

BESZÁMOLÓ

IMPACT Conference: „OCR in Mass Digitalisation” Challenges between Full Text, Imaging, and Language

6-7 April, 2009, Koninklijke Bibliotheek, The Hague

IMPACT Konferencia : „OCR a tömeges digitalizálásban” Kihívások a digitalizálás, fulltext és nyelv köréből.

2009. április 6-7., Koninklijke Bibliotheek (KB), Den Haag

KÉSZÍTETTE: RENKECZ ANITA (EDK)

TARTALOM:

BEVEZETÉS	2
1. NAP (2009. ÁPRILIS 6.), BEVEZETŐ SZEKCIÓ	3
DIGITISATION OF CULTURAL RESOURCES: EUROPEAN ACTIONS AND THE CONTEXT OF IMPACT.....	3
IMPROVING ACCESS TO TEXT – CENTRE OF COMPETENCE IN MASS DIGITISATION OF HISTORICAL PRINTED TEXT	5
LIBRARY CHALLENGES FOR MASS DIGITISATION.....	9
1. NAP (2009. ÁPRILIS 6.), DÉLUTÁNI SZEKCIÓ	11
ADAPTIVE OCR – COLLABORATIVE CORRECTION –COST EFFECTIVE DIGITIZATION	11
A USER FRIENDLY PLATFORM FOR DOCUMENT IMAGE ENHANCEMENT AND SEGMENTATION.....	16
MEASURING THE OCR ACCURACY ACROSS THE BRITISH LIBRARY’S 2 MILLION PAGE NEWSPAPER ARCHIVE	21
VITAPANEL	24
2. NAP 2009. ÁPRILIS 7.	26
DÉLELŐTTI SZEKCIÓ	26
“MANY HANDS MAKE LIGHT WORK” –PUBLIC COLLABORATIVE OCR TEXT CORRECTION IN AUSTRALIAN HISTORIC NEWSPAPERS	26
FUTURE CHALLENGES FOR OCR TECHNOLOGY	27
DIGITAL RESTORATION AND LAYOUT ANALYSIS	29
HISTORICAL LEXICON BUILDING & HOW IT IMPROVES ACCESS TO TEXT	31
DECISION SUPPORT TOOLS	33
PÁRHUZAMOS SZEKCIÓK	34
IMPROVING ACCESS TO TEXT: CENTRE OF COMPETENCE IN MASS DIGITISATION OF HISTORICAL PRINTED TEXT	34
2. CHALLENGES AND OPPORTUNITIES IN MASS DIGITISATION – HOW TECHNOLOGY CAN MEET LIBRARIES’ NEEDS	35
3. MASS DIGITISATION AT THE KB	36

Bevezetés

Az OSzK jelenleg nem tagja az IMPACT-projektnek, de mivel a TELPlus-ban részt vesz, bekerült azon intézmények körébe, akik a TELPlus Workpackage 1 (OCR) partnerekként meghívást kaptak az IMPACT rendezvényére. A kiutazás költségeit a TELPlus keretből finanszíroztuk, a projektmenedzsment engedélyével. Az OSzK képviselőjében kiutazó munkatársak: **Horváth Ádám** (Informatikai Igazgató) és **Renkecz Anita** (EDK).

A beszámolót a szokottnál részletesebben készítettem el, mivel az előadások döntő többsége rendkívül informatív volt, és olyan információkat tartalmazott, melyek más csatornán nem jutnának el intézményünkbe. A dokumentum kizárólag belső használatra készül, mivel helyenként az elhangzott prezentációkból kiemelt tartalmat közöl.

1. nap (2009. április 6.), Bevezető szekció

LEVEZETI: HILDELIUS BALK (KB)

DIGITISATION OF CULTURAL RESOURCES: EUROPEAN ACTIONS AND THE CONTEXT OF IMPACT

A kulturális örökség digitalizálása: Európa tevékenysége és az IMPACT kontextusa

PAT MANSON – EURÓPAI BIZOTTSÁG (UNIT FOR CULTURAL HERITAGE & TECHNOLOGY ENHANCED LEARNING); i2010



Pat Manson

i2010

Az IMPACT az **i2010** keretrendszer részeként indult projektum. Az i2010 program eredete az EU Bizottság 2006-os ajánlására megy vissza (*Digital Libraries: digitisation, accessibility and preservation of cultural and scientific information*); melyben a következő célokat jelölik meg:

- mindenki számára elérhető digitális tartalom
- közös, többnyelvű felület
- három kulcsfolyamat összehangolása (digitalizálás, online elérhetőség, digitális megőrzés)
- Europeana mint működtető és szolgáltató platform

A tagállamok felelőssége:

- nemzeti szintű stratégia és államközi együttműködés javítása
- részvétel az európai digitális könyvtár létrehozásában

Az EU Bizottság felelőssége az együttműködés, jogharmonizáció előmozdítása, a feltételek megteremtésének ellenőrzése.

Az i2010 fő kihívásai

- Szerzői jogi kérdések
 - o Jelenleg rugalmas megoldásmodelleket próbálnak megfogalmazni, különös tekintettel a következő problémakörökre:
 - o 20. századi művek
 - o árva művek
 - o out-of-print művek
- Anyagiak
 - o redundancia és digitális duplikátumok szűrése
 - o prioritások tisztázása
 - o együttműködés a piaci szférával (tapasztalatok és szponzoráció)
- módszertan javítása kompetenciaközpontok létrehozásával a források gazdaságosabb felhasználása érdekében.

Az EB-nek 2008-ban kellett először elszámolnia az ET felé. A jelentésben kiemelt fő probléma az, hogy a tagországok többsége nem tud számszerű adatokat biztosítani az i2010 által célzott tevékenységekről, valamint nem követhető a célzott pénzügyi tervezés és a források hatékony felhasználása. Egyes országokban beszélhetünk nemzeti digitalizálási központról (i. e. közintézményi téren), máshol ez nem valósult meg.

A Bizottság mérési tanulmányt dolgozott ki, melyet 2009 májusában oszt meg a tagországokkal. Ezzel a digitalizálás illetve a digitalizált anyag mérhetőségéhez és a költségek felméréséhez biztosítanak egységes módszertant.

A Bizottság 2006-os ajánlása előírta a tagországoknak, hogy létesítsenek és tartsanak fenn nagyteljesítményű digitalizáló eszközparkokat, melyekből összeállhat az európai kompetenciaközpontok hálózata, *s melyek folyamatosan lehetővé teszik a tömeges digitalizálást.* Ennek támogatására indult három nagyléptékű projektum (Large Integrated Project), három fő közgyűjteményi tartalomtípus alapján:

- Audiovizuális archívumok: **PRESTOSPACE**; PRESTOPRIME
- Múzeumok, régészeti lelettárak: **3d-COFORM**
- Nyomtatott szöveg: **IMPACT**
- Teljes EU támogatás: **37 millió €**

IMPROVING ACCESS TO TEXT – CENTRE OF COMPETENCE IN MASS DIGITISATION OF HISTORICAL PRINTED TEXT

IMPACT – kompetenciaközpontok a régi szövegek tömeges digitalizálásában

HILDELIES BALK, PHD – KB, A EUROPEAN PROJECTS VEZETŐJE (KB R&D) IMPACT
PROJEKTKOORDINÁTOR



Hildelies Balk

Az IMPACT projekt lényege:

Technikai szint: jelenleg nem áll rendelkezésre megfelelő eszköz a közintézmények számára igazán jelentős, nem jelenkori nyomtatott szöveganyag feldolgozásához. A kereskedelmi szoftverek modern anyag konverziójához vannak optimalizálva, régebbi anyag esetében nem produkálnak elfogadható minőséget. A közintézményi szférát tömeges szinten azonban épp a régi (legalább 70 évvel ezelőtti) művek érintik. A jelenlegi szoftverek nem kezelik jól ezek nyelvét, szerkezetét, fizikai formátumát, nyomdai jellegzetességeit. Az így keletkezett anyag kézi javításra szorul, ami veszteségessé teszi a digitalizálási projekteket.

Stratégiai szint: általában hiányzik a szakértelem és a tapasztalat, és ezzel ad-hoc megoldások és helytelen gyakorlatok születnek.

Célok:

Az IMPACT tárgya elsősorban a modern kor megelőző („historical”) nyomtatott szöveg, a következő célkitűzésekkel:

- az OCR és nyelvi elemző technológia továbbfejlesztése
- a szakértelem megosztása és a feldolgozó-kapacitás növelése Európa-szerte
- a projekt lezárulta után is fennmaradó eszközök és szolgáltatások kifejlesztése

Az IMPACT-konzorcium

- 15 partner, a közintézményi és a privát szféra jelentősebb képviselői
 - o könyvtárak (8): Koninklijke Bibliotheek, The British Library, Bibliothèque nationale de France, Deutsche National Bibliothek, Bayerische Staatsbibliothek, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Österreichische Nationalbibliothek, Universitätsbibliothek Innsbruck
 - o egyetemek, kutatóintézetek (6): Instituut voor Nederlandse Lexicologie (INL); Εθνικό Κέντρο Έρευνας Φυσικών Επιστημών "ΔΗΜΟΚΡΙΤΟΣ"; University of Salford; Ludwig-Maximilians-Universität München(CIS csoport), Leopold-Franzens-Universität Innsbruck (InfMath csoport); University of Bath
 - o piaci partnerek (2): IBM (Haifa Research Lab); ABBYY (Moszkva)
- Alapítási időszak: 2006. december-2007. március
- 2008-ban már sikereket jelentett vissza az EB-nak (FP 7 ICT Work Programme 2007.4.1 Digital Libraries and technology-enhanced learning)
- A tényleges munka 2008 januárjában kezdődött.
- Koordinátor: **KB**
- Teljes költségvetés **15,5 M €**; ebből az **EU 11,5 M €**-t ad
- Futamidő 4 év, hosszabbítás lehetséges
- Kb. 100 közvetlen munkatárs

