamegoldásnak ezt a heurisztikus, ismeretalapú megközelítését ragadják meg a szűkebb szakterületen.

Az angol „knowledge based system”, illetve „knowledge based approach” kifejezést tévesen „tudásalapú rendszernek”, illetve „tudásalapú megközelítésnek” fordítják. Ez azért helytelen, mert számítástechnikailag sohasem a tudatban létrejövő tudás, hanem csak annak társadalmilag egyeztetett, nyelvileg kifejezett formája, az ismeret kezelhető. Ezért helyesen „ismeretalapú rendszerről” kell beszélni magyarul. A „knowledge engineering”, „knowledge engineer” és a „knowledge organization” kifejezések esetében értelemszerűen „ismerettechnikát”, „ismerettechnikai mérnököt” és „ismeretszervezést” kell mondani magyarul.
A „tudás” és az „ismeret” fogalmának megkülönböztetésével kötetünk elején, *Bertram Claude Brookes* szemelvényében is foglakozunk.

Általában további elemek is szükségesek ahhoz, hogy létrejöjjön a kommunikáció a szakértői rendszer és a használó között. Először az intelligens használói interfész biztosítja az interakció lehetőségét. Ez kéri a releváns információt a kérdezőtől, és pontosítja az input adatokat. Másodszor, van egy magyarázó modul, amely a szakértői rendszer által elért következtetések igazolására való. Lehetővé teszi a használó számára, hogy kövesse az okfejtést vagy a szabályokat, amelyek alapján meghozták a döntéseket. Végül a segítő modul támogatja a rendszer használatát. A szakértői rendszer ritkán helyettesítheti az emberi szakértőt. Inkább tanácsadóként működik az adott szakterületen, amely az ismeretanyagban rejlő tudásra épít a beépített heurisztikával. Két klasszikus példa a MYCIN, a fertőző vérbetegségek diagnózisát és terápiáját szolgáló szakértői rendszer, illetve az INTERNIST, egy belgyógyászati szakértői rendszer.

A mai szakértői rendszerek közös vonásai a következők:

Először: a választott területen a döntéshozatal során számos lehetséges megoldást kellett figyelembe venni. A diagnosztikai rendszerben például az orvosnak sok tényezővel kell számolnia.

Másodszor: a szakterület problémáinak többségére nem adható egyszerű igen vagy nem válasz. Az orvosi diagnózis eredménye inkább csak bizonyos valószínűségi szinten adható meg, nem megfellebbezhetetlen tény.

Harmadszor: mindezekből következik, hogy a szakértőrendszerek számolnak az ilyen bizonytalanságokkal, mert olyan adatokkal dolgoznak, amelyek nem teljesek, szubjektívek vagy éppenséggel következetlenek.

Negyedszer: a szakértői rendszerek általában nyelvileg és nem pontos számszerű értékekkel kifejezett fogalmakkal operálnak.

Végül: a szakértői rendszerbe foglalt ismeretek típusa abból a szempontból is értékes, hogy hasznos lehet azok számára, akik nem tudnak könnyen szakértőkhöz fordulni.

A fentiekből könnyen kitalálható, hogy az információkeresés folyamata könnyűszerrel utánozható a szakértői rendszerekben. Az információkeresőközösség számára a határokat jelenleg az olyan szakértői rendszerek tervezése jelöli ki, amelyek az információkeresés bizonyos vetületeit utánozzák. A kereső szakemberek által követett megközelítés heurisztikus, és a szakmai tudás viszonylag körülhatárolt részéről van szó. A számítógép a keresőszakemberrel együtt például hatékonyan használható arra, hogy a használónak értelmes ötleteket adjon a keresőkérdést tisztázó megbeszélés során, amelyben az információkereső interakcióba lép egy intelligens számítógéprendszerrel. A keresőnek is hasznos kifejezéseket sugallhat a keresőkérdés megfogalmazása közben. A szakértői rendszerek technikájának az információkereső rendszerekre való alkalmazása még gyermekcipőben jár. A legtöbb projekt még kísérleti stádiumban van. Van, amelyben már elkészült a prototípus. A dokumentumkeresésnek három területe határozható meg: a feldolgozás, a tájékoztatás és a keresés közvetítése. A szakértői rendszereknek e területeken való felhasználásáról esik szó a továbbiakban.

Szakértői rendszerek az információkeresésben

Borko 1987-ben írta, hogy mivel „az osztályozás végül is szakterületünk alapvető tevékenysége, s ha a szakértői rendszer képes orvosi eseteket osztályozni és diagnosztizálni, akkor olyan szakértői rendszer is tervezhető, amely a könyvtári dokumentumokat osztályozza és katalogizálja.” A University of California, Los Angeles kebelében térkép-katalogizáló szakértői rendszer, a MAPPER, kifejlesztésén dolgozik. A rendszer szigorúan követi az AACR2 katalogizálási szabályait és a MARC térképformátumot, de igénybe veszi a térképet katalogizálók segítségét. A MAPPER négy egymáshoz kapcsolódó programhalmazból vagy modulból épül fel:

A Felhasználói Interfész modul teszi lehetővé, hogy a használó mezővezérelt utasításokat adjon a katalogizálandó térképekre vonatkozó releváns információk előhívásához. Úgy tervezték, hogy eltérő számítógépes és/vagy katalogizálási szinten lévő emberek is tudják használni.

Az Ismeretanyag modul egy sor, a térképek katalogizálása szempontjából jelentős döntési szabályt tartalmaz. Ezek a HA/AKKOR formátumot követik, amennyiben ha egy HA feltétel teljesül, az AKKOR következtetéshez vezet. Ezeket a szabályokat az AACR2 szabályaiból merítették.

Az Interferencia modul veszi fel a katalogizálótól származó információt, alkalmazza azt, és kiválasztja az ismeretanyagból a megfelelő szabályt, amelynek alapján a főtételt előállítja. Azt is ellenőrzi, hogy a bevitt adat megfelelő-e a HA mellékmondatban foglalt feltétel teljesítésére. Ha nem, a rendszer további információt kér.

A Magyarázó modul lehetővé teszi a katalogizáló számára, hogy megértse, miért választanak bizonyos szabályokat, és követheti a választás menetét. E modul elsődlegesen azt a célt szolgálja, hogy megnyerje a használó bizalmát a rendszer használatához. Arra is módot ad, hogy a használó további alternatív szabályokat javasoljon, vagy felülbírálja a rendszer választását. A MAPPER IBM személyi számítógépeken működik.

Az interaktív ismeretbázisú rendszerek közül az egyik legigényesebb az Indexing AID Project keretében készül, prototípusát a National Library of Medicine fejlesztette ki. A MI eljárásokat használták az orvosbiológiai folyóirat-irodalom indexelésének javítására és az indexelés következetességének biztosítására. Ezt a rendszert képzett MEDLINE indexelők számára tervezték.

A tájékoztató munkában a PLEXUS intelligens kertészeti forrástájékoztató rendszer lehet a szakértői rendszerek példája. Ezen jelenleg dolgoznak a University of Londonban. A rendszer intelligens interfész segítségével vezeti el a használót a legvalószínűbb információforráshoz. A kérdésre adott válaszban a rendszer azokat a személyeket, könyveket, helyeket, intézményeket vagy társaságokat nevezi meg, amelyek speciális kertészeti probléma megoldásában a legnagyobb valószínűséggel segíthetnek. Nagyra törő program ez. Bár úgy tűnik, hogy a tárgykör leszűkítése a kertészetre jól körülírt területet eredményez, a szükséges ismeret összetett és sokrétű. A legfőbb eredmény az, ahogyan a PLEXUS szemantikai feldolgozást végez. Kifinomult rendszert dolgoztak ki az ismeretek reprezentálására. Ahhoz, hogy a PLEXUS „megértse" a természetes nyelven feltett kérdéseket, szemantikai kategória-együttest hoztak létre, amelyek hierarchikus alkategóriákra oszlanak. Ezek a kategóriák képviselik e szakterület releváns fogalmait. Szótárt építenek a használói input követésére, a kifejezések osztályozására, szabályozására, és szemantikai kategóriákba sorolására. A rendszer képes a szinonimák és homográfok kiszűrésére. A kifejezések osztályait összekapcsolják. Komplex hálózatot hoznak létre. Több szemantikai kontextust létesítenek minden kifejezéshez, még akkor is, ha nincs lehetőség kimerítő felsorolásra. Összefoglalva, a PLEXUS gazdag tudásanyaggal és kifinomult szemantikai hálózattal rendelkezik.

Egy másik érdekes tájékoztatási szakértői rendszer, amelyet nem tartalmaz az *Information Processing and Management* 1987-es kiadása, az ANSWERMAN, amelyet a National Agricultural Library fejlesztett ki. Ez segíti a használót a mezőgazdasági kérdésekre keresett válaszokat tartalmazó kézikönyvek megtalálásában.

A szakértői rendszerek harmadik típusa az intelligens keresőinterfész. Számos ilyen rendszer működik. A kereskedelmi forgalomban kapható gateway

szoftverektől (PRO–CITE, SCI–MATE stb.) a kidolgozottabb CONIT-ig terjed a skála, amelyet az alábbiakban ismertetünk. Egyik sem tekinthető valóban *intelligensnek.* Az on-line keresés bonyolult mentális folyamatokat és feladatokat tételez fel. Ezek közül többen megkísérelték rendszerükbe beépíteni a kereső stratégiák meghatározott alhalmazait. A CONIT (COnnected Network for Information Transfer) automatikus keresésközvetítő. *Marcus* fejlesztette ki a MIT-nél többéves munkával. A rendszerfejlesztés elején járva automatikusan tud különböző keresőrendszerek számára logon eljárásokat végezni. Ezek után *Marcus* az adatbázis-kereső képességeket fejlesztette bármely témában. Bár nem igazán szakértői kereső, a munka elmélyítette a keresési folyamat összetettségének megértését.

Amint a MI kutatás a következő szakaszába ér, számos probléma merül fel. A MI közösségben egyre jobban átlátják az emberi szakértők ismereteinek és tapasztalatának megragadásában felmerülő nehézségeket. Az ismeretszerzés nehéz, kevéssé megfogható és költséges folyamat. A MEDINFO 86 (Fifth Conference on Medical Informatics) bevezető előadásában *Edward A. Feigenbaum*, a Stanford University elismert MI kutatója azt jósolta, hogy az ismeretszerzés és az ismeretek reprezentálása lesznek a MI kutatás következő témái. Megjegyezte, hogy a kísérletek megmutatták, hogy sikeres szakértői rendszerek építhetők. Jelenleg ismert, hogy nem az interfészeszköz, hanem az ismeretanyag az, amely meghatározza a szakértői rendszer minőségét és erejét. Ezt nevezte ismeretelvűségnek. Egyszerűen így fogalmazott: „Ha azt akarjuk, hogy egy program jól működjék, sokat kell tudnia a ‚világról', amelyben működik. Ismeretek hiányában mit sem érnek a következtetések.". Azt várta, hogy az ismeretrendszerekben közhely lesz a rendszer és a használó közötti természetes nyelvű, szóbeli kommunikáció – a klaviatúrán keresztüli input mellett –, mivel jól ismertek azok a jellegzetességek, amelyek a rendszer használata során vonzóak a használók számára. Álláspontja szerint a valódi kihívás annak megértése, hogy az emberek hogyan szereznek ismereteket, és hogyan strukturálják az információt az ismeretszerzés javítása érdekében egészen addig a pontig, amelyen a realisztikus ismeretanyag eléri a szakértői színvonalat. Az ismeretek reprezentációja és szerkezete az információkeresés kutatói számára is központi kérdés. Azt találták, hogy az információ reprezentálása és strukturálása szükségképpen dinamikus szemantikai hálózatok kialakításához vezet, amelyekben az egymással kapcsolatban álló fogalmak és a kifejezések között kapcsolatok épülnek ki. Mivel az információkeresés kutatása során komoly tapasztalatokra tettek szert a természetes nyelvek feldolgozásában, az ismeretanyag és a szemantikai háló komplex vizsgálata igen érdekes az informatikusok és a szakértői rendszerekben keresők számára.

Az értékelés másik fontos terület. Bár a rendszerek értékelésével foglalkoztak a kutatók, még korántsem oldottak meg minden kérdést. A legfőbb

probléma, hogy mi az a mérce, amelyhez a szakértői rendszer teljesítményét viszonyítani lehet? Másként megfogalmazva, normalizálható-e a szakértők magatartása?

A könyvtári informatika alapszintjén az informatikusok mostanában is birkóznak az információfogyasztót körülvevő problémákkal. A pszichológiából és a kognitív tudományokból származó ismereteket szabadon használják annak megértésére, hogyan fogalmazódnak meg a kérdések és mi megy végbe az emberi elmében a kérdést pontosító megbeszélés folyamán. A gyakorlat szintjén kezelhető problémákat megoldják. Ezt példázza azoknak a jellemzőknek a vizsgálata, amelyek hozzájárulnak az on-line rendszerek felhasználóbarát jellegének fokozásához. E vizsgálatok alapján lépnek előre az on-line rendszerek tervezésében és konfigurálásában.

A szakértői rendszerek egy-egy szakterületre korlátozódtak. De az információkeresés nem korlátozódik egyetlen témára vagy kontextuális területre. A tervezési problémák húsba vágóak, mert ezen a területen a legfrissebbek az alkalmazások. Mások szerint a MI eljárásainak nem szakterületre korlátozott alkalmazása valószínűleg hosszú távú cél lehet csak, még ha keresztül vihető is, mert a piac még nem elég erős a hatalmas és nagy bonyolultságú kiegészítő feldolgozó rendszerek jelentős fenntartási költségének fedezésére. Ismét mások meg vannak győződve arról, hogy lehetetlen szakértői keresőrendszereket kifejleszteni. Az on-line keresésben meggyőző bizonyítékok vannak arra, hogy különböző szakértők eltérő keresési eredményeket produkálnak. Régebben a szakértői rendszerek olyan területekre készültek, amelyeken emberi szakértők is vannak. Ha a keresési szakemberek nem tudnak megegyezni az adott keresésben követendő stratégiát illetően, nehezen lehet megragadni magát a szakértelmet.

Jelenleg köztes megoldásokkal próbálják optimalizálni a keresőrendszereket. Több rendszert ruháztak fel valamilyen „intelligens” jellemzővel, amely az on-line keresőrendszer használóját támogatja. A CIT/NTM az egyetlen működő rendszer, amely fejlett eszközökkel támogatja a keresőt. Elkészíti a kifejezések on-line listáját azoknak a keresőkifejezéseknek a szótöve alapján, amelyeket a használó bevitt, és további szintaktikai elemzést is végez, hogy hasznos keresőkifejezéseket tudjon megjeleníteni. A kereső ezután felhasználhatja ezeket a kérdés átfogalmazásához, hogy jobb eredményeket érjen el. A PaperChase rendszer felhasználóbarát interfésszel támogatja a gyakran kért folyóirat-hivatkozások részhalmazának keresését. Számos, a keresést segítő tulajdonságot építettek be a rendszerbe. A syracuse-i SIRE keresőrendszer a hivatkozásokat valószínűsített relevanciájuk alapján rangsorolja. A CONIT (COnnected Network for Information Transfer system) automatizált keresést közvetítő rendszer, amely hasznos on-line feladatok hosztjaként működhet.

Pao eljárást alakított ki a kifejezésekkel történő keresés és a hivatkozás keresés összekapcsolására, így kívánván javítani a teljességet és a pontosságot. A meglévő keresési jellemzőket még további kiegészítésekkel optimalizálják, és így javítják a mai rendszerek keresési eredményeit.

Végül *Feigenbaum* víziójának ködös messzeségeiben az információkeresés a „jövő könyvtára":

> *„Most képzeljük el a könyvtárakat aktív, intelligens »ismeret-szervernek«. Ez a ‚szerver' összetett ismeretstruktúrákban tárolja a tudományok ismeretanyagát (talán olyan formában, amelyet még fel sem fedeztek). Képes az ismereteket felhasználni következtetésekre, ha a használók igényei így kívánják. Ezeket az igényeket természetesen élő beszédben fejezik ki. S a rendszer természetesen keres és megmutat (elektronikus tankönyv). Összegyűjti a releváns információt; összefoglal; kapcsolatok mentén halad tovább. Meghatározott kérdésekben tanácsadóként működik, bizonyos megoldási módokat javasol, és ezeket a megoldásokat hivatkozásokkal vagy érvekkel támasztja alá. Ha a használó áll elő megoldási javaslattal vagy hipotézissel, azt ellenőrzi, sőt, továbbfejleszti. De bírálhatja is a használó felfogását egyetértésének vagy egyet nem értésének részletes kifejtésével."*

INFORMÁCIÓKERESÉS AZ INTERNETEN, AVAGY A VILÁGMÉRETŰ HOZZÁFÉRÉS A TÖMEGEK SZÁMÁRA

Az adatbázisokkal, akárcsak a hagyományos könyvtári katalógusokkal nemcsak a szakembereknek, hanem az alkalmi felhasználóknak is boldogulniuk kell. Az on-line információkereső rendszerben a felhasználónak nem kell ismernie magát a keresőrendszert, ahogy a személyautó vezetőjének sem kell értenie a járműve szerkezetéhez. Olyan felhasználói felületet kell létrehozni, amelynek szemiotikai struktúrája az adott kultúra megszokott, mindennapi jelrendszerének felel meg, azaz a természetes, laikus gondolkodásnak. A felhasználónak erre a „végfelületre" (end-user interface) van szüksége, és ha a szolgáltatók ezt biztosítják, akkor a könyvtártudomány negyedik ranganathani törvényének tesznek eleget: **„Kíméljük az olvasó idejét!"**

Az idevezető fejlődés már közvetlenül a háború után elkezdődött (lásd a kötet „Az információkereső gondolkodás kezdetei" című bevezető fejezetét) és a hipertext (hypertext) feltalálásán keresztül vezetett el az interneten megvalósult globális információkereséshez. (A fejlődést kizárólag a tartalom szerinti – például jelentéssel bíró szavak alapján végzett – információkeresés szemszögéből tárgyaljuk, és ezért a hálózati rendszerekkel és az internet egyéb vonatkozásaival nem foglalkozunk.)[1]

Vannevar Bush, aki a háború alatt az amerikai tudósok „hadseregét" irányította, és ezzel a győzelem egyik fontos, bár jobbára ismeretlen alakjává vált, 1945-ben fogalmazta meg először, hogy az információkeresés folyamatának (akkor még ezt a kifejezést – information retrieval – nem használták) az asszociatív kapcsolatokon kell alapulnia. Az „Úgy, ahogy gondolkodunk" és az „Endless horizons" („Végtelen láthatárok"), majd a húsz év múlva újrafogalmazott „Memex revised" („Módosított Memex")

1 Az internetre vonatkozó fontosabb testületi forrásokat, kronológiát stb. az OMIKK Virtuális Könyvtárának internetoldalai tartalmazzák: Az OMIKK Virtuális Könyvtára [on-line]. Szerk. Válas Gy., Horváth P. [1999.08.16.] <http://www.omikk.hu/omikk/virkonyv/inet.htm>

című tanulmányaiban ő használta először az összekapcsolt szövegblokkok fogalmát, ő vezette be a „link" – ebben az esetben a releváns szöveghelyekre utaló egyszerű kapcsolatjelölő (csatoló, kapocs, utaló, mutató, hivatkozás, ugrópont) – és a nyomvonal, valamint a háló kifejezéseket a textualitás új elképzelésének a leírására.[2] Koncepciója a gépesített, határtalan kapacitású, mindenféle dokumentumokat tartalmazó iratgyűjtemény és könyvtár, amely a felhasználó számára gyors, asszociatív keresést biztosít. A fél évszázada megálmodott elektronikus, hálózati könyvtár feltételei napjainkra értek meg.

Munkássága nagy hatással volt *Douglas Engelbartra* (az egér és az ablaktechnika feltalálójára) és a hipertext későbbi úttörőire, mint *Theodor Holm Nelsonra,* és a Brown University Információ- és Tudománykutató Intézetének (Institute for Research of Information and Science; IRIS) kutatócsoportjára, az Intermedia megalkotóira. *Nelson* eszméjének lényege, hogy a lineáris szövegfolyamon belül kisebb szövegrészeket kapcsolt össze. Ezek a kapcsolatok a szöveget keresztül-kasul behálózták, az olvasó maga határozhatta meg, milyen legyen az általa tanulmányozott szöveg szerkezete. Megszületett a nem-lineáris szöveg eszméje. *George P. Landow,* a Brown University angol irodalom és művészettörténet professzora a hipertext és az internet keletkezéstörténetével foglalkozó művében a következőképpen világítja meg a hipertextes, „középpont nélküli" technikának a gyökereit:

„Amikor az olvasók szövegek hálójában vagy hálózatában haladnak előre, folyamatosan változtatják kutatásuk vagy tapasztalatuk középpontját – s ezáltal a fókuszt vagy a szervező elvet is. Más szóval, a hipertext olyan korlátlanul újra középpontozható rendszerként szolgál, melynek ideiglenes fókuszpontját az olvasó jelöli ki, akiből ennek ellenére más értelemben válik valódi aktív olvasó. A hipertext egyik alapvonása, hogy egymással összekapcsolt (*Roland Barthes* által lexiáknak nevezett) szövegtestekből áll, melyek nem egyetlen főszervező tengely mentén kapcsolódnak. Más szóval, a metaszövegnek vagy dokumentumsornak – annak a dolognak, entitásnak, mely a nyomdai technikában meghatározza a könyvet, a művet vagy a szöveget –

2 As we may think [Úgy, ahogy gondolkodunk]. In: Atlantic Monthly, 1945, July, No. 176, p. 101–108. Magyarul: Út az új gondolkodás felé. In: Klaniczay Júlia, Sugár János: Hypertext + multimédia. Oktatási segédanyag. – Budapest: Artpool, 1996. p. 3–14.

Endless horizons – Washington: Public Affairs Press, 1946.

Memex revised. In: Sciences is Not Enough. – New York: Villiam Morrow, 1967. p. 75–101. továbbá <http://www.csi.uottawa.ca/ ~dduchier/misc/vbush/awmt.html>

További, az Internet történetével foglalkozó források: Hobbe's Internet timeline, v1.1 [1994.03.12.]: <http://info.isoc.org/guest/zakon/Internet/History/html>, valamint Kristula, D.: The history of the Internet [Az Internet története] 1997: <http://www.davesite.com/webstation/net-history.shtml>

nincs középpontja. Igaz ugyan, hogy a középpont hiánya problémát okozhat az írónak és az olvasónak is, ám a hipertextet használva mindenki saját érdeklődését teszi meg kutatása pillanatnyi de facto szervező elvének (vagy középpontjának). A hipertextet olyan rendszerként tapasztaljuk meg, mely korlátlanul középpont nélkülivé tehető és újra középpontozható részben azért, mert a hipertext átmeneti középponttá, a tájékozódást és a továbbhaladást segítő könyvtári katalógussá alakít bármely dokumentumot, mely egynél több kapcsolódással – csatolóval (hiperlinkkel) – rendelkezik.

A nyugati kultúra jóval a számítástechnika előtt ismerte már a hálózatba kapcsolt valóság félig-meddig mágikus kapuit. A bibliai tipológia, mely olyan fontos szerepet játszott az angol kultúrában a XVII. századtól a XIX. századig, a krisztusi elrendelés típusainak és előjeleinek kategóriáiban gondolta el a bibliai történetet. Vagyis Mózes, aki a saját jogán létezett, létezett Krisztusként is, aki beteljesítette a próféta jövendölését. Számtalan XVII. századi és viktoriánus prédikáció, traktátus és szövegmagyarázat demonstrálja, hogy bármely személy, esemény vagy jelenség mágikus ablakként szolgált az emberi üdvözülés isteni rendjének összetett szemiotikájában. A jelentős eseményeket és jelenségeket egyidejűleg több valóságban vagy valóságszinten megjelenítő bibliai típushoz hasonlóan az egyes lexiák is szükségszerűen utat nyitnak a kapcsolatok hálózatába. Feltéve, hogy az evangélikus protestantizmus Amerikában megőrzi és továbbfejleszti a bibliai szövegmagyarázatnak ezt a hagyományát, cseppet sem meglepő, hogy a hipertext első alkalmazásai között ott volt a Biblia és az exegetikai tradíció. […]

Valamennyi hipertextrendszer lehetővé teszi, hogy az olvasó maga válassza ki a kutatás vagy a tapasztalat középpontját. A gyakorlatban ez az elv azt jelenti, hogy az olvasó nincs bezárva semmiféle szerkezetbe vagy hierarchiába."[3]

Theodor Holm Nelson 1965-ben írta le a hipertext nevet és határozta meg – a felhasználó szemszögéből – a fogalmát:

„Írott vagy képi anyagok olyan komplex összeköttetése, amit papíron nem lehet kényelmesen megalkotni. Összefoglalókat és térképeket tartalmazhat a benne szereplő anyagokról és ezek egymáshoz való viszonyáról; és tartalmazhatja az anyaggal foglalkozó tudósok megjegyzéseit és lábjegyzeteit is."[4]

---

3 George P. Landow: Hypertext and critical theory [Hypertext és kritikai elmélet]. New York: The Johns Hopkins University Press, 1992., továbbá <http://twine.stg.brownedu/projects/hypertext/landow/ht/contents.html>

Magyarul: Hypertextuális Derrida, posztstrukturalista Nelson? In: Klaniczay Júlia, Sugár János, id. mű. p. 28–29.

Ebben a hézagpótló szöveggyűjteményben Vannevar Bush, Theodor Holm Nelson és a hipertexttel foglalkozó számos más szerző tanulmányának fordítása található.

4 Nelsont idézi Klaniczay Júlia, Sugár János, id. mű. p. 56.

Nelson egyben elkezdte a Xanadu nevű, máig meg nem valósult, maximalista hálózati hipertextrendszerének és az általa Egységes Adatstruktúrának (Unified Data Structure) nevezett formátumnak a tervezését is. Ahogy *Vannevar Bush* a felhasználóbarát számítógépes végfelületek, az ablaktechnika és a hipertext feltalálóinak körét, úgy *Nelson* Xanadu terve és Egységes Adatstruktúrája programozók kis, de lelkes körét befolyásolta tartósan. (Áttételesen még a bibliográfiai adatcsere-formátum létrehozását is inspirálhatta.)[5]

A Xanadu név Coleridge egyik költeményéből származik: az „irodalmi emlékezet mágikus helyét" jelöli, ahol minden megőrződik. Ahogy *Sugár János* fogalmazott: „A névválasztás is jelzi a hipertext eredendő és mély irodalmi gyökereit. Talán a hipertext az első civilizációs vágyálom, melyet a rohamosan fejlődő számítástechnika valósít meg." Figyelemre méltó, hogy csak amikor az internet hálózata kialakult, kerülhetett sor a hipertext Nelson által elképzelt alkalmazására.

Miközben Nelson a Xanadu megalomániás tervét kergette, Andries van Dam 1967-68 között elsőként ténylegesen működő hipertextrendszert készített. 1969-ben pedig az Egyesült Államok hadügyminisztériumának rendelésére megszületett az ARPANET (Advanced Research Project Agency Network), az internet őse. Annak érdekében hozták létre, hogy atomcsapás esetén se szakadjon meg a kommunikáció az amerikai kormány- és katonai szervek között. Az internet elve egyszerű: a hálózatnak nincs központja, a részei egymástól függetlenül működhetnek, mivel minden csomópont egyenrangú. A megcímzett adatcsomagok útja teljesen közömbös, csak az eredmény számít: a csomagok csomópontról csomópontra vándorolnak, míg el nem érik a címzettet. Hiába semmisült volna meg számtalan csomópont, a küldemények a háló megmaradt csomópontjait érintve járhatták az útjukat. Mivel idővel egyre több nem katonai intézmény is csatlakozott a hálózathoz, a nyolcvanas évek elején katonai részét különválasztották, és ami megmaradt, ahhoz teljesen szabaddá tették a csatlakozást. Az internet diadalútja elkezdődött.

A nyolcvanas évek végére már csak a könnyen kezelhető, grafikus kezelőfelület hiányzott, olyan felhasználóbarát „műszerfal", mely a legostobább végfelhasználó számára is lehetővé teszi az interneten a keresést. Ekkor jelent meg a színen *Tim Berners-Lee*, aki 1989-ben az Európai Részecskefizikai Laboratóriumnak (CERN) – saját bevallása szerint – a Xanadu inspirációjára javasolta a World Wide Web tervét. (A Xanadu programot viszont, mint annyi más úttörő, számos irreális vonást tartalmazó kezdeményezést, 1992-ben, miután közel 5 millió dollárt költöttek rá, az AutoDesk

5 Nelson életútját és a Xanadu történetét részletesen leírta Szakadát István: Xanadu. A Xanadu és Ted Nelson. <http://www.uniworld.hu/netskills/tudas/HTML/Xanadu.htm>

Company feladta.)[6] Akárcsak *Nelson, Berners-Lee* sem gazdagodott meg találmányából, mivel szabadalmi és kopirájtigényeinek bejelentését mellőzve, eredetileg csak a tudományos közösség számára akart olyan eszközt létrehozni, mellyel a hipertextes közleményeket korszerű szerkezetben lehet megjeleníteni és olvasni az interneten.

Az internet jelentősége, hogy távolsági on-line hozzáférést biztosít a laikus „tömegek" számára. Ebből nem következik, hogy nincs már szükség a speciális szakterületekre vonatkozó, elsősorban pénzért szolgáltató on-line adatbázisokban végzett kereséskor a részletesebb információkereső szakmai ismeretekre. Mint minden fejlődésben, itt is differenciálódás játszódik le: az internettel a használat újabb szintje jelent meg anélkül, hogy a korábban kialakult használati módok érvényüket vesztenék. Ahogy nem szűnik meg a nyomtatott dokumentumok használata sem az elektronikus dokumentumok megjelenésével (ennek ellenkezőjét legfeljebb az internet terjedésében érdekelt nyomásgyakorló csoportok tagjai terjesztik).

Mivel az internet több szakterület (adatátvitel, programozás, katalogizálás, osztályozás, információkeresés) metszéspontjában fekszik, melyeknek mind önálló szakmai nyelve van, az internettel átfogóan foglalkozó szakirodalomban az egyes szakterületek terminológiáját olykor felszínesen vagy önkényesen használják. A könyvtártudományban és dokumentalisztikában, később meg az on-line információkeresésben történetileg kialakult terminológia értelemszerűen fölhasználható az interneten végzett tartalmi feltárásra és keresésre, de ez csak lassan valósul meg.[7]

Az internet méretei

Kiszámították, hogy két, a jelenleg 800 millió weboldal közül kiválasztott tetszés szerinti HTML-dokumentum legfeljebb 19 hivatkozási ugrásnyira van egymástól. Bármit keresünk is a csatolók (hivatkozások, hiperlinkek) segítségével szörfölve, az átlagosan nincs messzebb mint 19 csatolóról csatolóra megtett lépés – legalábbis statisztikai szempontból. A számítógépek, melyek az internetet fenntartják, olyan szorosan összekapcsolódtak már, hogy a 800 millió dokumentum közül még a legtávolabbi is rövid idő alatt elérhető.

6 Szerényebb, önállósult keretek között ma is működik. Lásd: Project Xanadu. Founded 1960. The original hypertext project. [199.02.02.] <http://www.xanadu.net>

7 Az angol–magyar terminológiáról Drótos László állított össze a teljesség és véglegesség igénye nélkül szótárt. Drótos László: Elektronikus könyvtári értelmező szótár. 1998. <http://www.lib.uni-miskolc.hu/publ/ekszotar>

A bonyolult topológiájú véletlen hálózatok meglehetősen hétköznapiak a természetben, és segítségükkel olyan eltérő rendszerek is egyformán modellezhetők, mint az World Wide Web vagy a társadalmi és gazdasági rendszerek. Újabban az is kiderült, hogy a legtöbb rendkívüli nagyságú hálózat topológiai információi skálafüggetlen jellemzőikkel írhatók le. Megvizsgáltuk ezeknek az újabban ismertté vált skálafüggetlen modelleknek a skálatulajdonságait, melyek az említett rendszerek kisenergiájú eloszlásokon alapuló összefüggőségét (konnektivitását) megmagyarázhatják. A jelentésmező elméletet alkalmazva előre jelezhettük a gráf csúcsainak növekedési dinamikáját, és kiszámíthattuk a jelentésmező-elmélet alapján az összekapcsoltság eloszlását és a mértékfüggvényt. Az eredményeket a weben próbáltuk ki. […]

Olyan robotprogramot készítettünk, mely lényegében a web egy részét feltérképezte. Első lépésben adatbázisba gyűjtötte az egyes honlapok csatolóit, majd követte azokat a hivatkozott honlapokon és ezt folytatta. A kapott adatokat statisztikai módszerekkel értékeltük: meghatároztuk annak valószínűségét, hogy a vizsgált dokumentumra megadott számú honlapról hivatkoznak, illetve e honlap ugyanannyi oldalra hivatkozik. A véletlen hálózatokra jellemző valószínűségi eloszlást vártunk. Ez azt jelentette volna, hogy a legtöbb honlapon mondjuk 10–20 csatoló lesz majd más weboldalakra. Mi voltunk a legjobban meglepve, amikor ehelyett egészen mást találtunk, azt, hogy a csatolók számának eloszlása hatványfüggvény, ami önszervező rendszerekre jellemző, és arra utal, hogy sok honlap van, amelyről több ezer csatoló ki, és ugyanakkor rengeteg olyan honlap van, amelyre hihetetlenül sok más honlap mutat. Noha bármely weboldal-tulajdonos teljesen szabadon döntheti el, hány csatolót helyez el a honlapján, a teljes hálózat mégis általános törvényszerűségnek engedelmeskedik. […]

Az eredmények a keresőrendszerek tervezésében hasznosíthatók. A mai keresőprogramok helyett idővel talán kidolgozhatnak az új felfedezésen alapuló, intelligensebb keresési módszert, amennyiben kihasználják a háló összefüggőségét, és a felhasználó által kívánt információt az azonosított, legfeljebb tizenkilenc csatolót végigkövetve keresik meg.[8]

> A vizsgálatokra használt program tehát adott HTML-dokumentum összes csatolóiból kiindulva addig követte az adódó csatolókat, ameddig csak újabbat talált. A folyamatot rendkívül sokszor megismételték, majd statisztikai módszerekkel kiszámították két HTML-dokumentum között az átlagos csatolóugrások számát. Ez a háló két pontja közötti átlagosan legrövidebb „távolság", melyet a háló „átmérőjének" neveztek el. 800 millió HTML-dokumentum esetén eszerint két véletlenszerűen kiválasztott pont közötti átlagos távolság 19 hivatkozás.

8 Barabasi, A-L., Albert, R., Jong H.: Diameter of the World Wide Web [A World Wide Web átmérője]. In: Nature, 1999. 401. sz. szeptember 9, p. 130–130.

Ha a háló a jelenlegi méreteinek tízezerszeresére növekszik, akkor a hatványtörvény megmondja, hogy legfeljebb huszonegy csatolóra lesz szükségünk egy honlapról tetszőleges másik honlap elérésére.
Az eredmény egyben megvilágítja a web relatív kommunikációs korlátait. Az emberiség ugyancsak önszervező rendszernek tekinthető, melynek „összekapcsoltsága" kiszámítható. Barabási utal rá, hogy a ma élő hatmilliárd emberre vonatkozóan ezek a számítások kimutatták: legfeljebb hat lépésben minden embernek más embereken keresztül kapcsolata van bárkivel a világon. Azaz mindenki ismer olyan embert, aki megint csak ismer olyan embert stb., és a sorban a hetedikhez jutva az emberiség bármelyik tagjával áttételesen kapcsolatban vagyunk. A web kommunikativitása tehát az emberiségéhez képest alig harmadannyi.

A tartalom szerinti információkeresés az interneten

A keresőrendszerek története

Az internet forrásainak eléréséhez kezdetben meglehetősen körülményes, a laikus felhasználó számára nehezen vagy alig használható eszközöket alkalmaztak. Csak arra voltak jók, hogy a kapcsolat lehetőségére a szabványos feltételeket biztosítsák és elvégezzék az indexelést.

