

Beszámoló az utrechti konferenciáról

*Az európai kulturális örökség digitalizálása
Termékek – elvek – technikák
1999. október 21–23.*

Ezen a néhány szeles, hűvös, október végi napon 25 európai és Európán túli ország több mint 120 könyvtárosa és informatikusa gyűlt össze Hollandia egyik legszebb városába, hogy az európai kulturális örökség digitalizációjáról tanácskozzon, ismereteket cseréljen. A szervezők első ízben az európai digitalizációs projekteket szerették volna bemutatni az e területen túlnyomó többségben lévő amerikai projektek mellett.

A konferencia résztvevői – viszonylag kis létszámuk ellenére – Európa legtöbb országát képviselték. Örvedetes, hogy Nyugat-, Észak- és Dél-Európa országai mellett voltak itt néhány közép-kelet-európai országból is: Csehországból, Lengyelországból, Szlovákiából és Horvátországból. Talán tapasztalatgyűjtés végett néhány Európán kívüli országból is érkeztek vendégek: az USA-ból, Izraelből, sőt Dél-Afrikából is.

Catherine Lupovici, a Francia Nemzeti Könyvtár digitalizációs programjáról tartott részletes beszámolót. Ez már 1992-ben elkezdődött, párhuzamosan egy új épület és információs rendszer kialakításával. Digitális gyűjteményük 86 000 – eredetileg nyomtatott – könyvet, 250 000 képi file-t és több órnyi hanganyagot tartalmaz. Egy enciklopédikus francia könyvtár létrehozását tűzték ki célul különös tekintettel a régi vagy nehezen hozzáférhető dokumentumokra. A digitalizálás még jóval az Internet robbanása előtt megkezdődött. Az elektronikus állomány nagyobbik része csak helyben dedikált munkaállomásokon elérhető. A használat a képernyőn való olvasást vagy a nyomtatást jelenti. A digitális dokumentumok néhány megszokott szempont szerint (pl. cím, szerző, kiadó) visszakereshetők vagy böngészhetők. A dokumentumokat képi (images) formában, TIFF vagy PDF formátumban tárolják. A digitalizálás arányosan foglalja magába a különböző tudományterületek dokumentumait. A könyvtár szerződéseket kötött különböző kiadókkal azok kiadványai digitalizálásának, felhasználásának engedélyezésére. Közvetlenül azonban nem kapnak dokumentumokat elektronikus formában a kiadóktól. A francia kiadók többsége még nagyon konzervatív ebben a tekintetben – mesélte az előadó – és bizalmatlanok az Internettel, az elektronikus kiadással szemben az állomány kisebbik, public domain része 1997-től már nyilvánosan is elérhető a Gallica Web szerveren. (URL: <http://gallica.bnf.fr>)

Anne-Emmanuelle Cleulemans (Université catholique de Louvain) és Frans Wiering (Utrecht University) a „Thesaurus musicarum italicorum”-ról tartottak előadást és bemutatót. Vállalkozásuk lényege, hogy az 1650 előtti olasz zene teljes

corpusát digitalizálni fogják a kéziratoktól kezdve a kottáig. A munka azonban nemcsak a különböző művek valamilyen formátumban való digitalizálását jelenti.

1997-ben jelent meg a projekt első kiadványaként az a CD-ROM, amely Gioseffo Zarlino műveit tartalmazta. A szövegek feldolgozása SMGL alapon történt. Ez a technológia lehetővé tette, hogy a szerző egy művének – kéziratának – több, eltérő korabeli kiadását is rögzítsék úgy, hogy az fejlett eszközökkel összevethető, összehasonlítható legyen. Lényegében nem rögzítették külön-külön mindegyik kiadást, hanem csak a legutolsót és egyes SGML tag-ekkel jelölték, hogy a szöveg mely részei melyik kiadásban fordultak elő. Ennek alapján a megjelenítő program képes az egyes kiadásokat külön vagy együtt, a szövegváltozatokat más-más színnel jelölve megjeleníteni. A részletes SGML kódolás természetesen összetett keresésre is lehetőséget biztosított, pl. különböző névváltozatok esetén is megtalálja egy adott kortárs nevének összes előfordulását a szövegben.

A kéziratokban vagy mellettük korabeli kották is találhatóak, amelyek nemcsak megtekinthetők képi file formátumban, hanem MIDI file-ként a zene meg is hallgatható.

A vállalkozás jól szemléltette annak a lehetőségnek egy részét, hogy a digitalizálás milyen többletet képes adni a szövegek egyszerű elektronikus rögzítése mellett, hogy a tudományos kutatás egy-egy téma dokumentumait sokkal komplexebben dolgozhatják fel, mint hagyományos kiadásban. A CD-ROM mellett más Webes, Internetes megjelenítést is terveznek, ez azonban még fejlesztés alatt van. (URL: <http://candl.let.uu.nl/research/tmi/main.htm>)

Michael Pidd (University of Sheffield) egy másik példáját mutatta be annak, hogy a dokumentumok digitalizálása milyen lehetőségeket rejt magában a tudományos kutatók számára. A számítógépet a textuális kutatás szolgálatába állítva egy 15. századi angol művet, a Canterbury Meséket digitalizálták. A mű csak 84, korabeli kéziratban maradt fenn, és ennek alapján megpróbálták válaszolni arra az alapvető kérdésre, hogy mi is lehetett az eredeti mű. A különböző szövegváltozatokat képi és szövegformában is számítógépre vitték, majd SGML kódolással dolgozták fel. Ily módon fejlett eszközökkel tudták elemezni nemcsak a különböző szövegváltozatok eltérését, de a korabeli angol nyelv grammatikai szabályait, állapotát is.