Az IMPACT felépítése

Az IMPACT 22 work package-re épül, melyek 4 alprojektbe tömörülnek

- Technológiai és kutatási alprojektek
 - o TR Text Recognition –Karakter-felismerés (OCR) a képi dokumentumból a szöveg kinyerésére irányuló eszközöket fejleszt. Vezeti: Leopold-Franzens-Universität Innsbruck (könyvtár)
 - o EE Enhancement and Enrichment (language technology research) – Fejlesztés és minőségjavítás (nyelvészeti technológia) : a teljes szövegű digitalizálás kimenetének pontosságáért felelős eszközök Vezeti: Österreichische Nationalbibliothek
- Stratégiai alprojektek:
 - o OC Operational Context – Munkakörnyezet: a tömeges digitalizálás intézményi lebonyolítását támogatja eszközökkel. Vezeti: The British Library
 - o CB Capacity Building – Teljesítmény optimalizálás Az elért eredmények közintézményi világban való elterjedését hivatott biztosítani. Vezeti: University of Bath

Fejlesztett eszközök

1. OCR:

- Karakterfelismerés:
 - o Adaptív OCR-motor: elsősorban könyvtári igényekhez igazított szoftver-rendszer, mely menet közben alkalmazkodik a feldolgozott szöveghez. Alrendszerei
 - o Kép-optimalizáló eszköz
 - o Szerkezetelemző csomag
 - o Utókorrekciós eszközök
- Kísérleti és egyéb eszközök
 - o gépirat OCR-motor
 - o szófelismerő motor
 - o szótárgeneráló prototípus

2. Eszközök és tartalom optimalizálása

- Kollaboratív javítás
 - o Web-alapú javítószoftver, mely közvetlenül az OCR-motorral van összekötve
- Tartalom
 - o Köz- és tulajdonnévi szótárak (holland, német, angol, valamint további nyelv hozzáadását támogatja)
 - o Nyelv és írásrendszerek fejlődéséből eredő nehézség kiküszöbölését segítő eszközök
 - o Kollaboratív, web-alapú alkalmazás a szótári entitások közös kezeléséhez
- Strukturális metaadatok
 - o Functional Extension Parser: webalapú szolgáltatások, melyek automatikusan felismerik és leírják a strukturális elemeket

3. Stratégiai eszközök és szolgáltatások

- Projekt honlap: az európai digitalizálási kompetenciaközpontok hálózatának kiemelkedő pontja
- Döntés-előkészítési segédeszközök: a tömeges digitalizálási projektek tervezéséhez, lefolytatásához és pénzügyi felügyeletéhez használhatók
- Oktatási forráscsomag: a tömeges digitalizálással kapcsolatos irányelveket, és az egyéb eszközök használatának útmutatóit tartalmazza
- Képzés és támogatás:
 - o Help-desk rendszer, mely végfelhasználói kérdéseket közvetít a kompetenciaközpontok felé
 - o Tömeges digitalizálás kérdéseiben és technológiájával kapcsolatos oktatócsomag és dokumentáció

Ütemezés:

- Bevezető fázis: 2008-2009: IMPACT core consortium of 15 partners
 - o 15 partner
 - o kapcsolatok kiépítése a létező kompetenciaközpontokkal
 - o feladatok kiosztása
- 2. fázis 2009-2010:
 - o 3-6 demonstrátor hozzáadása
 - o 3 további nyelv hozzáadása az elkészült nyelvészeti eszközökkel (jelenlegi nyelvek: holland, angol, német)
 - o a nyelvfüggetlen eszközök tesztelése
 - o lehetséges új tagok kiválasztása és felvétele (utóbbi 2010 januártól) (50% önfinanszírozás)
- 3. fázis 2010-2011: a kompetenciaközpont létrejön és fennmarad
 - o Az IMPACT-közösség kibővítése
 - o Üzleti modell kialakítása
 - o Források megszerzése az esetleges folytatáshoz

A prezentációt követő **kérdések** között elhangzott: „A piaci forgalmazók hogyan szerezhettek licenszeket a fejlesztett eszközök kereskedelmi disztribúciójára?”

Válasz: Minden eszköz más licenz alatt jön majd ki, de ez még munka alatt van. Lesznek ingyenes eszközök is, a többi pedig elérhető lesz a piaci forgalmazók számára.

LIBRARY CHALLENGES FOR MASS DIGITISATION

A tömeges digitalizálás könyvtári kihívásai

ASTRID VERHEUSEN – KONINKLIJKE BIBLIOTHEEK, DIGITISATION DEPARTMENT (R & D)



Astrid Verheusen

A prezentáció első részében Astrid Verheusen ismertette a KB digitalizálási tapasztalatait tevékenységét a 1994-es kezdetektől máig. Néhány adat a jelenlegi, valamint tervezett munkáról:

Digitális könyvtári program: 2013-ig a teljes gyűjtemény 10%-át biztosítják digitálisan

Projektum	Oldalszám	Keret
A holland parlament iratai 1814-1995	2.300.000	M€ 10.5
Holland napilapok 1618-1995	8.000.000	M€ 12.5
Különgyűjtemények – 1800 előtt megjelent könyvek	1.300.000	M€ 3.0
Magazinok	1.500.000	M€ 1.8
Könyvek 1850-1950	12.000.000	M€ 1.0
Rádióhírek (gépelt, nyomtatott)	1.500.000	M€ 0.5
Metamorfoze - preservation imaging	4.500.000	M€ 18.0
Memory of the Netherlands	350.000	M€ 3.5
Összesen:	31.450.000	M€ 50.8

Átlagos költség: 0,3 – 1.5 € oldalanként

A könyvtári anyag digitalizálását a könyvtáraknak kell elvégezni, legalábbis management szintjén. Egyéb ágensek nem tartják szem előtt a minimális minőségi és tartalmi szempontokat. Példaként intézményi szórvány-digitalizálási ill. a Google Booksból hoz példát: mindkét esetben egy régi holland periodikum került ezekbe a digitális gyűjteményekbe – elfogadhatatlan minőségben, nem megfelelő leíró adatokkal. (*Vaderlandsche Letteroefeningen*). Az érintett források, ahol a lap digitális változata feltűnt: HATHI Trust; egy belga egyetemi gyűjtemény; Google Books; Rhino Research Center)

Érvek (a könyvtári digitalizálás maradjon a könyvtárak kezében). Egy piaci kezdeményezésű digitalizálás nem lesz tekintettel az alábbiakra:

- Minőség
- Teljesség (kollacionálás)
- Digitális megőrzés
- Hosszú távú elérhetőség
- Ingyenes hozzáférés
- Copyright
- Speciális gyűjtemények, sajátos dokumentum-típusok

Módszertan

- Szelekció, előkészítés
- Digitális duplikátumok szűrése – ez megoldható digitális masterek központi nyilvántartásával.
- Copyright figyelembe vétele
- Teljes gyűjtemények digitalizálása célszerű (ez szintén csak intézményi környezetben és szakértelemmel kivitelezhető)

Megvalósítás kérdései

- **Megőrzés: JPEG2000** bevált tárolóformátumként. A tapasztalatok fényében (pár millió oldal után mindenki be fogja látni, hogy a nyersanyag teljes megőrzése nem kivitelezhető). Bizonyos esetekben a *.txt lesz a master
- **Minőségellenőrzés:** erre nincs lehetőség a tömeges digitalizálási munkáknál. Olyan munkafolyamatot szabad csak nagyléptékben elkezdni, ami garantálja a minőséget, az utólagos ellenőrzés nem kivitelezhető. (Ld. erről a szerző 2008-ban a párizsi TELPlus Workshopon elhangzott előadását – tavalyi beszámolómban olvasható)
- **OCR:** speciálisan könyvtári eszköz kell, mely tekintettel van a nyomdai, történeti, műfaji sajátosságokra és nyelvi változatokra.
- **Költségek:**
 - o 30% munkaerő (in-house)
 - o 10% WARE
 - o 10% R&D
 - o 50% digitalizálás, OCR, metaadat (kiszervezve)
- **Keretek:**
 - o 80 ember dolgozik az érintett részlegekben
 - o projekt alapon dolgoznak, a tömeges digitális állomány a projektekből áll össze

1. nap (2009. április 6.), délutáni szekció

LEVEZETI: GÜNTHER MÜHLBERGER (UIBK)

ADAPTIVE OCR – COLLABORATIVE CORRECTION – COST EFFECTIVE DIGITIZATION

Adaptív OCR – Kollaboratív javítás – Költséghatékonyság digitalizálás

ASAF TZADOK – IBM HAIFA RESEARCH LAB



Asaf Tzadok

A prezentáció bevezetésekként ez az előadó is ismertette az i2010 kontextust és az IMPACT hátterét. Ezután tért rá az előadás fő témájára: ez az IMPACT kettes eszközcsoportjába tartozó „kollaboratív” webes alapú OCR-javító eszköz fejlesztése, melynek fő felelőse az IBM (Haifa Research Lab).

A jelenleg elérhető piaci eszközök a digitalizálandó tartalomnak csak nagyon kis, modernkori, vegyes fontbázisú szegmensének feldolgozására alkalmas

A munka célja:

- **Részfeladat:** automatikus OCR-eszköz fejlesztése, mely
- kollaboratív ellenőrző/javítókomponenst alkalmaz
- nyelvtörténetileg támogatott lexikonokkal támogatja a minőségjavító folyamatot
- **Teljes feladat:** adaptív OCR-eszköz, mely kifejezetten a közintézmények igényeit szolgálja ki, és számos komponensből áll (képporrektálás, szegmentálás, utójavító eszközök stb.)