– Az FTP (File Transfer Protocoll), az adatátvitel általános szabványa, s egyben program biztosítja, hogy a hálózat számítógépei között egységes formában jöjjön létre a kapcsolat.
– A WAIS (Wide Area Information Servers), az Archie, ill. a megfelelő későbbi programok (pl. Apache, INQUERY, Harvest) hozták létre a másodlagos adatállományokat (indexeket, relevanciaadatokat és leírásokat), melyek az internetre kerülő dokumentumokra utalnak. Ezek a szerveroldali indexelőrendszerek.

Mivel a háttérben működnek, démonoknak is nevezik őket. A használatuk bonyolult, és hiányzott mögöttük az intézményes háttér. Az áttörés két olyan rendszer megszületéséhez fűződik, amelyek bizonyos szempontból homlokegyenest az ellentétei egymásnak.

– 1991-ben először a Gophereket készítették. Ezek a szöveges, menüszerkezetű információs hálózatok leginkább a hierarchikus felépítésű tartalomjegyzékekhez hasonlíthatók. A Gopherben a fölhasználó szöveges (és csak szöveges) állományokat nézhetett meg és tölthetett le. Legismertebb keresőeszköze a Veronica integrált menülekérdező és indexelőrendszer (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives).

– Még ugyanebben az évben, *Tim Berners-Lee* kísérleteiből kiindulva megszületett a World Wide Web (web, WWW) hipermédia információs hálózat üzemszerű formája. Ebben a rendszerben a hipertext jóvoltából az információforrásokat hipertext csatolók (hiperlinkek) formájában teljesen szabadon lehet egymással összekapcsolni. A kapcsolódó és megjeleníthető információforrások nemcsak szövegek, hanem képek, hangok és mozgóképek is lehetnek. A rendszeren belül nem érvényesül semmiféle hierarchikus rendező elv, minden forrás egyszerre több más forrással is összefügghet és fordítva (azaz a kapcsolódások szerkezete M : N). A weben belül hamarosan kialakultak különféle *keresőszolgáltatások,* mint például az AltaVista, HotBot, Infosec, Magellan, excite, ill. Yahoo!, Magyarországon a HUDIR, ill. a Heuréka és az AltaVizsla. A web formájában végre megvalósult az általánosan hozzáférhető és az asszociatív gondolkodással összhangban álló felhasználói felület, melyet 1945-ben *Vannevar Bush* megálmodott. Benne minden addigi egységesítő (FTP), feldolgozó és keresőeszközt (WAIS stb.) integráltak.

A versenyből a Gopherrel szemben néhány év alatt a web került ki győztesen, de a Gopher–WWW kettősség nem véletlen jelenség, hanem a rendezőrendszerek kettős természetével függ össze (részletesebben a dichotómiára a továbbiakban még visszatérünk).

Az internet fejlődését másik kettősség – a kereslet-kínálaté – is meghatározta. Ennek következtében alakult ki a háló, s ezzel a web két „oldala”: a tartalomszolgáltatás és a keresőszolgáltatás.

A tartalomszolgáltatás és a belépőoldalak

Az interneten nem volna mit keresni, ha nem lenne ***tartalomszolgáltatás.*** Ezen egészen általánosan az információ (az adatok) elhelyezését értjük az interneten, főleg hipertextes HTML-oldalak (HTML-dokumentumok) formájában.

A távoli hozzáférésű források egyre nagyobb része a http kommunikációs szabályai alapján elérhető dokumentum. Közöttük ma szinte kizárólagos szerepe van a html-formátum szerinti szerkezetű dokumentumoknak, noha elképzelhető, hogy a jövőben a http-n belül más (pl. xml) szerkezetű dokumentumok is el fognak terjedni. A többi kommunikációs protokoll és szabály szerint elérhető távoli hozzáférésű forrásoknak (pl. gopher-, telnet-, ftp-dokumentum, elektronikus levél) alig van jelentősége, és ezért a továbbiakban, ha az internet dokumentumairól van szó, csak html-dokumentumokról beszélünk.

A kisebb vagy nagyobb tartalmi egység kezdő- vagy belépőlapját honlapnak (home page, ottlap) nevezik. Lényegében mindenki – akarva-akaratlan – tartalomszolgáltató, akinek honlapja van az interneten (így a személyes honla-

pok tulajdonosa is); a *keresőszolgáltatások* programjai elkerülhetetlenül indexelik az elérhető HTML-dokumentumokat, nem téve különbséget a kereskedelmi céllal végzett tartalomszolgáltatás és az egyéb (pl. intézményi, személyi) honlapok között.

A gyakorlatban tartalomszolgáltatáson a kereskedelmi célú információelhelyezést értik, mellyel erre szakosodott cégek foglalkoznak. A tartalomszolgáltatásban fontos szerepet betöltő webhelyek kezdőlapjára külön elnevezések születtek (honlap, ottlap, hálószem). Ha sokféle elérhetőséget biztosítanak a kezdőlapon (elsősorban kereskedelmi, szolgáltató vagy legalábbis professzionális, intézményi célból), akkor portálról (portáloldalról, portálszolgáltatásról) beszélnek[9].

Léteznek tartalomszolgáltatók, akik a különféle keresőszolgáltatásokról tájékoztatnak, gyakran értékelve is ezeknek minőségét. Ezekből kiderül, hogy a keresőszolgáltatások száma a századfordulón több száz volt, és lehet, hogy előbb-utóbb megközelíti az ezres nagyságrendet. Ilyen szolgáltatás portálja látható az 1. ábrán.[10]

A rendszereket működtető fizikai berendezések a „helyek" (site). A webszervergép például webhely, és eme igazi helyen belül afféle virtuális „részhelyek" a gépen található információforrások (pl. adatbázisok, hirdetőtáblák, honlapok). A „tér" a hálózat, a „helyek" pedig a kiindulási, fizikai szinten a gépek, a további szinteken pedig a különféle „részhelyeket" képviselő webhelyek, ftp-helyek, hírcsoporthelyek stb., melyek az információforrások szerepét játszhatják. A hálózat gépeit és azokon belül az információforrásokat (a „helyeket" és „részhelyeket" a „térben") a szabványosított formában írt IP-cím, ill. az azt egyszerűbb, megjegyezhetőbb formában kifejezett név, az URL (Unified Resource Locator) jelöli. Egy szervergépi helyen általában nagyon sok „részhely" (tartalomszolgáltató, honlap, azaz weboldal) található.

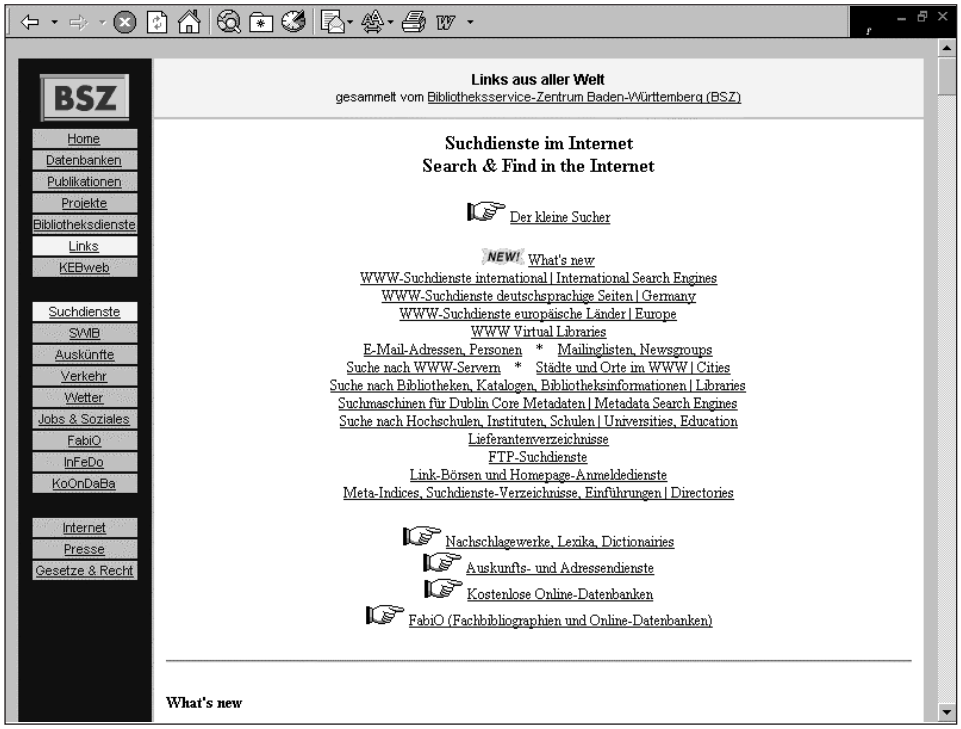
Keresőszolgáltatások és a rendezőrendszerek kettőssége

Meghatározás

Az elsődleges adatokat tartalmazó dokumentumokat csak a részben belőlük nyert (pl. cím, kiadó), részben intellektuálisan megállapított (pl. besorolási adatok, deszkriptorok, jelzetek) másodlagos adatok alapján lehet ke-

9 Ilyen például a Hungary.Network egyszerű, jól áttekinthető HUDIR portálja, <http://www.net.hu>, az AltaVizsla indexelőszolgáltatását is magába foglaló Origo összetett felépítésű portálja, <http://www.altavizsla.origo.hu>, és a Yahoo! 5. ábrán látható portálja.

10 A Configurable Unified Search Engine például csoportosítva és magyarázattal ellátva adja meg a különféle típusú keresőszolgáltatásokat, de még szótárakat is, és belőle kiindulva bármelyikben keresni lehet. <http://www.unix-ag.uni-siegen.de/search/#mp3>



1. ábra. Keresőszolgáltatások különféle típusairól tájékoztató, részben kétnyelvű tartalomszolgáltató portálja

reshetően tárolni. A másodlagos adatok egy-egy dokumentumra vonatkozó összessége alkotja a dokumentumleírást vagy dokumentumrekordot, könyvtári szabványoknak megfelelő formájuk a bibliográfiai tétel. Az interneten is meg kell különböztetnünk az elsődleges dokumentumok szerepét játszó HTML-oldalakat és a rájuk vonatkozó másodlagos vagy metaadatokból felépülő HTML-rekordokat. HTML-rekordon tehát a HTML-dokumentumról készült, a dokumentum másodlagos adatait tartalmazó információtételt értjük. Ez utóbbiakat az internetes keresőszolgáltatások hozzák létre annak érdekében, hogy a felhasználók keresni tudjanak. A HTML-dokumentumokhoz a hozzáférést biztosító keresőszolgáltatások jelentős része a hirdetésekből tartja fenn magát, és nagy részük ingyenes, kisebb részük használatáért (pl. az OCLC NetFirst) fizetni kell. A hirdetések a szolgáltató lapjain megjelenő csatolókon keresztül érhetők el; ezeket kiválasztva jut el az érdeklődő a hirdetés feladójának HTML-oldalára. Mennél többen használják az adott keresőszolgáltatást, mennél többen keresnek a segítségével, annál jobban vonzza a hirdetőket, annál több a jól fizető hirdetés. A szolgáltatások fejlődését ma elsősorban ez a

tény határozza meg. A keresőszolgáltatások érdekeltek abban, hogy a felhasználók mennél könnyebben és eredményesebben kereshessenek, ezért a választék növelése érdekében óriási adatbázisok alakultak ki. Így érik el, hogy sokan használják őket, ami kihat a hirdetések számára. Mindez emlékeztet a sajtó világára.

Tágabb értelemben keresőszolgáltatások a webinterfésszel rendelkező on-line nyilvános adatbázisok is. Ezek elsődleges dokumentumai nem HTML-formátumúak, a szolgáltatáshoz dinamikusan lefordítják őket erre a formátumra. Általuk valójában távolsági on-line információszolgáltatás valósul meg az interneten (a nagy on-line szolgáltatók webopciót biztosítanak a használatukhoz, mint amilyen például a DIALOG Web[11]). Az adatbázisok ezáltal a klasszikus (telnetes) adatátviteli hozzáférés mellett, ill. helyett a weben keresztül is elérhetők. E szolgáltatások az osztályozás és az információkeresés szempontjából változatlanul „hagyományosak", és többnyire nem ingyenesek. Túlnyomórészt ellenőrzött információkereső-nyelvi szótárt, (tárgyszójegyzéket vagy tezauruszt, ill. osztályozási rendszereket) használnak bennük, az információk dokumentációs egységek (dokumentumok leírásai), a tartalmi feltárást intellektuálisan végzik, és az alkalmazott dokumentációs adatbázis-kezelő rendszerek jóvoltából a keresési lehetőségek sokkal fejlettebbek, mint az internetes kereskedelmi keresőszolgáltatások kizárólag indexekre vagy katalógusokra alapozott keresési lehetőségei. (A távolsági on-line információkeresést az „On-line információkeresés elterjedése és a kézikönyvek" című fejezetben tárgyaltuk.)

Ugyancsak tágabb értelemben keresőszolgáltatásnak tekinthetők a speciális forrásokat egyetlen adatbázisból szolgáltató rendszerek, mint a webtelefonkönyvek, elektronikus menetrendek, elektronikus postai és egyéb címtárak stb., elektronikus szótárak, hirdetések, üzleti információk, humorlapok stb. Ezek sem HTML-dokumentumokat szolgáltatnak, hanem tényadatokat (faktografikus információkat). Jelentős részüket a globális keresőszolgáltatások segítségével is le lehet kérdezni. (Részletesebben a „Speciális adatbázisok" című fejezetben foglalkozunk velük.)

A továbbiakban csak a HTML-dokumentumokhoz való hozzáférést biztosító keresőszolgáltatásokkal foglalkozunk. Ezekben az osztályozás és információkeresés szempontjából jelentős új fejlemények figyelhetők meg.

A gyűjtőkör szempontjából a szolgáltatások többsége ***globális,*** azaz – legalábbis elvileg – az egész háló a gyűjtőkörre (pl. AltaVista, Yahoo!), kisebbik része ***nemzeti (állami), regionális vagy meghatározott nyelvre*** korlátozza a gyűjtőkörét (pl. AltaVizsla, HUDIR). Egy részük ***egyetemes,*** azaz minden fajta és mindenféle tartalmú HTML-dokumentum a gyűjtőkörükbe tartozik, másik részük csak ***speciális*** tartalmú (pl. a WWW Women csak

11 DIALOG Web: <http://dialog.krinfo.com>

női tárgykörű), vagy speciális típusú (pl. a MusicSearch csak zenei) HTML-rekordokat szolgáltat. Az utóbbihoz tartoznak a szakterületi információs kapuszolgálatok (subject based information gateways) is.
Egyes szolgáltatások a gyorsaságukkal, mások a keresési eszközök gazdagságával, megint mások a feldolgozott állományuk nagyságával tűnnek ki. Vannak a relevancia szempontjából megbízhatóbbak és kevésbé megbízhatóak. Mindez az erős versenyben állandóan változik. Gyakran jelennek meg elemzések a hálón, melyekből tájékozódni lehet az aktuális helyzetről, de még nem alakultak ki megbízható tudományos módszerek az értékelésre (a hatvanas évek elején a hagyományos információkeresés hatékonyságára vonatkozó cranfieldi vizsgálatokhoz hasonló jelentőségű elemzések még váratnak magukra – lásd e kötetünkben az „Információkeresés értékelése" című fejezetet).

Tájékozódás arról, milyen keresőszolgáltatások léteznek?

Bármelyik nagyobb keresőszolgáltatásnak feltehetjük a kérdést, hogy hol található értékelés a keresőgépekről? Az AltaVistának például megadhatjuk az „*evaluation of search engines"* vagy „*Bewertung von Suchmaschinen"* láncot, ill. összetett kereséshez a *(„search engines") AND evaluation* vagy *Suchmaschinen AND Bewertung* keresőkérdést, és válogathatunk az információk között. A szolgáltató rendszerek minőségéről mindig akadnak naprakész vizsgálatok, melyeket az interneten publikálnak.

A szolgáltatások közötti nagy különbségek miatt nem szerencsés rangsorolni a teszteredményeket. A különféle keresési célokra különféle induló szolgáltatások vehetők igénybe. A MetaCrawler működtetői által végrehajtott vizsgálat azt jelezte, hogy pillanatnyilag egyetlen nagyobb keresőszolgáltatás sem képes a források 45%-ánál többet feltárni. Alig akad tehát olyan kérdés, melyre egyszerre több szolgáltatással végzett keresés nélkül érdemben válasz kapható.[12]

A rendezőrendszerek kettőssége az interneten

Az információkeresés és osztályozás szempontjából a keresőszolgáltatások két fő típusa alakult ki: az ***indexelőszolgáltatások („keresőgépek")*** és a saját adatbázist kezelő, a piacon az előbbinél valamivel korábban

12 Koch, Traugott: Suchmaschinen im Internet [Az Internet keresőszolgáltatásai]. Vortrag auf der 1. INETBIB-Tagung, Dortmund, 1996. márc. 11. [Előadás az 1. INETBIB-konferencián]. <http://www.ub2.lu.se/tk/=demos/DO9603-manus.html>

megjelent ***internetkatalógusok (vagy böngésző szolgáltatások).*** Számos szolgáltatásban egyre inkább mindegyik típussal találkozhatunk.

A két rendszer jóformán egyidejű megszületésében és ellentétében sajátos, az osztályozási rendszerek (és egyben az információkereső nyelvek) korai történetére emlékeztető logika köszön vissza, mely időközben a Gopher és a web kialakulásával kapcsolatban is megfigyelhető volt. (A korai időszakkal részletesen első kötetünk elején foglalkozunk. A kettősség elméleti kérdéseivel ebben a kötetünkben a „Dualitás elve" és a „Kettősség elve az osztályozáselméletben" című részletekben Jurij Šrejder foglalkozott.) Az indexelőszolgáltatások az analitikus (individualizáló, nem hierarchikus, posztkoordinált), az internetkatalógusok pedig a szintetikus (generalizáló, hierarchikus, prekoordinált) tartalmi feldolgozást és keresést teszik lehetővé.
1876-ban, az amerikai könyvtártörténet „csodálatos évében" ugyancsak szinte egy időben született meg az a két rendszer, mely lényegében alapja lett a modern osztályozásnak és információkeresésnek. Dewey Tizedes Osztályozása és folytatása, az ETO a Gopherhez hasonlóan hierarchikus szervezettségű volt, és a ráépülő szakkatalógusok ugyancsak fölfoghatóak egyetemes léptékű, korlátozott számú csúcsfogalommal rendelkező óriási tartalomjegyzékeknek, mint a mai internetkatalógusok. (Az egyik – eredetileg Gopher-menüként született – szolgáltatás, az 1989-ben született CyberDewey nevében is utalt erre a rokonságra.) Az internetkatalógusok is az ETO szellemiségén alapulnak: meghatározott, könnyen áttekinthető számú főosztály és a belőlük kiinduló alosztályok alá–fölérendeltségi szerkezete jellemzi őket.

Ezzel szemben Cutter természetes nyelven alapuló tárgyszórendszerének elvileg végtelen sok hierarchiacsúcsa lehet, akárcsak a web indexelőrendszereinek. Az összetett tárgyszavakon belül ugyan érvényesült kezdetben valamiféle hierarchikus szervezettség, de a tárgyszavakat mind szabadabban kezdték egymással kombinálni, és idővel az összetett tárgyszavak használatáról is eltekintettek. Fokozatosan kialakultak – Taube „uniterm" rendszerének hatására – a deszkriptoros információkereső nyelvek és szótáraik, a tezauruszok. Szerkezeti felépítésük nagyon emlékeztet a hipertexten belüli kapcsolódásokra, mert a tezauruszok deszkriptorai tetszés szerinti deszkriptorral összekapcsolhatók és fordítva (azaz a kapcsolódások szerkezete itt is M : N, akárcsak a hipertext esetén). Az indexelőszolgáltatásokban az egyedi szavakat tartalmazó indexek alapján végezhető a természetes nyelven alapuló keresés, és egyre gyakoribb, hogy ezt kötött keresőszótárak, tezauruszok alkalmazásával támogatják.

A szerver- és kliensoldali keresés

A születés lázában és nem utolsósorban a nagyobb hírverés kedvéért olykor hangzatos, olykor meg többjelentésű megnevezéssel találkozik a felhasználó:
A navigálás szempontjából az internetet – William Gibson 1984-ben írt Neuromancer című fantasztikus elbeszélése[13] nyomán – afféle virtuális „kibertérnek" (cyber space) nevezik. E „térben" a weben folytatott kereséshez külön szoftvereket használnak a szerver-, és külön szoftvereket a kliensgépeken, melyeket szerver-, illetve kliensoldali „navigációs eszközöknek", egyszerűbben szerver-, illetve kliensprogramoknak is neveznek. (A még egyszerűbb tolvajnyelvi „kliens" és „szerver" elnevezést a bennfentesek bizonyára a kezdők elriasztására használják, nehogy egyértelmű legyen számukra, mikor van szó gépről és mikor programról.)
Meg kell különböztetni a kétfajta keresőszolgáltatáshoz szükséges szerveroldali és kliensoldali keresőeszközöket. A **szolgáltatói vagy szerveroldalon** találhatók az indexelőszolgáltatások és az internetkatalógusok (böngészőszolgáltatások) programjai és adatbázisai, a **felhasználói vagy kliensoldalon** pedig visszakereső- („nézegető", viewer, browser) rendszerek.

Az ***indexelőszolgáltatások*** szervergépein a következő szoftvereket használják:

a1 a „leszedőnek" (krauler, crawler, spider, wanderer, gatherer, scooter) vagy „robotnak" nevezett szoftver, mely afféle webvándorként járkál a kibertérben, és a HTML-oldalakon beágyazott hipertext csatolókat kihasználva mozog egyik webdokumentumról a másikra, hogy felhasználói beavatkozás nélkül egyetlen mutatóba gyűjtse össze a HTML-dokumentmok kulcsszavait;

a2 ezt egészíti ki az indexelő (indexkészítő) szoftver (pl. WAIS, Archie, INQUERY, Apache, Glimpse, Harvest), mely a felkutatott, indexelt HTML-dokumentumok másodlagos adatait (leírásait) adatbázisokba rendezi. Az indexkifejezéseket automatikusan generálják, különös figyelemmel a HTML-oldalak címfejében szereplő másodlagos (meta-) adatokra.
Az adatbázis tartalma az URL, kulcsszavak, webcím, rövid tartalmi kivonat, teljes szöveg első sora stb. Ezek alkotják a másodlagos információtételeket vagy rekordokat, az indexelt HTML-oldalak pedig az elsődleges dokumentumok. (A másodlagos információkat hálózati és digitális könyvtári környezetben metaadatoknak neveik, az elsődleges információkat pedig digitális objektumoknak.)

13 Gibson, William.: Neuromancer. Neurománc. Budapest: Valhalla Páholy, 1992. 344 p.

Az adatbázisban tárolt információk frissítése kumulatív vagy reprodukáló szokott lenni. Az előbbi esetben az új rekordok hozzáadódnak a meglévőkhöz, az utóbbi esetben pedig időközönként az új rekordokkal a teljes adatbázist újjászervezik. (Rekordon itt a keresőszolgáltatások által összegyűjtött másodlagos információtételeket értjük, elsősorban HTML-dokumentumok leírásait);

a3 a „leszedő" és az indexelőprogramot integráló egységet, mely egyben elvégzi a kliensoldalról közvetített szerveroldali keresést is, összefoglalóan „keresőgépnek" („keresőmotor", „keresőmű", search engine), szerényebb megnevezéssel keresőrendszernek (search system) nevezik; előfordul, hogy „keresőgépen" nem szerveroldali szoftvert, hanem azt a számítógépet értik, melyet a keresőrendszer futtatására állítottak üzembe. Még gyakoribb, hogy magát a szerveroldali teljes keresőszolgáltatást nevezik „keresőgépnek", „keresőmotornak", „robotnak" (search engines, bots, robots, Suchmaschinen, Roboter), noha e megnevezések csak a szerveroldali leszedő, indexelő és kereső programrendszerek együttesét, robot esetében pedig a leszedőt jelölik és nem a teljes szolgáltató rendszert, melybe beletartozik még a felhasználói felület és a szolgáltatott tartalom is. (Azt mondják, hogy az AltaVista „keresőgép", holott az AltaVista a teljes keresőszolgáltatás neve, melyen belül – többek között – leszedők, indexelő- és keresőprogramok működnek.)

Az ***internetkatalógusok (böngészőszolgáltatások)*** szervergépein a következő szoftvereket használják:

b1 a ma még szinte kizárólag intellektuálisan osztályozott HTML-dokumentumok másodlagos adatait (leírásait) kezelő adatbázis-kezelő rendszer, melybe az alkalmazott rendezőrendszert is integrálták (ez tehát nem indexelést végző „keresőgép", noha indexelőprogramok kiegészítő alkalmazása is mind gyakoribb);

b2 szükség esetén a felhasználóbarát megjelenítést biztosító előtétprogram.

A felhasználói kliensgépeken – a kliensoldalon – fut az ablakos, felhasználóbarát vizuális felülettel ellátott „nézegetőnek", „böngészőnek" (viewer, browser) nevezett szoftver (mint amilyen például a Mosaic, a Netscape Navigator, az Internet Explorer stb.). Ezek a programok valójában nem keresnek, hanem a felhasználó által kijelölt keresési parancsokat közvetítik a szerveroldali automatikus keresést elvégző „keresőgépnek", ezért is neveztük őket feljebb visszakereső-rendszereknek, mivel egyszer már kikeresett rekordokra irányulnak. Hívják őket közvetítőknek is.
A szerveroldali szoftverekkel a felhasználó mindig csak a kliensoldali „nézegetőn" keresztül kerül kapcsolatba. A „keresőgépek", adatbázis-kezelők stb. a háttérben automatikusan működnek.

A szerveroldali szoftverek tehát olyan kliensoldali szoftvereket igényelnek, melyekkel az előbbiek szolgáltatásai realizálhatók a felhasználóknak. Az utóbbiak fogadják a keresőkérdéseket, megteremtik az összeköttetést a szerverprogramokkal és ezeknek a válaszát megfelelően „kiszerelve" közvetítik a felhasználónak. Ezt az egymást feltételező szoftverszerkezetet nevezik kliens-szerver rendszernek.
A nézegetők mindinkább az internet felhasználói rendszerei lesznek. Nekik köszönhető, hogy a nagy jelentőségű, ám nehezen hasznosítható számítógépes kapcsolatokból informatív és könnyen kezelhető kommunikációs eszköz lett. A végfelhasználó nem is veszi észre, hogy a nézegető használatakor indexelőfolyamat eredményében részesül, mert eltakarja előle a felhasználóbarát, „természetelvű" felhasználói felület.

Keresés URL alapján

A különféle típusú (web, gopher, hírcsoport, ftp stb.) internetforrások helyét meghatározó egységes forrásazonosító (URL; Unified Ressource Locator) ismerete alapján elvileg minden HTML-dokumentum megtalálható. A gyakorlatban ez nem mindig van így. Számos szolgáltató honlapjáról több lépésben hozzáférhetők értékes dokumentumok, melyeket szándékosan helyeznek el úgy, hogy közben több lépést is meg kell tenni, miközben hirdetésekkel találkozik az érdeklődő. E dokumentumok pontos azonosítója alapján közvetlenül a dokumentum mégsem érhető el, mert a tartalomszolgáltató ezt megakadályozza. Azt akarja, hogy csak a honlapján található csatolókon keresztül jusson el a felhasználó a dokumentumhoz, mert eközben kénytelen a hirdetését is megismerni.

Indexelőszolgáltatások („keresőgépek")

Meghatározás

Az indexelőszolgáltatások „keresőgépeket" alkalmazó szolgáltatások (robot generated indices), melyek adatbázisa a „keresőgépek" által indexelt HTML-dokumentumok rekordjait (másodlagos adataiból álló leírásait) tartalmazza. Bennük természetes nyelvű szavakkal végezhető a lekérdezés.
Az ismertebb globális rendszerek közé tartozik például az AltaVista, excite, Hot Bot, Infoseek, Lycos, Northern Light. A magyarországi webhelyeket 1996 óta a Heuréka (Hungary.Network) dolgozza föl, 1998 után pedig megjelent az AltaVizsla (MATÁV) is.

A szolgáltatások leszedői éjjel-nappali üzemben, csatolóról csatolóra haladva indexelik a HTML-dokmentumokat. Jelentős részük a teljes szöveget indexeli, de közülük sokan a teljes szövegből csak meghatározott számú sort (pl. az első húsz sort) és a metaadatokat veszik figyelembe (pl. a Lycos). Léteznek szolgáltatások, melyek eleve csak a HTML-dokumentumok metaadatait vagy kis részét dolgozzák föl (pl. a WWW Worm).
Hibásan az egész szolgáltatást „keresőgépnek" nevezik, holott a „keresőgép" a szolgáltató rendszernek csak egyik része.

Indexelés, „begyűjtés"

Az indexelőszolgáltatások fontos jellemzője a gyűjtőkör és kiválasztási–indexelési módszer. Az elsőre szerencsés esetben már a szolgáltatás nevéből következtetni lehet, és mindig található a belépőlapon olyan csatoló (pl. Magunkról, About Lycos), melyet működtetve a szolgáltatás céljáról tájékozódhatunk. A másodikról csak közvetett információk állnak rendelkezésre, a belépőlapról kiindulva e tekintetben semmiféle érdemleges adathoz nem lehet jutni. Számos vizsgálat a szolgáltatások közvetlen megkérdezésével készül el.

Az adott „begyűjtési stratégia" (gathering, harvesting) és forrásfelkutatás (resource discovery) dönti el, hogy milyen szervereket talál meg a keresőgép és azon belül milyen dokumentumok indexelését részesíti előnyben. Az indexelt egységek száma szolgáltatásonként különböző, néhány tízezertől (Harvest Home Page Broker) a tizen- és huszonmilliókig terjed (AltaVista, Lycos, Northern Light). De hogy mit tekintenek egységnek, az ugyancsak szolgáltatásonként változó. Van, amelyik – mint a Lycos – minden elért URL-t számol, noha a dokumentumainak csak töredékét indexeli, az Open Text annyiszor számolja az URLT-t, ahányszor az a legkülönbözőbb dokumentumokban előfordul, az Inktomi viszont csak a teljes szövegükben indexelt dokumentumokat számolja.

Az „először átfogóan" (breadth-first) indexelőstratégiát alkalmazó rendszerek gyűjtőköre nyilván nagy lesz, az „először mélyen" (deapth-first) indexelőstratégia eredménye pedig a részletesen indexelt, de kevés dokumentum, egyben kevés begyűjtött szerver lesz.[14]

A szervergépen kezelt adatbázisba betárolt adatok az indexek alapján kérdezhetők le a kliensoldali nézegetőkkel. A találatokat elemzik és többnyire relevanciavizsgálatnak is alávetik. Az indexelőszolgáltatások szempontjá-

14 Koch, T., id. mű.

ból a HTML-dokumentumok internetforrások, meghatározott összességük a keresőszolgáltatások „gyűjtőköre".
Vannak olyan keresőszolgáltatások is, melyek katalógusokból (is) készítenek indexeket a lekérdezéshez (pl. ALIWEB, Yahoo! Search[15], InterCat OCLC), és számos kereskedelmi szolgáltatáshoz ingyen be lehet jelentkezni.

Az indexelőszolgáltatásoknak be is lehet küldeni HTML-dokumentumokat, amit szívesen vesznek, mert bővíti a választékot. (Az internetkatalógusok kisebbik része kizárólag ezen az alapon működik.) A manuálisan gyűjtött, intellektuálisan feldolgozott indexek előnye a tartalmi ellenőrzöttségben rejlik. Olyan tételek indexei ezek, melyeket vagy a szolgáltatás szakembere, vagy a szerző maga dolgozott föl. Hiába állnak rendelkezésre jól szerkesztett bejelentkezési űrlapok, pl. az ALIWEB tapasztalatai alapján ezeket többnyire felületesen töltik ki. A nem szöveges dokumentumok esetén nyilván mindig szabványosított beviteli űrlapokat kell használni.[16]

Avulás és frissítés

A HTML-rekordok hamar avulnak, mert a HTML-dokumentumok megszűnhetnek, átalakulhatnak. A feldolgozott állomány frissítése az indexelőrendszerek többségében elvileg könnyebben megoldható, mint az internetkatalógusokban, ahol intellektuálisan osztályozzák a HTML-dokumentumokat, és az automatikus frissítés hiányában kialakulnak a zsákutcás, halott tételek (dead links).

A keresőgépeken alapuló szolgáltatások legnagyobb előnye, hogy a körülményekhez képest rendkívül gazdagok. Noha elvileg adott a gyakori frissítés lehetősége, a valóságban ezekben a rendszerekben a tételek kb. 15-20%-a már nem létezik. Mivel az esetek többségében az eredeti források lényeges részeit, sokszor a teljes szöveget indexelik, nagy a valószínűsége annak, hogy rendkívül speciális információk is megtalálhatók. Éppen ez a tény indokolja, hogy előbb-utóbb érdemes lesz jobb eszközöket is rendelkezésre bocsátani az információkereső stratégiához.

Az aktualizálás gyakorisága a hetenkénti (pl. Lycos, WebCrawler) a félévenkénti, sőt évente egyszeri gyakoriság (WWW Worm) között mozog. A

15 Nem tévesztendő össze magával a Yahoo! Internetkatalógussal. A Yahoo! Search e katalógus állományát indexelő rendszer, vele a katalógusban nem szisztematikusan (azaz nem böngészve), hanem egyedi szavakkal lekérdezve lehet keresni.

16 Koch, T., id. mű.

legtöbb szolgáltatás nem közöl erről semmit. Ugyanannak a HTML-oldalnak különféle változataiból közelítőleg megállapíthatók az erre vonatkozó adatok. Mennél nagyobb a szolgáltatás, annál kisebb frissítési gyakoriságra lehet számítani.[17]

Az elemzések szerint a nagyobb szolgáltatások többségében a halott tételek (dead links) száma megközelíti a 20-30%-ot. A kisebb szolgáltatásokban a helyzet ennél lényegesen jobb (lásd. az 1. táblázatot).

Keresőszolgálat	nem élő tételek %-a
Lycos	29%
AltaVista	18%
Northern Light	16%
MSN Web Search	14%
Yahoo!, Inktomi	13%
Snap!	11%
Infoseek	8%
HotBot	4%
Google!	0%
Excite	0%

1. táblázat. Lekérdezés eredményeként kapott nem élő tételek száma 1999. 03. 05-én[18]

Elvileg az internetkatalógusokban is automatikusan elvégezhető volna a frissítés azáltal, hogy megfelelő program törli a már nem előhívható HTML-dokumentumok kapcsolatait az indexadatokhoz, de erről nincs információ.

A mennyiségi teljesítmények lenyűgözőek: az AltaVista keresőrendszerében pl. naponta kb. 10 millió HTML-dokumentumot néz át a leszedő, ez kb. tizede a több mint 120 millió indexelt tételnek, melyet a rendszer adatbázisa tartalmaz.[19] A 2. táblázatban néhány keresőszolgálat adatbázisának mérete látható.

17 Koch, T., id. mű.

18 Notess, Greg R.: Search engine showdown. The user's guide to the web searching [Keresőgépek áttekintése. Felhasználói kalauz a webkereséshez]. [1999.03.05.] <http://www.notess.com/search/stats/dead.shtml>

19 1998-as adat. <http://www.privat.schlund.de/B/BesserSuchen/t-sucher.htm>

Keresőszolgálat	Rekordok száma
Northern Light	128,540.264
AltaVista	106,169.808
HotBot/Anzwers	99,409.035
Schnäppers	98,638.820
Google!	71,065.137
Infoseek	59,700.192
MSN Web Search	39,589.032
Excite	32,896.723
Lycos!	22,781.237

2. táblázat. Keresőszolgáltatások adatbázisainak mérete 1999. 03. 05-én[20]

Keresési módszerek és stratégia

A szolgáltatások általában arra törekszenek, hogy a teljesség (recall) legyen nagy, ezért pontosságról eleve nincs szó. Az alkalmazható módszerek, stratégia meglehetősen változatos. Az alapértelmezésen túlmenő lehetőségek (részletes keresés, advanced search) a szolgáltatások jelentős részénél nincsenek előtérben, a laikus sokszor nem is veszi észre őket.

A kereséshez egyedi szavakat adhatunk meg.

Hozzáértők választhatnak más Boole-operátorokat és helyzeti (távolsági/közelségi) operátorokat.

Alkalmazható a „szólánccal" végzett keresés (string-search), hol idézőjelek közé téve a láncot, hol legördülő mezőn minősítve.

