HEKTÁR: Hazai elektronikus könyvtári rendszerek összekapcsolása

Kiss Gergő, Kovács László, Micsik András
{gergo.kiss, laszlo.kovacs, micsik}@sztaki.hu
Elosztott Rendszerek Osztály
MTA SZTAKI
Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet

Moldován István
moldovan@oszk.hu
Országos Széchényi Könyvtár

A HEKTÁR[1] nevű futó ITEM projekt keretén belül az MTA SZTAKI Elosztott Rendszerek Osztálya és az Országos Széchényi Könyvtár az Open Archives Initiative (OAI) ajánlásainak alkalmazásával különböző digitális könyvtári rendszereket kapcsol össze. Az OAI két alapvető fogalma az adatszolgáltató (data provider) és a szolgáltatási pont (service provider). Az adatszolgáltató egy szabványos protokollon keresztül lekérdezhetővé teszi metaadatállományát a külvilág számára. A szolgáltatási pontok ezt a protokollt felhasználva a kiválasztott adatszolgáltatók számára értéknövelt és egységesített szolgáltatásokat valósítanak meg.

A projekt során az MTA SZTAKI nyílt forráskódú példa-implementációt készített az OAI metaadat begyűjtő protokollra (OAI-PMH), amelyet a Magyar Elektronikus Könyvtárhoz (MEK) és más gyűjteményekhez illesztettünk hozzá. Az így létrejött adatszolgáltatókat egy saját fejlesztésű közös kereső szolgáltatással kapcsoltuk össze, melyben a cikk írásakor 14,000 tétel között lehet keresni. Az OAI alkalmazása megkönnyíti különböző telephelyek, könyvtárak összekapcsolását, egyszerűvé teszi újfajta közös szolgáltatások bevezetését, és nem utolsó sorban az alkalmazókat bekapcsolja a dinamikusan növekvő nemzetközi OAI közösségbe is. Az OAI kulcsszerepet játszik az NDA (Nemzeti Digitális Adattár) kezdeményezésben is. Ezért érezzük fontosnak az OAI és a kapcsolódó technológiák (pl. Dublin Core) elterjedését Magyarországon, melyet e projekttel is szeretnénk elősegíteni.

Az OAI-PMH protokollról

Az Open Archives Initiative (OAI)[2] a sikeres preprint mozgalmak példájából kiindulva a digitális archívumokban található információk hozzáférhetőségének, ismertségének növelését tűzte ki célul. A Dienst elosztott digitális könyvtári rendszer[3] vagy a Z39.50 alapú közös lekérdezési felületek üzemeltetési tapasztalatai alapján már látható volt, hogy a párhuzamos keresés több archívumban, majd az eredmények összefésülése, mint közös keresési elv nem igazán tartható. Ha ezen elv szerint zajlik a keresés, akkor a lassan válaszoló (pl. hálózat értelemben véve távoli) archívumok miatt lelassulhat a válaszadás, az éppen elérhetetlen archívumok miatt a találatlista csonka maradhat.

Ezért az OAI az archívumok “kinyitására” és összekapcsolására egy másik elvet vezetett be: a közösített szolgáltatások a kapcsolódó archívumoktól lemásolják, begyűjtik (harvesting) az adattartalomra vonatkozó metaadatokat, és így a felhasználói kéréseket helyben, hálózati forgalom nélkül tudják kiszolgálni. Érdekességképpen megemlítjük, hogy ezt az elvet 1996 óta az

NCSTRL (Networked Computer Science Technical Reports Library) elosztott könyvtári rendszer is alkalmazta, melynek közép-európai regionális csomópontját osztályunk üzemeltette. Az NCSTRL még a párhuzamos keresési elv alapján működött, de a keresés gyorsítására bevezették a regionális gyorsítótárakat, melyek már gyakorlatilag az OAI terminológia szerinti közösített szolgáltatásként (service providerként) működtek. Így egy rendszeren belül két fajta közös keresési elv is alkalmazásra lelt, ami azt is megmutatja, hogy itt nem versengő, hanem egymást kiegészítő elvekről van szó. A párhuzamos keresési elv most van megújulás alatt a ZING[4] protokollcsalád elterjedésével, mely a Z39.50 protokollt hivatott felfrissíteni. Napjaink The European Library (TEL)[5] projektje ismét mind az OAI-PMH mind a ZING protokollt felhasználja architektúrájában.