Kollaboratív javító eszköz:

Webalapú eszköz, mely tömeges önkéntes közreműködést (ellenőrzés és javítás) tesz lehetővé, és amely közvetlenül az OCR motorhoz kapcsolódik, s az így előálló visszacsatolás biztosítja az OCR tanulékonyosságát

Tanulás karakter ill. szó alapon

- Karakteralapon: adott készlet elemeit tanulja meg, előzőleg magkapott készlet-referencia alapján
- Szóalapon: szótöredék minták alapján
- Nagy terjedelmű, formailag homogén anyagon lesz legsikeresebb az alkalmazkodás

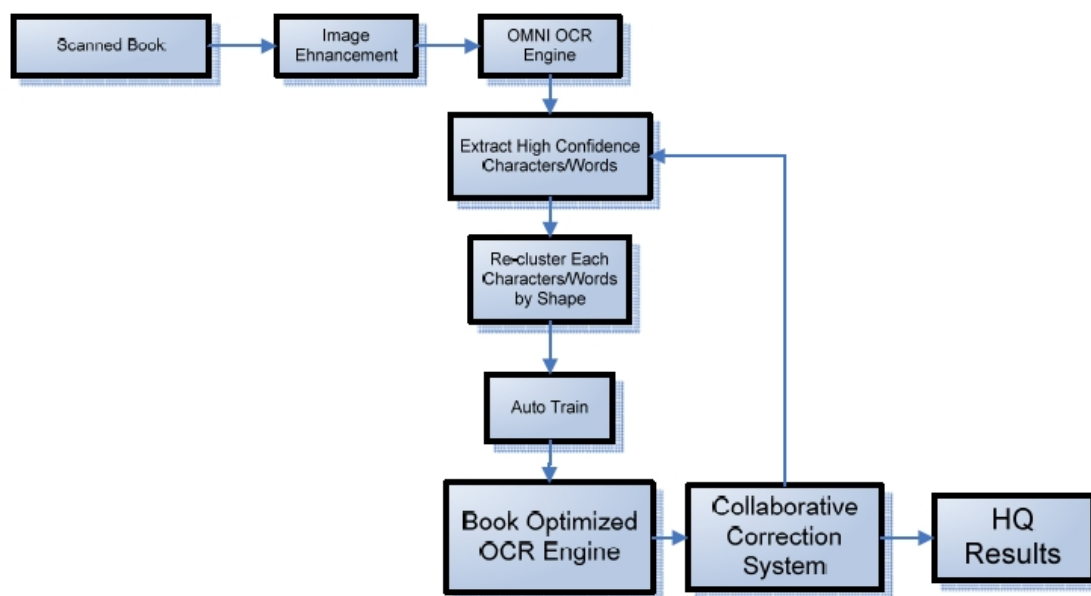
Működő minta a pl. a Gutenberg projektnél figyelhető meg, a megvalósítás azonban sok kívánnivalót hagy maga után, az eszköz fejletlensége miatt nehézkes a tanulás, valamint a javítás sem elég hatékony, ha két szövegdobozt kell párhuzamosan figyelni

„Intuitív” vs. referencia-alapú adaptivitás

A jelenleg elérhető szoftverek felismerési pontossága néha az 50%-ot is alig haladja meg nem modern anyagnál. Ennek oka, hogy elsősorban jelenkori eszközökkel előállított, formailag vegyes anyag feldolgozására születtek. Ez másfajta adaptivitást feltételez, mint ami közintézményi digitalizálás számára optimális, ahol inkább nagy mennyiségű, homogén anyag felismertetése a cél. Az adaptivitás itt azt jelenti, hogy az *ideális eszköz a lehető legjobban összehangolódik a feldolgozott dokumentumegység(ek) egyedi tulajdonságaival* (pl. betűkép, tördelés, kiemelések stb.)

Erre kézi ellenőrző/javító modul épül, melynek inputját az OCR-motor beépíti a felismerési algoritmusba. (Későbbiekben ezt a kézi beavatkozást is felválthatja automatizmus, de ahhoz még sokat kell fejlődnie.)

Az alkalmazkodás mértéke szembetűnő lesz a javítás folyamatában (észrevehetően kevesebb hibát talál a korrektor, ahogy előre halad a dokumentumban).



1. A teljes rendszer modellje

1. ADAPTÍV OCR - Az algoritmus felépítése (szó alapon)

Minden egyes új szó esetén

- Több jelölts csoportot határoz meg (közelítő csoportok képzése)
- Általános képanalízist végez
- Fuzzy illesztést végez a hagyományos (pre-adaptív) OCR eredményei és a kigyűjtött közelítő csoportok között
- Kiszűri a valószínű olvasatokat

A cél az, hogy ezek között már ott legyen a megfelelő olvasat (azt a korrekciónál mindenképp felajánlja)



1. Elkülönített csoportok, ill. egy csoport belülről

Az egyformaként azonosított szavakon

- részletes képanalízist végez (Detailed Comparator), így dönti el, hogy tényleg egyeznek-e
- Morfológia szűrő segítségével elkülöníti az esetleges téves illesztéseket

Ezen a ponton feltehetőleg nem lesznek tényleges hibák (tehát 2 különböző szó azonosként felismerve)

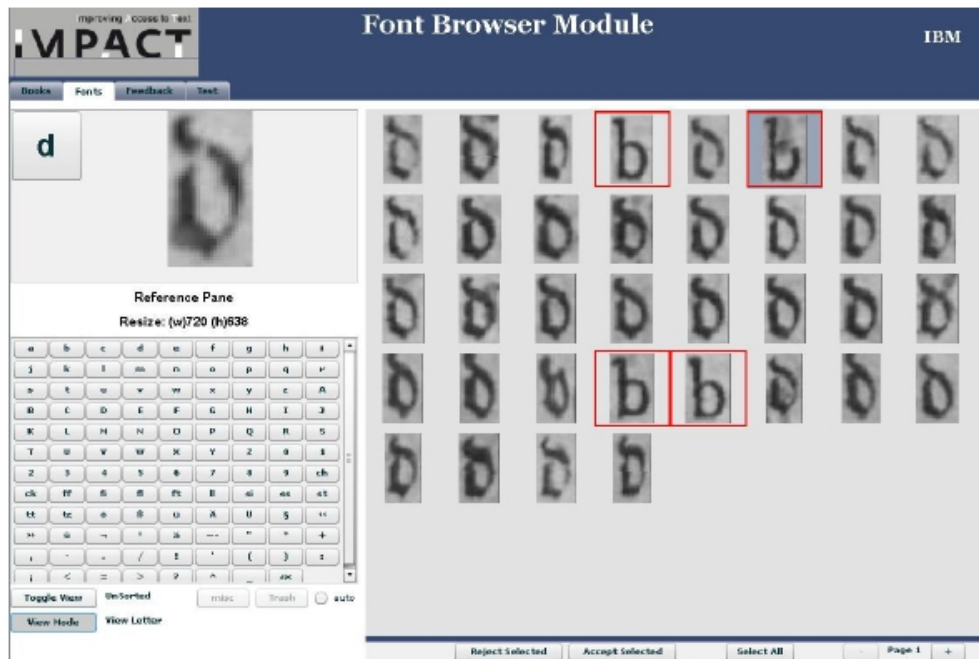
2. KOLLABORATÍV KORREKTÚRA

Web-alapú rendszer, mely önkéntesen nagy tömegét mozgósítja a jelenleg létező megoldásokhoz képest továbbfejlesztett környezetben (Pl.: Wikipedia, Gutenberg Project)

Az eddig látott szoftverek általában megjelenítenek egy szekvenciát, és mellette az annak megfelelő felismert változatot, szerkeszthető módon. Ez nem megbízható és fárasztó megoldás, ráadásul egy javító csak minimális mennyiségi előrelépést érhet el. Az IMPACT eszköze különböző szinteken különböző megközelítéssel végzett munkát tesz lehetővé, és különböző elérési módokon látja a javítandó tartalmat:

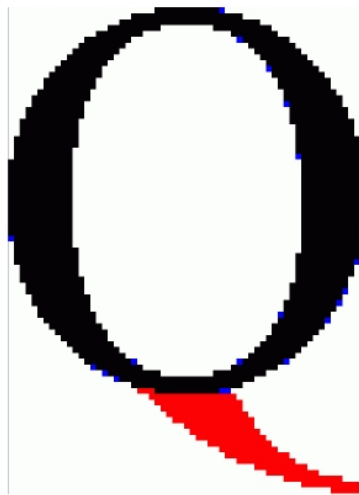
karakter/szimbólum azonosítás(„carpets”) ► karakterbevitel ► szavak javítása (kontextusban) ► oldal javítása

1. szint: Carpets – szimbólumazonosítás



1. Karakter-, szimbólumazonosítás

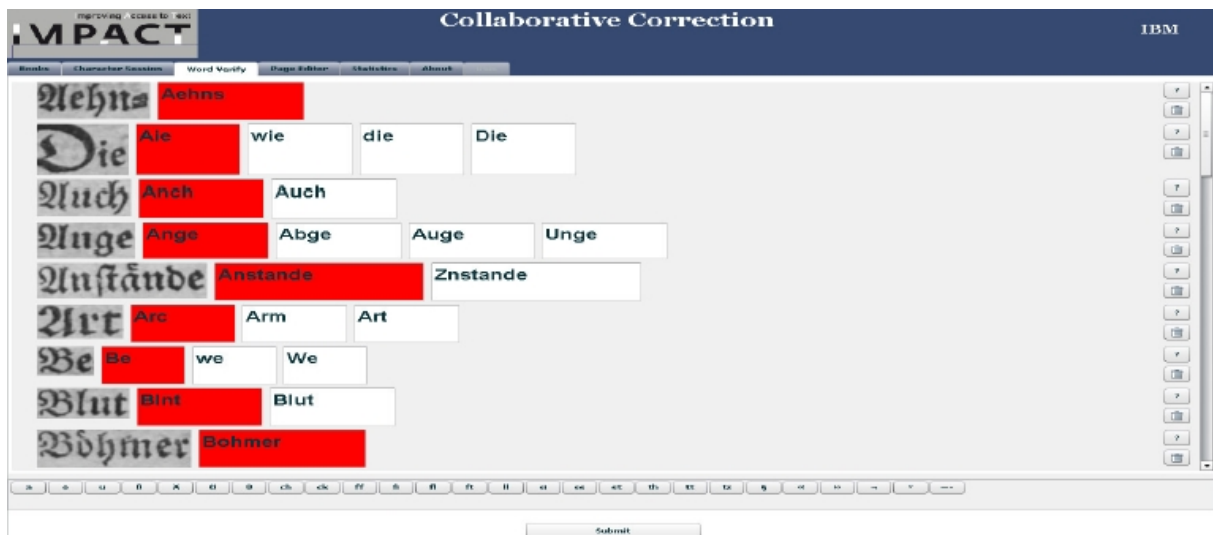
Alapja az absztrakt „**super-key**” referencia minden azonosítandó jelre. Ebben a munkafolyamatban a korrigáló az automatikus folyamatban előválogatott egyedi előfordulások megítélésével alakítja ki a releváns mintát. A minták tulajdonság-hierarchia alapján tömörülnek.