Többnyire megadható, hogy csonkoltan vagy pontosan értelmezendő-e a keresőszó. Egyes rendszerekben (pl. AltaVista) megkülönböztethető a kis- és nagybetű.

A szolgáltatások kis részénél (AltaVista, Excite, Lycos) megadható a nyelvi, sőt – mint a Lycos esetében – néhány dokumentumtípus szerinti szűkítés is, azaz kérhető csak meghatározott nyelvű vagy dokumentumtípusba tartozó találatok megjelenítése.

Néhány szolgáltatás, mint pl. az AltaVista „idézetes" keresést (citation indexing) is biztosít, azaz megadja azokat az összetett szavakat, melyekben az

20 Notess, Greg R., id. mű. <http://www.notess.com/search/stats/sizeest.shtml>

egyedi keresőszó előfordul, és ezeket fölhasználva szűkíthető a keresés (pl. a „műanyag" kifejezéssel keresve felajánlja a „hőre lágyuló műanyagok", „ipari műanyag burkolatok" stb. kifejezéseket is a kereséshez).

Van olyan szolgáltatás, amelyben kiköthető, hogy a keresés csak a HTML-címben, az összefoglalásban vagy a teljes szövegben szereplő szavakra korlátozódjék, vagy elvétve kiköthető, hogy a dokumentumból mely oldalak jelenjenek meg.

Van olyan szolgáltatás (Highway 61), amelyben megadható, hogy a kereső milyen színvonalú lekérdezést igényel.

Olykor az elvégzett keresés eredményhalmazán végezhető másodlagos keresés (relevanz feedback, find similar pages). Ez annyit jelent, hogy a találathoz lekérhetők az adott találathoz „hasonló" tartalmú (similar, related topic) tételek.

Az Infoseek már „specifikus keresést" is biztosít: az átfogó jelentésű kifejezéshez a | (vonal, pipe) jellel megadható a specifikus (pl. „tánc | tangó" esetén a „tánc" alapján kiválasztott halmazból a „tangóval" jellemzett rekordokat kapjuk meg).

A fejlettebb rendszerekben (pl. MetaGer) az is kérhető, hogy ellenőrizzék, élnek-e még egyáltalán a talált tételek? Ilyenkor a végeredményre valamivel tovább kell várni.

Az egyik nagy probléma a keresőszolgáltatások túlnyomó részében, hogy a keresés nem korlátozható mezőkre (formátum szegmensekre), és a találatok csak néhány szolgáltatásban rendezhetők különféle szempontok (dátum, hely stb.) szerint. Létezik néhány kivétel: a Lycosban például kiköthető, hogy a keresés a teljes szövegben, a címben, vagy az URL-azonosítóban történjék; az Infoseek lehetővé teszi a találatok rendezését dátum szerint is. A viszonylag már elterjedt nyelvi szűkítés mellett olykor a regionális finomítás is lehetséges (Yahoo! Get Local).

A keresési végfelület (ablak) a legtöbb szolgáltatásban a végletekig egyszerű, általában semmiféle keresési segítséget nem tartalmaz. Ehhez a megfelelő, alig észrevehető csatolót kell megkeresni (részletes keresés, advanced search). Csak kevés szolgáltatás adja meg az eszközök választékát az első oldalon. Noha általában a „legostobább felhasználóra" számítanak, ehhez képest a keresési segítség, különösen pedig a keresési példák hallatlanul szegényesek, az on-line szolgáltatásokban természetes keresési stratégia és keresőkép (profil) fogalmai teljesen ismeretlennek – legalábbis egyelőre.

A keresés finomítása terén az internet indexelőszolgáltatásai még alulmaradnak az on-line szolgáltatásokkal szemben. Az utóbbiakban az alkal-

mazott hagyományos adatbázis-kezelő rendszerek jóvoltából az információkeresési stratégia teljes tárháza rendelkezésre áll. Ez a helyzet azonban rohamosan változik. A felhasználói komfort dolgában az indexelőszolgáltatások már ma nem egy vonatkozásban előbbre vannak.
Az üzleti szempontok következtében egész sor tájékoztatási komforttal látják el a felhasználót. Ilyen például a leggyakrabban használt keresőszavak százalékban megadott gyakorisága. Első helyen ugyan az erotikus információk keresettségére utaló kifejezések állnak, sokkal jelentősebb azonban, hogy ezután a közhasznú dolgokra (állás- és társkeresés, közintézmények, adattárak, telefonkönyvek, menetrendek) vonatkozó keresőszavak következnek. Ezt követik a rendkívül kis gyakoriságú speciális szakkifejezések (az „epitaxiától" az „aloe veráig", az elméleti matematika kifejezéseitől a teológiai fogalmakig).
Mindebben az a fontos, hogy gyakorlatilag a szaknyelv minden elképzelhető szavát használva tesznek föl keresőkérdéseket. A komoly keresők részéről tehát rendkívül nagy és differenciált igények jelentek meg.

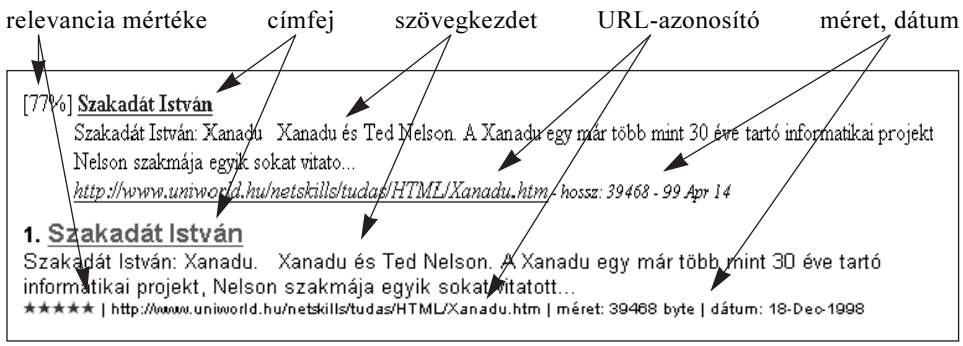
A 3. táblázatban ilyen gyakorisági jegyzék nagyon leegyszerűsített kivonata látható.

Keresőszó	gyakoriság
szex	8,91%
sex	7,59%
porno	2,38%
erotika	2,19%
magyar	0,98%
társkereső	0,55%
társkeres	0,54%
telefonkönyv	0,49%
önkormányzat	0,11%
tenzoralgebra	0,0098%
szikraforgácsolás	0,0022%
kontrakció	0,0019%
túlhűlés	0,0017%
Pragmatica Sanction	0,0008%

3. táblázat. Kivonat az AltaVizslában megadott felhasználói keresőszavak gyakorisági jegyzékéből 1998 májusa és augusztusa között

Találatmegjelenítés

A megjelenített másodlagos információtétel (HTML-dokumentum találati leírása) többnyire egyszerű, és szolgáltatásonként különbözik. Nincs szabványosított megjelenítési forma (egységes megjelenítés legfeljebb az itt nem tárgyalt on-line szolgáltatásokban fordul elő, de ott is nagyon ritkán felel meg bibliográfiai szabványoknak). Sokszor megadható, hogy egyszerre hány találat jelenjék meg, nagyon kevés rendszerben (pl. Infoseek) lehetséges nem csak relevancia, hanem dátum szerint is rendezni.
Az alábbiakban a Heuréka és az AltaVizsla keresőszolgáltatások egyszerű információtételei láthatók.

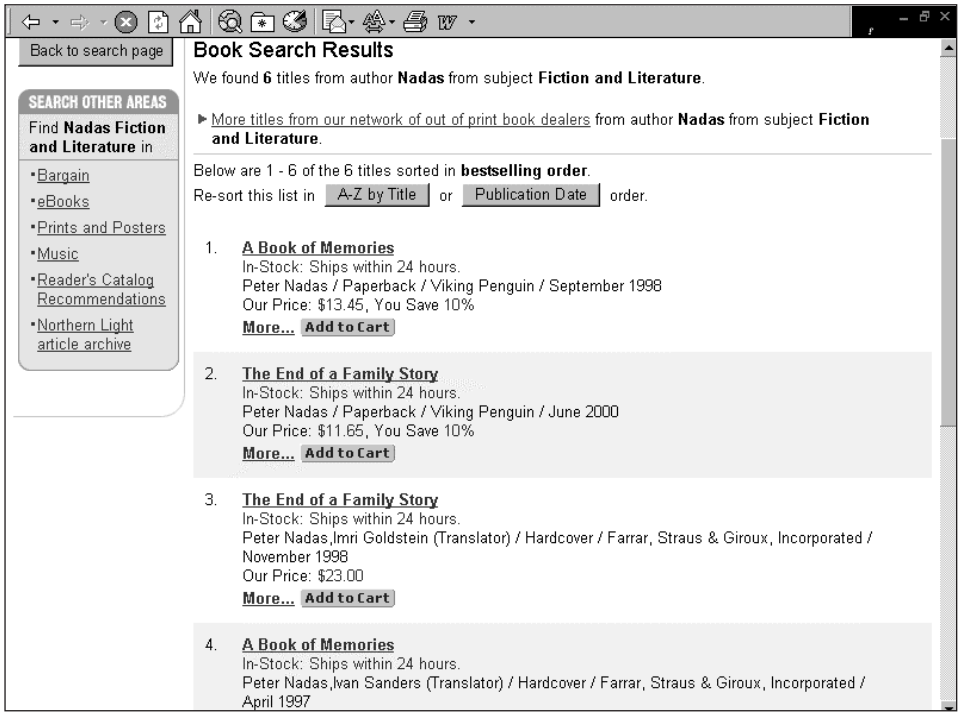


2. ábra. Ugyanannak a találatnak információtételei (HTML-rekordjai) a Heuréka, ill. az AltaVizsla keresőszolgáltatásokban

Ritka kivétel az olyan szolgáltatás, mint a Lycos, melyben a találatok leírása, azaz a másodlagos információtétel részletes és gondosan strukturált (3. ábra).

A találatok relevanciája

A relevancia mértékét egyrészt annak alapján állapítják meg, hogy a keresőszó a HTML-formátum címfejében (<Title>) szerepel-e, vagy csak a tartalmi kivonatban, ill. szövegben, és az utóbbin belül milyen gyakorisággal. Másrészt automatikus indexelési módszereket használnak. Az ismertebb indexelőprogramok közül pl. a WAIS a vektortérmodellt alkalmazza: az in-



3. ábra. A Lycos részletes találati leírásai

dexelt kifejezések alapján dokumentumvektort számít ki, és ezt hasonlítja össze a keresőkérdés vektorával, kiszámítva a kettő közötti távolságot egy n-dimenziós vektortérben. Minél kisebb a távolság, annál nagyobb a relevancia. Az INQUERY az interferenciahálók modelljét használja: e hálók a keresési folyamat határozatlanságát képzik le, melyből valószínűségi módszerek segítségével számítják ki a relevanciát. Ahhoz képest, hogy a dokumentációs célú automatikus indexelés és osztályozás terén néhány évtizeddel korábban milyen eredmények születtek, a keresőszolgáltatások relevanciavizsgálatai – egyelőre még – meglehetősen szegényesek (lásd kötetünkben az automatikus indexeléssel foglalkozó fejezetet).

A talált tételek rendkívül vegyes minőségűek. Gyakori, hogy már nincs is mögöttük élő tartalomszolgáltatás, sokszor ugyanannak a HTML-dokumentumnak különböző időpontokból származó változata jelenik meg, és a találatok túlnyomó többsége valójában teljesen irreleváns, mivel az indexelt szó nem a dokumentum tartalmát reprezentálja. Általában elmondható, hogy ha a relevancia mértéke az 50% alá csökken, a találat már teljesen irreleváns. Olykor maguk a kereskedelmi tartalomszolgáltatók is tovább rontják a találatok minőségét: HTML-oldalaik címfejébe olyan kifejezéseket is elhe-

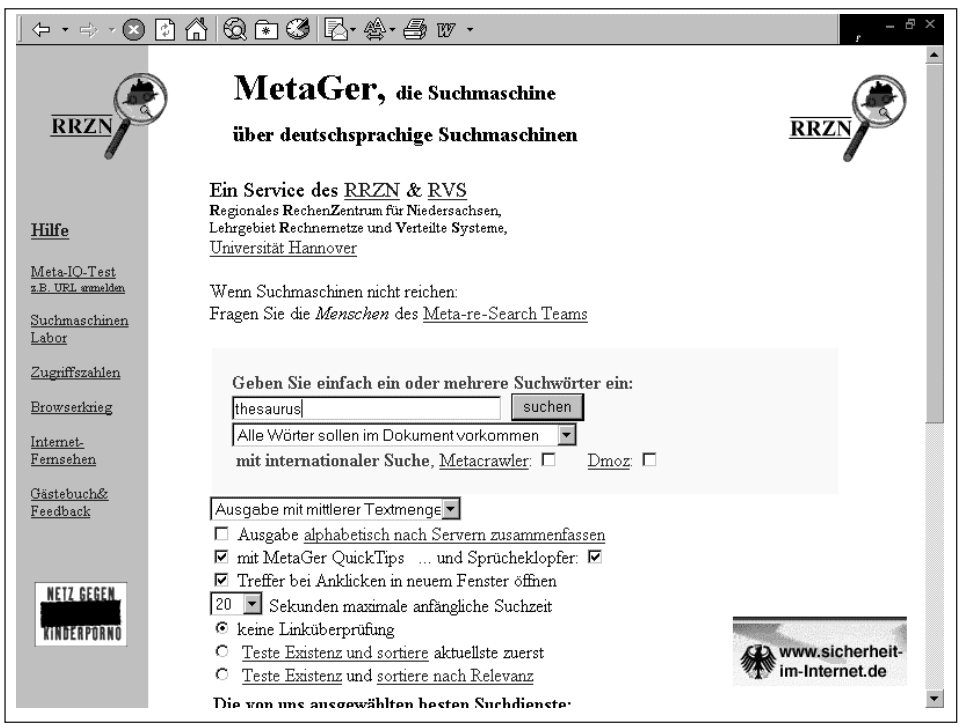
lyeznek, melyek valójában nem igazán jellemzik a lapjukat, de amelyekről tudják, hogy a gyakran keresettek közé tartoznak, hogy a leszedő – rájuk találva – különlegesen értékes találatként értékelje és a találati jegyzéken a legelső helyeken jelenítse meg őket.
Az indexelőszolgáltatások leszedőinek túlnyomó többsége kötött szótár nélkül válogatja ki a szövegszavakat (a hagyományos információkeresés nyelvén ez „szabad szövegen belüli keresés", „szabad kulcsszavas keresés"). Ritkán előfordulnak közöttük olyanok, melyekben kötött szótárat, néhányukban (pl. AltaVizsla, Engineering Electronic Library, Kolibri) tezauruszt is használnak. Ez utóbbi szolgáltatások információtételeinek relevanciája általában lényegesen nagyobb.

Gyűjtő és többszörösen indexelőszolgáltatások

Az indexelőszolgáltatások választéka ma már rendkívül nagy. A közel tucatnyi nemzetközileg ismert rendszeren kívül nagyon sok a speciális gyűjtőkörű (dedikált) rendszer, mely csak meghatározott típusú HTML-dokumentumokat dolgoz föl (pl. könyvkiadók legújabb kiadványait, műszaki folyóiratokat, folyóiratcikkeket, Usenet-cikkeket, cégeket), továbbá az olyan rendszer, amely csak meghatározott államon belüli webhelyek HTML-dokumentumait indexeli (a Heuréka és az AltaVizsla pl. csak a magyarországiakét). A szolgáltatások teljesítménye kisebb-nagyobb mértékben különbözik egymástól, ami a felhasználót arra kényszeríti, hogy a lehető legnagyobb teljesség érdekében több indexelő keresőszolgáltatást is igénybe vegyen, ami meghosszabbítja a keresést. Ráadásul nehéz összehasonlítani az eredményt, mert a találatok külön-külön jegyzékekben jelennek meg.
Ezen hivatottak segíteni a ***többszörösen, szimultán*** vagy ***meta-keresőszolgáltatások (multiple/parallel/meta search engines, Meta-Suchmaschinen).*** Velük egyszerre több indexelő keresőszolgáltatásban lehet keresni anélkül, hogy a felhasználónak az egyes szolgáltatásokkal külön foglalkoznia kellene. Ez körülbelül olyan, mintha valaki könyvet keres, mégpedig az összes magyarországi könyvtárban, és megkapja találatként, hogy az adott könyv milyen adatok kíséretében található meg az egyes könyvtárakban.
A többszörösen indexelőszolgáltatás leszedője a kijelölt indexelőszolgáltatásokat a keresőkérdések alapján párhuzamosan fésüli át és a találatokat közös listában jeleníti meg, ami a nagyobb választék mellett a jobb összehasonlítást is elősegíti. Az első ilyen rendszerek 1995-ben készültek. A hatékony működésüket az elosztott rendszerű működéssel fokozzák: a részműveletekre felbontott feladatokat egyszerre több számítógép

leszedő- és indexelőprogramjaira delegálják. A legismertebb ilyen többszörös feladatmegosztásra képes rendszer – a Harvest –, mely a legfejlettebbek közé tartozik, jelenleg már ingyen hozzáférhető.[21] Számos ismert indexelőszolgáltatásban használják.

Elég megadni a keresőkérdést, a többszörös indexelőszolgáltatás a profiljába fölvett szolgáltatásokat végignézve kilistázza a találatokat. A jobb minőségű rendszerekben a talált információtételeknél feltűntetik, hogy melyik indexelőszolgáltatásból származnak. Így gyakori, hogy ugyanaz a tétel többször is megjelenik, és összehasonlítható, melyik rendszer szolgáltatja a legfrissebb találatokat. A 4. ábrán a „Metager" keresőszolgáltatás belépőlapja látható (az angol „Metacrawler" belépőlapja nem annyira strukturált és informatív, ezért választottuk a német változatot).



4. ábra. Meta-keresőszolgáltatás belépőlapja. A „Quick Tip" (más rendszerekben a „Direct Hit") azokat a külön megjelenő találatokat adja meg, melyek a keresőszót az URL-névben tartalmazzák. Kérhető a találatok egzisztenciájának ellenőrzése, és a dátum vagy a relevancia szerinti rendezés. A táblázatban felsorolt, egyszerre lekérdezhető keresőszolgáltatások beállítása megváltoztatható, és kérhető, hogy a keresésben az angol „Metacrawler" is részt vegyen.

21 Harvest, version 1.5, University of Edinburgh, 1999. Hírcsoportja a következő címen található: <comp.infosystem.harvest>. A rendszerre vonatkozó publikációkat az alábbi címen archiválják: <http://www.mathematik.uni-osnabrueck.de/harvest/brokers/CIH>

Mára kialakultak ezeknek a szolgáltatásoknak a kritériumai:

– **párhuzamos keresés,** azaz egyszerre több szolgáltatás lekérdezése egyetlen keresési műveletben;

– **eredmény-összefésülés,** azaz a találatok megjelenítése egyetlen formátumban;

– **többszöröződések kezelése,** azaz ugyanaz a HTML-dokumentumot a rendszernek fel kell ismernie és jelölnie kell az egyes forrásokat, amelyből származik;

– **ÉS- meg VAGY-művelet**, mint minimális logikai keresési eszköz;

– **információveszteség nélküli működés** (ha pl. az egyik forrás tartalmi kivonatokat készít, azt át kell tudni venni);

– **forrásrendszerfedés,** [hiding] (a lekérdezett indexelőszolgáltatások tulajdonságai nem játszhatnak semmiféle szerepet a metarendszer szintjén, a felhasználónak semmit sem kell tudnia ezekről a specifikumokról);

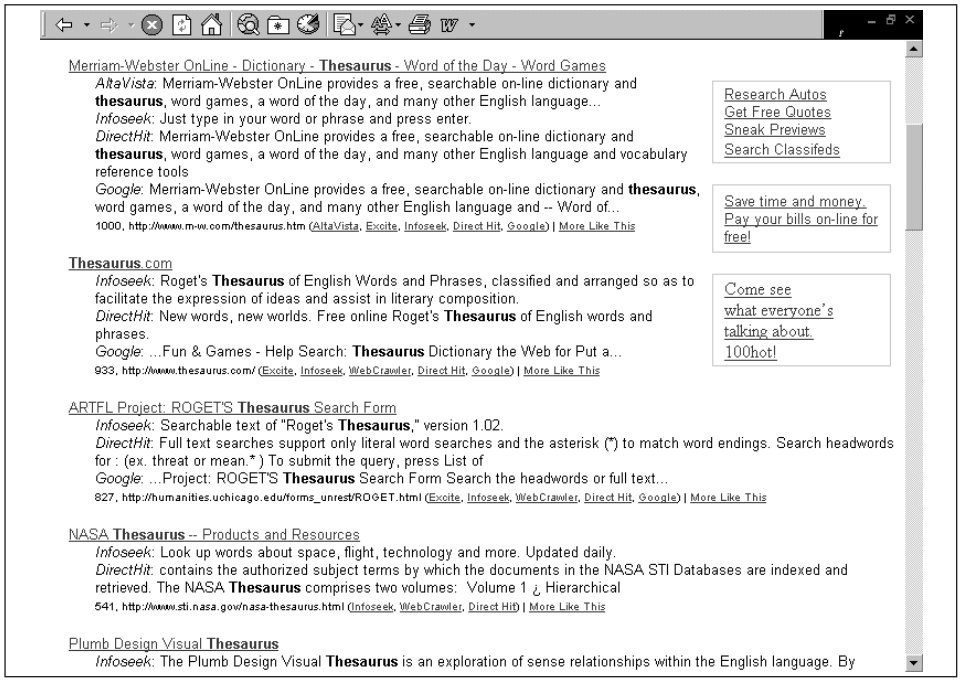
– **teljesség** (a keresésnek addig kell tartania, ameddig a lekérdezett szolgáltatásokból találatok nyerhetők).

E követelmények együttes kielégítése minden jel szerint komoly nehézségekkel jár, mivel jelenleg világszerte alig néhány metarendszer képes teljesíteni (MetaCrawler, MetaGer és Highway61). Az 5. ábrán a MetaCrwaler megjelenített találatai láthatók

Egyszerűbb, átmeneti típus a gyűjtőszolgáltatás (unified search interface [CUSI], all-in-one formular, sample service, Sammeldienst), amely felkínál több, olykor nagyon sok keresőszolgáltatást (mintegy gyűjteményként), de mindig csak egyet lehet kiválasztani a lekérdezésre (ilyen például az 1. ábrán látható BSZ, a CUSI különféle regionális változatai, Magyarországon pedig a HUDIR-t és a Heurékát fenntartó Hungary.Network gyűjtőszolgáltatása [Keresők gyűjteménye]).

A többszörös indexelőszolgáltatásokhoz hasonló, de más szervezettségű a karlsruhei könyvtár internetkatalógusa (Karlsuher Virtueller Katalog), melyen keresztül az összes nagyobb német – valamint a Kongresszusi, az angol, francia stb. nemzeti – könyvtár állománya egyetlen keresőkérdéssel elérhető.

Magyarországon a Közös Elektronikus Katalógusban (KözElKat), a Voyager adatbázis-kezelő rendszert (VOCAL) és a TINLIB adatbázis-kezelő rendszert (AROMO) használó könyvtárak közös katalógusában lehet egyetlen felületen elvégezni a keresést. Akárcsak a karlsruhei esetben valójában távolsági on-line adatbázis-hozzáférést biztosítanak összevont, közös webfelületen (ún. „brókeren").



5. ábra. A találatok megjelenítése a Metacrawler meta-indexelőszolgáltatásban. A keresőkérdés „Thesaurus" volt. Az egyes tételek legalsó sorának elején a relevancia mértéke látható (az első találatot a rendszer 100%-osnak értékelte), és a sor végén szerepelnek mindazok az egyedi keresőszolgáltatások, melyekben a találat szerepel.

Internetkatalógusok

Meghatározás

Az internetkatalógusok hierarchikus (ritkábban enumeratív) osztályozási rendszert alkalmazó keresőszolgáltatások, melyek adatbázisa a – túlnyomórészt intellektuálisan – osztályozott HTML-dokumentumok rekordjait (másodlagos adataiból álló leírásait) tartalmazza, valamint egyéb adatbázisok információtételeit. Bennük az osztályok alapján – elsősorban a katalógusban „lapozva" – végezhető böngészés.

Az ismertebb globális rendszerek közé tartozik például az excite review, Magellán, Yahoo!. A keresőszolgáltatásoknak ez a fajtája jelent meg először, valójában már a web előtt, a Gopherrel egy időben. Magyarországon 1995-től működik a HUDIR (Hungary.Network), 1999-től a Kincskereső (Elender), 2000-től pedig az AltaVizslának (Matáv) is van az indexelőszolgáltatás mellett saját katalógusa.

Nevezik ezeket böngészőszolgáltatásnak, tárgyszótárnak, tématárnak (subject directories, Themenverzeichnisse).

Forráskiválasztás

A manuálisan előállított internetkatalógusokra jellemző, hogy kisebb-nagyobb mértékben intellektuálisan sorolják be (osztályozzák) a HTML-dokumentumokat az alkalmazott osztályozási rendszerbe. Automatikus osztályozással működő rendszerekből alig van néhány (velük részletesebben T. Koch tanulmánya foglalkozik).
A feldolgozandó dokumentumok kiválasztását elvileg ugyancsak intellektuálisan végzik, de nagyon különböző színvonalon. A szolgáltatások egy részében semmiféle aktív kiválasztás nem zajlik, kizárólag olyan katalogizált tételeket tartalmaznak, melyeket önkéntesen adnak át a honlapok tulajdonosai, szerzői, akik többnyire az osztályozásról is gondoskodnak, vagy legalábbis szabad tárgyszavakkal, tartalmi leírással látják el a beküldött tételeiket.
A szolgáltatások többségében ugyan válogatnak, a kiválasztás kritériumai azonban alig ismerhetők meg. A különböző felmérések tanúsága szerint úgy fest, mintha a dokumentumok feltárását általában nem előzné meg határozottan körvonalazott gyarapítási tevékenység, csak afféle „spontán érkeztetés" zajlik.

„Maguk a szolgáltatások személyes megkérdezés esetén is csak nagyon kevés, ill. pontatlan információt közölnek kiválasztási kritériumaikról, a honlapjaikon pedig általában semmiféle tájékoztatás nem található róluk. Feltehető, hogy a kiválasztást sokszor nem valami tudatosan végzik, még ha olykor léteznek többé-kevésbé pontosan megfogalmazott követelmények. Többségükben szerkesztőket alkalmaznak, de nem ismerhető fel, miféle szelekciót végeznek: minden jel szerint nem annyira a kiválasztásra helyezik a hangsúlyt, mint inkább a tartalmi feltárásra. Egy tanulmányban a Yahoo!-ról ez szerepel:

Először összegyűjtik az új weboldalak URL-jeit. A legtöbb közülük drótpostán érkezik azoktól, akik a hálón szereplő oldalaikat szeretnék fölvetetni, a többit pedig a Yahoo! leszedője szállítja – egyszerű robot, mely új weboldalakat keresve csatolóról csatolóra ugrál. Ezt követően a húsz osztályozó valamelyike átnézi a weboldalt és elvégzi a besorolást.

Különösen a nyelvi vagy tematikus alapon szelektáló szolgáltatásokról nincs információ a kiválasztáskor figyelembe veendő tartalmi kritériumokról. Legfeljebb azt említik, hogy félig üres weboldalak nem jöhetnek szóba, az IK Web Library (a brit »nemzeti katalógus«) pedig bizonyos tartalmú (pl. trágár)

dokumentumokat kizár a gyűjtésből. Az általános gyűjtőkörű szolgáltatásokban az előbbiekhez képest inkább alkalmaznak tartalmi és formális kritériumokat.

A szerkesztőket alkalmazó szolgáltatásokban a döntéseket minden jel szerint intuitíve, a szakmai tapasztalatok alapján hozzák. (Magellan: *Minden szerkesztőnk szakember a maga területén, ezért a végső döntés mindig az ő kezében van*.) Részletezett, konkrét kiválasztási kritériumokat a 12 általános és globális szolgáltatás közül csak az Argus Clearinghouse, a NetFirst és a Webcrawler select közölt.

Részletesebben tájékoztattak a szolgáltatások, a feldolgozott weboldalak minősítési (rating) kritériumairól (átfogó és egyedi tartalmi, megjelenési és technikai/szoftverminősítés).

Alig van olyan szolgáltatás, melyben megkülönböztetnek feltétlenül betartandó és másodlagos kritériumokat, nem is súlyozzák ezeket. Argus Clearinghouse bizonyos metaadatok (szerzőség, dátum) létét elengedhetetlennek tekinti, a Lycos számára a más weboldalról származó hipercsatolók gyakorisága a legfontosabb kiválasztási feltétel.

Beszélni kell az itt felsorolt kritériumok operacionalizálásáról. Erről akkor van szó, ha a feltételeket mérhető adatokkal kapcsolják össze: melyek konkrétan a kizárandó és a fölveendő tartalmak: mennél nem régebbi weboldalak vehetők föl, milyen metaadat megléte elengedhetetlen stb. Az objektív felhasználhatóság érdekében az arra alkalmas kritériumokat operacionalizált formában kell megfogalmazni. A weboldal látogatási gyakoriságának, idézettségének (hipercsatoltságának) megkövetelt határértékeit például számszerűen is meg kell adni. Vizsgálatunkban a 19 megkérdezett szolgáltatás közül egyetlen egy sem említett operacionalizált feltételeket."[22]

A kritériumok a vizsgálatok alapján az alábbiakban foglalhatók össze (az aláhúzottak a feltétlenül betartandók, a többiek másodlagosak):

1. Stabilitási kritériumok:

1.1 a forrás könnyen és biztosan elérhető
1.2 a forrás előre láthatólag nem rövid életű
1.3 a forrás aktualizálására, karbantartására számítani lehet

2. Tartalmi kritériumok:

2.1 a forrás tartalma hihető, a létrehozója a tartalom vonatkozásában hiteles, megbízható testület vagy személy
2.2 a forrás időszerű
2.3 a forrás érdekes, közérdeklődésre tart igényt

22 Ohler, Angele: Browsingdienste im Internet [Böngésző szolgáltatások az Interneten]. Berlin: Freie Universität, 1996. <http://userpage.fu-berlin.de/~angele/bond/brows04.htm>

2.4 a forrás informatív, érdekes
2.5 a forrás jól szerkesztett, részletes, egyedi, tipikus, speciális
2.6 a forrás nem tartalmaz olyasmit, ami a mindenkori kizáró tényezők jegyzékében szerepel

3. Formai kritériumok:
3.1 a forrás nem régebbi, mint ...
3.2 a forrásnak megvannak a felsorolt metaadatai (cím, szerzőség/közreadó, tárgyszavak), html-szerkezete szabványos
3.3 a forrásban sok más forrásra vonatkozó csatoló van, különösen a teljes html-dokumentumokra, szolgáltatásokra utal
3.4 a forrásra gyakran utalnak más forrásokból
3.5 a forrást gyakran használják, sok a látogatója
3.6 a forrás nem túl kicsi (hacsak nem nagyon időszerű, közérdekű)
3.7 a forrás szép, látványos, különleges formatervezésű
3.8 a forrás ingyenes

Avulás és frissítés

Az internetkatalógusok állományai ugyanúgy avulnak, akár az internet többi állománya. Frissítésükre azonban még az indexelőszolgáltatásokban alkalmazott gyakoriságoknál is ritkábban kerül sor, mivel a katalógusok HTML-dokumentumait intellektuálisan dolgozzák föl, s nem mindig áll rendelkezésre olyan keresőgép, mely a frissítést végrehajthatná. Ezért az internetkatalógusokban sokkal több a zsákutcás HTML-rekord, melyből kiindulva az eredeti HTML-dokumentum már nem hívható elő.

Osztályozási rendszerek

Hagyományos osztályozási rendszereket alkalmazó internetkatalógusok

McKiernan, az iowai egyetem könyvtárosának mutatója, a Beyond Bookmarks, mely a hagyományos osztályozási rendszereket, tárgyszójegyzékeket és tezauruszokat használó keresőszolgáltatásokról tájékoztat[23], 2000

23 Beyond Bookmarks: Schemes for organizing the web [Beyond könyvjelző: a web szervezőrendszere]. <http://public.iastate.edi/~CYBERSTACKS/CTW.htm> (1999. 12. 01-i állapot).
Egy másik ilyen mutató: Koch. T.: DC subject. Thesauri and classification systems available in the WWW [TO tárgykör. Weben elérhető tezauruszok és osztályozási rendszerek]. 1996. [1999. 09. 27.]<http:/www.ub2.lu.se/metadata/subject-help.html>

elején 55 olyan internetkatalógust sorol föl, melyben hagyományos osztályozási rendszereket használnak. Ezen belül 22 Dewey Tizedes Osztályozását, 11 az ETO-t és 6 a Kongresszusi Könyvtárét.
A dokumentációs-könyvtári, vagy egyéb bevált hagyományos osztályozási rendszer alkalmazása elsősorban azokra a szolgáltatásokra jellemző, melyek fölhasználói köre tudományos és egyéb szakemberekből áll, és ezért elsősorban tudományos jelentőségű forrásokat dolgoznak fel. A feldolgozás kiválasztási kritériumainak itt lényegesen nagyobb a jelentősége. A hagyományos osztályozási rendszereket többnyire kisebb internetkatalógusok használják, egy részüket a könyvtárak hozták létre (pl. BUBL, NISS, WWW Virtual Library, NetFirst).
A hagyományos, bevált és tudományos igénnyel készült osztályozási rendszerek alkalmazóin belül külön csoportot alkotnak azok a szakterületekre specializálódott gyűjtőkörű katalógusok, melyekben minőségbiztosítási szempontokat alkalmaznak a kiválasztásban és feldolgozásban, részletes tartalmi és formai leírást készítenek, többek között annotációt, összefoglalásokat, és a munkákat a szakterület szakértőivel végeztetik el. Ezeket ***szakterületi információs kapuszolgálat***oknak (subject based information gateway) nevezik. Pl. az informatikai weboldalakat feldolgozó Ariadne, melyben az ACM számítástechnikai osztályozási rendszerét (Computer Classification System), vagy az Engineering Electronic Library (EELS), melyben speciális osztályozási rendszert és az EI tezauruszt használják. Ebben a körben jelennek meg az automatikus osztályozást alkalmazó internetkatalógusok is (Scorpion, GERHARD).

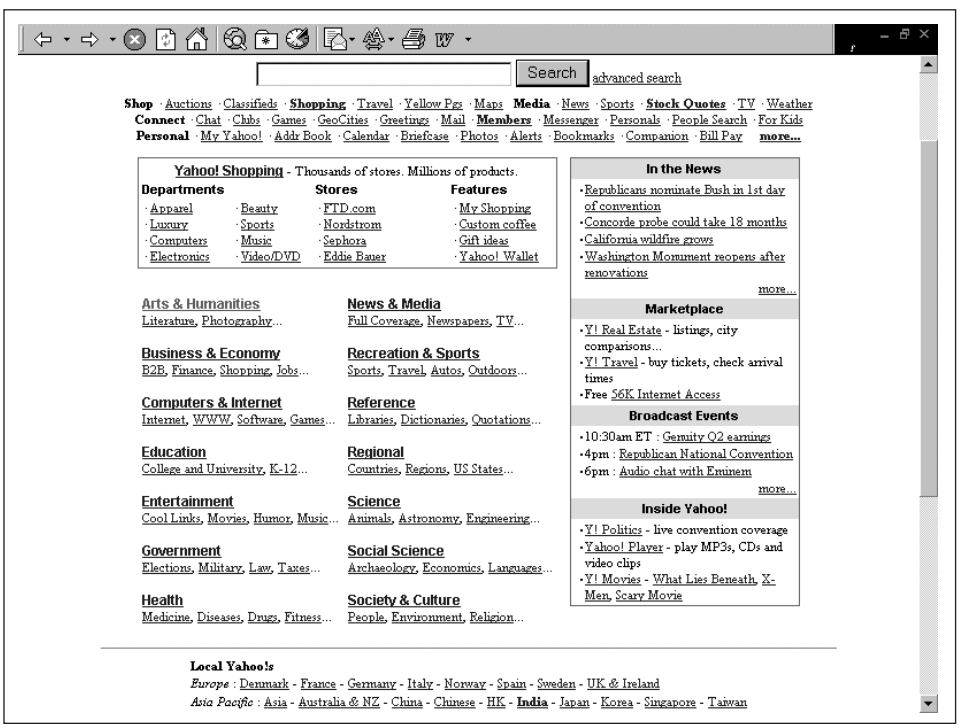
(Szemelvényeink közül T. Koch tanulmánya foglalkozik részletesen a hagyományos és automatizált osztályozási rendszereket alkalmazó internetkatalógusokkal.)