A metaadatok cseréjéhez, begyűjtéséhez szükséges protokollhívásokat az NCSTRL már tartalmazta, és ezek köszönnek vissza letisztultabb formában az OAI-PMH protokollban. Az OAI-PMH hat protokollhívást támogat, melyek az archívum és belső struktúrájának lekérdezésére, valamint a metaadatok begyűjtésére szolgálnak. Érdemes röviden felsorolni az OAI-PMH terminológiáját, mely általánosságra törekszik, ezért egyes szakmai körökben némely szakszó furcsán csenghet. A szolgáltatási pont (service provider) rendszeresen begyűjti (harvesting) a kapcsolódó tárolók (repository) metaadatait. A tárolók (vagy archívumok) belül készletekre, részhalmazokra (set) tagolódhatnak, akár hierarchikus módon is. Készletet képezhet például egy témakör, vagy egy intézményileg, fizikailag elkülönülő gyűjtemény. A tárolóban tételek (item) találhatók, és a tételekhez metaadat rekordok társulnak. A begyűjtést a következőkkel lehet még jellemezni:

- Mindig a begyűjtő kezdeményezi a begyűjtést, az adatszolgáltató készenlétben vár, és csak válaszol a protokollhívásokra.
- Az OAI elvű adatszolgáltatás során a tételek (az archívum tartalma) nem másolódnak le, a tárolóban maradnak, és azokról továbbra is teljes körűen a tároló rendelkezik.
- A begyűjtés „kötelező nyelve” a Dublin Core, bár lehetőség van emellett más metaadatformátumok használatára is.

A metaadatok begyűjtése történhet szelektíven és inkrementálisan is, azaz lehetőség van dátum és készlet szerint válogatni a metaadatok között, és nagyobb mennyiségű metaadatot több részletben, fokozatosan lehívni. Az OAI-PMH egyszerűsége mellett bonyolult architektúrák felépítésére is alkalmas. Egy szolgáltatási pont (service provider) a data provider szerepét is betöltve továbbadhatja a begyűjtött metaadatokat egy másik szolgáltatásnak, így többszintű hierarchia is kialakítható.

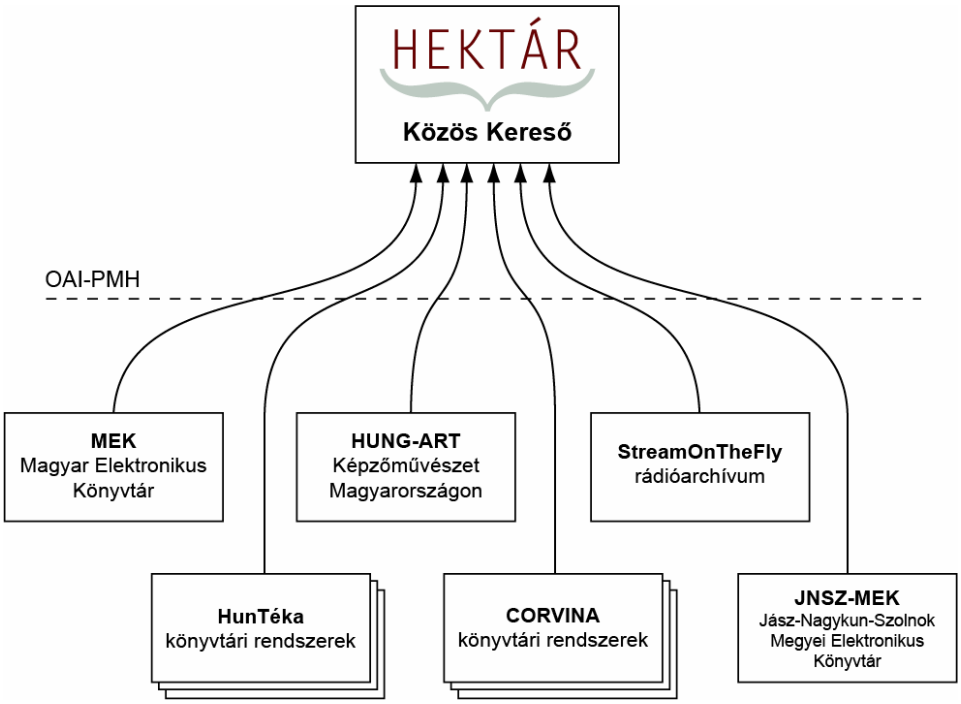
Tevékenységek a Hektár projektben

A projekt jelenlegi állása szerint 5 digitális archívum van összekapcsolva az OAI protokoll szerint:

- Magyar Elektronikus Könyvtár (MEK)
- Képzőművészet Magyarországon (www.hung-art.hu)
- StreamOnTheFly nemzetközi elosztott közösségi rádiós archívum (radio.sztaki.hu)
- A Földrajzi Társaság Könyvtára (HunTéka rendszeren keresztül)
- A Magyar Képzőművészeti Egyetem könyvtári képadatbázisa (Corvina rendszeren keresztül)

A Jász-Nagykun-Szolnok Megyei Elektronikus Könyvtár bekapcsolása még folyamatban van. A felsorolt tárolók közül az első három a saját fejlesztésű PHP nyelvű adatszolgáltató implementációval működik. Ez egy általános data provider megvalósítás, melyet az archívum belső adatszerkezetéhez, felépítéséhez kell illeszteni. Egy SQL alapú adatbázis illesztésére a kód tartalmaz példát. Az adatszolgáltató megvalósítása a projekt honlapjáról ingyenesen letölthető.

A fennmaradó két tároló két könyvtári rendszer, a Corvina és a HunTéka[6] illesztése. Mindkét esetben az OAI illesztőt a rendszer fejlesztői készítették, mely munkához szakmai útmutatást és segítséget adtunk, illetve az elkészült illesztőket ellenőriztük az OAI-PMH és a Dublin Core helyes használata szempontjából.



1. ábra: A közös kereső kapcsolatai

A felsorolt archívumok összekapcsolásának demonstrálására egy közös keresőt építettünk, mely április elején kezdte meg nyilvános működését. A közös szolgáltatás teljes egészében saját fejlesztés, tartalmaz egy általános begyűjtő (harvester) implementációt, egy felhasználó-kezelő rendszert, böngésző és kereső funkciókat. A begyűjtött metaadatokra nézve a szolgáltatás a következő funkciókat valósítja meg (ezek némelyike csak 2004 májusától lesz nyilvános):

- Böngészés: a metaadat rekordok archívumok, gyűjtemények szerinti listázása, a rekordok megjelenítése
- Egyszerű keresés: szavakra, szórészekre történő keresés, többféle listázási sorrendben, a keresett szavak környezetének kiemelésével a találati listában.
- Összetett keresés: elemi kifejezésekből tetszőleges összetett kereső kifejezés állítható össze, ezeket regisztrált felhasználóink elmenthetik, és később újra lefuttathatják, vagy módosíthatják. Az elemi kifejezéseknél bármelyik Dublin Core mező tartalmára lehet keresni, a kifejezések és/vagy műveletekkel lehet összekapcsolni. Gyakorlatilag a diszjunktív

vagy konjunktív logikai normálformának megfelelő bármilyen kereső-kifejezés felépíthető ilyen módon.
- Adminisztrátori felület: a tárolók begyűjtésének monitorozása, tárolók felvétele és módosítása, kézi beavatkozás az automatikus begyűjtési folyamatokba.

A közös kereső szolgáltatás a hektar.sztaki.hu címen bárki által használható. A szolgáltatást a projekt végéig (2004 június) folyamatosan fejlesztjük, javítjuk, és a szolgáltatást törekszünk hosszabb távon is fenntartani a projekt zárása után is. Lehetőség szerint segítjük további digitális gyűjtemények, archívumok kapcsolódását a közös keresőbe.

Tanulságok, tapasztalatok

A HEKTÁR közös keresőjébe eddig több mint 14,000 metaadat rekordot gyűjtöttünk be. A jobb minőségű keresőszolgáltatás elérése érdekében harmonizálni kellett a Dublin Core használatát a különböző adatszolgáltatókban. Az alap Dublin Core még túl tág és szabad formátum ahhoz, hogy jól használható közösített szolgáltatásokat lehessen építeni ez alapján. Gyakorlatilag nincsenek megkötések, csak ajánlások az egyes DC elemek tartalmára nézve.