1. Super Key – a kék terület a lényegtelen, a piros a releváns sajátosságokat jelöli

2. szint: Szó-alapú korrekció

Morfológiai, nem lexikai alapon azonosít. A kétséges szavak estén a javító lehetséges szavak listáját kapja (ez az első, automatikus fázisban jön létre), melyekből választhat, csak szükség esetén kell begépelnie a helyes alakot.



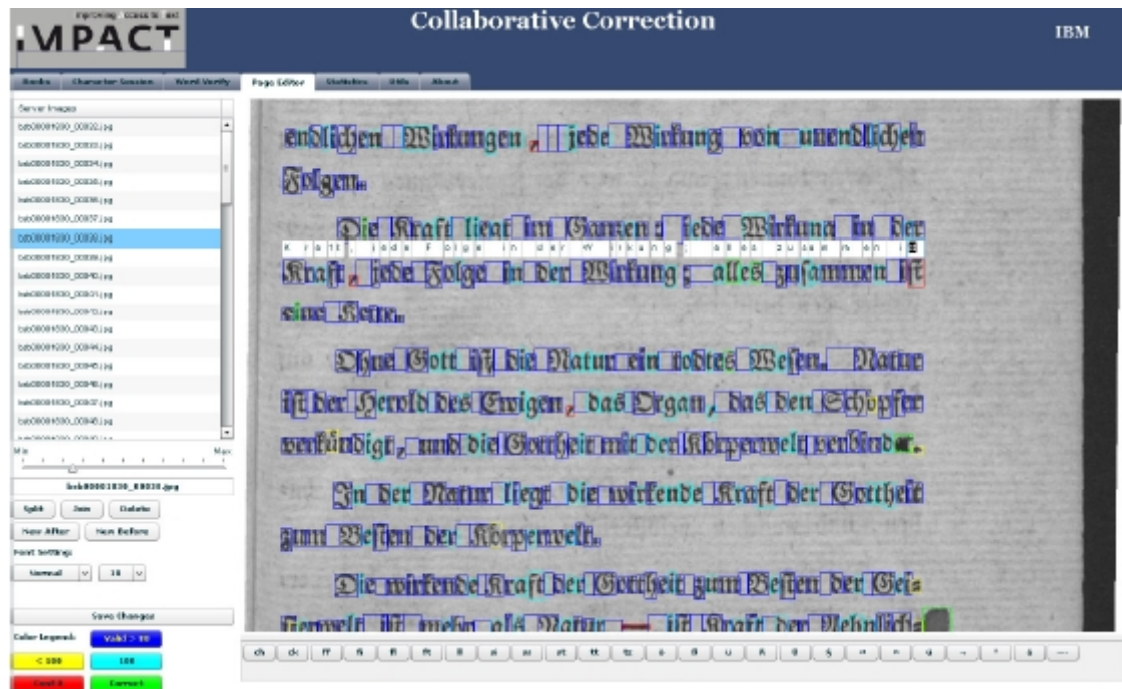
Szó-alapú javítás

3. szint: Oldalszintű javítás

A szegmentálási problémák kiküszöbölésére, valamint a tartalom teljességének ellenőrzésére szolgál (pl. a hiányos szegmenseket így eldobhatók.) Page OCR Editor/Verifier

Az egyéb szintű azonosítás előkészítéséhez fontos műveletek

- jelek összevonása/szétbontása
- felismerési polygonok korrigálása
- hiányok kipótolása



Javítás oldalszinten

KITEKINTÉS: a fejlesztés kilátásai

- Super Key : hasonló szimbólumok egyesítése automatikusan
- Valószínűségi alapú keresés OVR által
- Automatikus eBook előállítás
- Automatikusan vektorizált könyvfüggő fontok
- Kontextuális azonosítás
- Oldaltorzulás kiküszöbölése

A USER FRIENDLY PLATFORM FOR DOCUMENT IMAGE ENHANCEMENT AND SEGMENTATION

Felhasználóbarát platform a képkorrekcióhoz és szegmentációhoz

BASILIS GATOS – COMPUTATIONAL INTELLIGENCE LABORATORY, INSTITUTE OF INFORMATICS AND TELECOMMUNICATIONS ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΕΡΕΥΝΑΣ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ "ΔΗΜΟΚΡΙΤΟΣ";

Az athéni székhelyű kutatóintézet (*National Center for Scientific Research "Demokritos"* – Institute of Informatics and Telecommunications – Computational Intelligence Laboratory, <http://iit.demokritos.gr>) fő tevékenységi/kutatási területe:

- mintaosztályozás és -felismerés
- multimédia feldolgozás
- dokumentumszerkezet-analízis és OCR
- régi dokumentumok feldolgozása és felismertetése

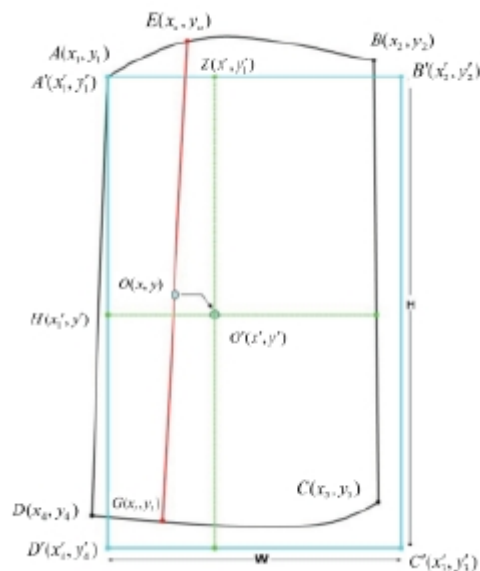
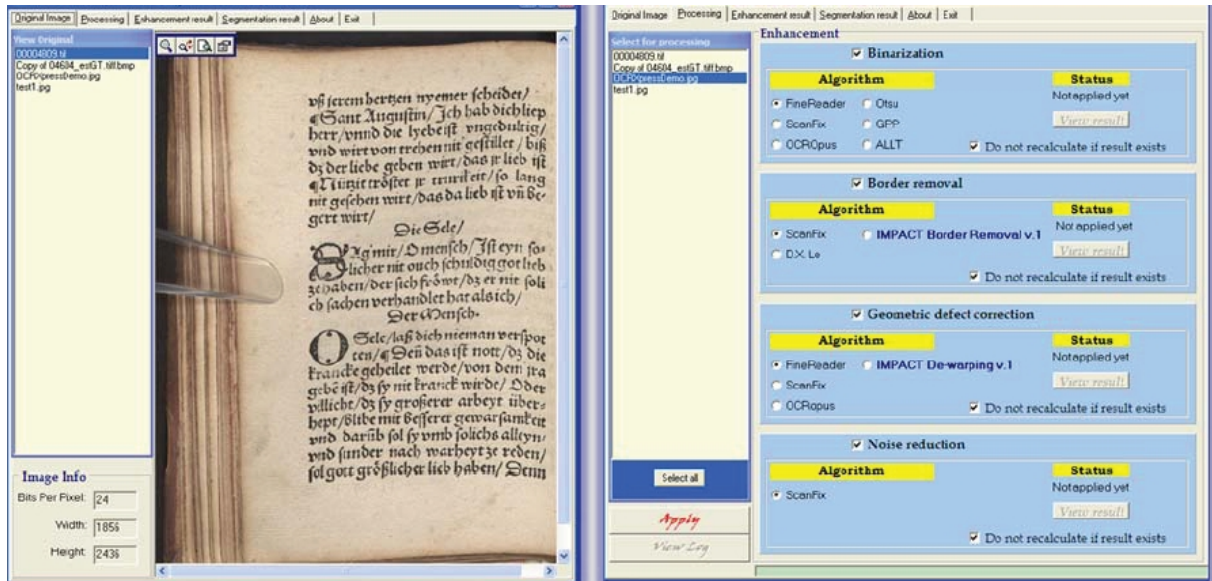
Az intézet a következő projektek számára végez jelenleg fejlesztéseket:

- **POLYTIMO** – A System for Processing, Managing, and Providing Access to the Content of Valuable Historical Books and Handwritten Manuscripts (<http://iit.demokritos.gr/cil/Polytimo/>)
- **D-SCRIBE** – A System for Digitization and Processing of Greek Manuscripts (<http://iit.demokritos.gr/cil/dscribe/>)
- **KT-DigiCult-BG** project – Knowledge Transfer for Digitisation of Cultural and Scientific in Bulgaria
- **BOEMIE** – Bootstrapping Ontology Evolution with Multimedia Information Extraction (<http://www.boemie.org>)
- **IMPACT**

IMPACT fejlesztések

- TR Text Recognition (Szövegfelismerés) alprojekt:
- TR1 képtimalizálás
- TR2 szegmentálás
- TR4 kísérleti OCR-motor

Ezen kívül még részt vesznek az Operational Context alprojekt OC3 (Evaluation & resources) és OC5 (Technical framework and integration) munkacsoportjaiban



TR1-2: IMPACT Enhancement & Segmentation Platform (Optimalizáló és szegmentáló platform)

Az eszközt a következő jelenleg elérhető technikák kiértékelésével állították össze:

- WiseBook; BookRest; ScanFix; FineReader; ALLT; GPP; SAUV; OTSU

Jelenleg elérhető funkciók

- képtimalizáló eszköztárak:
- dewarping (egyenesítés)
- border removal (lapszél eltávolítás)
- szegmentálás:
- homogén blokkok elkülönítése (kép, szöveg stb.)
- nyomtatott sorok ► szavak ► karakterek izolálása

Az egyes műveletek batch-módban is futtathatók.