Önállóan kialakított osztályozási rendszert alkalmazó internetkatalógusok

Ezek alkotják az internetkatalógusok túlnyomó többségét.
A legfelső szinten néhány jól áttekinthető és főleg közismert szakterület (főosztály) jelenik meg. Az osztályozási rendszerek többnyire ismeretterületeket tartalmaznak, de vannak földrajzi, időrendi, dokumentumtípusok stb. szerinti rendszerek is.
A nagyobb, nemzetközi internetkatalógusokban szinte mindenütt saját fejlesztésű egyetemes osztályozási rendszereket használnak, melyeket túlnyomórészt a hagyományos osztályozási rendszerektől teljesen függetlenül, feltehetően azok ismerete nélkül, elsősorban kereskedelmi szempon-

tokat figyelembe véve alakítottak ki. A főosztályok kiválasztása és rendezettsége messzemenően a köznapi nyelvhasználat, gondolkodás és tájékozódás igényeit tükrözi. Ez egyben friss látásmód is az osztályozási rendszerek alapvetően konzervatív világában és előbb-utóbb számolni lehet megtermékenyítő hatásával a könyvtári-dokumentációs osztályozásra. Ugyanakkor számtalan következetlenség, dilettantizmus és rövidlátó prakticizmus forrása. Ezekben az osztályozási rendszerekben számos rendkívül rugalmasan alkalmazott megoldásra bukkanunk, jelentős részük a web körülményei között akkor is beválik, ha logikailag ellentmondásos, de gyakoriak a rendszer koherenciáját gyengítő megoldások is, melyek a későbbi fejlődés során bonyodalmakat okozhatnak.

A 6. ábrán az egyik leismertebb internetkatalógus, a Yahoo! portáloldalán megjelenő osztályozási rendszer legfelső hierarchiaszintje látható.



6. ábra. A Yahoo! internetkatalógus belépőlapjának részlete, melyen az osztályozási rendszer legfelső szintje látható. A legfelső három sorban az osztályozási rendszer hierarchiájától elkülönített osztályok kifejezései láthatók, melyek egy-egy adatbázis (pl. Shopping [Bevásárlás], Classifieds [Apróhirdetések]) vagy szolgáltatások (pl. My Yahoo! [a Yahoo! átszabása személyes igényeknek megfelelően]) belépőpontjai.

A nagy keresőszolgáltatások ma mintegy internetes húzóágazatként működnek, jelentőségüket nem lehet eléggé felbecsülni. Egyetemes igényű osztályozási rendszereiknek futtában végzett készítési és fejlesztési körülményeire fényt vet az alábbi interjúrészlet, melyben a Yahoo! osztályozási rendszerének szerzője a következőket nyilatkozza:

„Négy hónappal ezelőtt Srinavasan közölte velem, hogy további kategóriákat vett föl és szinte minden nap változtat valamit az ontológián.”[24]

Az internetkatalógusok osztályozási rendszereinek az osztályait, függetlenül azok szintjétől, a szolgáltatók általában „kategóriáknak” nevezik. Ez, és sok más elnevezésbeli eltérés a hagyományostól feltehetően éppen abból ered, hogy a készítőkben nem is tudatosult: olyan rendezőrendszert terveztek és használnak, melynek osztályaiba besorolják az információtételeket, azaz a rendszer segítségével osztályoznak. Innen nézve nem a rendszer logikai/filozófiai (kategoriális), hanem besoroló, „tartalmazó” szerepéről van szó, azaz dolgok (HTML-rekordok) osztályairól (nem pedig HTML-rekordok „kategóriáiról”). Az osztályozási rendszer sem „ontológia”, noha ugyanúgy létezik, akár a sertéscsülök, mivel az ontológia (a létről szóló tan) a filozófia egyik ága, tehát tudomány, az osztályozási rendszer viszont nem tudomány, hanem konkrétan létező termék. A hierarchikus osztályozási rendszerek korántsem olyan „nyitottak”, mint a tárgyszójegyzékek vagy tezauruszok, s ezért teljesen alkalmatlanok arra, hogy konzisztenciájuk összeomlása nélkül naponta változtatgassanak rajtuk.

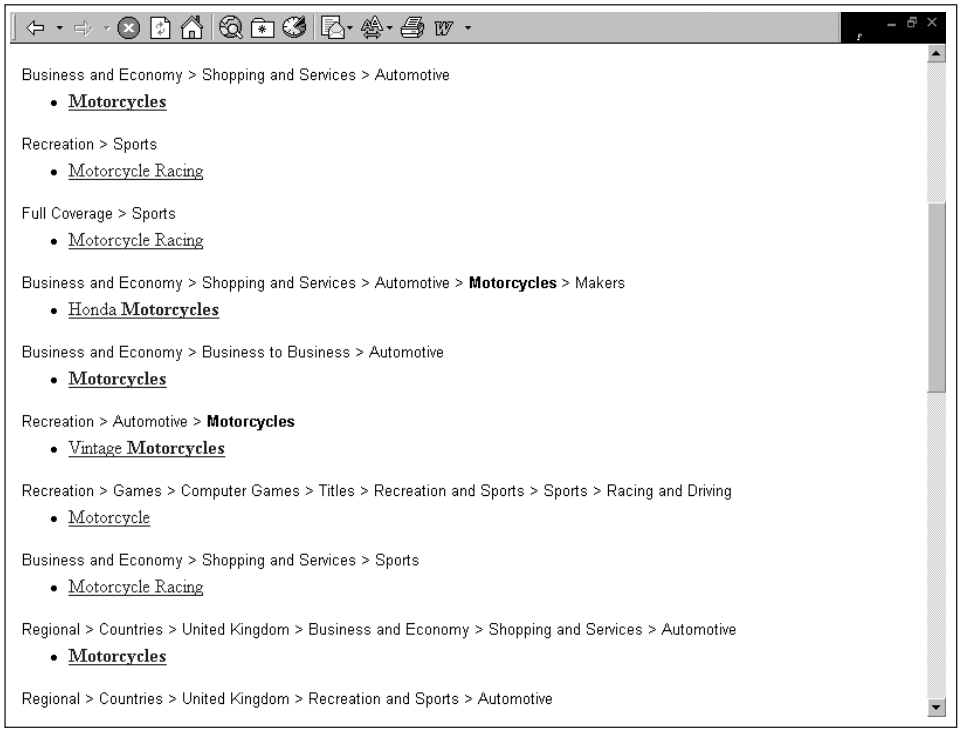
A tervezők osztályozási hagyományoktól való érintetlensége abban is megmutatkozik, hogy az egyes szinteken az ilyen típusú rendszerek többségében az osztályokat nem szisztematikusan, hanem betűrendben jelenítik meg. Indokaik kétségtelenül nyomósak: a lehető legkevesebb szellemi erőfeszítést szeretnének okozni a végfelhasználónak. A legfelső szinten még nem annyira feltűnő, hogy a hierarchikus rendszer adott szintjén a betűrend miatt össze nem tartozó osztályok kerülnek egymás mellé, mert ezen a szinten minden keresőszolgáltatásban a lehető leggyorsabb áttekintésre törekszenek: egy pillantással lehessen fölmérni, hogy a rendszer lényegében mit és hol tartalmaz. Az alsóbb szinteken azonban szokatlan találkozások adódnak. A Science (Tudomány) második szintjének több mind 60 osztálya például így kezdődik: Acoustics (Akusztika), Agriculture (Mezőgazdaság), Alternative (Alternatív techni-

24 Steinberg, Steve G.: Seek and ye shall find (maybe) [Keresni, és – ha csak lehet – találni is, jé]. In: Wired, 4 (1996) 5., p. 108–114, 172–182.

kák), Amateur science (Amatőrök által művelt szakterületek), Antropologhy and Archeology (Embertan és régészet), Artificial Life (Mesterséges élet), stb.

A hierarchikus rendszer nem különösen „mély": alig 3-4 szintet tartalmaz. Ezért jelenik meg a második és a harmadik szinten olykor nagyon sok osztály. A szerkesztők valószínűleg nem mernek a már széles körben megismert főszerkezeten változtatni; ilyen változtatás nélkül azonban nem oldható már meg, hogy az egyes szinteken az osztályok számát csökkentsék. Az egész emlékeztet a termeszek hangyaboly-építményeire: a fejlődés szerves és nagyon gyakorlatias, mindig kizárólag a lehetőségekhez igazodik, soha sem elvekhez. Kétségtelen, hogy az elvek alkalmazásának vannak praktikus határai. De az is igaz, hogy a prakticitás túlfeszítéséből is adódnak határok. Van, amikor már nincs megtévesztőbb, mint a realitás. Az eddig megjelent átfogó internetkatalógusok egyetemes célú osztályozási rendszereit nem jellemzi a felosztási szempontok következetessége. Érezhető, hogy kereskedelmi szempontok érvényesülnek az osztályok fölvételében: az a felfogás, hogy „mi van azon a szakterületen eladható információ". Ez határozza meg, milyen osztályokat vesznek föl a rendszerbe. Csak feltételezzük, hogy a keresőszolgáltatások gépei által feldolgozott információtételek mennyiségének növekedésével a rendszerek finomszerkezete tartalmilag fokozatosan koherensebbé válik. Ugyanakkor az alkotók szakmai érintetlenségének előnyei is vannak: friss szemmel vágtak neki a világ rendszerező célú felosztásának, s ez hosszabb távon nem maradhat következmények nélkül a hagyományos könyvtári és dokumentációs osztályozásra sem.

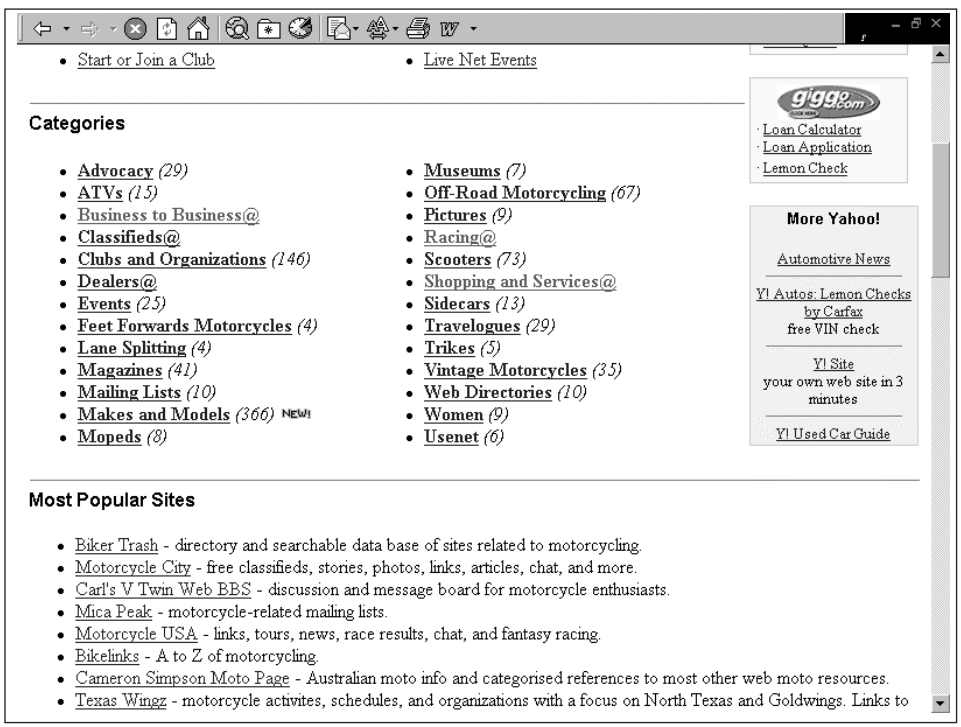
Különösen hasznos megoldások születtek az ilyen osztályozási rendszerek hierarchiáncai között. Ennek alapja, hogy a hipertext ***a*** kereszthivatkozások eszményi rendszere, és ezt hasznosítják a hierarchikus szerkezeten belül is. Itt is létrehoznak keresztirányú összefüggéseket. Ez abban nyilvánul meg, hogy egy-egy osztály egyszerre több magasabb szintű osztály alárendeltje is lehet, az osztályozási rendszerek tehát – szemben a hagyományos egyetemes könyvtári rendszerekkel – polihierarchikusak. Ez olykor rendkívül bonyolult, néha már lehetetlennek tűnő struktúrákat eredményez, de a felhasználót nagyon jól szolgálja, mert az ismétlődések következtében a hierarchikus rendszer redundáns.
A 7. ábrán azt láthatjuk, hogy például a Motorcycles (Motorkerékpárok) hány különféle hierarchiáncon belül jelenik meg. Mindig van „gazdaosztály" („szülőosztály"), melyhez a polihierarchikusan alárendelt alosztály kapcsolódik (a többi előfordulást a megjelenítésben a @ jellel jelölik).



7. ábra. A Motorkerékpárok (Motorcycles) polihierarchikus előfordulása a Yahoo! osztályozási rendszerében

A 8. ábrán a Motorkerékpárok osztály alatti utolsó előtti hierarchiaszint látható. Megjelenítettük az első néhány találatot is azok közül az információtételek közül, melyeket az átfogó Motorkerékpárok osztályba soroltak, és nem az ennél speciálisabb alosztályok valamelyikébe.
Kerek zárójelek között az osztályhoz tartozó találatok száma látható. Azokat az alosztályokat, melyek alapvetően nem ide tartoznak, noha itt is föltüntették őket, a @ jelöli.
A helyzet azonban ennél bonyolultabb. A szerkesztők friss szemléletét minden jel szerint nyelvészeti szempontok sem kötik gúzsba; nem sokat foglalkoznak például a homonimák megkülönböztetésével. Gyakori, hogy ugyanazzal a névvel a rendszeren belül másik helyen másik osztályt is jelölnek amelynek vagy nem ugyanaz a terjedelme (nem azonosak a hozzá besorolt információtételek), vagy nem ugyanahhoz a felosztáshoz tartozik (nem azonosak a fölötte megjelenő alosztályok). Például a Motorkerékpárok osztályai között vannak olyanok, amelyek a Recreation főosztály fokozatos alosztásaiból keletkeztek. A „Recreation–Automotive–Motorcycles” és a „Recreation–Hobbies–

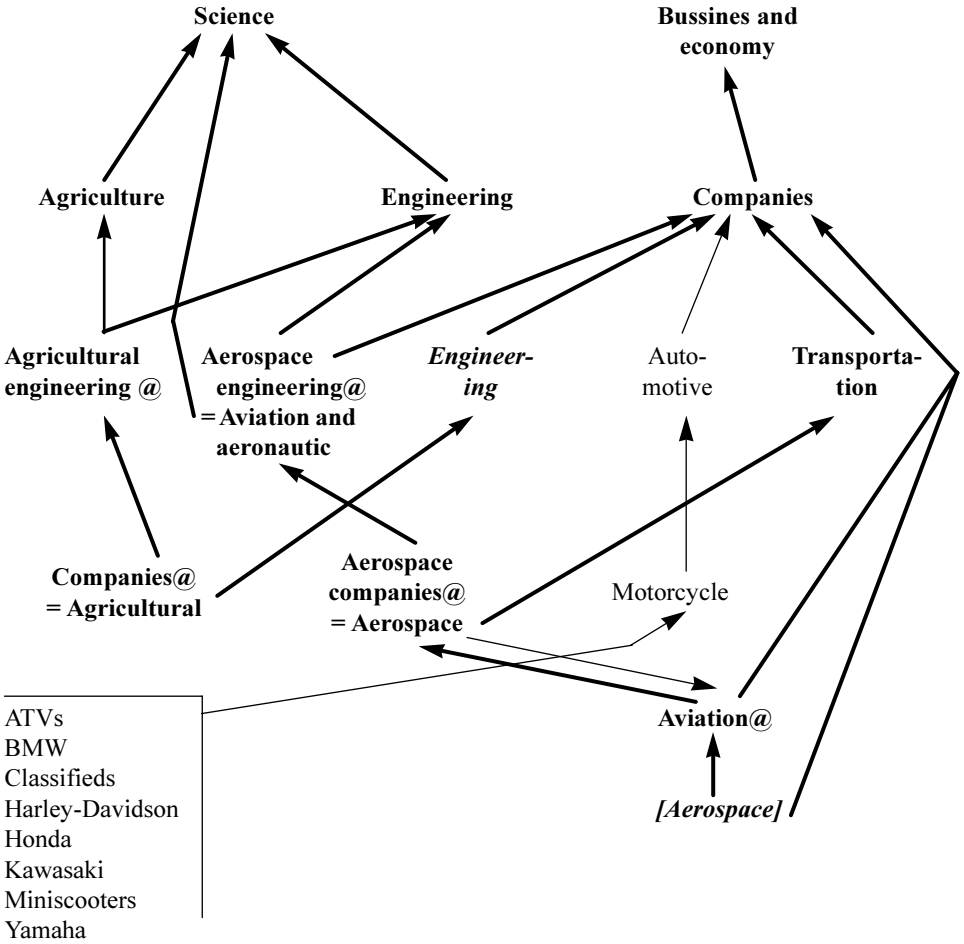
Models–Motorcycles" láncban a Motorkerékpárok osztálya nem ugyanaz az osztály-előfordulás a rendszeren belül, mint mondjuk a Bussines and Economy–Companies–Automotive–Motorcycles láncban szereplő Motorkerékpároké. Ezért az előbbi két osztálylánc Motorkerékpárok osztályát a következőképpen kellene megkülönböztetni a többi, ugyanilyen nevű osztálytól: „Motorkerékpárok (a szabadidő és a barkácsolás szempontjából)". A szerkesztők nyilván abból indulnak ki, hogy maga a hierarchialánc is definiálja a jelentést. Hozzá kell azonban tenni, hogy „adott esetben". Más esetekben ugyanis eltérő hierarchialáncokban ugyanaz az osztály szerepel (pl. Motorkerékpárként), azaz az eltérő hierarchialánc nem definiál eltérően.



8. ábra. A Motorkerékpárok osztályának alosztályai és a Motorkerékpárok osztályba sorolt találatok jegyzékének eleje

A struktúrák gazdagsága

Hogy ezeknek az osztályozási rendszereknek a rejtett szerkezeti bonyolultságát jobban lássuk, a 9. ábrán a Yahoo! osztályozási rendszerének egy részletét kiemeltük, és címkézett irányított gráffal ábrázolva mutatjuk meg.



9. ábra. A Yahoo! polihierarchikus osztályozási rendszerének részlete címkézett, irányított gráf formájában

Az előbbiekben tárgyalt Motorkerékpárok osztály összefüggéseit a jobb elkülöníthetőség kedvéért nem félkövéren jelenítettük meg.

A gráf alapján a következők ismerhetők fel.

Az Agricultural engineering (Agrotechnika) egyrészt az Agriculture (Mezőgazdaság), másrészt – @ jelöléssel – az Engineering (Mérnöki tudományok/ Technika) alosztálya.

Az Aerospace engineering (Repüléstechnika) az Engineering és a Companies (Cégek), továbbá Aviation and auronautic (Légügy/Repüléstan) néven a Science (Természettudomány) alosztálya, mely utóbbinak ugyanakkor tranzitív alárendeltje.

Az, hogy ugyanazt az osztályt más néven a tranzitív fölérendelt alá rendeljék, hajmeresztő a hagyományos osztályozási rendszerek ismerőjének (olyan ez, mintha a Kutyát egyrészt alárendelik a Háziállatnak, ugyanakkor Eb néven az Állatnak, melynek ugyanakkor a Háziállat a közvetlen alárendeltje). A piaci viszonyok terén iskolázott rendszertervező viszont abból indulhatott ki, hogy a Természettudományok felől nézve jobban fest az általánosabban megfogalmazott osztálymegnevezés (Légügy…), nem pedig a Repüléstechnika, amely viszont a Technika felől nézve adekvátabb osztálynév.

Azt is észre kell venni, hogy az Aerospace engineering az Engineering alá rendelve valójában olyan osztályt képvisel, amely a repüléstechnikára vonatkozó ***információk*** tételeit tartalmazza, a Companies alá rendelve pedig azt, amely a repüléstechnikával foglalkozó ***cégek*** információit tartalmazza. Ennek a példának az esetében nincs a Yahoo!-ban különbség a két osztály terjedelme (információtételei) között.

Az Engineering esetében azonban van. Ebből ugyanis két osztályt találunk, de ez a két osztály nem ugyanaz: a Cégeknek alárendelt osztály ugyanis – melyet dőlt betűvel jelenítettünk meg – csak a műszaki tevékenységeket végző cégek információtételeit tartalmazza, a Természettudományoknak alárendelt Engineering ezzel szemben minden, a technikára és a műszaki tudományokra vonatkozó információtétel osztályozására való.

A dőlt betűvel megjelenített *Engineering* alárendeltje az Agricultural (Mezőgazdasági) [így, jelzősen], mely az agrotechnikai cégek információtételeit tartalmazza. Ugyanennek az osztálynak az Agrotechnika alárendeltségében viszont Companies (Cégek) a neve. Ha belegondolunk, ez egész logikus: az Agrotechnika felől nézve cégekről, a műszaki cégek felől nézve meg „mezőgazdaságiról", azaz Agrotechnikai (cégekről) van szó.

Talán a legmerészebb húzás, amikor ugyanazt az osztályt alárendelik egy másiknak, ugyanakkor fölérendelik neki. Ez a helyzet az Aerospace (= Aerospace companies @) és az Aviation között. De ha meggondoljuk, hogy ezekben az osztályozási rendszerekben egyáltalán nincs pontosan meghatározva, hogy mit is értünk tulajdonképpen azon a relación, amely az egyes osztályokat összekapcsolja, ez a megoldás korántsem olyan hajmeresztő, mint ahogy logikai szempontból látszik. Eddig ugyanis abból indultunk ki, hogy az internetkatalógusok osztályozási rendszerei hierarchikusak, és alapvetően csak alá–fölérendeltségi kapcsolatokat tartalmaznak. Valójában azonban olyan rendezőrendszerekről van szó, melyekben nincs egyértelműen definiálva a kapcsolat: lehet hierarchikus (az esetek többségében), de van, amikor egyszerűen csak annyit jelent, hogy „lásd még". Az Aerospace és az Aviation között valójában az utóbbi összefüggésről lehet szó, és ez logikailag teljesen megengedett. Más lapra tartozik, hogy ezekben az osztályozási rendszerekben a mindenkori, definiálatlan relációt csak az jelöli, hogy „az egyik következik a másik után". Ha a tezauruszszabvány szerint pontosan jelölnénk a tárgyalt esetet, a 10. ábrán látható szócikkeket kapnánk:

Transportation	Aerospace	Aviation
A Aerospace	F Transportation	F Transportation
Aviation	X Aviation	X Aerospace

10. ábra. Yahoo! összefüggések szabványos tezauruszcikk formában[25]

Az osztályozás

A HTML-dokumentumok tartalmi leírása egyrészt abból áll, hogy besorolják a megfelelő osztályba, és az osztály dokumentumhoz kapcsolt megnevezése vagy jelzete egyben „leírás" is. Ez a tartalmi leírás azonban formális adatok (szerző, cím, kiadó, annotáció stb.) nélkül használhatatlan, mert nincs, ami a dokumentumot egyértelműen azonosítaná (az URL kivételével).

Az internetkatalógusokban intellektuálisan dolgozzák föl a HTML-dokumentumokat, ezért nem készül keresőprogrammal („keresőgéppel") automatikusan formális dokumentumleírás (mint amilyenek az indexelőszolgáltatások vonatkozásában a 2. és 3. ábrán láthatók). A formális leírásokat tehát szintén manuálisan kell elkészíteni, hogy létrejöjjön a metaadatokat (szerző, cím stb.) tartalmazó teljesebb másodlagos információtétel. Ezeket az esetek jelentős részében maguk a beküldők, tehát laikusok készítik el.

Az önkéntesen beküldött tételek számos katalógusban többségben vannak, de jóformán minden kereskedelmi célú szolgáltatásban rendelkezésre állnak bejelentési űrlapok. A Yahoo!-ban pl. az egyes osztályok lapjának alján található „Suggest a site" , (másutt „Add a site here", „Add URL" stb.) csatolóval hívható be. Bennük megtalálhatók a rovatok az osztályozás, a cím (Title), URL, tartalmi kivonat (Description) stb. számára.

„A beküldött űrlapok adatait elvileg a szerkesztők felülvizsgálják. A tapasztalatok arra utalnak, hogy ez annál nehezebb, mennél szabadabban adhatók meg az adatok, annál nagyobb munka az egységesítésük. Mivel a mennyiségi növekedés miatt egyre kevésbé képesek a szolgáltatások saját erőből elvégezni a leírásokat, a metaadatok megállapítását igyekeznek a beküldőkre bízni. Ennek érdekében részletező űrlapok szükségesek, hogy a laikus mindent jól értsen (jól példázzák ezt a Magellan és az NISS űrlapjai). A metaadatok előrehaladt nemzetközi szabványosítása, különösen pedig a Dublin Core metaadat szabvány az

25 Ungváry Rudolf: A tartalom szerinti információkeresés az Interneten. In: Tudományos és Műszaki Tájékoztatás. 10. évf. (2000) 1. sz. p. 3–19.

internetkatalógusok információtételeiben a leírások egységesülését segíti elő. A fejlettebb katalógusokban, mint amilyenek a szakmai információs kapuszolgálatok, részletesebb és színvonalasabb rekordleírási szabályzatok alakulnak ki.

A tételek megjelenítése és találati értékelése szempontjából különösen a tartalmi kivonatnak van nagy jelentősége. Számos katalógusban ez még csak egyetlen mondat. A részletesebb leírásokat szolgáltató katalógusokban a tartalmi kivonatot szemlének (review) is nevezik, de ezek sem lépik túl a hagyományos annotációk terjedelmét.

Különösen az igényesebb szolgáltatásokban előfordul, hogy az osztályozási rendszer valamelyik osztályába besorolt dokumentumhoz még tárgyszavakat vagy deszkriptorokat lehet rendelni. Mivel számos internetkatalógusban nemcsak böngészni lehet az osztályozási rendszer hierarchikus szerkezete mentén, hanem természetes nyelven is le lehet kérdezni az állományt, a tárgyszavak és deszkriptorok kereshetőbbé teszik a HTML-rekordokat. A Beyond Bookmarks[26] szerint 2000 elején 20 szolgáltatásban használtak szabványosított természetes nyelven alapuló szótárt, ezen belül 13 tezauruszt. Az Engineering Electronic Library például a hierarchikus osztályozási rendszere mellett saját tezauruszt is használ. A NetFirst a Kongresszusi Könyvtár osztályozási rendszerének (LCC) dokumentumtipológiája szerint is osztályoz.

Vannak internetkatalógusok, melyekben intellektuálisan értékelik a dokumentumokat (pl. Argus Clearinghouse, Lycos/Point Top 5%, excite Reviews és Magellan's Reviews). Többnyire 1 és 5 közötti skála értékeit adják meg pontokban."[27]

Lekérdezés az internetkatalógusokban és a kereső- és böngészőszolgáltatás egyesítése

Általános jelenség, hogy az internetkatalógusokban nemcsak a hierarchikus osztályozási rendszerben lapozgatva lehet böngészni, hanem megadható külön ablakban természetes nyelven a keresett szó. Ha ez megegyezik a rendszer valamelyik osztályának nevével, vagy nevének részletével, akkor a kereső rögtön az adott osztálynál találja magát (így kérdeztük le pl. a 6. ábrán a „motorkerékpár" kifejezést a Yahoo!-ban).

E nem különösen szellemes segítségen kívül azonban megfigyelhető tendencia, hogy a katalógusokat integrálják az indexelőszolgáltatásokba. A katalógusok adatbázisainak mérete lényegesen kisebb, mint az indexelőszolgáltatásoké. Mivel többnyire intellektuálisan osztályoznak, a teljességre eleve nem törekedhetnek. Annak érdekében, hogy még több releváns

26 Beyond Bookmarks, id. mű.
27 Oehler, A., id. mű.

adatot szolgáltassanak, hogy ők legyenek a „legjobbak a weben" („the Best of the Web") „keresőgépet" is alkalmaznak, és az így megvalósítható lekérdezést szorosan vagy kevésbé szorosan összekapcsolják a böngészéssel. Általános gyakorlat, hogy az osztályozási rendszer bármelyik pontjából mind az osztályozási rendszer megnevezései, mind pedig a „keresőgép" által indexelt állomány lekérdezhetők. A szorosabb integrációra jellemző példa az excite meg a Magellán, melyben kiválasztható, hogy az egész adatbázisban, a katalógus intellektuálisan feldolgozott és értékelt tételei (rated and reviewed sites) között, vagy a gyerekek számára is megengedhető „zöld" tételek állományában („green light sites") kívánunk keresni. A puszta egymás mellett létezésre is számos példa akad (mint a Lycos német változatában).

Regionális katalógus változatok

A nagyobb internetkatalógusok egyre több nemzeti/nyelvi változatot is létrehoznak. Ezek jelentős része valójában teljesen önálló, csak éppen átveszi a know how-t. Bennük csak az adott ország, régió forrásait dolgozzák föl. A Yahoo! jelenleg már tucatnyi nemzeti változatban létezik, de a Lycos se nagyon marad le mögötte. Az előbbiben a World Yahoo! osztály alatt találhatók meg az egyes nyelvi változatok, melyek nem pontos másolatai az angolnak, hanem az adott ország körülményeihez alkalmazkodó fejlesztések (van már kínai nyelvű is).
A tendencia – kevésbé erőteljesen – az indexelőszolgáltatások terén is megfigyelhető, jellegzetes példa erre az AltaVista magyar változata, a MATÁV AltaVizsla indexelőszolgáltatása, vagy a nemzetközi Metacrawler és annak német változata, a MetaGer.
Nem tévesztendők össze a nagyobb keresőszolgáltatók regionális változatai az önálló nemzeti jellegű keresőszolgáltatásokkal. A magyar Hungary.Network HUDIR internetkatalógusa például teljesen önálló fejlesztés, noha korai változatában a Yahoo! mintáját követte, az első magyar indexelőszolgáltatás, az ugyancsak Hungary.Network által fenntartott Heuréka pedig az AltaVistától teljesen függetlenül jött létre.

Speciális adatbázisok

Mind az indexelőszolgáltatásokra, mind az internetkatalógusokra jellemző, hogy a keresőprogramokkal („keresőgépekkel") végzett lekérdezést, ill. a hierarchikus katalógusaikban végezhető böngészést különféle kisebb adatbázisokkal és szolgáltatással is kiegészítik, melyek többsége

önálló, tágabb értelemben vett, nagyon specializált keresőszolgáltatásnak is tekinthető. Afféle miniatűr on-line szolgáltatókká válnak. A nagyobb piaci részesedés és a reklámbevétel növelésének reményében létrehozott kiegészítő adatbázisokra jellemző, hogy általános érdeklődésre tarthatnak számot, ingyenesek és könnyen kezelhetők. Ezek az adatbázisok a hierarchikus rendszertől elkülönített osztályok (Bevásárlás, Apróhirdetések, Szótárak stb.) formájában jelennek meg a portállapokon. Az osztályozáselmélet szemszögéből felsoroló, enumeratív osztályozási rendszert alkotnak. (A Yahoo! esetében ilyen enumeratív rendszert képviselnek a 6. ábra legfelső három sorának osztályai.) Könyvtárszervezési szempontból azt mondanánk, hogy ahány osztálytípus, annyiféle gyűjtőköri forrástípus.

Az osztályok (adatbázisok) típusai

Szakterületek, tudományok, tevékenységi körök

Arts & Humanities (Művészet és társadalomtudomány)
Bussines & Economy (Kereskedelem és gazdaság)
Computers & Internet (Számítástechnika & internet)
Education (Oktatás–művelődés)
stb.

Ezek az osztályok felelnek meg a dokumentumok hagyományos osztályozási rendszereiben alkalmazott osztályoknak, de itt is lépten-nyomon érheti az embert meglepetés: valamelyik szakterületen belül felbukkanhat apróhirdetéseket tartalmazó osztály, vagy tényadatokat tartalmazó osztály stb. (A 4. ábrán a felső vízszintes elválasztó vonal alatti hierarchikus rész ezekből az osztályokból épül fel.)

Kereskedelmi jellegű osztályok:

Shopping (Bevásárlás)
Travel Agent, Travel Finder (Utazási irodák)), Book a hotel (Szállodafoglalás)
Buy a car, Buy a home (Autóvásárlás, Lakásvétel)
Classified (Apróhirdetések, üzleti)
Personals (Apróhirdetések, személyi)
Careers, Jobs (Álláshirdetések)

Ezek elsősorban arra valók, hogy az adás-vételt támogassák. Az osztályok erősen válogatott, csak a rendelésfeladás szempontjából szóba jöhető szakterületek. Ezeken belül a besorolt információtételekből kiindulva megrendelhetők árucikkek, utazáshoz jegyek, elérhetők a hirdetések feladói.

Adattárak, címek, helyek osztályai

Community (Közérdekű és igazgatási információk)
Yellow Pages (Szakmai telefonkönyv), White Pages (Betűrendes telefonkönyvek)
People Search (Drótpostacím és személykeresés), WhoWhere (Ki kicsoda)
Search for Missing Children (Eltűnt gyerekek)
Books (Könyvek)
Auctions (Kiállítások, árverések)
Maps (Térképek)
Pictures & Sounds (Képek, Hangdokumentumok)
Photo Finder (Fényképek)
Dictionaries, thesauri (Szótárak, tezauruszok)
Airlaine Tickets (Repülőjegyek), Menetrendek

Ezekben az osztályokban fehér és sárga telefonkönyvek, cégek, személyek adatait tartalmazó információtételek, egyéb céginformációk találhatók. Elmondható, hogy a segítségükkel az internethez már kapcsolódó országok túlnyomó részében szinte minden cím megtalálható. A térképek esetében helyek azonosíthatók vizuálisan. A szótárak valamint a tezauruszok egy része többnyelvű. Egyes keresőszolgáltatások felveszik a közlekedési vállalatok menetrendjeit is az enumeratív osztályozási rendszerükbe. Különlegesség – például az Infoseekben – a személyes honlapokat tartalmazó adatbázis.

Hírek, tényadatok

Today`s news (Aktuális hírek)
Stock Quotes (Tőzsdehírek)
Sports (Sporthírek)
Weather (Időjárás-jelentés)
TV (Tévéműsor)

Ezekben az osztályokban tényadatok szerepelnek.

Segítségek, gondűzők

Calendar (Naptár, események)
Horoscopes (Horoszkópok)
Games (Játékok)
Pager (Letöltő)
My Yahoo! (Testre szabható Yahoo!)
Yhooligans (Gyerekek Yahoo!-ja)
E-mail (Drótposta bejelentkezés)
Funny Site (vicckereső)

Ezekben az osztályokban a mindennapokban hasznos eszközök és játékok találhatók. Többségük valójában nem is osztály (nem információtételeket tartalmaznak), hanem speciális szolgáltatások belépőpontjai.