A Dublin Core-nak azóta megjelent minősített változata (Qualified Dublin Core) már sokkal fejlettebb ebből a szempontból, mivel itt már az elemek jelentését finomítani lehet (pl. dátum esetén készítés, nyilvánosságra hozatal, módosítás, stb.), illetve az elem értékét megadott formátum szerint vagy listákból kiválasztva lehet megadni (pl. dátum formátum, típus elem értékére DCMI Types). A DCMI (Dublin Core Metadata Initiative) igyekezett összegyűjteni a hasznos értékmegadási sémákat (pl. DDC, LCSH, TGN, W3C-DTF), és ahol szükséges volt, maga készített ilyen sémákat (pl. DCMI Box, DCMI Period).

Gyakorlatilag a Qualified Dublin Core önmagában már nem értelmezhető, használható, az különféle elemkészleteknek, sémáknak, szabályoknak összessége, melyeknek csak egy része van a DCMI gondozásában, a többi részét más testületek kezelik, mint például a Library of Congress, World Wide Web Consortium (W3C), IEEE, stb. Egy ilyen séma-halmaz logikai modelljéről és belső szabályairól a CORES[7] európai projektben végzett munkánk kapcsán írunk [8].

Az OAI-PMH saját, úgynevezett oai_dc sémát használ, amely az 1.1 verziójú Dublin Core elemkészletre épül, és gyakorlatilag csak az elemeket és előfordulásaik számát definiálja. Az OAI tervezi az áttérést a minősített Dublin Core-ra, de ez egy hosszabb folyamat lesz a meglévő OAI adatszolgáltatók nagy száma miatt. Az oai_dc tehát nem ad semmilyen megkötést vagy támpontot az elemek tartalmára, kitöltésére vonatkozóan. A DCMI ajánl egy Using Dublin Core című dokumentumot[9], amely minden egyes elemhez az alapszabványnál jóval bővebb kommentárt és használati példákat is tartalmaz.

A DC és OAI Magyarországon

Az oai_dc használatakor tehát csak két támpontot találunk: a Dublin Core ajánlást és a Using Dublin Core ajánlott dokumentumot. Sajnos egyik dokumentumnak sincs még elfogadott magyar fordítása. Nagyon fontosnak tartanánk, hogy az ilyen alapvető dokumentumoknak legyen hivatalos és nyilvános magyar fordítása, sőt megszülessen ezeknek a hazai használatra szolgáló

adaptációja is. Ezért elkészítettük az OAI-PMH ajánlás magyar fordítását is, amely szintén a projekt honlapjáról érhető el.

A szabványok szó szerinti alkalmazása sajnos nem oldja meg a hazai problémákat, ezért is van szükség adaptációra. Az elsődleges kérdés az, hogy a metaadatokat milyen nyelven adjuk meg? Az OAI nem csak a hazai, hanem a nemzetközi közönség számára is elérhetővé teszi a metaadatokat. Ha magyar nyelven adjuk csak meg a metaadatokat, akkor a külföldi közönséget zárjuk ki, ha mondjuk csak angol nyelven, akkor pedig a hazait. Logikus érvelésnek tűnik, hogy csak azon a nyelven adjuk meg a metaadatokat, amilyen nyelven az alapobjektum íródott. Viszont képek, számadatok, szobrok és más nyelvhasználattól függetlenül érthető ábrázolásoknál ez az elv nem követhető. Véleményünk szerint minél több nyelven legyenek elérhetők a metaadatok. Ez viszont két új problémát vet fel:

- Ki állítsa elő a metaadatokat több nyelven? A metaadatok fordítása nagy szakértelmet igénylő és fáradságos feladat.
- Hogyan ábrázoljuk több nyelven a Dublin Core metaadatokat? Erre az oai_dc nem ad könnyű megoldást.

Az első probléma lehetséges megoldásait itt nem részletezzük. A második kérdés egyik megoldása az, hogy a DC mezőket annyiszor ismételjük meg, ahány különböző nyelven a mező tartalma rendelkezésre áll.