Balra: Klaus U. Schulz

Közreműködők: Annette Gotscharek, Uli Reffle, Christoph Ringlstetter(CIS MUNICH); Katrien Depuydt, Jesse de Does (INL Leiden)

Az előadó a szótár-alapú OCR megközelítés fejlesztési lehetőségeit vázolta fel. A jelenlegi technikák passzív (token-alapú) előfordulás/hiány alapon társítják a felismert elemeket a rendelkezésre álló szótári egységekkel. Ez nyelvészeti megközelítéssel nagyban továbbfejleszthető, további szabályok bevezetésével, úgynevezett tudásalapú megközelítés alkalmazásával – azaz nyelvészeti és társtudományi ismeretekből kiszűrhető szabályok alkalmazásával.

Mindennek alapja a dokumentumról, annak nyelvéről, illetve az OCR-folyamatról birtokunkban lévő tudás rendszerezése, és a sematizált ismeretek beépítése a szótárhasználatba.

Nyelv viselkedési szabályainak figyelembevétele:

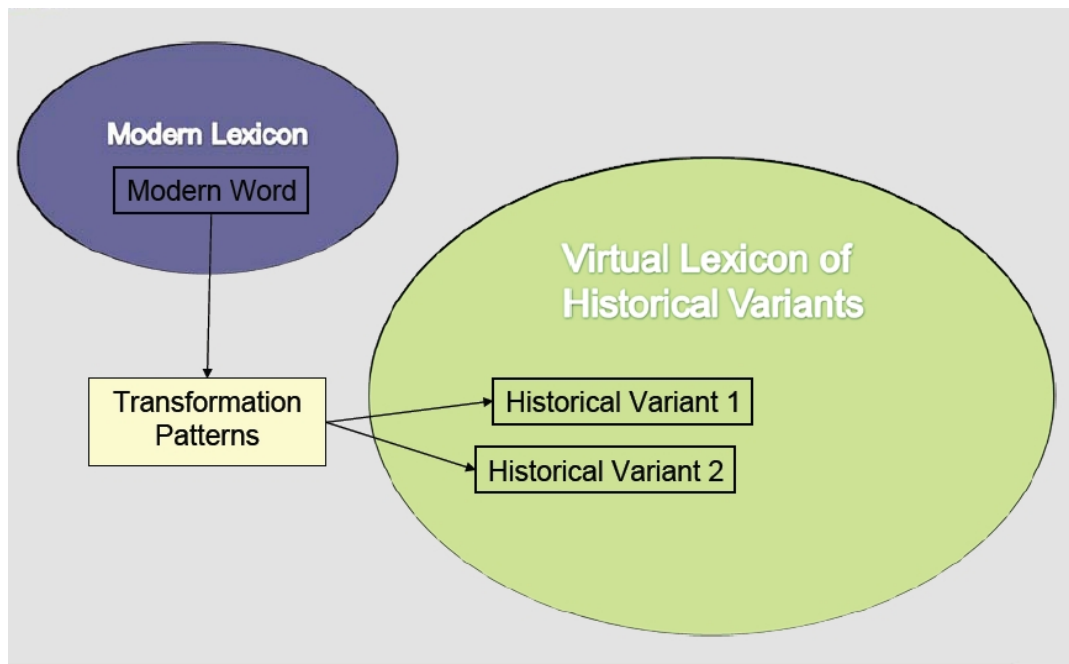
- kontextus
- szógyakoriság
- nyelvtörténeti alakok
- a dokumentumról való ismeretek figyelembevétele
- tipikus/valószínű szavak
- speciális nyelvváltozatok
- ismert történelmi szóalakok

Az OCR-folyamat sajátosságainak figyelembevétele

- várható pontosság
- tipikus hibák

Szótár-alapú optimalizálás:

A tudott információ alapján a megfelelő szótárbázis összeállítása, virtuális szótár építése az alkalmazható létező korpuszok, nyelvi transzformációs szabályok segítségével („virtuális szótár” felállítása ABBYY FR 8.0 SDK segítségével).



Transzformációs szabályokat támogató ismeretek:

- dokumentum nyelve
- korszecifikus nyelvváltozat
- korszecifikus változatok
- nyelv belső transzformációs szabályai
- szógyakoriság
- kontextus

Mindezek kinyerhetők a következő forrásokból:

- nyelvi analízis
- OCR-profil készítése
- dokumentumprofil készítése

A fentiek nem tényjellegű, hanem relatív ismeretek, és túllépnek a legtöbb részleten, de figyelembe vételük így is drasztikus OCR-javulást eredményez.

MEASURING THE OCR ACCURACY ACROSS THE BRITISH LIBRARY'S 2 MILLION PAGE NEWSPAPER ARCHIVE

Az OCR pontosságának felmérése a British Library 2 millió oldalas újságarchívumában

SIMON TANNER – KING'S DIGITAL CONSULTANCY SERVICES



Simon Tanner

King's Digital Consultancy Services referenciák:

- Skót nemzeti könyvtár digitalizálási stratégia
- Brit Nemzeti archívum – digitalizálási tanácsadás
- Tate Archive – digitalizálási üzleti terv
- Oxfordi Egyetem Digitális Könyvtár – szöveg-felismerési tanulmány
- Mellon Foundation projektek (USA)
- Felmérések 100 művészeti múzeumban
- Ír nemzeti könyvtár: digitális könyvtári stratégia
- Holt-tengeri tekercsek digitalizálása
- Digital Futures Academy

British Library 19. századi újságok digitalizálásának minőségellenőrzése (jelen prezentáció alapja)

- Üzleti modell: Digital Divide Data
- (Simon Tanner a *Digital Futures* c. könyv szerzője)
- British Library: 2 millió oldal 19. századi újság OCR kiértékelése
- Az OCR eredmények pontosságát mérték, nem magát a folyamatot, nem a beszállító eszközeit és módszereit, csakis az eredményt.
- Az OCR-rel előállított XML formátumú anyagot vetették össze a bemenet pontos szövegével, amit a mintaoldalak újragépelésével állítottak elő.

Példák:

A karakter alapú egyezés általános aránya félrevezető, nem árul el igazán sokat az OCR pontosságáról. Ha pl. egy hibás karakter van minden ötödik szóban, az nem utal ijesztően rossz arányra a karakterek arányában mérve, de valójában azt jelenti, hogy jó eséllyel minden ötödik szó hibás, ami viszont nem elfogadható.

Másik példa (Tipikusan angol nyelvű szcenárió):

- 1000 szó; 5000 karakter (átlagosan 5 karakter/szó. leszámítva a szóközöket)
- Ennél a 90%-os karakterpontosság = 4500 pontosan felismert karakter
- Ez jelenthet:
 - o Maximum 900 pontos szót (90% szavak szintjén)
 - o Minimum 500 pontos szót (50% szavak szintjén)
- Az igazság a kettő között van.

(**N.B.** ugyanez az arány magyar nyelven sokkal rosszabb eredményt jelentene)

Tényezők, melyek pontosabban meghatározzák az OCR sikerességét

- Keresések és találatok megjelenítésének pontossága
- Találatok relevanciájának tényleges tükrözése
- A megfelelő minőség eléréséhez szükséges utókorrekció mennyisége

Tényleges értékelőmunka szempontjai (az ismertett BL projektben)

- karakterek pontossága
- szóhelyesség
- fogalomszavak pontossága
- tulajdonnevek pontossága
- számsorok pontossága

Mindenhol a *legmagasabb* kimutatható pontossági értéket igyekeztek meghatározni, nem a legalacsonyabbat.

Mintavétel az ellenőrzéshez:

- az összes oldal ~ 1%-a
- a mintából oldalanként 2 blokk újragépelése (ezt végezték a Digital Divide Data keretében)
- az újragépelte blokkok OCR megfelelőinek kigyűjtése a BL XML-kimenetéből
- tiltólista segítségével a mellőzendő szavak kigyűjtése (a fogalomszavakat vizsgálták)
- A két szöveg összevetéséhez használt eszközök:
 - o **diff** (<http://en.wikipedia.org/wiki/Diff>)
 - o **Levenshtein távolsági algoritmus**
(http://en.wikipedia.org/wiki/Levenshtein_distance)
 - o **Excel** táblázat használata a végeredmény elemzésére
 - (> 40,000 összehasonlított sor; > 4,000,000 szó; > 25,500,000 karakter)

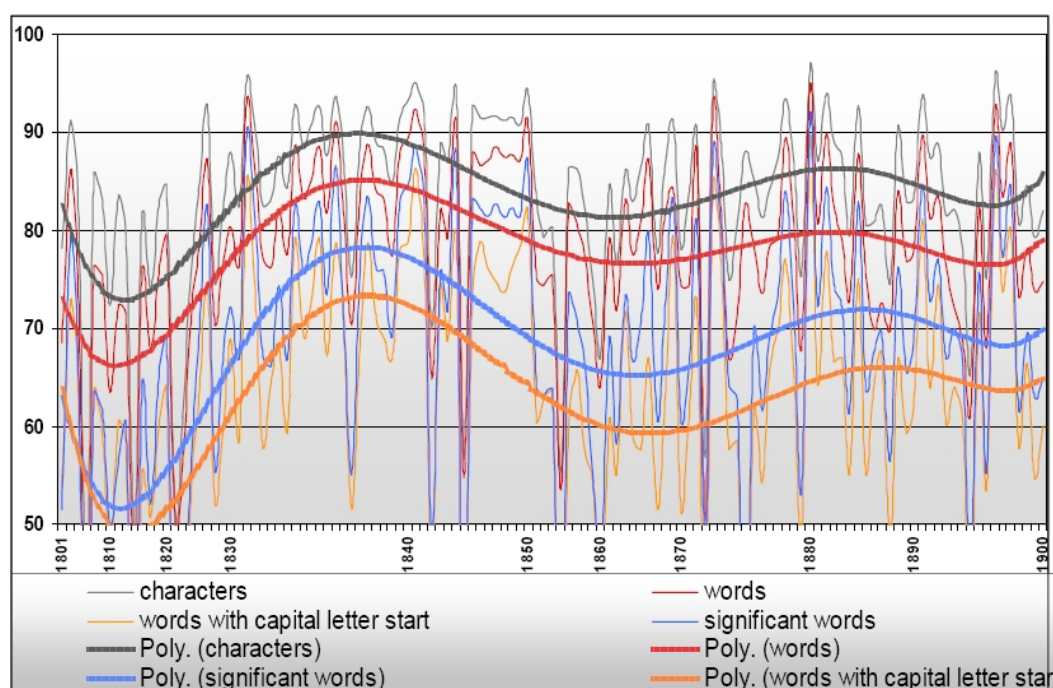
Összesített eredmény (BL) projekt

- Helyes karakter = 83.6%
- Helyes szó = 78%
- Helyes fogalomszó = 68.4%
- Helyes tulajdonnév = 63.4%
- Helyes számsor = 64.1%

N.B. Az egyes címeket külön-külön is mérték.