Szolgáltatásként – a szótárakat és tezauruszokat kiegészítendő – feltűnnek az automatikus fordítórendszerek is; velük tetszés szerinti szöveg gépi fordítása végeztethető el a nagyobb világnyelvek között, az URL megadásával egész honlapok is lefordíthatók.[28]

Terminológia

Ha az internetkatalógusokban, osztályozási rendszerek alapján végzett keresésről, azaz „szisztematikus lapozásról", vagy „strukturált gyűjteményekben való navigálásról" van szó, mindig böngészésről beszélünk. Az angol és német szakirodalomban túlnyomórészt „browsing" a neve.
Az internetes indexelőszolgáltatásokban természetes nyelvű kifejezésekkel, tárgyszavakkal, deszkriptorokkal és a boole-műveletek segítségével végzett keresésre az általános keresés vagy a lekérdezés szót használjuk (searching, scanning, Suche).
Ha a dokumentumok szövegén belül hipercsatolók felhasználásával – tehát nem szisztematikus rendszer mentén – kutakodunk, „szörfölésről" (surfing, Surfen) beszélünk. Az utóbbival összefüggésben beszélnek olyan keresésről, melynek során értékes dolgok fedezhetők föl kevéssé valószínű helyeken is (serendipitous discovery); ezt nevezzük „innovatív vagy felfedező keresésnek". Ellentéte a hagyományos eszközökkel végzett böngészés és lekérdezés, melyekre összefoglalóan angolul (a „tunnel vision" = csőlátás analógiájára) a nem túl hízelgő „tunneled searching" kifejezést használják (magyarul „kötött pályás keresés").
A böngészés, lekérdezés és szörfölés, ill. a kötött pályás és az innovatív keresés szakterülete az információkeresés (information retrieval). E szakterülethez tartozik az automatikus indexelés és osztályozás is.
Hagyományos körülmények között a szörfölésnek a könyv teljes szövegében végzett lapozás, a böngészésnek a tartalomjegyzékben, a lekérdezésnek a név- és tárgymutatóban végzett keresés felel meg.

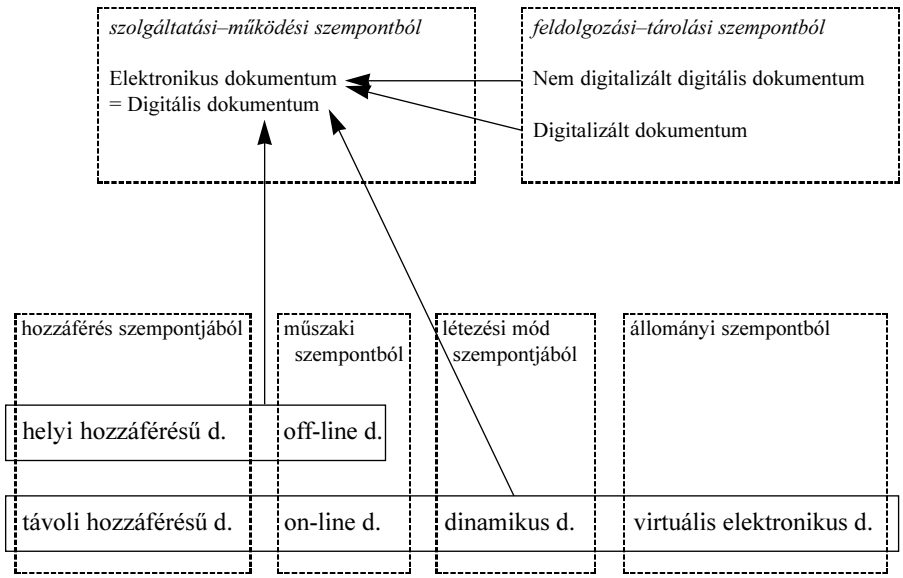
28 A leginkább elterjedt Systran fordítórendszert alkalmazza az AltaVista Translator <http://world.altavista.com/> és a Go translator service <http://translator.go.com/>

Internetes dokumentumformátumok

A digitális és a virtuális dokumentum fogalma

Az internet különféle dokumentumai alkotják a virtuális könyvtár potenciális gyűjtőkörét. E gyűjtőkör dokumentumai túlnyomórészt nem kerülnek a könyvtár fizikai értelemben vett állományába, a könyvtári tárolás szempontjából ezek a dokumentumok virtuálisak.
A digitális (csak digitális formában létező) és digitalizált (eredetileg nem elektronikus formában készült) dokumentumok a digitális könyvtár gyűjtőkörét alkotják. Ezek a dokumentumok lehetnek az internet html-dokumentumai is, de olyanok is, melyek fizikai értelemben is a könyvtár állományába tartoznak, tehát tárolási szempontból nem virtuálisan, hanem fizikailag léteznek (pl. CD-ROM kiadványok). Az elektronikus könyvtár lényegében a digitális könyvtár szinonimája (egyes szakemberek szolgáltatási–működési szempontból elektronikus, feldolgozási–tárolási és adatszervezési szempontból digitális könyvtárról beszélnek).
Az egyes könyvtárak által feldolgozott, de állományba nem vett html-dokumentumok az adott könyvtár szempontjából virtuálisak, melyek távoli hozzáféréssel érhetők el (szemben az állományba vett elsődleges dokumentumokkal, melyek helyi hozzáférésűek).
Tágabb értelemben virtuális minden olyan dokumentum, mely nem tartozik az adott könyvtár állományába, de a könyvtárban katalogizálták, és ezért a másodlagos információk (mint pl. a katalógustétel adatai) alapján elérhető. A gyakorlatban csak az elektronikus dokumentumok esetében van a virtuális jellegnek jelentősége. A távoli hozzáférésű elektronikus dokumentumok mind virtuális dokumentumok, melyeket dinamikus dokumentumoknak is neveznek.
Mindezek alapján a virtuális könyvtár a digitális könyvtár egyik fajtája (a másik fajtája pl. a CD-ROM könyvtár). Fordítva ez nem igaz: nem minden digitális könyvtár virtuális.
Digitális, de elsősorban virtuális könyvtári környezetben a dokumentum fogalma problematikussá válik, ezért inkább digitális objektumokról beszélnek. Ezek megfelelnek a hagyományos könyvtárak állományi egységeinek (könyvek, térképek, zeneművek stb.). Mind a digitális és digitalizált, mind a hagyományos könyvtári dokumentumok elsődleges adatokat tartalmaznak, és maguk is elsődleges dokumentumok.
Beszélnek még off-line és on-line elektronikus dokumentumokról. Az előbbiek az adott könyvtár állományában vannak (pl. CD-ROM típusú dokumentumok), az utóbbiakat csak külső on-line hozzáféréssel lehet használni. Az off-line dokumentumok a helyi hozzáférésű elektronikus dokumentumok, az on-line elektronikus dokumentumok pedig a távoli hozzáfé-

résűek. Az utóbbiak felelnek meg a virtuális elektronikus dokumentumoknak. A terminológiát a 11. ábrán címkézett, irányított gráffal szemléltetjük. A közös szaggatott keretbe foglalt kifejezések közös szempontból megfogalmazott megnevezések. A közös folytonos keretbe foglalt kifejezések egymás szinonimái. Az azonos jelentés könnyen ellenőrizhető: ha pl. minden „elektronikus dokumentum" „digitális dokumentum", és minden „digitális dokumentum" „elektronikus dokumentum", akkor a két megnevezés ugyanazt a dokumentumot jelöli, tehát szinonim.



11. ábra. Az elektronikus dokumentumok átfogó tipológiája

Az elektronikus (digitális/digitalizált és virtuális) dokumentumok és a hagyományos dokumentumok között az alapvető különbség, hogy az előbbieknek mind a tárolása, mind az olvashatósága ugyanabban a gépi keretrendszerben játszódik le. (A hagyományos dokumentumokat nem gép tárolja, noha géppel [be/le]olvashatók.) A digitálisan feldolgozott dokumentumot a számítógép mintegy „belülről" ismeri, azaz minden adatához funkcionálisan hozzáfér. Ebből következik, hogy az elektronikus dokumentumszövegek gépi kezelési szerkezetének funkcionális szempontú szintaktikai–szemantikai egységesítése közérdek: ilyen módon válik ugyanis lehetővé, hogy a dokumentumokat (objektumokat) a legkülönfélébb információs szervezetek nehézségek nélkül kezelni tudják, amikor arról van szó, hogy szolgáltatni kell őket.

Formátumok

Elsődleges dokumentumok formátumai

Ebből a célból születtek meg az elektronikus dokumentumok formátumszabványai, melyek alapján a digitális/digitalizált szöveg bizonyos szerkezeti egységei egységesen kódolhatók (minősíthetők). Rendeltetésüket tekintve nagyon hasonlóak azokhoz az adatcsere-formátumokhoz, melyeket a másodlagos adatokra vonatkozó dokumentációs és könyvtári adatok számára alakítottak ki. Lényeges különbség, hogy elektronikus dokumentumok esetében a szabványosítás a közvetlen számítógépes kezelhetőség és olvashatóság következtében már az elsődleges dokumentumra vonatkozóan megvalósítható, ill. megvalósítandó. (A könyvtári-dokumentációs bibliográfiai adatcsere-formátumokkal kötetünk korábbi részében részletesen foglalkozik *Mirna Willer.*) Mivel a nyomtatott dokumentumok ma már számítógépek igénybevételével készülnek, létezik elektronikus változatuk, melyek előbb-utóbb bekerülnek a tárolandó és kereshető állományok világába.

Az elsődleges elektronikus dokumentumok szerkezetét az elsődleges dokumentumon belül leíró metaadatszabvány az 1986-ban elfogadott (ISO 8879) SGML (Standardized General Mark-up Language; Szabványos Általános Jelölő Nyelv). Készítői az egyszerűbb és a tényeknek megfelelőbb „formátum" vagy szabvány helyett a „nyelv" megnevezést használták, noha nincs szó olyan értelemben mesterséges nyelvről, mint amilyenek a programnyelvek (hiszen a formátum, akárcsak az űrlap vagy a könyv, nem nyelv, hanem valamilyen nyelven kifejezett információ, adat, esetünkben szabvány). Az SGML szabvány elsődleges feladata ugyanaz, mint a MARC formátumoké, szintaktikai és szűkebb értelemben szemantikai szabályokat határoz meg a szöveg hierarchikusan rendeződő elemeinek formális leírásához. Alapvető különbség a MARC formátumokhoz képest, hogy az SGML segítségével ugyanazt a dokumentumot különféle – konkurens – szerkezetben is le lehet írni. Az adott, ténylegesen használt leírás neve a Document Type Definition, a DTD.

A HTML (Hypertext Markup Language; Hipertext Jelölő Nyelv) a web közismert adatformátuma, valójában SGML-alkalmazás, vagyis egy lehetséges DTD, melyet a World Wide Web Consortium (a W3C) definiált. A webnézegetők valójában olyan SGML-olvasók, melyek csak meghatározott – viszonylag egyszerű – DTD feldolgozására alkalmasak. A HTML DTD elsősorban olyan alkotóelemeket tartalmaz, amelyek a képernyőmegjelenítést szabályozzák, vagyis minimális mértékben határozza csak meg az adat logikai–szemantikai szerkezetét, hierarchiáját. Mint ilyen, kevéssé alkalmas a jól visszakereshető, strukturált digitális objektumok rög-

zítésére. A kliens-szerver szerkezetű dinamikus keresőszolgáltatások megjelenése fokozatosan megváltoztatja ezt a helyzetet, melyről Lou Burnard, az SGML szintaxison alapuló szemantikai rendszer, a TEI (Text Encoding Initiative) egyik szerkesztője így ír:

„Mégis, miért használjuk a HTML-t? A gazdasági, politikai és szociológiai érvek mellett van még egy eddig figyelmen kívül hagyott szempont: a web tartalmának jelentős része eredendően tiszavirág-életű. Ezek az anyagok csak itt és most kívánnak hatni, például terméket eladni vagy egyszerűen szenzációt kelteni. Ebből következően semmi értelme ezekre több energiát pazarolni, mint a hasonló papírbrosúrákra. A gondot inkább az okozza, hogy éppen úgy a HTML-t kell használunk, ha fontos kézikönyvet digitalizálunk, mint ha éppen üdítőitalt reklámoznánk.

Valójában azonban még az értékesebb művek rögzítésénél is csak akkor tűnik föl a HTML gyengesége, ha a szerző vagy a kiadó szempontjából vizsgáljuk a helyeztet. Ha a képernyőkép tetszetős, az olvasó számára végső soron mindegy, hogy a korszerű objektumorientált adatbázis-kezelőből, postscript fájlból vagy pedig fekete mágiával előállított HTML-fájlból származik-e… A HTML-nek, mint szerveroldali formátumnak van néhány nyilvánvaló hátránya. Noha a kezdeti költségek kicsik, HTML-dokumentumokkal aligha tanácsos komolyabb, hosszabb távú szolgáltatást indítani. A csatolók konzisztenciájának megőrzése már viszonylag dinamikus állomány esetében is rendkívül sok fejfájást okozhat."[29]

A megoldást minden jel szerint a tényleges SGML és a kurrens HTML-változat ötvözése jelenti, mindegyiket arra használva, amire való: valódi SGML formátumot használni a szerveroldalon és HTML-t a kliensoldali megjelenítéshez. A gyors fejlődés jele, hogy a World Wide Web Consortium 1998 február elején adta közre az XML (Extensible Markup Language; Kiterjeszthető Jelölő Nyelv) webszabvány első változatát, mely az SGML lényegesen egyszerűsített változata, többféle dokumentumtípus rögzítéséhez használható szabvány, szemben a régi HTML-lel, mely a multimédiás környezetben is megállja a helyét.[30]

Mivel elvileg nincs akadálya annak (csupán megfelelő konvertáló programok kérdése), hogy a HTML és az XML formátumon belül a dokumentum típusát meghatározó leírást (ez a DTD nevű rész) a MARC-formátumot használók áttegyék a saját formátumukba, ezért csak idő kérdése,

29 Burnard, L.: SGML on the Web: too little too soon, or too much too late. In: Computers Texts, 1997. aug. No. 14. <http://info.ox.ac.uk/ctitext/publish/comtxt/ct15/burnard.html>

30 XML (Extensible Markup Language) O'Reilly <http://www.xml.com; továbbá: The XML Caver Pages <http://www.oasis-open.org/cover/sgml-xml.html>

hogy az elektronikus dokumentumokat a könyvtárak automatikusan is átvegyék és a saját igényeik szerint kezeljék. Az elektronikus dokumentumformátumok kialakulása arra az ismeretelméleti felismerésre utal, hogy az internet (és dokumentumainak) megjelenésével ugyanaz fejlődik tovább, mint ami az írott történelem kezdetén a könyvtárakban elkezdődött.

Másodlagos adatok formátuma (metaadat-formátum)

Az elsődleges dokumentumokra vonatkozó adatok a másodlagos adatok. Ilyenek a bibliográfiai leírás szabványosított adatelemei, továbbá minden, a dokumentumok tartalmi leírására felhasznált információkereső nyelvi/osztályozási adat (kulcsszó, tárgyszó, deszkriptor, osztályozási jelzet). Digitális könyvtári környezetben ezeket az adatokat metaadatoknak nevezik, ilyen adatokat határoznak meg az előbb ismertetett formátumszabványok. Segítségükkel az elsődleges elektronikus dokumentumok egységes gépi kezelése valósítható meg.

Metaadatok tehát az internetforrások intellektuálisan vagy automatikusan létrehozott másodlagos adatai, melyeket vagy magába az elsődleges dokumentumba ágyaznak be, vagy csatolókkal kapcsolnak hozzá. Korántsem olyan nagy a választékuk, mint a bibliográfiai formátumokban rögzített adatelemeké, és nem olyan komplexek, mint az utóbbiak.

Szükségesnek bizonyult maguknak a metaadatoknak az egységes elektronikus kezelése is. Ide tartozik a metaadatoknak az elsődleges dokumentumokból (digitális objektumokból) való kinyerése vagy kiszámítása, a dokumentumok számítógépes leírása. Ezek az adatok a funkcionálisan strukturált (pl. SGML) dokumentumok esetében rendkívül könnyen kinyerhetők, noha erre alapul szolgálhat az elektronikus dokumentum teljes szövege is. A sokféle metaadatelem léte hívta életre a Dublin Core (DC; dublini alap[mag]-metaadatok) formátumát, melynek 1.1 változata 15 metaadatelemet tartalmaz az elektronikus dokumentumok egységes leírására (és tegyük hozzá: eme adatelemekből felépülő rekordok cseréjére is). Ez a viszonylag egyszerű formátum független attól a szintaxistól, amelyben az elektronikus dokumentumot funkcionálisan strukturálták (elvileg tehát alkalmazható nemcsak SGML dokumentumokra is). Minden adatelemnek több értéke lehet (ismételhető) és opcionális.

A DC metaadatelemei az elektronikus dokumentumok katalogizálását, és ezáltal a leghatékonyabb, gyors visszakeresését hivatottak biztosítani. Közöttük van például a „Tárgy” (<Subject>) azonosítójú metaadatelem, melynek ismételhető értékei kulcsszavak, tárgyszavak, deszkriptorok, osztályozási jelzetek lehetnek.

Mivel szükség van a DC formátumot kiegészítő információkra is (pl. a felhasználás feltételeire), született erre vonatkozó átfogó ajánlás (architektúra, container architecture), melyet Warwicki forrásleíró keretmegállapodásnak (Warwick Framework, Resource Description Framework) neveznek.
A fejlődés iránya, hogy a HTML-rekordok valamilyen formátum szerint egységesüljenek. A fejlődés a DC formátum irányába mutat.
A metaadat-szabványosítás terén két irányzat küzdelme figyelhető meg: a minimalisták szemében csak az a fontos, hogy a keresést megkönnyítsék (ezért legyen a lehető legegyszerűbb a formátum); a strukturalisták fontosnak tartják, hogy a digitális dokumentumnak legyen valamilyen azonosító jellegű, a bibliográfiainak megfelelő leírása is, hogy adatcsere esetén tudni lehessen, miről is van szó a tételek esetén.
A DC elsősorban a web számára kialakított szabványos formátum. A digitalizált (tehát eredetileg nem digitális) dokumentumokra nem alkalmazható kifogástalanul. A keresés szempontjából például a „Dátum" és a „Kiadó" adatelemek okoznak problémát, melyek a szabvány szerint nem az eredeti mű, hanem a digitalizált dokumentum adatai. Márpedig képzőművészeti alkotás vagy szépirodalmi mű esetében az eredeti mű dátuma és kiadója sokkal fontosabb, semmint hogy elhagyható lenne. Bibliográfiai szempontból a „Cím" is rendkívül problematikus, melyre semmiféle egységesítést nem írnak elő.

* * *

Az alábbi első szemelvényben *Traugott Koch* a hagyományos osztályozási rendszereket alkalmazó internetkatalógusokkal foglalkozik.
A következő szemelvényben *Peggy Zorn* és társai a web néhány tartalmi keresőrendszerét ismertetik. A tartalmi információkeresés területén a web várhatóan még viharos fejlődés előtt áll, és a lehetőségek tovább differenciálódnak, hiszen az internetet a végfelhasználók többsége nem a maga gyönyörűségére akarja igénybe venni, nem többnyire előre ismert címek alapján akar egy információs hálózatot használni, hanem azért, mert valamit nem tud, valaminek a forrását vagy tartalmát nem ismeri. És ezért alapvetően valamilyen tárgy, tárgykör – és nem forrásazonosítók (URL-ek) – szerint akar keresni.
A speciális internetdokumentumok hivatkozási „dokumentumleírása" még nincs szabványosítva, noha ezeknek a hivatkozásoknak a jelentősége növekedni fog. Ezzel foglalkoznak a harmadik szemelvény szerzői.
Utolsó szemelvényünkben az amerikai könyvtáros, *Carla List* megpróbálja az őt megillető „helyére tenni" az internetet.

TRAUGOTT KOCH (1950)

A Lundi Egyetemi Könyvtár főkönyvtárosa, Traugott Koch saját honlapja szerint digitális könyvtártudós (Digital Library Scientist). Azon könyvtáros szakemberek közé tartozik, akik már munkásságuk kezdetétől fogva az elektronikus könyvtári kérdésekkel, majd az internetes indexelő- és keresőrendszerekkel foglalkoztak. Az On-line & CD-ROM Review és az Ariadne szakfolyóiratok szerkesztőbizottságának tagja, több elektronikus könyvtári projekt (WoPEc, BIOME), digitális és hálózati kérdésekkel foglalkozó konferencia szervezője. Mint a Dublin Core Initiative felügyelőbizottságának tagja közreműködött a hálózati dokumentumokra vonatkozó metaadatelem-szabvány kidolgozásában.
Itt közölt tanulmánya jelzi, hogy az internet megjelenésével lassan megint időszerűek lesznek mindazok a hagyományos – különösen pedig egyetemes – osztályozási rendszerek, melyek elméletével az első kötetünkbe fölvett szemelvények szerzői foglalkoztak, s amelyek a távolsági és helyi on-line információkeresés elterjedésétől kezdve fokozatosan a háttérbe szorultak.

Internetforrások tökéletesebb leírásához, szervezéséhez és kereséséhez alkalmas osztályozási rendszerek használata[31]

1. Bevezető

Az internet világában egyre nagyobb jelentősége van az osztályozásnak, noha ezt a kifejezést ebben a környezetben nem valami gyakran használják. A keresés céljára rendelkezésre bocsátott számtalan többé-kevésbé szisztematikus rendezettségű internetes jegyzék, táblázat valójában a feldolgozott források osztályozási rendszere, függetlenül attól, hogy ezzel a szolgáltatók tisztában vannak-e vagy sem. Naponta találnak föl új osztályozási rendszereket, többnyire egyre rosszabbakat, mivel fogalmuk sincs arról, hogy a nagy mennyiségű információ strukturált tálalása nem jelent semmiféle új problémát. Noha a szolgáltatók és felhasználók többsége tudja, mi fán terem a könyvtár – legalábbis a papíron rögzített dokumentumok vonatkozásában –, az ilyen intézményekben használt módszerek nem mindig váltottak ki maradandó pozitív benyomásokat.

31 Traugott Koch: Nutzung von Klassifikationssystemen zur verbesserten Beschreibung, Organisation und Suche von Internet Ressourcen. In: Buch und Bibliothek, 1998, 5. sz., p. 326–335.

A Yahoo! globális internetkatalógus különösen jól példa: ez az egyik leggyakrabban látogatott webhely, 1988 januárjában 27 millióan keresték föl, többször annyian, mint az ugyancsak nagy exice és AltaVista keresőszolgálatokat, messze megelőzve az összes többit. A leggyakrabban használt 10 keresőszolgáltatás közül 8 rendelkezik strukturált katalógussal.

A Yahoo! a legnagyobb kihívás is a hagyományos osztályozással szemben. Készítői kezdettől fogva teljesen önálló rendezőrendszer mellett döntöttek. Ez a rendszer mára 30 000 osztályt és áttekinthetetlen számú osztályok közötti kapcsolatot tartalmaz és rendező elvei meglehetősen nehezen láthatók át.

A Yahoo! nemzetközi szinten is kiválóan példázza az internetkatalógusok válságát. Méreteit tekintve messze a robotok által naprakészen tartott indexelőszolgáltatások mögött kullog (az AltaVista a maga 100 milliós dokumentációs egységével a legnagyobbak közé tartozik); szerkesztői nem képesek már ilyen nagyságrendben a tartalomszolgáltatók kínálatát szisztematikusan átfésülni, és a megjelenő források legfeljebb 25-30%-át dolgozzák föl, azt is meglehetős késéssel.

A szisztematikus rendezettségű információgyűjtemények (internetkatalógusok) iránti hatalmas igényt az is nagyon jól megvilágítja, hogy a nagyobb, kezdetben csak indexelő keresőszolgáltatások többsége ma már katalógusokkal is jelen van, holott a kettő kombinációja eleinte legfeljebb kivételesen fordult elő. E katalógusok osztályozási rendszerei a belépőlapon többnyire vizuálisan is az előtérben állnak, miközben a keresőkérdések mezője szinte eltűnik a portálon. Legutóbb az AltaVista vette föl a keresőgépei által működtetett indexelőszolgáltatása mellé a LookSmart katalógusát.

Az utóbbi időben egyre nagyobb számban megjelenő speciális, szakmai vagy regionális keresőszolgáltatások is többé-kevésbé strukturált, hierarchikus osztályozási rendszereken alapuló katalógusokkal jelennek meg, hogy a forrásgyűjteményükben az osztályozási rendszer alapján végzett böngészést megkönnyítsék.

> Böngészésről akkor beszélünk, ha osztályozási rendszer kapcsolatait felhasználva végezzük a keresést. Indexelőszolgáltatások segítségével végzett, egyedi szavakat vagy keresőprofilokat használó kutakodás esetén keresésről vagy lekérdezésről, szövegen belüli nem szisztematikusan rendszerezett csatolókat felhasználó kereséskor pedig szörfölésről beszélünk.

Az alábbiakban főleg a bevált könyvtári osztályozási rendszerek alkalmazásának pillanatnyi helyzetével foglalkozunk az internet keresőszolgáltatásaiban. Áttekintjük alkalmazásuk formáit, előnyeiket és hátrányaikat, és konkrét felhasználási példákat ismertetünk.

Az automatikus osztályozási módszerek internetes alkalmazására két létesítmény összefüggésében térünk ki. Az automatikus osztályozás megoldás lehet az internet mennyiségi problémáira és a Yahoo!-hoz hasonló keresőszolgáltatások válságára.

2. Alkalmazás

Hangsúlyozzuk, hogy a szisztematikus osztályozási rendszerek – még általánosabban az ismeretek strukturálása – a tartalmi feltárásnak csak az egyik eszközei. A másik eszközt az ellenőrzött szótárak (tárgyszójegyzékek, tezauruszok) képviselik. A velük végzett tartalmi feltárás nagy részével itt egyáltalán nem foglalkozunk. Ugyancsak figyelmen kívül hagytuk e két tartalomfeltáró és információkereső nyelvi eszköznek a gyakorlatban rendkívül fontos együttes használatának a kérdését is.

Az osztályozás az ismeretek szervezésének egyik módszere a sok közül. Nem öncél, hanem szerves része a dokumentumtárolás ügyvitelének és a gyűjteményekben végzett keresésnek.

Az osztályozással tartalmilag írható le a dokumentum. Ez az osztályozás elsődleges feladata. Az osztályozási adatok a dokumentumok leírásaihoz tartoznak, s mint ilyenek metaadatokat (másodlagos információkat) képviselnek. A DC alapadatelem-készlet (Dublin Core Element Set), a jelenleg legkidolgozottabb átfogó metaadatszabvány a Tárgy (<Subject>) adatmezőt tartja fenn az osztályozási adatok számára. Ehhez társul az alkalmazott osztályozási rendszer (séma) adata. A hálózati dokumentumok osztályozási adatainak tehát már van helye ellenőrzött és újra hasznosítható metaadat-formátumban.

Ennek nyomán már születtek konkrét nemzeti metaadat-formátumok is, melyek a tartalmi feltárás eredményeinek rögzítésére fölhasználhatók. Például a „Nordic Metadata Project" keretében elkészült, a Dublin Core ajánlásokra épülő formátum (Metadata Creator) (http://www.lub.lu.se/cgi-bin/nmdc.pl). Ebben a formátumban megfelelő mezők találhatók a felhasznált osztályozási rendszerek jellemzőinek rögzítésére is. Az általunk ismert, és az interneten belül szabadon hozzáférhető osztályozási rendszert és tezauruszt, továbbá egyéb szótárt e formátum segédletében soroltuk föl (http:/www.ub2.lu.se/metadata/subject-help.HTML). Ezek közül több automatikusan, a Javascript segítségével segédmezőbe töltődik, ha a metaadat-formátumban a megfelelő mezőben az osztályozási rendszert megadtuk.

Ha az osztályozási adatokat szabványos metaadatok formájában kapcsoljuk a dokumentumokhoz, akkor biztosítható, hogy ezeket az adatokat a keresőszolgáltatásokban és egyéb célokra tovább hasznosíthassák. Számos, kapcsolatokat (linkeket) tartalmazó jegyzékben és adatbázisban az osztályozási adatokat nem kapcsolják össze tartósan az adott dokumentummal, ill. külső

szolgáltatások ezeket a kapcsolódásokat nem hasznosíthatják. Ezáltal a további használat számára a tartalmi feltárás lényeges adatai vesznek el. Ha nem alkalmazzák valamelyik szabványosított osztályozási metaadat-formátumot, nincs lehetőség sem az ellenőrzött szótáras keresésre, sem pedig a különféle katalógusok osztályozási rendszereinek integrálására, további felhasználására.

Az osztályozás második feladata, hogy szisztematikus katalógusokat szerkeszthessünk a források böngészésére, összhangban a választott metaadat-formátummal. Ha az egyes dokumentumokhoz hozzákapcsolták az osztályozási adatokat, maga az osztályozási rendszer HTML-oldal szerkezete automatikusan rekonstruálható, azaz újabb dokumentumok fölvételekor aktualizálható ez a szerkezet az átvevő rendszerében, hogy az osztályozásra igénybe vehessék. A felhasználónak számos különféle megjelenítési formátum (ún. nézet, „view") készülhet, azaz a többnyire hierarchikus osztályozási struktúrák vizualizálhatók. Megvalósítható az osztályozási rendszer automatikus konvertálása is. Az osztályozási rendszerhez különféle belépési lehetőségek adhatók, például a hierarchikus szerkezetben végzett lapozással, címkézett kereséssel, vagy a talált dokumentumok környezetében végzett szisztematikus továbbkereséssel. Az osztályozásnak ez a fajta alkalmazása az interneten belül messze a leggyakrabban fordul elő, és az 5. fejezetben felsorolt alkalmazások mindegyikére jellemző.

A harmadik alkalmazási területe az osztályozásnak a lekérdezés és szörfölés támogatása. Fejlettebb alkalmazásokban összefüggő navigációs tér áll rendelkezésre, melyben a szisztematikus böngészés az indexek alapján végzett lekérdezéssel kapcsolható össze. Tetszés szerinti belépési pontokból kiindulva felváltva végezhetők szisztematikus böngésző- és szabad lekérdezőműveletek, lapozni lehet az osztályozási struktúrában, deszkriptorokkal és kulcsszavakkal lekérdezés végezhető és szörfölve véletlenszerűen továbblapozhatunk a talált dokumentumok környezetében. Ma már vannak olyan, minőségbiztosítási szempontok alapján készült szakmai információszolgáltatások (pl. a svéd Engineering Electronic Library, http://www.ub2.lu.se/eel/), melyekben a fentiekben leírt navigációs eljárások az osztályozás segítségével megvalósíthatók. Ezekben, de például a Yahoo!-ban is az osztályozási rendszer szűrőként használható, azaz a keresés az adatbázis meghatározott szisztematikus részterületére korlátozható.

3. Előnyök és hátrányok

Mielőtt ebben a fejezetben az interneten már használt osztályozási rendszerekre és alkalmazási példákra kitérnénk, a keresőszolgáltatásokban használt osztályozási rendszerek előnyeivel, de hátrányaival is foglalkoznunk kell.

Az információk keresését csatolókkal ellátott jegyzékekkel, szakmai információs kapuszolgálatokkal vagy ellenőrzött szótáras szakmai tájékoztatók-

kal támogató keresőszolgáltatásokban (internetkatalógusokban) nem volna szabad megelégedni a rendszerezetlen, betűrendes vagy lapos enumeratív osztályozási struktúrákkal. A szisztematikus, hierarchikus felépítésű, már bevált osztályozási rendszereknek velük szemben egész sor előnyük van.

3.1 Előnyök

a) Jelentősen könnyebb a szisztematikus, böngésző keresés. Különösen a tapasztalatlan, vagy a szakterületet, annak szerkezetét, terminológiáját nem eléggé ismerő felhasználóknak nagy segítség a logikusan strukturált információs kínálat, és jól is teszik, ha ezt részesítik előnyben a szokásos indexelőszolgáltatásokkal szemben.
   Mennél bőségesebb az on-line kínálat, annál fontosabb a navigáció támogatásában az áttekinthető, logikus és hierarchikusan kialakított osztályozási szerkezet. Az interneten manapság még sokkal elterjedtebbek a források véghetetlen betűrendes, vagy mechanikusan sem rendezett jegyzékei.
b) Az alkalmazott osztályozási rendszer különféle módon generálható és reprezentálható a keresés közbeni navigáció támogatására. Egyszerre több hierarchiaszint osztályai jeleníthetők meg ugyanazon a képernyőn különféle részablakokban belépőpontként és annak érdekében, hogy a keresést beszűkíthessék (ilyen lehetőséget kínál például az OCLC által fenntartott NetFirst és annak segítsége, a Mr. Dui's Topic Finder: http://www.oclc.org/oclc/fp/mrdui/mrdui.htm).
c) A már bevált osztályozási rendszerek előnye, hogy különféle felhasználói körökben már ismertek. A hagyományos osztályozási rendszerekkel a könyvtárak rendszeres látogatói többnyire megbarátkoztak. A szakemberek általában ismerik azokat a rendszereket, melyeket a szakterületükön a rendszerezésre használnak.
d) Néhány osztályozási rendszer többnyelvű változatban létezik, őket használva különféle nyelvű gyűjteményekhez és forrásokhoz lehet hozzáférni. Akárki is legyen, aki a többnyire nyelvektől független jelzetekkel az osztályozáskor a dokumentumokat leírta: a más nyelvű fölhasználók a saját keresőnyelvükön férhetnek hozzá az összes, különféle nyelven írt dokumentumhoz.
   Ehhez a katalógustételek semmiféle módosítása nem szükséges. E tételeket számos szolgáltatás esetén amúgy is elosztva tárolják és központilag nem is lehet elérni őket. A többnyelvű osztályozási rendszer jelzetei kapcsolónyelvként funkcionálnak.
   Az OCLC már említett böngészője, a Mr. Dui's Topic Finder angolul, spanyolul, franciául és oroszul kínálja föl Dewey Tizedes Osztályozá-

si rendszerét, melynek megjelenítése különféle jelkészletekkel, fontokkal valósítható meg.

A német akadémiai webindex, a GERHARD az ETO olyan változatát használja, mely angol, francia és német nyelvű (http://www.gerhard.de).

e) A bevált osztályozási rendszerek lehetővé teszik a különféle információs rendszerek együttműködését (az „interoperativitást"), amelyben az egyes internet szolgáltatások, könyvtári on-line katalógusok, referáló- és indexelőszolgáltatások, cégek osztályozási adatai kölcsönösen fel- és továbbhasználhatók.

Optimális esetben, ha több szolgáltatás is ugyanazt vagy kompatibilis, ill. szabványosan konvertálható osztályozási rendszert használja, a böngészés és a tárgyi keresés bármelyik szolgáltatásban végezhető anélkül, hogy központosított szolgáltatásra volna szükség. Ehhez még megegyezés vagy formális együttműködés se szükséges.

A hierarchikus böngészőstruktúrák interoperativitása a heterogén információszolgáltatások, köztük a mindenekelőtt a hagyományos könyvtári és interneten alapuló digitális szolgáltatások integrációjának fontos feltétele.

A műszaki megvalósítást megkönnyíti, ha létezik szabvány (protokoll) a ténylegesen közös (osztott) keresésre és mindenekelőtt a keresőkérdés továbbítására (query routing, query forwarding). Ilyen célból javaslat szintjén már léteznek szabványok (IETF Draft, CIP [Common Indexing Protocoll], 1997).

A közös (osztott) katalógusokban végzett böngészést a W3C (World Wide Web Consortium) 1999-ben elfogadott Forrásleíró Keretmegállapodásában (Resource Description Framework) megadott metaadat-szintaxis alkalmazása tette lehetővé (az erre vonatkozó javaslatot lásd Kirriemuir [et al]).

f) Az osztályozási rendszerek alkalmazása lényegesen hozzájárul ahhoz, hogy az információkeresés lehetőségei bővüljenek. A keresőkérdések ellenőrzött módon (az adott hierarchikus szerkezet mentén) bővíthetők vagy szűkíthetők. A felsőbb hierarchiaszint bevonásával (pl. ha „Kutya" keresése esetén bevonjuk az átfogóbb „Állat" kifejezést is a keresésbe) a teljesség növekszik, a hierarchiaszinten lejjebb lépve (a keresést szűkítve a „Kutya" helyett a „Puli" kifejezéssel) a pontosságot fokozhatjuk. A hamis találatok száma csökkenthető, és a homonimák (az azonos alakú, de különböző jelentésű kifejezések) problémája is jórészt megszüntethető, ha az osztályozási rendszer alkalmazásával mintegy „megszűrjük", behatároljuk a keresést.