A szöveg feldolgozása során nyert tapasztalatok, kifejlesztett technológiák, programok más, hasonló esetekben is a kutatók rendelkezésére állhatnak majd.

A munka eredménye ugyancsak CD-ROM-on jelent meg, amely jó szolgálatot tesz a visszajelzések alapján a tudományos kutatók számára. A projekt 1986-ban kezdődött, és az első CD-ROM-ot 1996-ban adták ki. (The Humanities Research Institute URL: <http://www.shef.ac.uk/uni/academic/D-H/hri> – The Canterbury Tales Project URL: <http://www.shef.ac.uk/uni/projects/ctp/index.html>)

Bas Savenije (Utrecht University Library) előadása a könyvtári szolgáltatások fejlődését tekintette át az új informatikai kihívások kapcsán, kitérve a digitalizáció tendenciájára, értelmére, hasznára.

Az új informatikai környezetben megváltoznak a könyvtári szolgáltatások, a könyvtári magatartás, a stratégiai célok is. A gyűjteménygyűjtés mellett egyre nagyobb jelentőséget kap az információkhoz való hozzáférés biztosítása, pl. az online előfizethető elektronikus folyóiratok esetében. A korábban helyi olvasók



egy része online „userre” válik, az egyetemi hallgatók, oktatók már a saját szobáikból is hozzáférhető információs szolgáltatásokat igényelnek. Ebből következik, hogy az információs források felhasználásának ismereteit egyre inkább be kell építeni az egyetemi oktatásba is.

Az információs váltás kihatással van a tudományos kommunikációra is. Egyre nagyobb szerepet kapnak a távolról is elérhető digitális archívumok, és a könyvtáraknak támogatniuk kell az elektronikus publikálást is. A hagyományos információs lánc sokkal komplexebbé válik, amelyben a könyvtár szerepe már korántsem olyan evidens, mint eddig. A digitalizálás kétségtelenül nemcsak fontossá, de hasznossá is válik a könyvtári politikában. Felgyorsul a dokumentumok szolgáltatása, az elektronikus dokumentumok visszakereshetővé válnak, és a szövegek feldolgozásának az eddig nem ismert módja gazdagíthatja azok felhasználását. A digitalizált dokumentumok révén a könyvtárak raktári helyet takaríthatnak meg, megőrizhetik kulturális örökségünket a fizikai romlás veszélye ellenére.

Ezekből következően az egyik alapvető kérdés a digitalizálás folyamán a „mit digitalizáljunk” kérdése. A digitalizálandó dokumentumok kiválasztását szorosan az oktatási, tudományos igényekkel kapcsolatban kell elvégezni, összehangolva az oktatási programmal. Lényeges szempont továbbá, hogy a tudományos kutatásban olyan szövegeket kell előnyben részesíteni, amelyek digitalizálása gazdagíthatja azok kutatását, tudományos elemzését. A kulturális örökség szempontjából fontos dokumentumok digitalizálását kormányzati szinten kell ösztönözni és főként finanszírozni. A könyvtáraknak elsősorban a közvetlen felhasználóik igényeit kell elsőbbségben részesíteni.

Franz Klaproth (Göttingen Digitalization Center) a Göttingeni Egyetemi Könyvtáron belüli digitalizációs projektekről és munkákról számolt be. 1996-ban hozott létre a Német Kutatási-egyesülés (Deutsche Forschungsgemeinschaft) egy új

könyvtári alapítványi programot „Elosztott Digitális Kutatókönyvtár” („Verteilte Digitale Forschungsbibliothek”) néven. Ennek a programnak a keretében kezdtek meg néhány területen (pl. matematika, útleírások) a régebbi dokumentumok retrospektív digitalizálását. Az előadó részletesen bemutatta a dokumentumok digitalizálásának, feldolgozásának és kezelésének technológiai adatbázis modelljét.