Az eredmények idő szerinti bontásban

(szürke = karakterek; piros = szavak; :kék = fogalomszavak; narancs = tulajdonnevek; Poly = összesített eredmény)



Összefoglalás

Címenkénti mérés eredményei alapján:

- 80% helyes *karakter*: a lapok **2/3**-ánál
- 80% helyes *szó* csak a lapok **felénél**
- 80% helyes *fogalomszó*: csak a lapok **negyedénél**

80% szóhelyesség elfogadható, de csak *fuzzy* keresési algoritmusok alkalmazásával biztosítható a megfelelő kereshetőség, ami >95-98% *keresési* pontosság (nem összekeverendő az OCR pontosságával)

(80% alatt az állomány nem minősül kereshetőnek)

Általános értékelés:

- A BL vizsgált anyaga a felhasználói igényekhez mérten *nem kielégítő*
- Ennek oka: nem állt rendelkezésre megfelelő eszköz az ilyen jellegű anyag feldolgozására.

Tervezés:

Az előadó ismertetett OCR-értékelési metódus használatát ajánlja **projekt-tervezéshez** is. Ezt a mintavételt a tömeges digitalizálási munka megkezdése előtt is el lehet végezni, és így megjósolható annak sikeressége és minősége.

Vitapanel

OCR kihívások

Simon Tanner bemutatója után panelbeszélgetésre került sor „OCR-kihívások” címmel

MODERÁTOR: GÜNTHER MÜHLBERGER (UIBK)



Günther Mühlberger

Résztevők:

- Asaf Tzadok (Haifa IBM Labs)
- Simon Tanner (King's College, London)
- Claus Gravenhorst (CCS – Content Conversion Specialist GmbH)
- Claus Schulz (CIS – Münchener Egyetem)
- Basilis Gatos (Demokritos)
- Jupp Stöpetie (ABBY)

A beszélgetés résztvevői a hallgatóság által felvetett kérdéseket vitatták meg.

Többek között az alábbiak kerültek szóba:

- **Keleti nyelvek és írásrendszerek:** Nem megnyugtató a keleti nyelvek OCR-támogatottsága. Az ABBY jelenleg a távol- illetve közel-keleti írásrendszerek és nyelvek támogatásán dolgozik, de a jelenlévő képviselő állítása szerint jelenleg még „nincs a piacon használható szoftver”
- **Legacy:** Az egyik legérdekesebb kérdés azt fogalmazta meg, hogy mi a teendő a felgyűlt legacy anyaggal. Van-e értelme ezeket fejlettebb eszközökkel újra feldolgozni, vagy újra kell-e kezdeni a folyamatot, optimalizált bemenettel? A szakértők szerint a helyzet az, hogy amit ezek a tevékenységek termelnek, az elkerülhetetlenül legacy anyag, csak az a kérdés, hogy mennyi idő múlva hívjuk annak. Ezért azt javasolják, hogy csak akkor álljunk neki tömeges digitalizálásnak és OCR-nek, ha biztos, hogy naprakész eszközökkel, a lehető legjobb minőségben tudunk dolgozni. Ha ezt nem látjuk biztosítottak, inkább várjunk meg a

feltételek javulását – rossz minőségben, kapkodva nincs értelme a munkának. Mindemellett fontos, hogy a master a lehető legjobb minőségű, és részletgazdag legyen. (Ehhez hasznos olvasnivaló: *Deciding whether Optical Character Recognition is feasible* http://www.odl.ox.ac.uk/papers/OCRFfeasibility_final.pdf)

- **Greyscale:** A CCS a hágai anyagnál szürkeárnyaltos JPG masterekkel dolgozott. Az került szóba, hogy ez mennyire követhető. A szakértők szerint egyáltalán nem, a greyscale csak kivételes esetben, nagy biztonságú anyagnál alkalmazható. A JPG szintén nem ideális, legalábbis JPEG 2000-ban kell dolgozni, de inkább ajánlják a TIFF-et.



2. nap 2009. április 7.

Délelőtti szekció

LEVEZETTE: ALY CONTEH (BRITISH LIBRARY)



Aly Conteh

“MANY HANDS MAKE LIGHT WORK” –PUBLIC COLLABORATIVE OCR TEXT CORRECTION IN AUSTRALIAN HISTORIC NEWSPAPERS

Nyilvános, kollaboratív OCR-javítás régi ausztrál újságokban

ROSE HOLLEY – AUSTRALIAN NEWSPAPERS DIGITISATION PROGRAM

Rose Holley prezentációja volt a konferencia egyik legérdekesebb és legterjedelmesebb bemutatója. Az ausztráliai gyakorlat bemutatása a kollaboratív korrekció alkalmazásának egy modelljével kapcsolódott a konferencia tematikájához.

Az ismertetett munkafolyamat egyéb részletei azonban sokkal jelentősebbek a web2.0-ás fejlesztések filozófiájára nézve, így erre tekintettel az előadást bővebben nem ismertetem, javasolom mindenki olvassa el, mert talán a legjobb létező könyvtári web2.0-ás és PR-modell. Teljes terjedelemben elérhető:

http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf

(Igény szerint külön összefoglalom, de nem ebben a beszámolóban.)

FUTURE CHALLENGES FOR OCR TECHNOLOGY

Az OCR technológia várható kihívásai

CLAUS GRAVENHORST (DIRECTOR STRATEGIC INITIATIVES, CONTENT CONVERSION SPECIALISTS GMBH)



1. Jobbra: Claus Gravenhorst

A **Content Conversion Specialists** (továbbiakban CCS) 1976-ban alapított német székhelyű cég, az intézményes OCR egyik úttörője az európai piacon. Az IMPACT-hoz több szalon kötődnek, például mint a KB beszállítója az újságdigitalizálási projekt kapcsán.

CR-Történet

Az előadás kezdetén a CCS történetével párhuzamosan végigkíséri az karakterfelismerő technológia evolúcióját

- **ICR**-technológia (vektoralapú karakterfelismerés)
 - o 1976 Raymond Kurzweil olvasógépe (Kurzweil Reading Machine, KRM), vakok számára készült olvasó-felolvasógép, hangzó kimenettel
 - o 1978 Kurzweil Data Entry Machine (KDEM), nyomtatott szöveg ipari újrafeldolgozására
 - o Instituut voor Nederlandse Lexicologie (szótárak)
 - o Kluwer Rechtswetenschappen (jogi adatbázis)
- **OCR** (optikai karakterfelismerés)
 - o Omnifont karakter/szó/szótár-alapú tanítható szoftver (csak a berendezéssel együtt)
 - o először a szkennerek váltak ki önálló gépként
 - o OCR koprocesszorok jelentek meg
 - o végül megjelent az önálló OCR-szoftver, mely SDK-környezet segítségével szolgálja ki a felhasználókat (ld. ABBYY)

Jelenlegi OCR-technológia

- kb. 170 nyelvhez létezik támogatás
- még mindig oldal-alapú a felismerés

33 év elteltével sem sikerült a technológiának megközelíteni a humán olvasást

A legjelentősebb CCS-fejlesztés: strukturált OCR: METAe–The Metadata Engine

Strukturális metaadatokat tároló OCR kimenet, mely a nyomtatott anyagot strukturált XML-objektummá alakítja.

- 2000-2003 EU-támogatott FP5 kutatási projekt
- 14 partner, koordinátor: Leopold-Franzens-Universität Innsbruck
- technológia: CCS, ABBYY
- eredmény: 2003-ban licenszált docWORKS/METAe nevű termék (szoftver, későbbi mutációk: newsWORKS, mediaWORKS)
- alkalmazott szabványok: METS, DC, NISO MIX, ALTO

OCR új rétege: **ISR = Intelligent Structure Recognition** (intelligens struktúra-felismerés)

METAe referenciák

- Norvég Nemzeti Könyvtár
- Finn Nemzeti Könyvtár
- Dán Királyi Könyvtár
- Harvard Egyetem Könyvtára
- Stanford Egyetem Könyvtára
- British Library
- Koninklijke Bibliotheek

Az OCR munkafolyamat

Jellemző problémák:

- Oldalon belül változó fényerő
- Előtér/háttér elkülönítése problémás (alacsony kontraszt)
- Egyenetlen, kötésnél behajló sorok
- Egyenetlen, függőlegesen ferde újsághasábok
- Átütő tinta, pizok (zaj), hiányos karakterek (kopás, festékhány)
- Formázási kiemelés (a legfontosabb információ szedése a legproblémásabb: címek, dőlt, félkövér, ritkított kiemelés)
- Hiányzó szótárak egyes nyelvekhez, nyelvváltozatokhoz
- Technikai, tudományos szimbólumok azonosítása problémás