A keresési eredmények az osztályozási rendszer struktúrájának megfelelően csoportosíthatók és megjeleníthetők, ezáltal lényegesen javul a kapott találatok tartalom szerinti áttekinthetősége.

g) A bevált osztályozási rendszerek struktúrája és szókincse fölhasználható a tudásbázisok (ismeretbázisok, knowledge bases) kialakítására. Az automatikus osztályozásnak és a színvonalas információkeresésnek ez előfeltétele. Ez egyben visszahathat az osztályozási rendszerre, melyet a tudásbázis alapján javítani lehet.
Az osztályozási rendszerek fejlesztéséhez különféle lehetőségek adódnak: ha rendelkezésre áll a gyűjtőkör számára osztályozási rendszer, akkor ennek dokumentumaiból időszerű és az egyes osztályok vonatkozásában releváns kiegészítő szókincs válogatható ki, például statisztikai eljárásokkal. Az együttes előfordulások (co-occurrence) elemzése alapján a tezauruszokból és bizonyos bibliográfiai adatbázisokból és online katalógusokból ellenőrzött szóállomány kapcsolható össze szisztematikusan az osztályozási rendszerrel. Ez valósul meg például az OCLC Scorpion projektjében (OCLC ExTended Concept Trees): a Kongresszusi Könyvtár tárgyszavait (Library of Congress Subject Heading, LCSH) és osztályozási jelzeteit (Library Congress Classification, LCC) összekapcsolják Dewey Tizedes Osztályozásával.

h) Az osztályozási rendszerek használata ráadásul a szolgáltatók számára időmegtakarítással is járhat. A rendszereket általában különböző szervezetek együttműködve tartják fenn és gondozzák, nincs tehát szükség arra, hogy egyedül kelljen foglalkozni ilyen rendszer kialakításával és fenntartásával. Ha változásokra kerül sor, átfogó konverziós szabályokat is közösen fogalmaznak meg. Sok osztályozási rendszer áll már géppel olvasható formában rendelkezésre, egy részük szabadon hozzáférhető az interneten (lásd pl. az előzőkben már említett segédletet: http:/www.ub2.lu.se/metadata/subject-help.html).
Mindezek az osztályozási rendszerek dokumentáltak, segédletek tartoznak hozzájuk. Az egyénektől függetlenül és hosszú ideig léteznek.

3.2 Hátrányok

Éppen a könyvtárosok ismerik a legjobban a hagyományos osztályozási rendszerekkel szemben megfogalmazott bírálatokat. Egyes országokban és szakterületeken az osztályozási rendszerek használata átmenetileg jelentősen visszaesett. Az internetes keresőszolgáltatások működésének első éveiben is sok szolgáltató érvelt azzal, hogy az osztályozási rendszereket és a velük összefüggő metaadatokat (a jelzeteket, osztálymegnevezéseket, deszkriptorokat) a fejlődés – azaz a teljes szövegek indexelése – már maga mögött hagyta, és nem használják már őket.

a) Az osztályozási rendszerek egyik hátránya, hogy elválasztanak egymástól logikailag összefüggő forráscsoportokat. Tény, hogy a hagyo-

mányos osztályozási rendszerekkel osztályozva gyakran eldurvítják a tényleges dokumentumok közötti rendkívül gazdag összefüggéseket. Különleges gyengéje ezeknek a rendszereknek, hogy sokszor nem vehetők figyelembe interdiszciplináris szakterületek. Mégis: ha megfelelő „lásd még" utalásokat alkalmaznak és lehetővé teszik összetett jelzetek képzését – ami lényegében az ÉS-kapcsolatoknak felel meg –, sőt ezen túlmenően minősíthetők is lennének még a felhasznált jelzetek közötti kapcsolatok, az ebben a pontban tárgyalt hátrányok jelentős mértékben kiküszöbölhetők lennének.

Szöveggyűjteményünk első kötetének „Jason L. Farradane, Jean Martin Perreault, avagy a szintaktikai relációk" című fejezetében találhatók szemelvények, melyekben az ETO jelzetei közötti szintaktikai relációkkal foglalkoznak.

b) Néhány osztályozási rendszerben valóban korszerűtlenül, logikátlanul rendezik el az osztályokat, vagy eleve nem alkalmaznak jól átgondolt rendszerezést. Ez természetesen jelentősen korlátozza a böngészés eredményességét.
c) Mivel a legjelentősebb hagyományos osztályozási rendszereket nehézkesen működő nemzetközi szervezetek gondozzák, gyakran lassan reagálnak a tudományokban és egyéb szakterületeken, de még a mindennapi életben is bekövetkező változásokra. Az internet megszületését követő első években a hálón tényleg megjelenő dokumentumok jórészt néhány olyan szakterületet reprezentáltak, melyek a hagyományos osztályozási rendszerekben nem voltak részletesen képviselve, s ezért nehezen lehetett volna őket kellő mélységben tartalmilag feltárni ezeknek az osztályozási rendszereknek a segítségével. Többek között ez is magyarázza a sok önerőből kialakított osztályozási rendszer – lásd a Yahoo! és a nagyobb internetkatalógusok – létrejöttét. A régiók és szakmák fölötti, lényegében egyetemes osztályozási rendszerek iránti igény csak most válik felismerhetővé, hogy jelentősen könnyebbé vált a forrásokhoz való hozzáférhetőség és elterjedt az interneten belüli együttműködés.
   Kimutatták például, hogy a Yahoo! osztályainak nagy része Dewey Tizedes Osztályozásában is megtalálható (Vizine-Goetz, D.).
d) Az előbbi hátránnyal párhuzamos, hogy sokszor a terminológia is elavult, mellyel a hagyományos rendszerekben az osztályokat megnevezik. Az internetkatalógusok új, önerőből készített osztályozási rendszerei kiváló segítséget nyújthatnak ahhoz, hogy a hagyományos osztályozási rendszerek terminológiai elöregedését felszámolják.
   Például az OCLC hálózati böngészőjében (NetBrowse Prototype) Dewey Tizedes Osztályozásának hagyományos kifejezéseit az aktuális

terminológiával bővítik ki, és az eredeti osztályozási rendszer nyelvét a könyvtári szaknyelvből „végfelhasználói" nyelvre ültették át. Mind a terminológiát, mind pedig egyes osztályok kiemelését hozzá lehet ezáltal igazítani a használati gyakorisághoz és fontossághoz.
Az elidegenítő numerikus és alfanumerikus jelzeteket a háttérben, a felhasználótól rejtve lehet használni.

4. Osztályozási rendszerek és alkalmasságuk

Akárcsak a hagyományos osztályozás esetében, a keresőszolgáltatások számára is mindenekelőtt a gyűjtőkör és a felhasználói kör határozzák meg, milyen típusú osztályozási rendszer használata jöhet szóba. A teljesen új fejlesztés mellett a választék egyetemes, szakmák fölötti és szakmai (nemzetközi vagy nemzeti/nemzeti nyelvű) osztályozási rendszerekből áll.

További kiválasztási kritérium, hogy mekkora az információs rendszer által lefedett szakterület és milyen igényeket kell támasztani ebben a szakmai körben az osztályozási rendszer minősége iránt, továbbá milyen, a böngészést elősegítő tulajdonságai vannak ennek a rendszernek. Természetesen előny, ha a rendszer digitalizált formában is rendelkezésre áll, az interneten már alkalmazzák, és jól bevált a hagyományos és egyéb on-line szolgáltatásokban.

Célszerű figyelembe venni, hogy mennyire kiépített a rendszer, mennyire fejleszthető, és megvannak-e az előfeltételei a többnyelvű használatnak. Ugyancsak fontos szempont a kiválasztáshoz, hogy a választandó osztályozási rendszer más rendszerek ellenőrzött szókincsével összekapcsolható-e, más forrásokból származó tartalmi leírások osztályozhatók-e vele. Nem utolsó sorban ügyelni kell a jogi kérdésekre és a költségekre.

A legfontosabb osztályozási rendszerek előnyeit és hátrányait, és az interneten alkalmazott rendszerek leírását részletesen tárgyalja a Európai Unió által támogatott DESIRE beszámoló, melyben minden, általunk ismert keresőszolgáltatásban alkalmazott osztályozási rendszerre kitértünk (Koch, T., Day, M.).

A főbb hagyományos osztályozási rendszerek és az interneten használt önálló fejlesztések előnyeinek és hátrányainak részletes elemzését két hálózaton is elérhető dokumentum tartalmazza[32]. Az adatok az 1997-es állapotot tükrözik, de az osztályozási rendszerek terén szerencsére nem kell számolni olyan gyors avulási folyamattal, mint amely egyébként az internetre általában érvényes.

---

32 Koch, T. és Day, M.: The role of classification schemes in Internet resource description and discovery (EU Project DESIRE. Deliverable D3.2.3) <http://www.ub2.lu.se/desire/radar/reports/D3.2.3>

Beyond Bookmarks: Schemes for Organizing the Web. <http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm>

5. Intellektuálisan használt – „kézi" – osztályozási rendszerek

Az interneten manapság használt osztályozási rendszerek a legkülönfélébb típusokhoz, szakterületekhez és nyelvterületekhez tartoznak. Az említett DESIRE mellett létezik másik összeállítás is az internetkatalógusokban alkalmazott osztályozási rendszerekről, melyet McKiernan készített, s ezáltal megfelelő ellenőrzés és összehasonlítás végezhető el (McKiernan, G.).

A fejlődést főleg a szolgáltatások fajtái és a már létező példák befolyásolták. 1993 körül, a Gopherek idején az egyetemes rendszerek uralkodtak és az első alkalmazások közé – hagyományos amerikai elterjedtségének köszönhetően – Dewey Tizedes Osztályozása tartozott. Ma ez a rendszer uralkodó szerepet játszik, legalábbis az egyetemes, globális szolgáltatások világában, a nemzetközi szakmai osztályozási rendszereket pedig a speciálisabb szakterületeken tevékenykedő szolgáltatások használják.

Ugyanazt a rendszert meglehetősen eltérő módon használják föl a különböző internetkatalógusokban. Az áttekinthetőségre többnyire nagyobb súlyt helyeznek, mint a hierarchia mélységére. Ritka, hogy a keresés a nyelvek vagy más mezők szerint szűkíthető volna. Sok esetben kisebb-nagyobb változtatásokat is eszközöltek az átvett rendszer szerkezetén és szókincsén, a jelzeteket elrejtették stb. Információs rendszerek közötti – interoperatív – együttműködések még nem alakultak ki, és feltehető, hogy az említett átalakítások ezt hátráltatják is.

5.2 Egyetemes rendszerek

5.2.1 Dewey Tizedes Osztályozása

Dewey Tizedes Osztályozása (Dewey Decimal Classifiation, DDC, Magyarországon elterjedt neve Tizedes Osztályozás, a továbbiakban TO)[33] rendkívül elterjedt, és lényegesen gyakrabban aktualizálják, mint bármelyik vele összehasonlítható osztályozási rendszert. 30 nyelvre fordították le. 2000 elején 22 olyan keresőszolgáltatást találtunk, mely a TO-t használta.

A TO jelzeteit gyakran összekapcsolják a Kongresszusi Könyvtár tárgyszavaival (LCSH) és osztályozási jelzeteivel (LCC), többek között a USMARC adatcsere-formátumot használó állományokban.

A böngészést jó súgók támogatják, az első három hierarchiaszint (Summary of DDC 221) a hálón szabadon hozzáférhető (DDC: http://www.oclc.org/oclc/fp/about/ddc21sm1.htm).

33 A Tizedes Osztályozással első kötetünk első fejezetében foglalkozunk.

Az OCLC kutatási eredményei arról számolnak be, hogy a TO mind az on-line nyilvános katalógusokban, mind pedig a hálón böngészőstruktúraként beváltnak tekinthető (Markey, K., 1989; Vizine-Goetz, D.).

Alkalmazási példák

Három keresőszolgáltatás globálisan, minden szakterületre kiterjedően alkalmazza a TO-t:

- Az OCLC saját internetkatalógusa, a NetFirst, melyet meglehetősen későn tettek hozzáférhetővé és használatáért fizetni kell. (NetFirst: http://www.oclc.org/oclc/netfirst/). Készült Dewey nevét parafraizáló böngésző, a „Mr. Dui's Topic Finder", mellyel kísérleti jelleggel jobban kihasználható a TO (http://www.oclc.org/oclc/fp/mrdui/mrdui.htm).
- A BUBL LINK Angliában működik, globális akadémiai szolgáltatás, itt 1997-ben figyelemre méltó módon áttértek az ETO-ról a TO-ra. Egyedi gyűjteményeket is osztályoznak (http://link.bubl.ac.uk/ISC2).
- Az Internet Public Library On-line Texts Collection 6500 szöveget katalogizál jelenleg a TO-val (http://www.ipl.org/reading/books).

Három globális, ill. regionális keresőszolgáltatásban viszonylag mélyen a TO részeit alkalmazzák:

- Az ottawai Kanadai Nemzeti Könyvtárban a Canadian Information by Subject (http://www.nlc-bnc.ca/caninfo/esub.htm).
- A bristoli kereskedelmi főiskola könyvtárának internetkatalógusában (Biz/ed: http://www.bized.ac.uk/roads/htdocsbrowse.htm).
- Az aberystwyth-i könyvtári és informatikai főiskola könyvtárának internetkatalógusában (PICK: http://www.aber.ac.uk/~tplwww/e/contents.html).

Még két további alkalmazást érdemes említeni:

- Az amatőr csillagászok számára szolgáltató Expanding Universe (a classified search tool for amateur astronomy, Metropolitan Toronto Reference Library) http:/www.mtrl.toronto.on.ca/centres/bsd/astronomy/index.html
- WWLib, Interface Catalogue of UK Web Pages. Peter Burden, Wolverhampton (WWLib: http://www.scit.wlv.ac.uk/wwwlib/browse.html).

5.2.2 Egyetemes Tizedes Osztályozás (ETO)

Az ETO ugyancsak átfogó osztályozási rendszer, 60 000 osztályból és segédtáblázatokból áll. Négy nyelven készült teljes kiadás, és további 20 nyelven léteznek rövidített, ill. részleges kiadások. Elsősorban Európában kerül sor internetes alkalmazására (2000 elején 11 szolgáltatásban használták).

– GERHARD (German Harvest Automated Retrieval and Directory) internetkatalógus és indexelőszolgáltatás a német internetforrások számára. A keresőgépek által létrehozott adatbázis összes dokumentációs egységét számítógépes nyelvészeti és statisztikai módszerek felhasználásával osztályozzák a zürichi műszaki főiskola ETO változatával. A böngészéshez az osztályozási rendszer több nyelven használható. Ez az egyetlen keresőszolgáltatás, mely források százezreit automatikus osztályozással sorolja be 70 000 osztályt tartalmazó osztályozási rendszerébe (http://www.gerhard/de/). A továbbiakban a rendszerrel még részletesebben is foglalkozunk.
– NISS szakmai információs kapuszolgálat (Information Gateway, Directory of networked resources) mind az ETO hierarchiája szerint, mind pedig lineáris rendezettségben (ahogy az ETO szerinti szabadpolcos rendszerekben szokás) szolgáltatja a dokumentációs egységeit.
– A társadalomtudományokra szakosodott információs kapuszolgálat, a SOSIG (The Social Science Information Gateway) keretében az ETO-val osztályozzák a dokumentumokat, de a keresőszolgáltatást csak betűrendes mutatók formájában lehet igénybe venni.

5.2.3 A Kongresszusi Könyvtár Osztályozása (LCC)

Az LLC talán a legtöbb hagyományos könyvtári felhasználóval rendelkező rendszer, de sajnos túlságosan enumeratív (felsoroló) jellegű, és az Egyesült Államokra koncentrálódik. Hiányoznak a fordítások. A viszonylag kevés internetes felhasználó felszínesen használja a rendszert, viszonylag kis gyűjtemények katalogizálásához. Két keresőszolgáltatást érdemes említeni (2000 elején összesen 6 szolgáltatásban használták):

– McKiernan rendszere, az indianai állami egyetem könyvtárában működtetett Cyberstacks az LCC „Természettudományok" és „Technika" főosztályait alkalmazza (http://www.public.iastate.edu/~CYBERSTACKS/).
– A Scout Report Signpost a katalógusához igényesen alkalmazza az LCC-t (http://www.signpost.org/signpost/index.html).

5.3 Átfogó nemzeti rendszerek

– Holland Alaposztályozási Rendszer (Nederlandse Basisclassificatie, BC). A DutchESS (Dutch Electronic Subject Service) keresőszolgáltatás német és angol nyelven is használható katalógusa minden szakterületről gyűjti az internetes forrásokat. A közös katalogizálás elve alapján működik a Holland Királyi Könyvtárban (KB) és számos holland egyetemi könyvtárban (http://www.konbib.nl/dutchess/nbc_main.html).

– Svéd Nemzeti Könyvtári Osztályozási Rendszer (Sveriges Allmaenna Biblioteksfoerening, SAB).
  A rendszer a svéd iskolai hálózat számára készült, intellektuális minőségi katalogizálást folytat és két további nagyobb könyvtárban is felhasználják az internetes dokumentumok osztályozására (Laenkskafferiet: http://www.ub2.lu.se=/skolverket/sab_top.html).
– Finn Közművelődési Könyvtárak Osztályozási Rendszere (Classification for Public Libraries, Finland) (http://wwwkirjasto.dci.fi/lindex.htm)
– Észak-Rajna-Westfália főiskolai könyvtárainak internetes információs rendszerében (Internet Basierte Informations System Nordrhein-Westfälischer Hochschulbibliotheken, IBIS) a Regensburgi Közös Katalogizálási Rendszerben osztályoznak (http://www.hbz-nrw.de/ibis/ibis-tree.html, továbbá:http://www.ub.uni-bielefeld.de/netacgi/nph-browse?query=root).

5.4 Nemzetközi szakmai rendszerek

– National Library of Medicine (NLM).
  Az NLM osztályozási rendszerét az angol OMNI (Organising Medical Networked Informati) keresőszolgáltatásban alkalmazzák. A korábban párhuzamosan végzett osztályozását az ETO-val megszüntették, abban a reményben, hogy interoperatív keretek között a nem orvosi adatokhoz más úton is hozzájuthatnak majd (OMNI: http://omni.ac.uk/browse/).
– Engineering Information (Ei) Classification Codes.
  A szolgáltatáshoz tezaurusz is tartozik. Svéd és angol használója van (EELS [Engineering Electronic Library, Sweden]: http://www.ub2.lu.se/eel, továbbá EEVL [Edinburgh Engineering Virtual Library]: http://eevl.icbl.hw.ac.uk).
– Mathematic Subject Classification (MSC).
  A szakterületen domináns szerepet játszó osztályozási rendszert az összes érdekelt szakmai szolgáltatás alkalmazza. Távlatilag bibliográfiai és preprint adatbázisokkal integráltan is használható lesz.
  Az American Mathematical Society (AMS) honlapján a dokumentumokat már ebben a formában szolgáltatják (AMS: http://www.ams.org/mathweb/mi-methbyclass.html).
  Ugyancsak használják az MSC-t az osnabarücki és müncheni reprint-szerverekben és a göttingai Subject Guide Math szakinformációs rendszerekben (SGG: Math: http://www.sub.uni-goettingen.de/ssgfi/math)
– ACM Computing Classification System (CCS).
  Különös, hogy a rendkívül kurrens szakterület és nagyszámú dokumentum ellenére ezt az osztályozási rendszert csak elvétve alkalmazzák. A

MeDoc keresőszolgálat Ariadne rendszere, melyben régóta használják, minden jel szerint a túlélésért küzd (Ariadne: http://ariadne.inf.fu-berlin.de:8111/cgi-bin/navigate.cgi).

5.5 Egyéb rendszerek

Nemzeti szakmai rendszerek ritkán fordulnak elő az interneten, talán mert a nemzeti és nemzetközi tudományos tevékenység nehezen választható el egymástól. Példa azért akad, mint a Danish Veterinary and Agricultural Library Classification, dán és angol nyelven (http://www.dvjb.kvl.dk/dvjbfag/umnet02.htm#start).

Mivel a bevált hagyományos könyvtári osztályozási rendszerek internetes alkalmazására összpontosítottunk, nincs helyünk az önerőből készült osztályozási rendszerek – mint amilyet a Yahoo! is alkalmaz – tárgyalására, noha rendkívül nagy számban fordulnak elő.

A szakmai információ osztályozásának gyakori formája, hogy az adott szervezetben előforduló szakterületeket, vagy a szakreferensek feladatköreit tekinti főosztályoknak, és az ennek alapján kialakított rendezőrendszer szerint katalogizálnak. Az alosztályok többnyire formai kategóriák (pl. Szerszámok, Gépek, Anyagok) vagy dokumentumtípusok. Ennek első látásra előnyei lehetnek, az intézményhez kötődő felhasználó számára praktikus, de még a saját szervezeten belül is lehetetlen a továbbfelhasználás, nem beszélve arról, hogy más szakmai szolgáltatások a tartalmi feltárásnak eme korlátolt szemléletű módszere miatt képtelenek átvenni az osztályozás eredményeit. Ezáltal veszendőbe mennek az internet által kihasználható racionalizáció és együttműködés potenciális előnyei.

5.6 Több osztályozási rendszer együttes használata a keresőszolgáltatásokban

Van már példa olyan keresőszolgáltatásokra, melyekben a forrásokat különféle osztályozási rendszerek jelzeteivel látják el, és egyszerre több osztályozási rendszer használatát is biztosítják. Többnyire akkor kerül erre sor, ha több forrásból gyűjtik az osztályozási adatokat, vagy ha az on-line nyilvános katalógus több rendszerrel működik együtt.

Példa rá a göttingeni SSG GeoGuide, melynek forrásai TO, LCC és BC jelzetekkel rendelkeznek (GeoGuide: http://www.sub.unigoettingen.de/ssgfi/geo/).

Konvertálás vagy információveszteség nélkül nem nagyon lehet ebből egységes böngészőrendszert kialakítani.

6. Automatikus osztályozás

Ha az interneten rendkívül sok dokumentumot kell osztályozni, és hozzájuk megfelelő hierarchikus böngésző struktúrákat kialakítani, vagy ha a gyűjtőkör forrásai vagy egyedi dokumentumai gyakran változnak, elvileg csak az automatikus osztályozás segíthet, hacsak nem mondunk le az osztályozásról, hogy kizárólag keresőgépeket alkalmazva indexelt állományokat biztosítsunk a felhasználóknak.

A 4. fejezetben már említett, hálózaton is elérhető dokumentumban (DESIRE Report) részletesen áttekintettük az automatikus osztályozási módszereket, alternatívákat és projekteket. E helyen csak két rendszert mutatunk be, mindkettő bevált könyvtári osztályozási rendszert használ.

6.1 Scorpion

A fejlesztési szakaszban levő Scorpion (http://purl.oclc.org/scorpion/) már két éve működik és a TO-t használja az internetes források automatikus osztályozására. A FirstNet szolgáltatás néhány tízezer dokumentumát egyben intellektuálisan is osztályozták és indexelték. A Scorpion ugyanezt az állományt osztályozza automatikusan, működésének eredménye tehát az intellektuális osztályozás eredményeivel való összehasonlítás alapján ellenőrizhető, és a rendszer folyamatosan tökéletesíthető. Azt akarták bizonyítani, hogy a dokumentumklaszterálás alapján működő automatikus osztályozás összekapcsolható hagyományos könyvtári osztályozási rendszerekkel, azaz automatikus osztályozás segítségével ezeknek a rendszereknek az osztályaiba végezhető el a besorolás.

A projekt támogatására a TO-n alapuló tudásbázist építenek, mely osztályozási rendszer szókincséből, jelzeteiből és hierarchikus meg egyéb relációiból áll. A tudásbázis magját az Electronic Support System (ESS) alkotja, melyet az OCLC Forest Press Electronic Dewey for Windows CD-ROM termékében használnak.

A tudásbázisba jelenleg egyre nagyobb számban építik be a TO osztályok és az LC tárgyszavak közötti relációkat. Ezeket a megfeleltetéseket részben felelős bizottságok állapítják meg, részben statisztikai módszerek segítségével létező on-line nyilvános katalógusokból és adatbázisokból veszik át, mindenekelőtt az OCLC világkatalógusából (World Cataloge).

Az új szókincs és a hagyományos TO osztályok közötti asszociált kapcsolatokat az ExTended Concept Trees nevű rész tartalmazza, melynek forrását az olyan alternatív rendszerek alkotják, mint a Kongresszusi Könyvtár tárgyszavai (LCSH) és osztályozási jelzetei (LCC). A munkát maga a Scorpion végzi el olymódon, hogy a két alternatív rendszerrel osztályozott dokumentumokat a TO alapján automatikusan újraosztályozza, s ezáltal a keletkező TO osztályok és a két rendszer tárgyszavai, ill. jelzetei összekapcsolódnak.

A Scorpion az automatikusan osztályozandó dokumentumot kérdésként adja át a TO tudásbázisnak. Ennek során – dokumentumtípusokként szétválasztva – a címek, fejezetcímek, metaadatok, valamint a dokumentum teljes szövege is a keresőkérdés szerepét játsszák.

Az így elvégzett keresés eredményeként TO osztályok rangsorolt jegyzéke jelenik meg, mintegy a dokumentum automatikus osztályozási ajánlataként. Ez automatikusan véglegesíthető (hozzárendelhető a dokumentumhoz), az osztályozási adatok a HTML címfej metaadatmezőibe bevihetők. Opcionálisan választható az is, hogy intellektuális osztályozó az eredményt a tökéletesebb osztályozáshoz felhasználja.

A vizsgálatok kimutatták, hogy a TO jól alkalmazható a fogalmak meghatározására, mivel osztályai egyértelműek. Az LC osztályozási rendszere erre kisebb mértékben alkalmas. Az utóbbival a könyvek 46–86%-át lehetett megfelelően automatikusan osztályozni. Ilyen esetekben minden jel szerint a félautomatikus, tehát részleges intellektuális közreműködést igénylő eljárás a megoldás (Thompson [et al]; Larson, RR.).

6.2 GERHARD

A GERHARD (German Harvest Automated Retrieval and Directory) az oldenburgi könyvtári információs rendszer (Bibliotheks Informationssystem, BIS) projektje (http://gerhard.de). Az indexelő szolgáltató része nyilvánosan is hozzáférhető.

A GERHARD keresőgépet használ, mellyel – hasonlóan számos más ország gyakorlatához – a világhálón megjelenő német tudományos (akadémiai) publikációk teljes szövegéből webindexet állítanak elő. A különbség az, hogy a GERHARD a több százezer forrásdokumentumot egyidejűleg többnyelvű hierarchikus katalógusban is rendelkezésre bocsátja a böngészés céljára. A hierarchikus osztályozási rendszert automatikus osztályozással állítják elő. A keletkező hierarchiarendszer és vele az internetkatalógus rendkívül átfogó és mély, és az automatizált előállítás következtében meglehetősen naprakész. Mindenesetre nagyobb, lényegesen aktuálisabb, rendszerezettebb és lényegesen kisebb ráfordítással készül, mint pl. a Yahoo!.

Voltak már korábbi próbálkozások is, hogy az automatikus osztályozást, összekapcsolva a hagyományos, bevált könyvtári osztályozási rendszerekkel az interneten fölhasználják, de ezek lényegesen szűkebb szakterületekre terjedtek ki, heterogénebb dokumentumtípusaik voltak és kevésbé korszerű matematikai eljárásokat használtak.

A GERHARD-ban számítógépes nyelvészeti módszereket ötvöznek statisztikai eljárásokkal, hogy a dokumentumok természetes nyelvű szövegét automatikusan összevethessék a választott könyvtári osztályozási rendszerrel.

Osztályozási rendszerül a zürichi műszaki egyetemen karbantartott, kibővített és többnyelvű ETO-változatot használják. A kb. 70 000 osztályból álló rendszer német, francia és angol változatban létezik, szerkezetét a számítógépes alkalmazás céljára átdolgozták.

Az automatikus osztályozásnak két részfolyamata van: az egyik a gépi nyelvészeti és statisztikai műveleteket tartalmazza, a másik pedig az eredmény összehasonlítását a preparált ETO-táblázatokkal. A HTML-dokumentumok természetes nyelven megfogalmazott tartalmát megfelelő egységekbe, szavakba és mondatokba szegmentálják és az ETO-ból kialakított szókészlettel vetik össze. Az eredmény egy-egy dokumentumra vonatkozó ETO-jelzetek sorozata, melyet a gyakoriság és a dokumentumok szerkezeti egységeinek jelentősége szerint rangsorolnak. A dokumentum címe alapján kapott osztályozási jelzeteket például többre értékelik a súlyozáskor. Végül hozzárendelik a dokumentumokhoz a legdominánsabb klaszterből következő jelzeteket és a többszörös találatokat eredményező tartalmasabb jelzeteket. Egyben törlik ugyanannak a jelzetnek hierarchikusan átfogóbb (rövidebb) változatait, meghagyva a legspecifikusabb (leghosszabb) jelzetláncot. Egy-egy dokumentum átlagosan 6-7 jelzetet kap.

A dokumentum tartalmának releváns részeit indexelik és az ORACLE által kezelt adatbázisban hozzáférhetővé teszik. Az indexbe az osztályozási jelzetekhez tartozó kifejezések is bekerülnek. Jelenleg a német webszerverek kb. 950 000 HTML-dokumentumát osztályozták már ilyen módon. Ezzel párhuzamosan a böngésző keresés számára dinamikusan generálható az osztályozási rendszer hierarchiája, melyben a felhasználó lapozhat.

A keresőgépekkel létrehozott indexet és a böngészésre alkalmas hierarchikus osztályozási struktúrát ezáltal egyetlen rendszerbe integrálják. A felhasználó az indexelőszolgáltatásban az egyes találatokból kiindulva áttérhet az osztályozási rendszerre, melyben nagyszámú hasonló tartalmú dokumentumra akadhat, azaz az analitikus tartalmi feltárás kiegészíthető a szintetikus feltárás eredményével. A folyamat természetesen fordítva is elvégezhető: az osztályozási rendszerből is át lehet térni az indexelőszolgáltatás használatára, az egyes katalógus oldalakon időzve, természetes nyelvű szavakkal fogalmazhatók meg keresőkérdések az indexelőszolgáltatás számára, azaz a katalógusból átléphetünk az indexelőrendszerbe.

Jelenleg még gondot okoz, hogy az ETO természetes nyelvű mutatója (szótára) tartalmaz többértelmű, ill. redundáns kifejezéseket is, és logikailag sincs kifogástalan struktúrája. Az ETO hierarchia néhány elszigetelt láncát a jobb logikai felépítés érdekében a GERHARD-ban máshová helyezik el a struktúrában.

A többnyelvűség következtében a homonimák is több problémát okoznak (vö. pl. az angol „windows” kifejezés különféle jelentéseit más nyelveken).

Mind a Scorpion, mind a GERHARD azt bizonyítja, hogy az automatikus osztályozás előtt nagy jövő áll az interneten. E téren ők képviselik a kezdetet. A jelenlegi változatokat még tovább kell javítani.

7. Összefoglalás

Noha már többéves tapasztalatokra tekinthetünk vissza, az alkalmazások, mindenekelőtt pedig a módszerek száma nem elég nagy ahhoz, hogy teljesen egyértelmű megoldásokról beszélhessünk.

Néhány gyakorlati tanács azonban megfogalmazható.

a) Egyedi objektum (digitális dokumentum) esetén a tárgyszavas (deszkriptoros) indexelés fontosabb, mint a hierarchikus rendszerrel végzett osztályozás.

b) Áttekinthető számú csatolt objektum esetén elegendő a házilagos rendszerezés vagy egyszerűbb szakmai felosztás a dokumentumtípusok részletes megkülönböztetésével.

c) Nagy mennyiségű csatolt objektum esetén, legyen az egyetemes vagy meghatározott szakterületé: célszerű megfelelő osztályozási rendszert választani.

d) Más keresőszolgáltatásokkal és adatbázisokkal együttműködő keresőszolgáltatás esetén: az együttműködő szolgáltatások és adatbázisok alkalmazkodjanak valamelyiknek az osztályozási rendszeréhez.

e) Egyetemes gyűjtőkörű vagy nemzeti internetkatalógus a nagyközönség számára: célszerű a Yahoo! vagy egy ennek megfelelő súlyú katalógus osztályozási rendszeréhez alkalmazkodni. A felhasználók bizonyos rendszerekhez már hozzászoktak.

f) Intellektuálisan ellenőrzött (minőségbiztosított) szakmai információszolgáltatások esetén: mindenképpen valamelyik bevált osztályozási rendszert kell használni, adott esetben különféle nézetekkel. Mindegyik dokumentumot le kell írni szabványosított metaadatokkal, megadva a szabványt is, amelyet alkalmaznak, hogy a visszakereshetőséget és újrafelhasználást globálisan biztosítsák.

Természetesen nem minden digitális dokumentumokat tartalmazó gyűjtőkörben kell és lehet többé-kevésbé hagyományos módon intellektuálisan osztályozni. A fontosságukat hosszú távon is megőrző dokumentumok, továbbá a hagyományos dokumentumok esetén viszont éppen ez a tartalmi feltárás legjobb módja. Az automatikus osztályozási eljárások továbbfejlesztésének pedig,

amellyel majd rendkívül sok dokumentum is feldolgozható, elsődleges jelentősége van.

Mindennél fontosabb azonban, hogy a dokumentumokat lehetőleg jó minőségű tartalmi leírással kell ellátni szabványos metaadat-formátumokat fölhasználva, és biztosítani kell a hozzáférhetőségüket mind az internetkatalógusokban végezhető böngészéshez, mind pedig az indexelőszolgáltatásokon keresztül a végezhető lekérdezéshez.

8. Irodalom

CIP, Common Indexing Protocol. „The Architecture of the Common Indexing Protocol (CIP)" 11/21/1997.(ftp://ftp.nordu.net/Internet-drafts/draft-ietf-find-cip-arch-01.txt)

Kirriemuir, J. (et al.): Cross-Searching Subject Gateways. The Query Routing and Forward Knowledge Approach. In: D-Lib Magazine, Jan. 1998. (http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/january98/01kirriemuir.html)

Koch, T., Day, M.: The role of classification schemes in Internet resource description and discovery (EU Project DESIRE. Deliverable D3.2.3). 1997. (http://www.ub2.lu.se/desire/radar/reports/D3.2.3)

Larson, R. R.: Experiments in automatic Library of Congress Classification. In: Journal of the American Society for Information Science, 1992, 43, (2), p. 37–48.

Markey, K.: Subject searching strategies for on-line catalogues through the Dewey Decimal Classification. In: Hildreth, C. R. (ed.): The on-line catalogue: developments and directions. London: Library Association, 1989, p. 61–83.

McKiernan, G: Beyond Bookmarks: Schemes for Organizing the Web. (http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm)

Thompson, R., Shafer, K., Vizine-Goetz, D.: Evaluating Dewey concepts an a knowledge base for automatic subject assigment. Dublin, Ohio: OCLC, 1997. (http://orc.rsch.oclc.org:6109/eval_dc.html)

Vizine-Goetz, D.: Using library classification classification schemes for Internet resources (Position Paper). Proceedings of the OCLC Internet Cataloging Colloquium, San Antonio, Texas, January 19, 1996. Dublin, Ohio: OCLC, 1996. (http://www.oclc.org/oclc/man/colloq/v-g.htm)

PEGGY ZORN [ET AL.]

Peggy Zorn rendszerelemző, Mary Emanoil referensz könyvtáros és Lucy Marshall rendszeradminisztrátor kutatóintézeti könyvtárosok (Parke-Davis Pharmaneutical Research Library), Mary Panek pedig kutatóintézeti munkatárs (United Technologies Research Center).

Keresés a hálón haladóknak: szakmai fogások[34]

Összefoglalás: A végső felhasználók egyre inkább támaszkodhatnak az információs szakemberekre, ha a hálón összetett keresésre[35] van szükségük.

A World Wide Weben keresni nem több, mint hogy ráklikkelünk a Netscape-en a Net Search gombra, beírunk néhány, a kérdést reprezentáló keresőszót, elküldjük ezeket a Submit gombbal és néhány perc múlva már kész is a válasz. Vagy mégsem?

Az alábbiakban mindig olyan keresőrendszerekről van szó, melyek a szerveroldali navigációs eszközök feladatát látják el.

Az utóbbi hónapokban a háló keresőrendszerei gyorsan szaporodtak és fejlődtek, de nem nagyon foglalkoztak azokkal az összetett megközelítési módokkal, amelyekhez a keresőszakemberek és könyvtárosok más on-line információforrásoknál már hozzászoktak.

Az újsütetű webhasználók és az információs szakemberek gyakran átsiklanak az olyan részletinformációk felett, mint

- mit is keresnek egy bizonyos webkeresőrendszerben,
- hogyan indexelték az adatokat,
- hogyan keresi a keresőgép az adatokat,
- milyen eljárások állnak rendelkezésre (szótávolsági operátorok, zárójelekkel tagolt kérdések, a kérdéshalmazok manipulálása és összekapcsolása, az ismétlődő találatok, átfedések kiszűrése stb.).