A nyelvi problémákon kívül majdnem minden DC mező lehetőséget adna külön hazai használati egyezségek bevezetésére. Például a HEKTÁR projekt keretében a következő szabályokat rögzítettük az oai_dc használatára (ezek nagy része összhangban van a Using Dublin Core javaslataival):

- A személyneveket eredeti, magyar használatuk szerint adjuk meg a DC mezőkben
- A dátumok megadására a 2003-12-31 típusú jelölést használjuk, melyből a nap, hónap elhagyható
- A formátum megadásánál, amennyiben lehetséges, az IMT (Internet Media Types) sémát használjuk (pl. text/html, image/jpeg)
- A típus megadásakor a DCMI Type Vocabulary-t használjuk, emellett megadhatók a szűkebb magyar/angol nyelvű típusnevek is (pl. regény, szakdolgozat, stb.)
- A tétel nyelvét ISO-639-2 (hárombetűs) kódolással ábrázoljuk
- Ha a tétel tartalma a hálózaton hozzáférhető, akkor az Azonosító mezőben a tartalomra mutató URL-t is közöljük

Általános alapelvként kikötöttük még azt, hogy ha egy mezőhöz több érték tartozik, azokat a mező ismétlésével adjuk meg, nem egy mezőn belül. A mezőkben ajánljuk a hivatkozások URL-lel történő megadását, főleg a Kapcsolat és Forrás elemekben, mivel ezek növelik a böngészhetőséget. Lehetőség lenne a minősítők megadására szabad szöveges formátumban ilyen módon: Creator = Zöld Béla (riporter). Sem ezzel, sem egységes témafa vagy tezaurusz használatával nem volt módunk foglalkozni, amely a Téma (Subject) elem kitöltésénél segíthetett volna.

A Qualified Dublin Core a fenti problémák egy részére már megoldást ad: lehetőség van a mezőtartalom nyelvének, illetve a kitöltéskor alkalmazott séma leírására. Így a Qualified Dublin

Core rekordok sokkal strukturáltabbak és kevesebb bennük a szabad szöveges információ. Azonban még mindig szükséges a kitöltéshez használandó sémák összegyűjtése, illetve a hiányzó sémák megalkotása. Itt elsősorban a Téma mező kitöltésekor használható magyar nyelvű séma hiányzik. Véleményünk szerint ahhoz, hogy a Dublin Core és az OAI hazai használata sikeres lehessen, hazai ajánlásokat, egyezségeket kell megalkotni, amelyek a lokális és nyelvi sajátosságokból adódó problémákra megadják a választ. Végezetül, a hazai és nemzetközi ajánlásoknak való megfelelés vizsgálatára, ellenőrzésére is létre kellene hozni a megfelelő módszereket és eszközöket.

Érvek az OAI mellett

Magyarországon sokféle közös könyvtári kereső működik már (Mokka, ODR, KözElKat, WebKat, stb.), felmerül a kérdés, hogy miért kell még egy másfajta módon is foglalkozni ezzel a feladattal, és van-e lényegi különbség a hagyományos könyvtári módszerek és az OAI között?

A fő különbség az OAI és Dublin Core használatában van: ezek gyorsan terjedő, nyílt, nemzetközi ajánlások. Használatukkal nem csak a hazai közösített szolgáltatásokat lehet megvalósítani, hanem adatvagyonunk nemzetközi láthatósága is biztosítható, lehetőség nyílik külföldi tárolókkal való együttműködésre. A Dublin Core univerzális, tehát nem korlátozódik a könyvtárakra, bármilyen objektum, adat leírható vele. Így összekapcsolhatók könyvtárak és ipari adatbázisok, webes archívumok és tetszőleges más adatforrás.

Az OAI felépítéséből adódóan alkalmas rugalmas és önszervező architektúrák kialakítására. Az OAI adatszolgáltatókból, begyűjtőkből és szolgáltatási pontokból bonyolult kapcsolati rendszerek építhetők fel. Egy szolgáltatás a tárolókból begyűjtött metaadatokat továbbadhatja, mint adatszolgáltató. Egy adatszolgáltatótól több szolgáltatás is begyűjtheti a metaadatokat. Ily módon akár hierarchikus akár decentralizált rendszerek konstruálhatók, illetve a meglévő rendszerek menetközben bővíthetők, változtathatók.