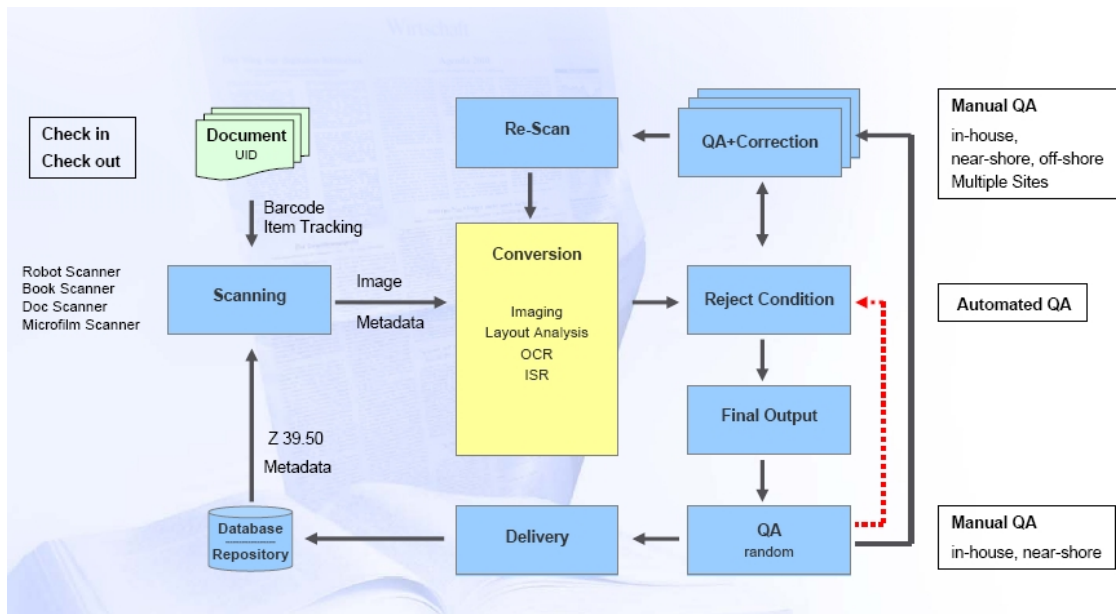
Mindezek kiküszöbölésével tudna szintet lépni az OCR-technológia.

Automatizált szegmentációs folyamat képanalízis alapján

- dokumentumsorrend az oldalszámok felismerésével
- dokumentumszerkezet felismerése a fejezetcímek, alcímek, képaláírások, jegyzetek azonosításával
- oldaltípus felismerése (címloldal, TOC, reklámloldal stb.)

Minőség-ellenőrzés

- automatikus korrekció a struktúra-felismerés után
- kézi javítás, de csak speciális jelentőséggel bíró strukturális elemeken: címek, oldalszámok képaláírások, jegyzetek, stb.



1. OCR munkafolyamat

Konklúzió: az OCR-nek közelíteni kell a tényleges olvasáshoz. Ezzel a technikai nehézségek kiküszöbölése mellett a nyelvészet, szemantika és statisztika terén kell elérni eredményeket.

DIGITAL RESTORATION AND LAYOUT ANALYSIS

Digitális helyreállítás és szerkezet-analízis

DR. APOSTOLOS ANTONACOPOULOS – PRIMA, UNIVERSITY OF SALFORD, UK

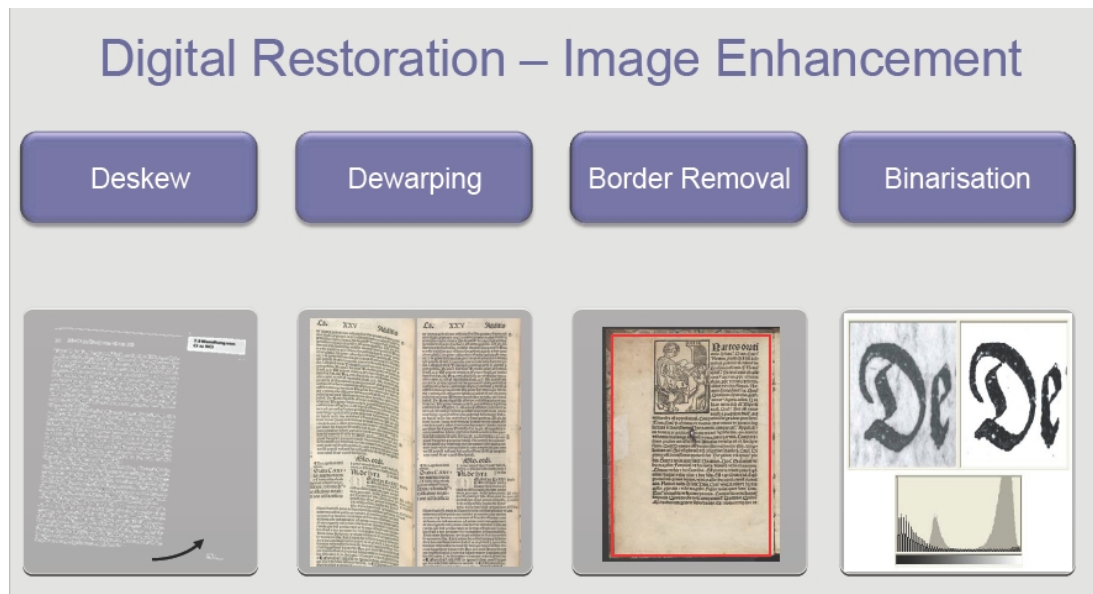


Apostolos Antonacopoulos

Az előadás az OCR előkészítés gyakorlati aspektusait mutatta be, elkülönítette a folyamatot alkotó részfeladatokat, és bemutatta, hogyan kezeli mindezt az IMPACT keretén belül készülő eszköz.

A bemenet optimalizálása a jó OCR alapja, mert minél több információt adunk be, annál pontosabb információ jön ki a folyamat végén. Az is fontos, hogy a bemeneti kép legalább a szabadszemes olvasás körülményeinek maximálisan megfelelően – ezzel egyben a Print-On-Demand funkcióra is felkészül a digitalizálási projekt. Minderre nemcsak a szkennelésnél, hanem az azt követő korrekciónál gondolni kell.

1. A képtimalizálás lépései



- geometriai korrekció
 - o dewarping (egyenésítés)
 - o deskew (forgatás)
 - o binarisation
- minta kiemelése, a következők korrigálásával
 - o túloldal átüt (bleed-through) vagy átlátszik (show-through) (nem ugyanaz)
 - o elkenődött festék
 - o változó tintasűrűség
- border removal (lapszél eltávolítása)

2. A szegmentálás lépései

Blokkok ► sorok ► szavak ► karakterek

Eddig elvégzett munka az alprojektben

- képkorrekciós és szegmentálási folyamatok kiértékelése
- könyvtári mintaanyag vizsgálata, a leküzdendő problémák kiszűréséhez
- ismerkedés a dokumentumok előállítás-történetével és módszertanával, restaurátorok, könyvtörténeti szakemberek bevonásával
- tömeges digitalizálási projektekhez készült ajánlások és korszerű módszerek tanulmányozása
- eszközprototípus kifejlesztése
- az első változat 2009 végére várható

Kiértékelési módszertan készül az IMPACT számára, mely közelebb hozza a tömeges digitalizálási projektek mérhetőségét

- közös könyvtári mintakészlet létrehozása, mely reprezentálja az általában előforduló problémákat
- a digitalizálási projektben előfordulható folyamatok és metrika, meghatározása, ground truth minden típusra és szintre nézve: *Aletheia*

HISTORICAL LEXICON BUILDING & HOW IT IMPROVES ACCESS TO TEXT

Nyelvtörténeti szótárak építése az OCR minőségjavítása érdekében

KATRIEN DEPUYDT, INL IMPACT TEAM

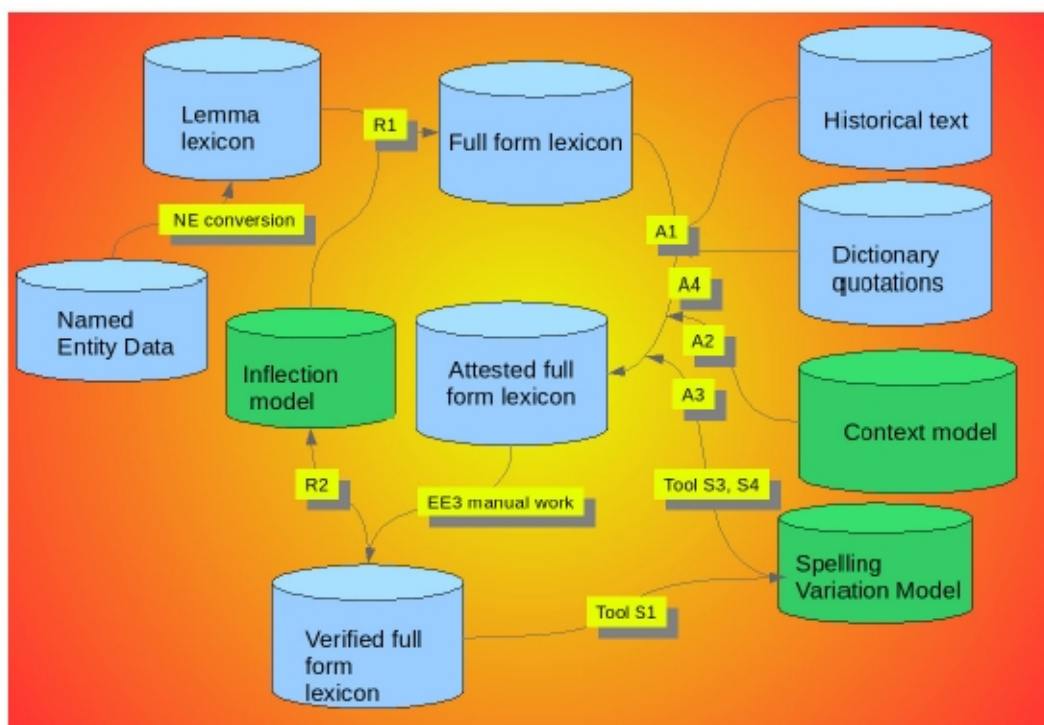
Az IMPACT EE (Enhancement & Enrichment) 2/3 workpackage célja a történeti nyelvkorlát kiküszöbölése

A történeti variánsok előállhatnak

- a helyesírási szabályok változásával
- változhat a lexikon tartalma
- változhatnak szón belüli fonémák/grafémák
- változhat a szó mint egység az egybe/különírási szabályok változásával
- az inflexió és egyéb morfológiai szabályok változásával

Ezek egy része történeti szótárak és nyelvi sémák útján kiküszöbölhető, a többire fuzzy illesztéseket kell készíteni.