Az egyszerű webkeresőrendszerekben, amilyen a Yahoo! és az Aliweb az indexelés mélysége és a keresőgép teljesítménye nem teszi szükségessé az összetett keresési eljárások ismeretét. A bonyolultabb rendszerek azonban előbb-utóbb elvezetnek ahhoz a fájdalmas felismeréshez, hogy a kereskedelmi forga-

34 Peggy Zorn [et al.]: Advanced Web Searching. Tricks of the trade. In: On-line, 1996, May/June, p. 15–28.

35 Az „advance search" kifejezést összetett keresésnek, a web-et hálónak fordítottuk, de ez utóbbi angol alakjával is éltünk (a szerk.).

lomban lévő on-line szolgáltatásokéhoz hasonló nagy találati halmazok miatt nem kerülhető ki a kulcsszavas keresés további szűkítése.

Ahogyan terjed a háló, és nőnek a web keresőgépeivel lekérdezett adatbázisok, egyre hatékonyabb információmegtaláló, indexelő- és információkereső eljárásokra van szükség, és egyre többet kell tudnia az információs szakembernek arról, hogyan kell hatékonyan keresni a hálóban. A tipikus végső felhasználónak általában nincs gondja az egyszerű témák keresésével, de az összetett keresőkérdések megfogalmazására kifinomult webkeresőgépek felhasználásával – már egészen más kérdés. Az összetett keresésekben a végső felhasználók egyre inkább rászorulnak az információs szakemberekre, ahogyan a kereskedelmi forgalomban lévő on-line adatbázisok esetében is történt.

A webkeresőrendszerek magasabb szintű lehetőségei

Nem minden webkeresőrendszer, és ezeknek nem minden keresőgépe és az ezekkel lekérdezett nem minden adatbázis készül egyformán. Ha a felvett URL címeket és az indexelés mélységét vizsgáljuk, csak néhányról mondhatjuk el, hogy a teljes hálóban használhatók. Alig van közöttük olyan, amelyik a könyvtárosok számára mérvadó összetett keresési lehetőségeket biztosít.

Az 1995. novemberi/decemberi ONLINE-ban *Martin Courtois, William Baer* és *Marcella Stark* kimerítő áttekintést adott a web számos keresőrendszeréről, és részletesen beszámoltak a mintakeresésekkel kapott teljesítménymérési eredményekről is. A szerzők többször is hangsúlyozták, hogy az összetett keresési lehetőséget biztosító rendszerek használati utasítását vagy megtalálni vagy megérteni nehéz. Jó hír viszont olyan tendencia megjelenése, hogy több keresési szempontot és az összetett keresésre vonatkozóan áttekinthetőbb dokumentációt kínálnak. Ezek az előrelépések azt ígérik, hogy javulni fog a webkeresés pontossága és hatékonysága.

A keresést finomító összetett keresési lehetőségek mellett az adatbázisok mérete és az indexelés mélysége is nagymértékben meghatározzák, milyen bonyolult keresést lehet adott keresőgéppel végrehajtani és milyen eredménnyel jár a keresés. Egyes gépek az internet helyek és az ezekben lévő adatbázisok teljes szöveges keresését támogatják. Mások, mint a Magellan és az Open Text, továbbmentek a mezőszintű indexeléssel (a cím, a szerző, a kivonat, a kulcsszavak stb. mélységéig). Ezek használnak a keresést és a visszahívást támogató deszkriptorokat és egyéb, a honalapokon (home page) nem szükségképpen hozzáférhető információkat. Ez a mezőszintű és deszkriptoros keresés és a kiegészítő indexelés a kereskedelmi forgalomban lévő on-line adatbázisok oszlopa, és egyaránt alkalmas a tág témák, illetve a webkeresésekkel visszahívott találatok szűrésére.

Az összetett keresési lehetőségek értékelése

Ebben a cikkben szeretnénk közelebbről megvizsgálni néhány összetett keresési lehetőségeket biztosító webkeresőrendszert, amelyekkel az internethelyeken átfogó, hiteles adatbázisokban keresni lehet. E két feltétel alapján az AltaVista, az InfoSeek, a Lycos és az Open Text rendszereket választottuk az értékelés és a vizsgálat tárgyául.

A keresési lehetőségek közül

- az összetett (zárójelekkel tagolt) Boole-logika,
- az ismétlődő rekordok, átfedések kiszűrése,
- a szövegösszefüggésben szereplő kulcsszó (kulcsszavak),
- a keresés leszűkítése mezőkre,
- a keresés a keresőszavak közelsége alapján (szótávolsági keresés) és/vagy összetett kifejezések alapján végzett keresés,
- a találatok relevancia szerinti rangsorolása,
- a találatok megjelenítésének beállítási lehetőségei,
- a keresési halmaz manipulálása,
- az (automatikus vagy a használó által megadott) csonkolás

érdekelt bennünket.

Nagy figyelmet szenteltünk az indexelés mélységének és a dokumentáció minőségének, valamint a keresőgéppel kezelt adatbázis méretének. Csak a 200 000-nél több internethelyre kiterjedő keresőgépeket vettük figyelembe.

Eleinte több más webkeresőrendszerrel és adatbázissal is fogalakoztunk, beleértve a Harvest, a Magellan, a NlightN és a Yahoo! rendszereket. Mind az NlightN, mind a Yahoo! fontos adatbázisokat kezelnek, de keresőgépeik nem elég erősek ahhoz, hogy a mi kérdéseinket hatékonyan kezelhessék. A Harvest és a Magellan viszont igen erőteljes keresőgépekkel rendelkeznek, amelyek biztosítják a fenti lehetőségek többségét, viszont adatbázisaik túlságosan kicsik ahhoz, hogy átfogónak nevezhetnénk őket.

A továbbiakban részletesen ismertetjük az általunk vizsgált négy webkeresőrendszert, ezek jellemzőit, hogyan használhatók, milyen teljesítményt nyújtottak a keresőgépek mintakereséseink során. Minden mintakeresés túlment az egyszerű kulcsszavas keresésen, és az összetett lehetőségek közül a lehető legtöbbet tartalmazza. Mind a három keresést lefolytattuk a négy keresőrendszerben, az egyes rendszerekben adott lehetőségeknek megfelelően átalakított szintaxissal.

A táblázatban áttekinthetjük a négy rendszer jellemzőit.

	AltaVista (altavista.digital.com)	**InfoSeek** (www.infoseek.com)	**Lycos** (www.lycos.com)	**Open Text** (www.opentext.com)
Az URL[35]-ek száma kb.	16 millió	1 millió	10,75 millió	1 millió
Dokumentáció	Kiváló; részletes, keresési példák mind az egyszerű, mind az összetett keresésre	Hosszadalmas; kissé nehéz megtalálni	Korrekt, nem olyan alapos, amilyen lehetne	Kiváló, minden lehetőséget részletesen ismertet
Átfedések kiszűrése	nincs	van	nincs	van
Keresés mezőkre	van	nincs	nincs	van, a mezőhatárok legördíthető menükből választhatók
Indexelés	teljes szöveg	teljes szöveg	URL-ek, a lapok és szövegek más részei	teljes szöveg
Többszörös kereső halmaz	nincs	nincs	nincs	nincs, de négy Boole- vagy szótávolsági operátor használható egy kérdésben
Zárójelezett boole-i keresés	van	nincs	nincs	nincs
Szótávolsági keresés	van, biztosítja, hogy mindkét szó (vagy kifejezés, ha idézőjelben áll) tíz szón belül forduljon elő	van	van, szoros megfelelés jellegű (laza, meglehetős stb.) legördíthető menü alapján	van, legördíthető menüből lehet operátort választani
Rangsorolás relevancia szerint	van, meghatározható mely kifejezések vezessék a keresési eredmények megjelenítési jegyzékét	van	van	van
Csonkolás	van, a három betűnél hosszabb szótövekhez csillag kapcsolható	nincs	van, automatikus	van, automatikus

Webkeresőrendszerek: Összetett Keresési Lehetőségek

35 URL = Uniform Resource Locator – általános forráshely-meghatározó, cím a hálón (a ford.).

Mintakeresések

#1 keresés: Keressünk információt a rákkutatás alapítványi vagy egyéb finanszírozási lehetőségeiről. A keresés tartalmazhat csonkolást a változó szóvégek visszahívása érdekében és több Boole-operátort a szinonimák és a fogalmak összekapcsolására. A hagyományos on-line szolgáltatásokban ez a keresés valahogy így festene:

(cancer or oncol*) and research and (grant* or fund*)

#2 keresés: Keressünk információt a Warner-Lambert-ről vagy gyógyszerészeti kutató részlegéről, a Parke-Davis-ről. A keresés a beágyazott írásjel miatt cseles. A probléma a keresőgép utasításszerkezetétől függően többféleképpen kezelhető. E fogalomra a tipikus on-line keresés esetleg így nézne ki:

(warner adj lambert) or (warner-lambert) or (parke adj davis) or (parke-davis)

#3 keresés: Keressünk információt a XI. nemzetközi AIDS konferenciáról. E keresés numerikus elemet is tartalmaz, és megköveteli a szinonimák és Boole-operátorok többszörös használatát. A téma keresőkérdése esetleg így fogalmazható meg:

(xi* or 11th or eleventh) and international and conference and aids

AltaVista http://altavista.digital.com

A Digital Equipment Corporation AltaVista rendszere nemrégen, 1995 decemberének közepén lépett a háló keresőszolgáltatásainak arénájába. Az AltaVista adatbázisában mintegy 16 millió URL teljes szövegének mutatóját tartalmazza, és igen nagy teljesítményű és rugalmas keresőgéppel végezhetők keresések. Az információs szakemberek számára az AltaVistában talán az a legüdítőbb, hogy a keresőutasítások és keresési lehetőségek nagyon hasonlítanak a hagyományos on-line adatbázis szolgáltatásokhoz. A Boole-operátorok (AND, OR, NOT, NEAR) alkalmazása, a zárójellel tagolt Boole-algebrai kérdések megfogalmazásának lehetősége, szükség esetén a csonkolás (amely azonban nem automatikus) ezek közé az ismerős eljárások közé sorolhatók. A keresési lehetőségek dokumentációját az AltaVistában részletes, számos példával illusztrálták.

Az AltaVista az adatbázisban végzett keresés két módját kínálja: az egyszerű és az összetett keresést. Az egyszerű keresés nem a Boole-operátorokra

épít, hanem különböző szintaktikai megoldásokkal fejezi ki, hogy a szavak kifejezéseket alkotnak, egymástól milyen távolságra vannak, és megkülönböztet szükséges és tiltott szavakat. A keresés korlátozható mezőkre is (cím, URL, host stb.), a csonkolás mindkét módban megengedett. A keresőkérdéssel az egyszerű keresési módban előhívott eredményhalmazt a használó relevancia szerint rendezett sorban kapja meg.

Az összetett kereséskor a szavak, kifejezések, helyettesítő karakterek és írásjelek meghatározására az egyszerű keresésével azonos szintaxist használunk. Az összetett keresés során *kötelező* a Boole-operátorok használata a szavak és kifejezések összekapcsolására és csoportosításukat zárójelekkel kell megoldani. A visszahívott halmazt tetszőlegesen lehet rendeztetni (meghatározhatóak azok a szavak vagy kifejezések, amelyek a jegyzékek vezérszavai lesznek), és időben is korlátozható.

#1 keresés: A mintakeresésekre összetett keresési módot használtunk. A rákkutatás támogatásáról szóló információ keresőkérdése eredetileg így festett:

(cancer or oncol*) and research and (grant* or fund*)

Hibaüzenet jelent meg, amely tudatta, hogy a „grant" és „fund" kifejezések csonkolt formái túl nagy halmazt eredményeznek, ezért a keresést a következő stratégiával fogalmaztuk újra:

(cancer or oncol*) and research and (grant or grants or fund or funds or funding)

Ezzel a kérdéssel több mint 20 000 dokumentumot hívtunk elő, amelyek legalább néhány kifejezésnek megfeleltek. A „cancer" és a „research" kifejezések relevancia rangsorolási kritériumként történő megadása javította az eredménylista élén megjelenő dokumentumok pontosságát.

#2. keresés: A Parke-Davisre vagy Warner-Lambertre vonatkozó keresés az AltaVista belső központozási konvencióját alkalmazta. A központozási jelek a szavakban vagy kifejezésekben elválasztójelként működnek, amelyek a szavakat elkülönített egységekre bontják. Ezen kívül a keresőkérdésben a nagybetűket és a kisbetűket külön kezelik. Általában a kifejezéseket kisbetűvel visszük be, és ezzel a kis- és nagybetűre érzékeny megfeleltetést jelezzük, ha nemcsak a nagybetűvel írt változatra van szükségünk. Így a keresési stratégia a következőképpen alakult:

„Parke Davis" or „Warner Lambert"

Ezzel mintegy 2000 találatot értünk el, amelyben az első tíz között szerepelt a Warner-Lambert honlapja.

#3 keresés: A XI. nemzetközi AIDS konferenciára vonatkozó információra vonatkozó keresés a következőképpen festett:

(xi or eleventh or 11th) and international and conference and aids

Az „aids", „international" és „conference" kifejezéseket adtuk meg a relevancia rangsorolás kritériumaként. Több mint 1000 találatot értünk el, és ezek élén a konferencia honlapja állt.

InfoSeek http://www2.infoseek.com/

Az InfoSeek keresőrendszert 1995 februárjában vezette be az InfoSeek Corporation, és ez volt a hálón az első olyan rendszer, amelyet csak térítés ellenében lehetett használni. Az InfoSeek jelenleg kétféle információkeresési lehetőséget kínál. A használók weboldalakon, hírcsoportokban, FTP[36] segítségével és gopherrel érhetik el a gyűjteményt, és 100 találatig a keresés ingyenes.

Az InfoSeek tagsági tervei hozzáférhetők azok számára, akik az InfoSeek további, térítés ellenében szolgáltatott adatbázisait – a Usenet News-t, Cineman Movie, Book, Music Reviews-t, a Hoover's Company Profiles-t, többféle gépi kapcsolatot (wire service), a CorpTech Directory of Technology Companies-t és az MDX Health Digest-et – is el akarják érni. A tagsági díj a keresésenkénti csekély 0,20 $ összegtől a havi 9,95 $ előfizetési díjig terjed. Az előfizetők számos cikk teljes szövegét is megkaphatják. Bár a keresési paraméterek azonosak az ingyenes és a térítéses adatbázisokban, csak az ingyenes szolgáltatásokat teszteltük.

Az InfoSeek jelenleg több mint egymillió internethelyet fed le. Adatbázisainak „benépesítésére" a hálót letapogató „robotok" és „leszedők" (kraulerek, crawlers) kombinációját alkalmazza, és él a használói bejelentkezésekkel is. A használók könnyedén felvetethetik lapjukat az InfoSeek adatbázisba. Nem kell kulcsszójegyzéket mellékelniük, de dokumentumaiknak mindenképpen tartalmaznia kell olyan szavakat, amelyek pontosan tükrözik a tartalmat, mert az InfoSeek az adatbázisába bekerülő valamennyi lap teljes szövegét indexeli.

A keresések eredményeit a szógyakorisági statisztikát és más szignifikancia és relevancia meghatározó eljárásokat alkalmazó program rögzíti. Bár az InfoSeek nemigen ismerteti rangsoroló eljárásait, hangsúlyozza, hogy e célra egynél több szógyakorisági számítást használnak. Ezen túlmenően az InfoSeek megpróbálja megakadályozni, hogy az URL-ek tételei többször forduljanak elő az adatbázisában, és egy URL-t állítása szerint soha nem indexel többször. URL-ek több címmel előfordulhatnak, de egy InfoSeek keresés során kevesebb vak találat születik, mint bármely más webkeresőrendszerben.

36 FTP = File Transfer Protocol = állományátviteli protokoll, fájlokhoz való hozzáférést és letöltést biztosító protokoll, eljárásgyűjtemény és szabvány (a szerk.).

Az InfoSeek az adatbázisaiban való keresést bemutató segítő képernyőt ad. Alapos magyarázatokkal és jó példákkal mutatja be az egyszerű keresést, az összetett keresésben az operátorok használatát, a keresési stratégiában követendő szintaktikai szabályokat és a keresési eljárások finomítását. Bőséges, de FAQ[37] (GYFK) fejezete is van, bár ezt kissé nehéz megtalálni.

Az InfoSeek többféle összetett keresési lehetőséget biztosít. Más keresőgépekkel ellentétben az InfoSeek nem kedveli a Boole-operátorokat, sem a helyettesítő karaktereket. Az InfoSeek használóinak az összetett keresésekben speciális szintaxist kell alkalmazniuk:

- az egymás mellett szereplő szavak alapján végzett keresésre kettős idézőjelet,
- a közvetlenül egymás mellett álló szavak alapján végzett keresésre kötőjelet,
- zárójellel kell közbezárni azokat a szavakat, amelyekkel tetszőleges sorrendben egymás mellett szereplő szavak alapján akarnak keresni,
- plusz jelet kell tenni azok elé a szavak vagy kifejezések elé, amelyek alapján egy az egyben akarnak keresni,
- mínusz jelet kell tenni azok elé a szavak vagy kifejezések elé, amelyek előfordulását bizonyosan ki akarják zárni az eredményül kapott valamennyi dokumentumból.

Az InfoSeek keresőgépe figyelembe veszi a kis- és nagybetűket. A nagybetűvel írt szavak csak a szó nagybetűs előfordulásait szűrik ki.

#1 keresés: Az InfoSeek elveti a Boole-operátorokat és a helyettesítő karaktereket, ezért a rákkutatás támogatására vonatkozó #1 kérdést nem lehetett úgy megszerkeszteni, ahogyan egy hagyományos on-line szolgáltatásban kellett volna. Első próbálkozásunkban a következő, kötőjelekkel elválasztott kifejezéseket vittük be:

cancer–research–grants

Ez a keresési stratégia csak három dokumentumot hívott elő, amelyek közül mindössze egyet találtunk relevánsnak. A kötőjel megköti, hogy a kifejezések egymáshoz nagyon közel jelenjenek meg, de megfeledkeztünk arról, hogy a kötőjelek a kifejezések sorrendjét is megkötik. Ezért a következő keresési stratégiában ugyanezeket a keresőkifejezéseket szerepeltettük, de egy zárójelben, hogy az egymás közelében tetszőleges sorrendben előforduló szavak alapján keressünk.

A zárójeles keresés 100 dokumentumot hívott vissza – ez az ingyenes keresésre megszabott felső határ –, és az eredmények pontosabbak voltak.

37 Frequently Asked Questions (FAQ) = Gyakran Feltett Kérdések (GYFK) (a szerk.).

#2 keresés: Nem tudtuk, hogyan kezeli az InfoSeek a beágyazott kötőjeleket, ezért a Warner-Lambert és Parke-Davis cégekre vonatkozó keresést kétféleképpen kíséreltük meg. Elsőnek így vittük be a kérdést:

„Warner-Lambert", „Parke-Davis"

Az idézőjelek olyan szavak alapján végzett keresést biztosítottak, amelyek egymás mellett szerepelnek. Ez a keresés 100 dokumentumot eredményezett. Meglepő, de a Warner-Lambert honlap nem a rangsor elején szerepelt. A 16. volt, néhány olyan dokumentum mögött, amelyek éppen csak megemlítették Parke-Davis-t mint korábbi munkaadót. A hatodik dokumentum első oldalán egyik cég sem szerepelt, és egyikre sem tartalmazott nyilvánvaló utalást. A cégek nevében szereplő kötőjelek jelentőségét kiderítendő, a második keresésben elhagytuk ezeket:

„Warner Lambert", „Parke Davis"

Ez a keresés ugyanazt a 100 dokumentumot hívta elő. Nyilvánvaló tehát, hogy az InfoSeek nem veszi figyelembe a beágyazott írásjeleket, legalábbis a kötőjeleket nem.

#3 keresés: Az 1996-ban rendezendő XI. nemzetközi AIDS konferencia meghirdetésére vonatkozó információ keresésére úgy fogalmaztuk meg a kérdést, hogy a teljes kifejezésből vettünk néhány kulcsszót. A kifejezések közelségét és sorrendjét meghatározandó, kötőjeleket tettünk a kifejezések közé:

11th–conference–AIDS

Egy dokumentumot kaptunk a kérdés segítségével, és ebben szerepelt egy rövid, a konferenciára vonatkozó hirdetés. A 11. lehetett volna „eleventh", „11th" vagy „XI" is. Mivel Boole-operátorok nem használhatók, három külön keresést kellett elvégeznünk, hogy a kifejezés eltérő írásmódjait megkapjuk. A második keresésre a következő, római számmal kezdődő kérdést használtuk:

XI–conference–AIDS

Ezzel a stratégiával 31 dokumentumot kaptunk, amelyek közül az első három közvetlenül elvezetett az 1996-os konferencia honlapjához. Egyetlen dokumentumot sem eredményezett a kiírt számnévvel (eleventh–conference–aids) folytatott keresés.

Lycos http://www.lycos.com/

Mint latin neve sejteni engedi (Lycosidae – farkaspókok), a Lycos (Wolf-Spider) spider-alapú[38] keresőrendszer. A School of Computer Science fejlesztette ki a Carnegie Mellon University-n, és jelenleg a CMG@Ventures és a Carnegie Mellon University közös vállalkozása, a Lycos Inc. tulajdonában van.

A Lycos naponta szisztematikusan újraépíti az adatbázisát, és ellenőrzi az aktív kapcsolatokat. A keresőgép súlyozott információkeresést biztosít az adatbázisból, és a használó kérdésére relevancia-rangsorban adja a találatokat. A rangsorolást a szónak a dokumentumban elfoglalt helye határozza meg. A címben vagy fejlécben szereplő kulcsszó magasabb súlyszámot kap, mint egy szövegszó. A keresési eredmények tartalmazzák a megfelelési pontszámot, a kapcsolatok számát, a dokumentum címét, a tárgyszavakat, mintareferátumot, az URL-t, valamint a dokumentum hosszát. A Lycos több mint tízmillió oldalas, világméretű indexszel büszkélkedhet – a hálónak több mint 91%-át indexeli. A Lycos katalógus heti 300 000 oldallal nő, és hamarosan a www 99%-át számba veszi.

A Lycos nemrégiben olyan kereső képernyőt állított össze, amely megkönnyíti a keresési és megjelenítési lehetőségek kihasználását. A Search Option gombbal választható az AND és az OR, és meghatározható, mennyire legyen pontos vagy megközelítő a megfeleltetés. A boole-i keresést annyiban módosította, hogy három keresőkifejezés közül kettő kombinálható. Ez a megfeleltetés nem olyan hatékony, mint az igazi Boole-algebrai keresés, de lehetővé teszi az összetett fogalmak és a szinonimák összekapcsolását. A Display Option gombbal a használó meghatározza, hány találatot akar, és milyen részletességet vár el az eredménytől (szabványos, tömörítvény vagy részletes). A rendszer automatikusan csonkolja a keresőkifjezéseket.

A keresés és a megjelenítés lehetőségeit nem túl részletesen, de dokumentálták. Leírták, hogyan kell néhány keresésfajtát elvégezni, de nem adnak magyarázatot arra, hogyan is folyik a keresés. Nem tisztázzák például, hogy a visszahívott találatok számán kívül mi is a különbség a „közeli" és a „laza" megfeleltetés között. Több e-mail kérést küldtünk a Lycosnak, amelyben további felvilágosítást kértünk, de ezekre nem válaszoltak.

#1 keresés: Az első keresést (három kifejezés megfelelésére beállított keresési opcióval) így vittük be:

cancer oncology research grant fund

Az oncology feltehetőleg nem jelenik meg, ha a cancer szerepel, és a fund és a grant is kizárják egymást. A keresés a vártnál pontosabb volt (146 dokumen-

38 A Spider szabad szövegen belüli keresőrendszer-család neve (a szerk.).

tum). Az első néhány találat néhány rendkívül releváns lapot tartalmazott, közöttük a NIH Grants Database-t és a FAA Research Grants Program-ot. Számos, a találati lista rangsorában hátrébb szereplő dokumentum tartalmazta a „research", „grant" és „fund" kifejezéseket, de nem szerepelt bennük a „cancer" és az „oncology". A Lycosban nem lehet valódi boole-logikai, zárójelekkel tagolt keresést lefuttatni, ezért különböző szinonimákkal és kombinációkkal több keresést kell elvégezni ahhoz, hogy átfogó eredményeket kapjunk.

#2 keresés: A Warner-Lambert-tel vagy Parke-Davis-szel kapcsolatos keresésben a Lycos sem a keresési stratégiában, sem a lekérdezéskor nem törődött a kötőjelezéssel. Ez azzal az előnnyel járt, hogy bár a Parke-Davis alapján végzett keresés jellemzően Parke Davis szerinti találatokat eredményezett, Parke-Davis szerinti találatok is képződtek. Ha mindkét kifejezésre szerkesztettünk keresési stratégiát, akkor ugyanazzal a problémával találtuk magunkat szemben, mint az első kereséskor, és a keresési opciót két kifejezésre kellett beállítanunk. A keresőgép e megkötése, hogy a kifejezéseket AND-del kapcsolja össze, gyakorta olyan eredményekhez vezet, amilyenre az egyik találatban bukkantunk: egy zenei lapra, amely Bobby Parker és James „Thunderbird" Davis munkásságával foglalkozik. Ez azt is példázza, hogy a Lycos automatikus csonkolása, amely jó a kimerítő keresést igénylő kérdéseknél, téves találatokhoz vezethet.

#3 keresés: A harmadik keresés a Lycos keresőgép egyik érdekes tulajdonságát hozta felszínre. Jelenleg nem keres számok alapján. Ebben a keresésben olyan példa szerepelt, amely világosan megmutatta, hogyan működik a leírásnak és az elvárásoknak megfelelően a Lycos Search Options. A XI. nemzetközi AIDS konferenciára vonatkozó információt keresve könnyen meghatározható volt olyan megfelelés, amelyben tetszőleges két kifejezésnek kellett megfeleltethetőnek lenni:

11 11^{th} XI eleventh aids

Nem valószínű, hogy egy dokumentumban mind az „eleventh", mind a „11^{th}" szerepel, bár a ténylegesen csak a „xi", „eleventh" és „aids" kifejezésekkel kerestek. Ez a Lycos keresés két találatot hozott, mindkettő a konferencia honlapjához vezető belső mutató (pointer) volt.

Open Text http://www.opentext.com

Az Open Text Corporation of Waterloo, Ontario és az UUNET Canada által kifejlesztett Open Text több mint egymillió weboldal, webhely és Gopher szerver teljes szövegének mutatója. Az Open Text „leszedőt" (kraulert, crawler) használ a weboldalak leolvasására, amely az Open Text 5 nevű saját indexelő-

szoftverje segítségével indexeli minden egyes weboldal minden szavát. Az eredményt az Open Text adatbázis tartalmazza. Az új lapokat naponta adják a rendszerhez, és a régi lapokat rendszeresen ellenőrzik, hogy felfrissítsék a rossz vagy megváltozott kapcsolatokat.

Az Open Text erőssége maga a mutató. A dokumentáció szerint az Open Text 5 több mint 40 fájltípust képes indexelni. Az oldal minden szavát indexeli (tehát nincsenek tiltott szavak a keresőkérdésekben), meghatározza, milyen mezőkben jelennek meg a szavak és többnyelvű. Ezen kívül több technológia is gondoskodik róla, hogy az Open Text 5 a nagy teljesítményű szöveges keresőrendszereket igénylő szervezetek számára használható legyen. Ha ezzel indexelik az internetet, az Open Text bemutathatja, mire képes, és hasznos szolgáltatott tesz az internetközösség számára is.

Az Open Text háromféle keresést biztosít: egyszerű, súlyozott és nagy teljesítményű keresést. Az egyszerű keresés alapvetően kulcsszavas keresés, amelyben lehet pontosan a bevitt kifejezésre, valamennyi szóra (rejtett AND) vagy bármely szóra (rejtett OR) keresni.

A nagy teljesítményű keresés árnyaltabb, és a használóra bízza, hol kell keresni az adott szó, vagy szavak alapján:

- bárhol (az interneten mindenhol),
- összefoglalóban (a címoldal, az első tárgyszó és az Open Text által fontosnak tartott szöveg),
- a címben (a weboldal szerzője adja),
- az első tárgyszóban (az oldal szerzője adja),
- csatolóban (a webhelyről kifelé vezető kapcsolatokban, hiperlinkben).

Az alapvető Boole-operátorokat és szótávolsági operátorokat lehet választani.

A súlyozott keresés segítségével a használó meghatározhatja a szó vagy kifejezés súlyát. Akárcsak a nagy teljesítményű keresésben, megszabható, hogy a weboldal mely részében folyjon a keresés. Ilyenkor egy, a fontosságot tükröző számot írnak a súly kockába, ahol a nagyobb szám nagyobb jelentőséget hordoz.

A három keresési típus közül a nagy teljesítményű keresés a legfejlettebb. A kereső könnyen használható felülettel találkozik, használhatja a Boole- és a szótávolsági operátorokat, és mezőkre szűkítheti a keresést. A mintakeresésekben is látható, hogy a zárójeles csoportosítási lehetőségek hiánya nem okoz gondot. Az on-line dokumentációban foglalkoznak az index korlátaival a keresőkérdések szintaktikai elemzésében. Nyilvánvaló, hogy nem megengedett a zárójel használata a használói interfészben, így a keresőkérdés végén szereplő kifejezések gyakran csak afféle függelékek. Az Open Text e probléma megoldására két külön keresés elvégzését javasolja, de ez a megközelítés nem határozza meg elég pontosan, milyen típusú információt keresnek. Az Open Textben hamarosan

szükség lesz két tulajdonság – a kereső halmaz manipulálása és a keresések mentése – bevezetésére.

Az eredmények mind az egyszerű, mind a nagy teljesítményű keresés nyomán relevancia rangsorban jelennek meg. A súlyozott keresésben a használó az előfordulási arányok (a keresőkifejezések előfordulásának száma szorozva a kifejezések súlyszámával), vagy a keresőkifejezések megléte/hiánya alapján (amelynek kiszámítása a kifejezések súlyszámából történik, tekintet nélkül a keresőkifejezésekkel való megfelelésre) jelenítheti meg a találatokat.

Az Open Text feltehetően képes bizonyos fokig elkerülni az átfedéseket, de beszámoltak hibákról. Tesztkereséseink során nem találkoztunk többször előforduló tételekkel. Minden megfeleltetett oldalon lehetőség van az oldal megtekintésére, a találatok böngészésére vagy hasonló oldalak kikeresésére. Kiváló on-line dokumentáció ad részletes felvilágosítást e lehetőségekről. A különböző keresési módok és az Open Text FAQ (GYFK) használati utasítása bőséges információval szolgál a legtöbb keresési típusra vonatkozóan. Az Open Text a dokumentációban nem érintett kérdésekhez visszacsatolási lehetőséget is nyújt.

#1 keresés: A rákkutatás támogatásáról a nagy teljesítményű és a súlyozott keresési módokban próbáltunk információt szerezni. Eredetileg a kifejezéseket a nagy teljesítményű keresésben így vittük be:

```
cancer or oncol and research and grant or fund
```

A csonkolás automatikus, ha nem követi szóköz a kifejezést, ez ugyanis megakadályozza a csonkolást. A keresések az első tárgyszóra, az összefoglalóra és a címre terjednek ki, és mindenhonnét kaptunk eredményeket, de relevanciájuk korlátozott volt. A keresés és az Open Text dokumentációjának elemzése után nyilvánvaló lett, hogy az „or fund” használata utolsó kifejezésként azt eredményezte, hogy a megelőző kifejezéseket összekapcsolta, majd ehhez kapcsolta az OR-ral a „fund”-ot.

A keresést ismét elvégeztük a „fund” kihagyásával. Tíznél kevesebb találat volt a három keresési mezőben, de csak az összefoglalók keresése vezetett kimondottan releváns kapcsolatokhoz. Az „oncol” elhagyása nem változtatott az eredményeken.

Mégis úgy látszott, relevánsabb lapoknak is lenniük kell, ezért a szótávolsági operátor használatával is megismételtük a keresést a következőképpen:

```
cancer near research near grant
```

Ugyanazokat az adattípusokat választottuk ki, a weboldal bármely részére kerestünk, és meglepő eredményekre jutottunk. Sem az első tárgyszavak között, sem a címek között nem találtunk egyetlen oldalt se. Az összefoglaló pedig csak

két oldalt hívott elő, a bárhol a weboldalon pedig 73 találatot hozott. Az összefoglalóban talált eredmények nagyon relevánsak voltak, és egyben a „bárhol" végzett keresésre kapott 73 tételes lista első két tételét adták. Érdekes, hogy az összefoglalóban megtalált lapok címében is szerepeltek a keresőkifejezések, így érthetetlen, miért nem találták meg őket a cím szerinti kereséskor.

Ugyanezt a keresési típust használtuk a súlyozott keresésben. Mivel csak négy kifejezés vihető be, a következők alapján kerestünk: cancer (20), research (10), grant (20), fund (5). A zárójelben a súlyszámok szerepelnek. Az eredmények az előfordulás sorrendjében rangsorolódtak. A bárhol a lapon végzett keresés 786 találatot eredményezett, míg az összefoglalókban, a címekben és az első tárgyszavak alapján folytatott keresés egyenként ötnél kevesebb tételt eredményezett. Egyik sem volt különösebben releváns.

#2 keresés: A Warner-Lambert or Parke-Davis keresését egyszerű és nagy teljesítményű kereséssel végeztük. A kifejezések kötőjellel összekapcsolva semmilyen eredménnyel nem jártak, így a kifejezéseket kötőjel nélkül vittük be. Az egyszerű keresés így festett:

warner lambert parke davis (any of these words)

Több mint 30 000 lapot találtak, de a Warner-Lambert honlap nem volt az első tíz között. Ugyanezt a keresést úgy is megpróbáltuk, hogy valamennyi kifejezést kértük, és így 35 oldal jött le, amelyek között nem szerepelt a Warner-Lambert lap. A nagy teljesítményű keresésben a kifejezéseket így vittük be:

warner followed by lambert or parke followed by davis

Ez a keresés 89 találatot eredményezett a „bárhol a weben" módban, amelyek közül sok releváns volt a Warner-Lambert Parke-Davis kérdésre, de a Warner-Lambert honlap ismét csak hiányzott. Az összefoglalóra, címre és első tárgyszóra korlátozva a keresést hétnél kevesebb találatunk volt minden mezőre, de a Warner-Lambert lap nem szerepelt közöttük. A keresés elemzése után nyilvánvalóvá vált, hogy az utolsó kifejezés, a „davis" nem kapcsolódik szervesen a kérdéshez. Ezt úgy próbáltuk kiküszöbölni, hogy a keresést az alábbi szerkezetben megismételtük:

warner followed by lambert or parke davis (where parke davis is a phrase)

Ez a keresés a címmezőben egyetlen találatot eredményezett, a Warner-Lambert honlapot.

#3 keresés: A XI. nemzetközi AIDS konferenciára is egyszerű és nagy teljesítményű kereséssel kérdeztünk. A konferencia számát 11th or Eleventh or XI formában vittük be, a többi kifejezés különböző kombinációival. Az ered-

mények helyek voltak, amelyek a konferencia honlapjához vezettek, sőt maga a honlap is előkerült.

Döntések összetett webkeresésekben

A kiválasztott négy keresőrendszer jellemzőit és a mintakeresések eredményeit mérlegelve megállapítottuk, hogy nincs olyan webkeresőrendszer, amely tényleg a „legjobb". A négy rendszer egyikének adatbázisa sem terjed ki az internet *egészére.* Nagyon nagy az eltérés az indexelésben és a felveendő weboldalak értékelésében, ezért egyetlen rendszer sem vallhatja magát teljesnek.

A felvett URL-ek számát tekintve talán az AltaVista és a Lycos a legátfogóbbak. Az URL-ek értéknövelő indexelése és kiválasztása azonban az AltaVistában, az InfoSeekben és az Open Textben az információkeresés relevanciája és pontossága szempontjából magasabbra rangsorolja őket. Az Open Text előnye, hogy a legfejlettebb keresési lehetőségeket biztosítja, jó a használói interfész, és kiváló a bonyolultabb funkciók dokumentációja.