Az OAI technológiailag jól illeszkedik az Internet és a WWW világába, a kurrens technológiákat használja: XML, HTTP, Dublin Core. A Z39.50 szabványcsalád frissítése, a ZING is ez irányba törekszik.

OAI adatszolgáltató létesítése

Tapasztalataink alapján egy OAI adatszolgáltató létesítése során az alábbi tervezési lépéseken kell túljutni:

- Az archívum alapadatainak meghatározása: mit mondjon az archívum magáról, mi legyen a neve, ki legyen a karbantartója, milyen opcionális OAI protokollelemeket támogasson?
- Kell-e az archívumot setekre, készletekre, gyűjteményekre bontani? Ha igen, milyen hierarchiája legyen e halmazoknak?
- Azonosító séma definiálása: az archívum minden tételének rendelkeznie kell egy globálisan egyedi azonosítóval. Erre létezik egy séma ajánlás, amely például ilyen azonosítókat támogat: oai:mek.oszk.hu:MEK-1234. Ebben az első rész az OAI-t azonosítja, a második rész a tárolót, a harmadik a tárolón belül a tételt.
- Támogatott metaadatformátumok kiválasztása: a kötelező Dublin Core mellett egy tároló támogathat más metaadatformátumokat is (pl. MARC). Ekkor minden tételhez többféle formátumú metaadat is rendelkezésre áll.

- Metaadatrekord előállításának megtervezése: az archívumban már rendelkezésre álló metaadatokat le kell képezni a Dublin Core mezőkre, lehetőleg minél több más archívummal, ajánlással összhangban, mivel ez növeli a metaadatok felhasználhatóságát. A fentiek közül ez a legnehezebb és legnagyobb feladat.

Ezután már csak az Interneten szabadon nagy számban hozzáférhető OAI adatszolgáltatók egyikét kell illeszteni a tároló szoftverrendszeréhez. OAI kapcsolat kiépítésével kapcsolatban további tanácsokért, segítségért a szerzőkhöz lehet fordulni.

Összegzés

Reméljük, hogy a HEKTÁR projekt elősegíti az OAI hazai elterjedését, mivel megmutatta, hogy az OAI elv és technológia olcsó, könnyen implementálható és alkalmas eltérő architektúrájú rendszerek összekapcsolására. Az OAI terjedése közvetve növelheti a Dublin Core hazai használatát is, amely még szintén gyerekcipőben jár Magyarországon.

A projekt során összekapcsolt archívumok az OAI data provider révén a világon bárhol láthatóak, nemzetközi, nyílt archívummá váltak.

A megvalósított architektúra és szoftverrendszer nem sokban különbözik a tervezett Nemzeti Digitális Adattártól (NDA). Az NDA már Qualified Dublin Core alapon fog működni, és a közös keresés ki fog bővülni egy közös tezaurusszal és tulajdonnévtárral.

Köszönet

Köszönjük az együttműködést és támogatást a különböző kapcsolódó archívumok és szoftverek gazdáinak; a MEK, a HunTéka és a Corvina fejlesztőinek, valamint Krén Emilnek (www.hung-art.hu) és Tóth Kornélnak (HunTéka).

[1] HEKTÁR projekt honlapja, http://hektar.sztaki.hu
[2] Open Archives Initiative, http://www.openarchives.org
[3] Kovács László, Micsik András: „Elosztott digitális könyvtári projekt Európában”, Networkshop 1996, Debrecen
[4] ZING, "Z39.50-International: Next Generation", http://www.loc.gov/z3950/agency/zing/zing-home.html
[5] The European Library (TEL), http://www.europeanlibrary.org/
[6] Tóth Kornél: „Elosztott könyvtári rendszerek megvalósítása a Z39.50 és az OAI protokoll használatával”, Networkshop 2004, Győr
[7] Rachel Heery, Pete Johnston, Csaba Fülöp, András Micsik: Metadata schema registries in the partially Semantic Web: the CORES experience, http://www.siderean.com/dc2003/102_Paper29.pdf
[8] Fülöp Csaba, Kovács László, Micsik András: „Metaadatsémák nyilvántartása szemantikus web alapon”, Networkshop 2004, Győr
[9] Diane Hillmann: „Using Dublin Core”, http://www.dublincore.org/documents/usageguide/