Az IMPACT EE2/3 feladata olyan eszköz előállítása, mely lehetővé teszi a történeti dimenzióban is bővülő szótárak és transzformációs szabályok kifejlesztését.



Elvárások a készülő szótárakkal szemben

- korszakspecifikus szótárak készüljenek (pl. a történeti modult ki lehet kapcsolni modern szövegek feldolgozásánál)
- modern keresőfelületeket szolgál ki – a jelenlegi nyelvváltozatban beírt kereső-kifejezéssel azonosítja a történelmi alakot és visszaadja releváns találatként

Elvárások az eszközzel szemben:

- nyelvfüggetlen
- nagy adatmennyiséget tud kezelni, minél több automatizálható folyamattal
- lemma + lemmata – releváns alakok és változataik kiszűrése és csoportosítása
- igazolt előfordulások kezelése (manuális megerősítés)
- nyelvfüggetlen modul a nyelvtani variáns alakok bevitelére

Automatikus folyamatok:

- azonos alakokból alapszoportok kigyűjtése
- variánsok csoportba sorolása lexikon és sémák alapján
- fennmaradó alakok közelítő besorolása

Manuális folyamatok

- téves csoportosítások javítása
- hiányzó, de létező variánsok bevitele

Decision Support Tools

A döntés-előkészítés eszközei

NEIL FITZGERALD – BRITISH LIBRARY



Neil Fitzgerald

A tömeges digitalizálási projektek nagyon sok aspektusban különböznek egymástól. Ennek a feladatnak a célja az, hogy kiszűrje azokat a közös pontokat, melyeken egyesíthetők a különböző projektek mérlegelési szempontjai.

A partnerek szempontjai:

- Meghatározható döntési pontok
- A bemeneti anyag és metaadat állapota
- Milyen formátumbeli és külső jellegzetességgel kell számolni a különböző dokumentumtípusok esetén
- Régi és speciális betűkészletek és írásrendszerek kezelése
- Nyelv
- Névterek és visszakereshetőségük
- Tárterület és megjelenítési módszerek
- Tervezett projekt méretének meghatározása, átméretezhető modellek alkalmazása
- Költségvetés és stratégia megtervezése
- Szerzői jog

Az IMPACT szempontjai

- Kerüljük az egyedi, barkácsolt megoldásokat
- Csak elméleti alapon tervezni veszélyes
- A megvalósítás részletkérdésit kell előtérbe helyezni az eszközök megválasztásánál
- Nyitottnak kell lenni mind az in-house mind a kiszervezett lebonyolításra
- A projekt kompetenciaközpontokban gondolkodik

A döntés-előkészítő eszköz

- három fázisban készül el, két köztes ellenőrzéssel.
- áttekinthető benne az általános digitalizálási gyakorlat
- a gyakorlati folyamatok támogatására létrehozott keretrendszert nyújt
- esettanulmányokat tartalmaz
- értelmes forrásokra hivatkozik

Párhuzamos szekciók

IMPROVING ACCESS TO TEXT: CENTRE OF COMPETENCE IN MASS DIGITISATION OF HISTORICAL PRINTED TEXT

IMPACT: Kompetenciaközpont régi szöveg tömeges digitalizálására

LEVEZETTÉK: HILDELIES BALK (KB) ÉS ALY CONTEH (BRITISH LIBRARY)

2012-re az IMPACT a tömeges digitalizálás stabil kompetenciaközpontjává válik Európában

- eszközöket és útmutatókat ad, illetve közvetít
- lehetővé teszi a tapasztalatok megosztását
- oktatási és workflow-vezérlő eszközöket biztosít
- a központ körül közintézményekből, kutatóhelyekből és magáncégekben álló közösség épül
- együttműködik más kompetenciaközpontokkal, megelőzendő az ismeretek elkallódását és a feleslegesen ismételt kutatást
- fennáll amíg el nem érte a célját: mikor a régi európai szöveganyag ugyanolyan formában elérhető, mint a digitálisan született dokumentumok

Közösségépítés az alábbi tényezők bevonásával:

- digitalizáló cégek
- kutatóintézetek
- közgyűjtemények
- szervezetek (CENL, LIBER, nemzeti szintű szervezetek)
- TEL/EDL közösség

Platformok

- Központ: IMPACT honlap és helpdesk
- Partnerek honlapjai és szolgáltatásai
- Szociális hálózatok (LinkedIn, FB)
- Egyedi résztvevők publikációi
- Szakmai s tudományok konferenciákon való részvétel
- IMPACT konferenciák (a jelenlegin kívül ebben a futamban egy záró-konferenciát terveznek)

Vita

A szekció hátralévő részében a jelenlevők több felvetett kérdésről vitáztak, pl.:

A készülő termékekkel és a projekt fenntartásával kapcsolatos üzleti modellek

- különszolgáltatásokból származó jövedelem (pl. oktatás, workflow szervezés)
- szponzorok, privát- közszféra együttműködések
- piaci licenszek
- tagdíjak
- állami és EB támogatás

(A szekción az OSzK-ból ***Horváth Ádám*** vett részt)

2. Challenges and Opportunities in Mass Digitisation – How Technology Can Meet Libraries’ Needs

LEVEZETTÉK: APOSTOLOS ANTONACOPOULOS ÉS STEFAN PLETSCHACHER (PRIMA, UNIVERSITY OF SALFORD, UK)

A szekción nem vettünk részt, a preuentáció a beszómolóírás pillanatában még nem elérhető. A szekcióhoz kapcsolódó kérdőívet kitöltöttük.)

3. Mass digitisation at the KB

3. Tömeges digitalizálás a KB-ben

LEVEZETTE: EDWIN KLIJN

Projektekből digitális könyvtár

A KB jelenleg több, párhuzamos, ill. egymást átfedő projekt keretében alakítja ki digitális könyvtárrá. A részfolyamatok hosszú távon egységes és teljes gyűjteménnyé állnak össze:

2011-ig lezáruló projektek:

- 41 millió oldal (könyv, folyóirat, újság)
- Keret: 54 millió €

A 2009-2013-as ütemezésű digitális könyvtári program célja 2013-ig digitalizálni a teljes holland dokumentumállomány (könyv, folyóirat, újság) 10%-át.

Workflow

1. Válogatás : KB és intézményi partnerek
2. Előkészítés, kollacionálás: KB
3. Szkennelés, OCR: kiszervezve (pl. CCS)
4. Hozzáférés (kereshető felület): KB

Példák az egyes projektekre

- Állományvédelmi digitalizálás: Metamorfoze <http://www.metamorfoze.nl>
- Újságok: Historische kranten in beeld <http://kranten.kb.nl>
- Magazinok online szolgáltatás: <http://www.hetgeheugenvannederland.nl>

A digitalizálás indokai

- a felhasználók gyakran keresik az adott anyagot
- a felhasználók valószínűleg gyakran fogják keresni az adott anyagot (i. e. ha elérhető lesz)
- meg kell védeni az eredetiket
- a felhasználók fizetnek a digitalizálásért

Nehézségek

1. Szerzői jog

Copyright-érintett helyzetek:

- Másolatkészítés (digitális vagy analóg)
- Zártláncú hálózaton való közzététel
- Interneten való közzététel

A jogtulajdonosok felkutatása minden esetben törvényi kötelezettség. A KB a közös jogkezelés lehetőségeiről tárgyal jelenleg az alkotókat képviselő ügynökségekkel (Commission Digitie)

2. Személyiségi jogok

Analóg formátumban egy adott személyt érintő adat közlése, illetve annak internetes utánközlése két különböző eset a személyiségi jogok szempontjából. Lehetséges, hogy megjelennek ilyen jellegű szabályozások, ezért a személyiségi jogi követelményeket időnként tekintetbe kell venni.

3. Költségek

- Digitalizálás
 - o „Quick & dirty” automata minőség: **0,3 €/oldal**
 - o „Quick & beautiful” képkorrekcióval, utókorrekcióval: **1.5 €/oldal**
- A gyűjtemények **fenntartása** külön költséget jelent. A KB esetében:
 - o **1 TB** tartalom fenntartása évi **7500 € költséget jelent**

Források:

- intézményi költségvetés
- célzott állami támogatás
- EU
- magánszektor

Digitalizálás irányelvei

- **JISC Digital Media** – <http://www.jiscdigitalmedia.ac.uk/>
- **Cornell University** – <http://www.library.cornell.edu/preservation/tutorial/>

Az elméleti bevezetés után az előadó megismertetett a munka gyakorlati hátterével. Mivel a KB a szkennelés és az OCR munkafolyamatokat kiszervezi (bár e döntését a vezetőség már megbánta); a könyvtár aktívan a kollacionálás folyamatában vesz részt a munkában.

Ennek során felderítik, hogy a több lelőhelyen elérhető példányok közül melyek az ideális jelöltek a digitalizálásra. Ezeken szükség esetén kisebb restaurálást hajtanak végre digitalizálás előtt. Kötetet nem bontanak meg, semmilyen anyagot nem vágnak fel vagy semmisítenek meg más módon a digitalizálás érdekében.

Jelenleg hetente kb. 50000 oldal anyagot adnak át az alvállalkozónak.

Az időszaki kiadványok esetében számonként tartják nyilván az átadott anyagot. A végeredmény cikkszinten azonban tagolódik, az előre meghatározott kategóriákba, melyek pl. (újságoknál): hírek, értesítések, reklámok, illusztrációk. Ezt a besorolást az előkészítés folyamán megkezdik.