Bizonyos keresési funkciók, például a keresőhalmaz manipulálása, az átfedések, ismétlődő rekordok kiszűrése, amelyeket minden on-line szolgáltatástól elvárunk, a webkeresőrendszerek többségéből még hiányoznak. A weboldalak teljes szövegének indexelése, beleértve a belső kapcsolatokat, a webhelyek olyan hatalmas adatbázisát eredményezi, amelyekben több száz ismétlődő tétel van, és ez megmagyarázza, miért van olyan sok átfedő találat a Lycos keresési eredmények között.

Minden valamirevaló kereső tudja, hogy a több Boole-operátort igénylő összetett keresések a legtöbb on-line keresőrendszerben egy kifejezésben is bevihetők. A kereső általában mégis jobban szereti a fogalmakat és az operátorokat több kereső állításra felbontani. Az Open Text kivételével minden rendszerben csak egy kereső ablak van a kifejezések bevitelére, így a fogalmak szétválasztása és a kifejezések összekapcsolása nem egyszerű. Még az Open Textben is, ahol látszólag a zárójelezés is lehetséges, az #1 keresés bebizonyította, hogy az OR használata a kérdés végén eltorzítja az egész keresést.

Szakértői (professzionális) keresés a hálón

Hogyan közelítsen a kereső szakember az összetett internet keresésekhez? Az adott eszközök birtokában a leghelyesebb talán a négy vizsgált keresőrendszer használata. Előfordul, hogy a kereső az egyes kereskedelmi on-line szolgáltatók több adatbázisában futtatja le a keresőkérdést, ha átfogó és pontos keresési eredményeket akar elérni, és ugyanez előfordulhat a webkeresés közben is. Az információs szakembereknek meg kell ismerkedniük e fejlett internet kere-

sőszolgáltatások jellemzőivel, lehetőségeivel, és igyekezniük kell ezeket beépíteni saját jövendő fegyvertárukba.

A szakértők kísérjék figyelemmel, hogyan alakulnak a keresési lehetőségek az interneten a következő hónapokban. A Microsoft engedélyezte a Lycos keresőgépet, a WebCrawlert az America On-line vezette be és egyre népszerűbbek a térítéses internet-keresőszolgáltatások, mint például az InfoSeek és a NlightN. Mindez a jelenlegi on-line keresési környezetéhez hasonló utat jelöl ki a webkeresőrendszerek számára. Néhány jelentős webkeresőrendszer üzletivé válása végső soron mind a szakemberek, mind a végső felhasználók számára a hatékonyabb, pontosabb és gazdaságosabb kereséshez vezethet.

G. FLETCHER–A. GREENHILL

A témával foglalkozó első könyv 1993-ban jelent meg[39], a publikációk száma azóta is folyamatosan gyarapszik. Számos formátumjavaslat készült, 1995-ben pedig megszületett a vonatkozó nemzetközi szabvány is[40] (a nem elektronikus dokumentumokra való hivatkozásnak már 1991 óta létezik magyar szabványa is). Az alábbi cikkben ismertetett formátumokat az ISO szabvány, továbbá a két legismertebb amerikai javaslat megoldásaival egészítettük ki.

Szakirodalmi hivatkozás internetforrásokra[41]

Az interneten elért szakirodalmi forrásokra való hivatkozásnak még nem alakult ki az egységes, konzisztens megoldása. Amíg ezen nem sikerül változtatni, az internetről elért forrásmunkák nem számíthatnak teljes értékűnek a tudományos diszkussziókban. Ennek folytán az ilyen módon publikáló kutatók sem számíthatnak a hagyományosan publikálókkal azonos elismerésre, így a kutatóhelyek megfosztják magukat az újdonságokra legfogékonyabb kutatóiktól.

---

39 Li Xia, Crane, N. B.: Electronic style: A guide to citing electronic information [Elektronikus stílus: elektronikus dokumentumok hivatkozási leírásainak útmutatója]. Westport, Mecklermedia, 1993. 65 p.

40 ISO/DIS 690-2. Excerpts from International Standard. Information and documentation. Bibliographic references. Part 2: Electronic documents or parts there of [ISO szabványajánlás. Bibliográfiai hivatkozások. 2. Rész. Elektronikus dokumentumok és részeik] [online]. ISO/TC 46/SC 9 [1999. 02. 10.] (1999. 12. 31.) Nyomtatott változata ISO, 1995. 26 p. <http://www.nlc-bnc.ca/iso/tc46sc9/standard/690-2e.htm>. Lásd még: Sipos M.–Ungváry R.: Hivatkozás távoli hozzáférésű HTML-dokumentumokra. In: Tudományos és Műszaki Tájékoztatás. 47. évf. 12. sz. p. 495–502.

41 Fletcher, G.–Greenhill, A.: Academic referencing of Internet-based resources. In: Aslib Proceedings, 1995, Vol. 47, No. 11–12, p. 245–252. (Ford. Válas György).

Az internetforrásokra való hivatkozás összhangban kell legyen az eddigi hivatkozási gyakorlattal. Ebben az írásban az internetdokumentum formátumából (állományformátumából) kiindulva mutatunk be javaslatokat a hivatkozás módjára. Elfogadása esetén szükségtelenné válik a számítógépes állománycímke megadása, amely a probléma megoldásának eddig elterjedt, de nem megfelelő módja volt, mivel a különböző számítógéprendszerek állományszerkezete annyira különböző, hogy önmagában az állománycímke nem nyújthat elegendő információt.

Az interneten megjelenő dokumentumok egyben fájlok, azaz elektronikusan tárolt állományok. Ezért sokan, köztük e tanulmány szerzői is, internetállományról, HTML-, Gopher-, FTP- stb. állományról beszélnek internet-, HTML-, Gopher-, FTP- stb. dokumentum helyett. Az „állomány" kifejezés indokolt, ha számítástechnikai, a „dokumentum", ha könyvtári-dokumentációs szempontból tárgyalják ezeket az entitásokat. A fordításban elsősorban a „dokumentum" kifejezést használtuk.

A hivatkozásnak olyan támpontot kell adnia, amelynek alapján a legkülönbözőbb kutatók ugyanazt az információt hívhatják le. Ehhez a hivatkozásnak tartalmaznia kell a hivatkozott dokumentum (állomány) fizikai és szimbolikus helyét, nevét és formátumát. Ezeket az igényeket jelenleg legjobban a World Wide Web (WWW) állományok megjelöléséhez használt *Uniform Resource Locator* (URL) rendszer elégíti ki. Ennek a szerkezete a következő:[42]

állománytípus://számítógép.részhálózat.országkód/
alkönyvtár/állománynév

A részhálózat, illetve az alkönyvtár neve több hierarchikusan egymás alá rendelt részből is állhat. Az alkönyvtár(ak) és az állomány nevében a megfelelő kis- és nagybetűk nem cserélhetők fel.

A URL némi redundanciát tartalmaz, ami bizonyos ellenőrzési lehetőséget jelent.

A klasszikus hivatkozási konvenció szerint a hivatkozás végére kerülő pont a URL után nem tehető ki, mert zavart okozhatna.

42 Az USA esetében az országkód helyén az intézménytípus (oktatási, kormányzati, profitorientált stb.) áll (a szerk.).

Hivatkozás internetdokumentumokra

A World Wide Web HTML (Hyper-Text Mark-up Language) formátuma az on-line tudományos folyóiratok legelterjedtebb formátuma.

A WWW dokumentumazonosítás alapja a URL. Előnyös, hogy emellett minden WWW dokumentumnak van „hétköznapi" címe is, amelyet a WWW böngészőprogram a képernyő címsávjában mutat meg. Az internetdokumentumnak tehát három kötelező adata van: állománytípusa, címe és URL-je. Emellett tartalmazhatja a szerző nevét, a közrebocsátó intézmény nevét és a létrehozás dátumát.

Ha a dokumentum tartalmazza a szerző nevét, nincs gond. Ha ez hiányzik, ami tudományos publikációk esetében ritka, a dokumentum vagy az abban elhelyezett kapcsolódási pontok valamelyike rendszerint utal egy intézményre. Gyakran szerepel a szerző elektronikus levelezési (e–mail) címe is. Ez olyan adat, amely a szerző neve helyett alkalmazva biztosíthatja a kereshetőséget. A szerző és a közrebocsátó intézmény neve elhelyezhető az internetdokumentum fejrészében (tehát egyszerű olvasáskor meg nem jelenő részében) is, ez a fejrész azonban sajnos nem kötelező.

Figyelembe kell venni, hogy az internetdokumentumoknak legalább két típusa létezik. A gyakoribb típus a gyűjtemény, amelynek nincs olvasható tartalma, csak kapcsolódási pontokat tartalmaz az olvasható dokumentumokhoz. Ezeknek a gyűjteményeknek rendszerint nincs feltüntetve a szerzőjük, viszont csak ritkán hivatkoznak rájuk. A ritkább típus elektronikus folyóiratot vagy adatokat tartalmaz, és nagy valószínűséggel tartalmazza az egyéni vagy az intézményi szerző nevét.

Az URL tartalmazza a közreadó nevét, de rejtjelezve. Vannak azonban interneteszközök, amelyekkel az URL-ből keresés végezhető a tényleges intézménynév segítségével is.[43] A későbbi biztonságos információkeresés érdekében ezt is fel kell tüntetnünk a hivatkozásban. Az intézmény nevével és a dokumentum címével ugyanis rendszerint akkor is megtalálható a dokumentum, ha azt időközben áthelyezték.

A dokumentumnak a szerző általi utólagos megváltoztatása a módosított új kiadáshoz hasonlítható. Ez azonban remélhetőleg ritkán következik be a tudományos dokumentumokban. Az on-line napilapok esetében azonban eltűnhet a dokumentum. Ezért helyes, ha a kutató saját archívumot rendez be az általa hivatkozott dokumentumokból.

Ha a dokumentumnak van nyomtatott változata is, hivatkozzunk inkább arra.

---

43 Ha a teljes URL-ből elhagyjuk az alkönyvtára(ka)t és az állománynevet tartalmazó részt, és a puszta állománytípussal és gépnévvel próbálunk bekapcsolódni (egyes esetekben még egy / jelet téve a végére), rendszerint olyan weblapra jutunk, amely tartalmazza a közreadó intézmény nevét (a szerk.).

Az internetdokumentumokra való hivatkozás legjobb stílusa a monográfiákra való hivatkozásból vezethető le. A hivatkozás a szerzővel kezdődjék. Ha ismert, szerepeljen a hálózatra történő első felvitel éve. Címként a dokumentum ablakának tetején szereplő címet szerepeltessük. A kiadónak az állományt kezelő intézmény neve felel meg, a kiadás helyének az URL (amelyet nem követhet pont, és amelyet elválasztani csak / jelek után szabad). Ha az elektronikus folyóirat kötet- és füzetszámozást tartalmaz, ez a folyóirat neve után helyezhető el. Ha az első felvitel éve nem deríthető ki, körülbelüli évet adjunk meg, pl. ≤1995 vagy ≅1995.

Példa az internetdokumentumra való következetes hivatkozásra:

Brenner, Anita 1995, The Murder Trial: Genre or Event-Scene?, C-Theory,
http://english–server.hss.cmu.edu/ctheory/e–murder_trial.html

Ha több weboldalra van elosztva a hivatkozott dokumentum, hivatkozzunk a részeire külön-külön, mint a monográfia különböző szakaszaira, fejezeteire.

Az ISO szabvány, és minden más érdemleges javaslat következetesen sorfolytonos megjelenítési formát ír elő, ill. ajánl. Ez összhangban van szerkesztőségek ama igényével, hogy a hivatkozások a lehető legkisebb helyet foglalják el. A bibliográfiai leírás sorfolytonos szerkezetét egyébként ugyanez az igény határozta meg.
Az internetdokumentumokon belül a szabvány megkülönbözteti az egyedi művek, a nem önálló részművek, az önálló műnek számító részművek és az időszaki kiadványok leírását. A legnagyobb nehézséget az okozza, hogy az elektronikus könyvek, folyóiratok és folyóiratcikkek kivételével az internetdokumentumok többségét kitevő ún. használati dokumentumoknak (közérdekű, szórakoztató és kereskedelmi jellegű html-oldalak) rendkívül hiányosak a kiadásra vonatozó adataik (az impresszumuk), és az is nagyon bizonytalan, mi tekinthető részműnek és mi nem.

Egyedi művek esetén a szabványos hivatkozási leírás (a kötelező és opcionálisan kötelező adatelemek félkövéren szerepelnek):

Elsődleges szerzőség:_Cím_[Adathordozó típusa]._Másodlagos szerzőség._**Kiadás._Megjelenési hely_:_Kiadó,_Megjelenési dátum._[Megújítás dátuma]**_ (Hivatkozás dátuma)_**Sorozat.**_ Megjegyzés._**<URL>**

Az alábbiakban összehasonlításként bemutatjuk az elektronikus dokumentum leírását a két legkimunkáltabbnak tekinthető amerikai javaslat – MLA (Modern Language Association) és APA (American Psychological Association) stílusú hivatkozásként szokták emlegetni őket – és az ISO 690–2 szabvány szerint.

APA
Burka, L. P. (1993). A hypertext history of multi–user dungeons. *MUDdex.* http://www.utopia.com/talent/lpb/muddex/essay/ (13. Jan. 1997).

MLA
Burka, Lauren P. „A hypertext history of multi–user dungeons.” *The MUDdex*. 1993. http://www.apocalypse.org/pub/u/lpb/muddex/essay/ (5 Dec. 1994).

Ugyanez és további példa a szabvány szerint:

ISO 690–2
Burka, Lauren P: A hypertext history of multi–user dungeons [on-line]. [1993] (1994.12.05.)·In: The MUDdex. <http://www.apocalypse.org/pub/u/lpb/muddex/essay>

További – teljes impresszumú – példa egyedi műre a szabvány szerint:

ISO 690–2
St. Thomas Aquinas: Summa Theologica [on-line]. Wheaton : Christian Classic Ethereal Library. [1999.05.27.] (1999.12.31.) Nyomtatott forma: Benzinger Bros, 1947. <http://ccel.org/a/aquinas/summa>

Részmű, ha nem önálló mű:

Gazdadokumentum elsődleges szerzősége._Gazdadokumentum címe_[Adathordozó típusa]._Kiadás._Megjelenési hely_: _Kiadó,_Megjelenési dátum._[Megújítás dátuma]_(Hivatkozás dátuma)_**Fejezet._Részdokumentum címe._Számozás a gazda-**

dokumentumon belül._Hely a gazdadokumentumon belül._ Megjegyzés._**<URL>**

ISO 690–2
Magyar Elektronikus Könyvtár [on-line]. [Budapest] : OSZK. (1999.10.18) Babits Mihály: Babits Mihály összegyűjtött versei, 5. rész. Talán a vízözön <www.mek.iif.hu/porta/szint/human/szepirod/magyar/babits/osszes/babits.05>

Részmű, ha önálló mű:

Elsődleges szerzőségi:_Címe_In:_Gazdadokumentum elsődleges szerzősége._Gazdadokumentum címe_[Adathordozó típusa]._Kiadás._Megjelenési hely_:_Kiadó,_Megjelenési dátum._[Megújítás dátuma]_(Hivatkozás dátuma)_ **Számozás a gazdadokumentumon belül._Hely a gazdadokumentumon belül.**_Megjegyzés._**<URL>**

ISO 690–2
Szakadát István: Xanadu. In: Uniworld. A virtual university [on-line]. Uniworld Közhasznú Egyesület, 1999. [1999.09.16.] [Az „internet gyakorlata és filozófiája” c.kurzus „Hálózott tudás” c. témájának „hypertext” csatolóján keresztül is elérhető ajánlott irodalma] <http://www.uniworld.hu/netskills/tudas/HTML/Xanadu.htm>

Elektronikus folyóirat:

Cím_[Adathordozó típusa]._Kiadás._Megjelenési hely:_Kiadó,_Megjelenési dátum._[Megújítás dátuma]_(Hivatkozás dátuma)_**Sorozat.**_Megjegyzés._**<URL> ISSN**

ISO 690–2
INCO Első magyar internetes folyóirat az információs korról [on-line]. [Budapest] : Harmadik Évezred Alapítvány Stratégiai Kutató Intézet, 1999–. (1999.12.31.) <http://www.inco.hu/>

Elektronikus folyóiratcikk:

Elsődleges szerzőség:_Cím._In:_Folyóirat címe_[Adathordozó típusa]._Kiadás._Keltezési/számozási adatok._[Megújítás dátuma]_(Hivatkozás dátuma)_**Hely a gazdadokumentumon belül.** Megjegyzés._**<URL>**

ISO 690–2
Nyíri Kristóf: Információs társadalom és nemzeti kultúra. In: INCO [on-line]. 1. évf. 1999. 1. sz. (1999.12.31.) Információs kor értékvilága. <http://www.inco.hu>

Példa használati html-dokumentumra (számos esetben nem állapítható meg megjelenési hely, kiadó, megjelenési vagy megújítási dátum):

ISO 690–2
Koster, Martijn: The Web Robots Pages [on-line] In: Martijn Koster. [1999.06.] (1999.12.31.) Martijn Koster's Projects <http://info.webcrawler.com/mak/projects/robots/robots.html>

ISO 690–2
Welcome to mnteverest.net. Dedicated. („...") [on-line]. (1999.12.31.) Mt Everest history and record <http://www.mnteverest.net/history.html>

ISO 690–2
Go translator. Systran translator software [on-line]. InfoSeek Corp., 1999. [Az InfoSeek honlapjáról is elérhető] (1999.12.31.) <http://translator.go.com>

ISO 690–2
Yahoo! [on-line] Yahoo! Inc. (1999.12.31.) Search Engines. Computers and Internet/Internet/World Wide Web/Searching the Web <http://www.dir.yahoo.com/Computers_and_Internet/Internet/World/Wide/Web/Searching the Web/Search_Engines/>

ISO 690–2

ELVIRA. Az intelligens vasúti menetrend [on-line]. MÁV Informatika Kft. (1999.12.31.)
<http://elvira.mavinformatika.hu/index.htm>

Hivatkozás Gopher-dokumentumra

Az internetdokumentumokra javasolt hivatkozási módszer alkalmazható a WWW böngészővel elérhető egyéb dokumentumokra (állományokra), így a Gopher-dokumentumokra is. Ezek URL-je a **gopher://** előtaggal kezdődik. Általában hosszabb, mint az internetdokumentumoké, de ugyanolyan kényes az elírásokra, és a nagy- és kisbetűk közötti különbségre.

Problémát jelent, hogy – ellentétben az internetdokumentumokkal, amelyek egyéni szerzője rendszerint meghatározható – a Gopher-dokumentumok többnyire inkább a közreadó intézményhez kötődnek, névtelenségben hagyva a szerzőt. Ilyenkor érdemes a lehető legkisebb meghatározható szervezeti egységet tekinteni szerzőnek.

Példa a Gopher-dokumentumra való hivatkozásra:

Library Services c. 1995, Internet User Glossary, North Carolina State University,
gopher://dewey.lib.ncsu.edu:70/7waissrc%3A/.wais/
Internet-user-glossary

A Gopher-szolgáltatások hanyatlásával csökken az ilyen forrásokra történő hivatkozások szükségessége, a Gophereken található információ átvándorol a WWW szolgáltatásba.

A korábban más formában létező dokumentum gépre vivőjének megemlítése a hivatkozásban hasonló a fordító vagy a szerkesztő megemlítéséhez. Ilyen vonatkozásban azonban az intézményt nem sok értelme van megemlíteni, mert az úgyis szerepel kiadóként, valamint az URL részeként.

Hivatkozás FTP-dokumentumokra

Az FTP (File Transfer Protocol) -gyűjtemény az internetpublikálás legrégibb formája. Általában szoftver vagy szöveg letöltésére használatos. Rendszerint egyéni szerzőhöz köthető, hivatkozásra alkalmas adatokkal. Legtöbbször hagyományosan publikált anyag digitalizált változata. URL-je **ftp://** előtaggal kezdődik.

Példa az FTP-dokumentumra való hivatkozásra:

Gaffin, Adam 1994, EFF's Guide to the Internet, v. 2.3, Electronic Frontier Foundation, ftp://nysernet.org/pub/resources/guides/bigdummy.txt

Aki nem WWW böngészővel éri el az FTP-szervert, az is könnyen átalakíthatja az elérési út leírását szabványos URL-formára.

Hivatkozás Usenet News dokumentumokra

Usenet News dokumentumokra történő hivatkozásra a URL-rendszer nem alkalmas, ezekre nem adható értelmes URL, hiszen a terjesztés természetéből következően a felhasználó gépén található a dokumentum. Ezért az ilyen dokumentumra periodika módjára érdemes hivatkozni, kihasználva a Usenet News csoportok hierarchikus szerkezetét. Általában szerepel a dokumentumokon valamiféle szerző, a Usenet News csoportok azonban nem zárják ki az álnév használatát. A dokumentumoknak általában van fejlécük, ez címként használható. A csoport kezelhető folyóiratként. A pontos dátum meghatározható, ez helyettesítheti a kötet- és füzetszámot. Például:

Graham, Adrian, 1995, 'Fishing in Mauritius', alt.fishing, 29th July.

Az Usenet-dokumentumokra való hivatkozás legnehezebb problémája a dokumentumok ideiglenessége. Nem minden dokumentumot archiválnak, és a ma még elérhető dokumentum holnapra köddé válhat. Bár sok csoport archivál, az archivált dokumentum megtalálása sokszor nem éri meg a fáradságot. Többnyire egyszerűbb megkeresni a szerzőt a dokumentum egy példányáért.

Hivatkozás levelezési csoport dokumentumára

A levelezési csoportok a manuálisan terjesztett folyóiratokhoz hasonlíthatók. A levelezési szolgáltató minden füzetből minden beiratkozott olvasó elektronikus postacímére küld egy-egy példányt. Szerencsére a levél fejrésze tartalmazza a hivatkozáshoz szükséges információ nagy részét. Szerepel benne a szerző neve, a dátum, és a „Subject"-mezőben a cím. A folyóirat neve-

ként a levelezési csoport neve szerepeltethető, kiadóként pedig a URL helyett a csoport beiratkozási címe.[44]

Hivatkozás elektronikus levélre

Az elektronikus levél az internet magánközleménye. Ugyanolyan fenntartással kell kezelnünk a hivatkozásokban, mint bármely más magánközleményt, ugyanúgy nem tekinthető valódi hivatkozásnak.

Ha nem tudjuk a levél küldőjének tényleges nevét, használjuk az elektronikus postacímében az @ jel előtt álló részt.

Következtetések

Az itt leírt módszer a World Wide Web dokumentumelérési módszerén alapszik.

Ma nem tekintik komoly dolognak az internetforrásra való hivatkozást. Ha az ilyen dokumentumokat a hivatkozásokban ugyanúgy kezeljük, mint a hagyományosakat, lassan megváltoztatható ez a negatív szemlélet.[45]

A cikk – és a szabvány – megszületése óta számos szakterületen a hivatkozások egyre nagyobb része vonatkozik távoli elérésű, elsősorban html-dokumentumra.

CARLA LIST

Az internet szentsége[46]

Az internet iránti érdeklődés, különösen a könyvtári világban, finoman szólva is széles körű. „A háló” az az új varázsige, amely azonnal műszaki kompetenciával itat át. A könyvtárakban *technofejek, szörfözők* és *ludditák* csoport-

44 Nagyon fontos, hogy a beiratkozási címet adjuk meg, ne a levelezési címet, mert ez utóbbi nem ad kulcsot az archív anyag visszakereséséhez. Mivel a legtöbb szolgáltató jól kezelt archívumot tart fenn a levelezési csoport leveleiből, megfontolandó, hogy a könnyebb visszakereshetőség céljából esetleg ne az eredeti levélre hivatkozzunk, hanem FTP-dokumentumként annak archivált változatára (a szerk.).

45 A referálólapok és bibliográfiai adatbázisok lassan kezdik bevonni a feldolgozott források körébe a lektorált elektronikus folyóiratokat. Ezt a Chemical Abstracts 1995-ben kezdte, az INSPEC (Science Abstracts) pedig 1996-ban kezdi. A szemlélet tehát kezd megváltozni (a szerk).

46 List, Carla: Sanctifying the Internet In: American Libraries, 1995, Vol. 78, No. 10, p. 1019–1022. (Ford. Zsadon Béla.)

jaira bomlott a szakma. A nézeteltérés világos és könnyen meghatározható: vannak könyvtárosok, akik jártasak a Hálón, és vannak, akik nem. A mostanában egyre többször hallható *luddita* terminus magában hordozza a technikától való félelmet és annak megvetését. A jelző azonban téves: az ebbe a kategóriába sorolt könyvtárosok nem félnek attól a technikától, amit az internet jelent, és nem is utálják azt. Egyszerűen csak nincsenek meggyőződve róla, hogy az volna munkájuk *egyedüli* eszköze.

Megerősíti az elkülönülést a *zsargon* bennfentes használata is. Aki nem tud „fingerelni" vagy „ftp-zni", az nyilvánvalóan a luddíták csoportjába tartozik. Ezeket aztán semmibe veszik vagy lekezelik és könnyen kapják meg a „hitetlen" minősítést. A vallásos hevület gyakran jár együtt az objektivitás hiányával; küldetéses emberekkel nagyon nehéz vitatkozni. Ebben a helyzetben az a szerencsétlen csavar, hogy a másik csoport tagjai sem merik megfogalmazni aggályaikat.

A legfőbb kérdés

A legfontosabb kérdés: *kinek az igényeit elégíti ki* az internetért folytatott hajsza? Vagyis ki húz hasznot abból a döntésből, hogy még több számítógépet szerezzünk be, de ne töltsünk be egy könyvtárosi állást? Vegyünk még több szoftvert, anélkül, hogy az azokat használó könyvtárosok kiképzésének költségeit előirányoznánk? Kínáljunk még több kapcsolati lehetőséget az olvasóknak, anélkül, hogy figyelembe vennénk azokat a gyakorlási lehetőségeket, amelyekre mi magunknak lenne szükségünk?

Kiknek jó ez valójában? A kérdésre több közkeletű válasz is van. Az első szerint a *vezetőségnek.* Milyen nagyszerű leendő egyetemistákat (és tandíjfizető szüleiket) elhozni egy számítógépekkel zsúfolt könyvtárba – íme, az egyetem élen jár a műszaki fejlesztésben. A könyvtár vezetői szeretik a magas szintű konnektivitást, mert demonstrálhatják, mennyire lépést tartanak a jövővel. Használhatnának ugyanerre egy *könyvtárost,* mondjuk egy „multi-tasking" bemutatón? Aligha.

Sok könyvtárban az elektronikus források viszik el a költségvetés nagyobbik részét. Ez nem mindig valami átgondolt fejlesztés része: a legtöbb pénz évről évre upgrade-ekre, új szoftverekre vagy jobb internetkapcsolatokra megy el. Néhány könyvtár tudatosan dönt úgy, hogy a költségvetésben az elektronikus forrásoké legyen a prioritás, megelőzve az állománybővítést és a személyi kiadásokat. A techno-dagály végigsöpört a vezetőkön, és a lövészárokbeli könyvtárosoknak fel kell ismerniük ennek következményeit.

Az tény, hogy az új technika lehetővé teszi a differenciált költségvetést. A könyvtár többé nem kell birtokolja a dokumentumot, csak a hozzáférésről kell gondoskodnia. A *hozzáférés kontra tulajdonlás* vita tart már egy ideje. Kezdetben a költséges feltáró, indexelőszolgáltatások álltak a középpontjában; ma inkább né-

hány alapvető referenszműről van szó. Minek megvásárolni az Oxford English Dictionaryt, hiszen az olvasó használhatja azt a Hálón. A vezetők és a könyvtárosok egyetértenek abban, hogy a hozzáférés meglehetősen különbözik a tulajdonlástól, de főként abban – és ez az egyetlen, amire a vezetők odafigyelnek –, hogy a technika olcsóbbnak látszik. A kulcsszó a *látszik.* Ki dolgozik majd ezekkel a számítógépekkel? A könyvtárosok, természetesen, bár számuk ebben a „karcsúsító" érában apadhat. A könyvtárosok rugalmasabbnak látszanak a gépeknél, és munkabírásukat a vezetők végtelenül tágíthatónak vélik. Bárki, aki a Hálóval dolgozik, tudja, hogy fejlődésének mostani szintjén a vele töltött idő túlnyomó része sokkal inkább a kompetencia fenntartására megy el, mintsem a használatához kellő szaktudás fejlesztésére. Mégis sok könyvtáros kapja feladatul, hogy sajátítsa el a Háló szakszerű használatát – anélkül, hogy feladatait arányosan csökkentenék. Sőt, feladatkörét az olvasók internetképzésével is megterhelik. És mi lesz a villanykönyvtáros asztalán rakásra gyűlő „off-line" munkával?

A Hálón eltöltött *idő* értékelése példa arra, hogyan szolgálja a Háló a vezetők szükségleteit, hiszen mércét kínál a személyzet teljesítményének mérésére. A Hálón való szörfözéssel eltöltött idő *hasznos* – üzenik a főnökök. A könyvtár jó PR-jéhez járul hozzá az a könyvtáros, aki elektronikus postán küldözget ismertetőket a legújabb csinos forrásokról. Következésképpen a vezetők szemet hunynak a fölött a tény fölött, hogy némely könyvtáros a szüntelen „hálózatozás" miatt *nem* végzi el a többi munkáját. De nem azért követelik meg a Háló használatát a referenszpultnál, mert minden kérdésre az a legjobb forrás, hanem azért, hogy technikailag kompetensnek lássanak. A munkaidő értékelésekor aztán a ludditának bélyegzett könyvtáros nehezebben szerez kiváló minősítést. Lehet, hogy a Hálómászó kitűnő munkát végezne a Háló nélkül is, de nem elektronikus munkája a könyvtárosság kevésbé látványos, kevésbé piacképes, kevésbé mérhető és magyarán szólva kevésbé csillogó-villogó területeihez tarozik; következésképpen sokkal könnyebb azt bagatellizálni.

Más személyzeti ügyekben is vonzó ürügy az internet. A tény, hogy a Hálón információt lehet elérni, megerősíti azt a vélekedést, hogy a referenszpultnál kevesebb könyvtáros is elég. Viszont hosszabb lehet a nyitvatartási idő. A csökkentett költségvetések korában a vezetés gyakran folytat takarékossági politikát, figyelemre méltó azonban, hogy a műszaki fejlesztéshez mindig sikerül forrásokat keríteniük.

Hullámlovaglás

Szóval kinek a szükségletei elégíttetnek ki? Egy másik válasz: a *sajátunk.* Az információ a hivatásunk – és amiről az internet mindenek előtt szól az az információ. Azért töltünk órákat elektronikus szörfdeszkáinkon, hogy naprakészek maradjunk. Az olvasó szemében ez persze nem árt az arculatunknak.

Minden hájjal megkent professzionalistának hatunk. Kinek jó mindez? Az olvasónak? Az egyetemi könyvtárakban a könyvtári ismeretek oktatása a könyvtár egyik legfőbb küldetése, gyakran szerepel valamelyik tanév anyagában. De szükségük van-e az angolszakos gólyáknak a könyvtárban az internetre? Vagy ez az a bizonyos ágyúval verébre eset? Vagy talán a professzornak van rá szüksége arra, hogy felmutassa, milyen „menő fej”? Vajon információkutatásra tanítjuk-e a diákokat, ha az internettel kápráztatjuk el? Jobban fog tudni kutatni a következő dolgozatához? Valóban az igazi igényeit elégítjük ki? A Háló fejlődésének jelenlegi szintjén – vagy pontosabban: szervezettségének jelenlegi szintjén – a válasz egy dörgedelmes *Nem*.

Az olvasókat sokszor lenyűgözi a technika, még ha csak egy on-line katalógusról van is szó. Kinek teszünk jót, ha erre az olvasóra rányitjuk az internet zsilipeit? Sok olvasó attól retteg, hogy tönkretesz valamit, ha hozzáér a billentyűzethez. Az ő számítógép-fóbiájuk leküzdéséhez úszóleckékre van szükség, mielőtt fejest ugratnánk őket a Hálóba. (Persze mint már említettük, ezeknek a leckéknek a feladatát már hozzácsapták a munkakörünkhöz.) A jó könyvtáros munkája sohasem merült ki abban, hogy információt keres az olvasónak. Inkább együttműködik vele, kideríti mire is van igazán szüksége, majd kielégíti ezt a szükségletet a *legmegfelelőbb forrásból,* függetlenül attól, hogy az elektronikus-e vagy sem. Könyvtárosi létünk rendeltetése az olvasó szolgálata – a neki legmegfelelőbb szinten.

Jó tanácsok

A hálóimádó villanykönyvtárosoknak:

tartsák észben, hogy változó időket élünk. Az interneten való jártasságnak sokféle szintje lehet. A szakmának szüksége van az olyan emberekre, akik követik a legújabb fejleményeket és feltárják azokat a többiek előtt. Buzdításra van szükség, nem exkluzivitásra. Ajánlatos tisztában lenni azzal is, mit *nem* végzel el, mialatt szörfözöl, ne növeld a többi kollégára háruló feladatokat!

A vezetőknek:

munkatársaikkal együtt kell kidolgozni azt a programot, ami meghatározza az internet helyét könyvtárukban. Világosan ki kell fejezni, milyen szerepet játszik a Hálón töltött idő a munkateljesítmény értékelésében. Közölni kell a munkatársakkal az elektronikus beszerzések indokait, ne érezzék, hogy a gépek viszik el az összes pénzt.

Az úgynevezett luddítáknak:

rázzák le magukról a jelzőt. Ismerkedjenek meg a Hálóval – olyan tempóban, amellyel hozzá tudják illeszteni munkájukat a technikához. A status quo egyszerű fenntartása nagyon gyorsan a szakmai

színvonal stagnálásához vezet. Kövessék és alkalmazzák a szakma fejleményeit! De a legfontosabb, ne csak azt kérdezzék meg a Hálón jártasabbaktól, hogyan kell használni az internetet, azt is kérdezzék meg *miért!* És kérdezzék meg újra és újra: kinek az igényeit *kell* kiszolgálni.

Talán ismerős *Dale Dauten* megjegyzése:

„Teljes kötetek minden témában: mindegyikben adatok megabájtjai. És valami ‚Dewey' nevezetűnek hála, minden könyv olyan jól szervezett rendszerben van, hogy bármelyik könnyen megtalálható. És ott vannak az úgynevezett ‚tájékoztató könyvtárosok' – igazi emberi lények, nem bábfigurák – arra várnak, hogy segítségedre legyenek. És mindezt ingyen."

MUTATÓ

ff = az adott tárgyban lényeges rész.
A gyakran előforduló kifejezéseknek csak a fontosabb előfordulásait adtuk meg.
A mutatószavak egy része az első kötet mutatójában is megtalálható.
Az alábbi központi fogalmak összefüggéseit — az első kötet mutatójához hasonlóan – külön is megadjuk.

információkeresés
lásd még speciálisabban
automatikus információkeresés fajtái
gépi információkeresés
on-line információkeresés
szabad szövegen belüli keresés
szerver- és kliensoldali keresés
lásd még egyéb összefüggésben
automatikus információkereső rendszer
keresési stratégia
keresési stratégia fajtái
információkeresés lélektani eszközei
keresési taktikák
képzelet taktikák a keresésben
keresőkép
keresőkép fogalma
keresőszó
információkeresés értékelése
információkeresés fogalma
információkeresés fogalma az interneten
információkeresés folyamata
információkeresés kezdetei
információkeresés kognitív modellje
információkeresés megszületése
információkeresési modellek

on-line információkeresés
lásd még egyéb összefüggésben
on-line információkeresés alapfogalmai
on-line információkeresés alapjai
on-line információkeresés helyi adatbázisban
on-line információkeresés kézikönyvei

információkereső rendszer
lásd még speciálisabban
automatikus információkereső rendszer
lásd még egyéb összefüggésben
információs rendszer
információs rendszerek fajtái

információkereső nyelv
lásd még egyéb összefüggésben
információkereső nyelv szókincse
szabványosított szótár
tezaurusz
tezaurusz és információelmélet
szabad tárgyszó
keresőszó
legtöbbet használt keresőszavak a weben
mesterséges nyelv
természetes nyelv

J

K

L

M

N

T

U

V

W