A dokumentumok kiválasztását kutatók és könyvtári szakreferensek (subject librarians) közösen végzik. A szerzői jogi problémák elkerülése végett csak régi kiadású könyvekkel és folyóiratokkal foglalkoznak. A digitalizált dokumentumokat képi (image) formában őrzik (TIFF), és a felhasználás során programokkal konvertálják ezeket szükség szerinti egyéb formátumokra (pl. GIF, JPEG, PDF). A képi dokumentumok karakterfelismerésével nem foglalkoznak. A dokumentumokhoz egy metaadat egység is kapcsolódik. A metaadat XML alapú, de a minimális Dublin Core elemeket is megadják. A metaadatot felhasználáskor konvertálni tudják a Web böngészők számára HTML-be, de importálni tudják online katalógusokba (pl. Pica) is. A metaadatok mellett XML technológiával feltárják a dokumentum struktúráját is, pl. tartalomjegyzék szinten. Az adatbázis modellben két külön rész kezeli a metaadatokat és az image-ekben tárolt dokumentumokat. A rendszer lekérdezhető, kapcsolatot fejlesztek ki mind az online katalógusok, mind a Web irányába.

Több kereskedelmi program kipróbálása után a Központ végül is saját program kifejlesztése mellett döntött. A kifejlesztett rendszert (AGORA) azonban forgalmazni kívánják, felhasználhatóvá tenni más könyvtárak számára. A digitalizált gyűjtemény szolgáltatása CD-ROM-on és online a közeljövőben fog beindulni, bár néhány dokumentum már megtekinthető az Egyesült Könyvtári Szövetség (Gemeinsamer Bibliotheken Verbund) OPAC-ján keresztül. (GDZ Homepage URL: <http://www.sub.uni-goettingen.de/gdz> – AGORA URL: <http://www.agora.de>)

Adolf Knoll (National Library of the Czech Republic, Prague) a cseh könyvtárak növekvő hálózati csatlakozásáról és részletesebben a „Memoria Mundi Series Bohemica” programról beszélt.

Csehországban a 90-es évek elejét az integrált könyvtári rendszerek bevezetése jellemezte. 1995 után erősödött meg az igény a retrospektív katalógusok konvertálására, miközben az első digitalizálási tevékenységek 1992–1993-ban kezdődtek. Ezzel párhuzamosan a nagyobb könyvtárak legfontosabb problémájává dokumentumaik megőrzése vált, különös tekintettel a gyakran használt régi napilapokra.

A Memoria Mundi Series Bohemica program 1992-ben kezdődött és része az UNESCO memory of the World programjának. (Adolf Knoll tanácsadóként részt is vesz a program egyik bizottságában.) A Memoria Mundi projekt célja régi iratok digitalizálása. Jelenleg kb. 3000 kézirat elektronikus változata már elkészült, 200 másik digitalizálása folyamatban van.

A dokumentumokat képi (image - TIFF) formában tárolják. A dokumentumokhoz metaadatok készülnek egy speciális SGML alapú formában, amelyet a könyvtárban dolgoztak ki (DOBM). A formátum segítségével külön-külön leírják a kéziratokat, az oldalakat és az oldalakon található képeket. A könyvek metaadatai azonban UNIMARC kompatibilisek is. A DOBM segítségével nemcsak a dokumentumról rögzítenek információt, de annak struktúrájáról, tartalmáról is. A dokumentumokról különböző képformátumot tudnak előállítani az Internet, a helyi felhasználók, egy kis méretű galéria céljára az eredeti nagyfelbontású képi file-okból.

A kidolgozott technológiát, tapasztalatokat szívesen megosztják más könyvtárakkal is. A gyűjteményt először 1993-ban adták ki CD-ROM-on, majd 1995-ben másodsor. Az adatbázis azonban elérhető az Interneten is.

A DOBM SMGL formátumot 1999 januárjában standardként elfogadta az UNESCO a Memory of the World program számára. Teljes változata elérhető az Interneten: <http://digit.nkp.cz/rare99/enter.htm>

A Nemzeti Könyvtár másik jelentős digitalizációs programja a mikrofilmekre terjed ki, amelynek keretében főként a napilapokkal foglalkoznak. Itt is a már kidolgozott DOBM formátumot használják.

A program eredményei a közeljövőben kerülnek nyilvánosságra.

Digitalizálási tevékenységeik során azonban katalógusok szkennelésével is foglalkoztak, pl. a Nemzeti Könyvtárával (kb. 5 millió katalóguscédula TIFF formátumban. – <http://katif.nkp.cz/>) vagy az arab kéziratok katalógusáéval (http://digit.nkp.cz/samples/arabica/Arabika_djvu.html).

Ez utóbbit az AT&T a Hónap Site-jának választotta idén augusztusban (<http://djvu.research.att.com/examples/som.html>). MEMORIAE MUNDI SERIES BOHEMICA URL: <http://digit.nkp.cz>

Patricia Alkhoven (Koninklijke Library) a Holland Nemzeti Könyvtár kézirat-digitalizálási tevékenységéről tartott bemutatót. Ez egyike a Nemzeti Könyvtár számos digitalizálási projektjeinek, amelyek régi kéziratokra, könyvekre, a könyvekben található képekre és a rájuk vonatkozó szakirodalomra terjednek ki. E könyvtárnak van Hollandiában a legnagyobb középkori kéziratgyűjteménye. A projekt képi (images) formátumú digitális dokumentumok nagy tömegű kezelését tűzte ki célul. Egy másik cél ennek a digitális állománynak az integrálása a könyvtári rendszerbe.

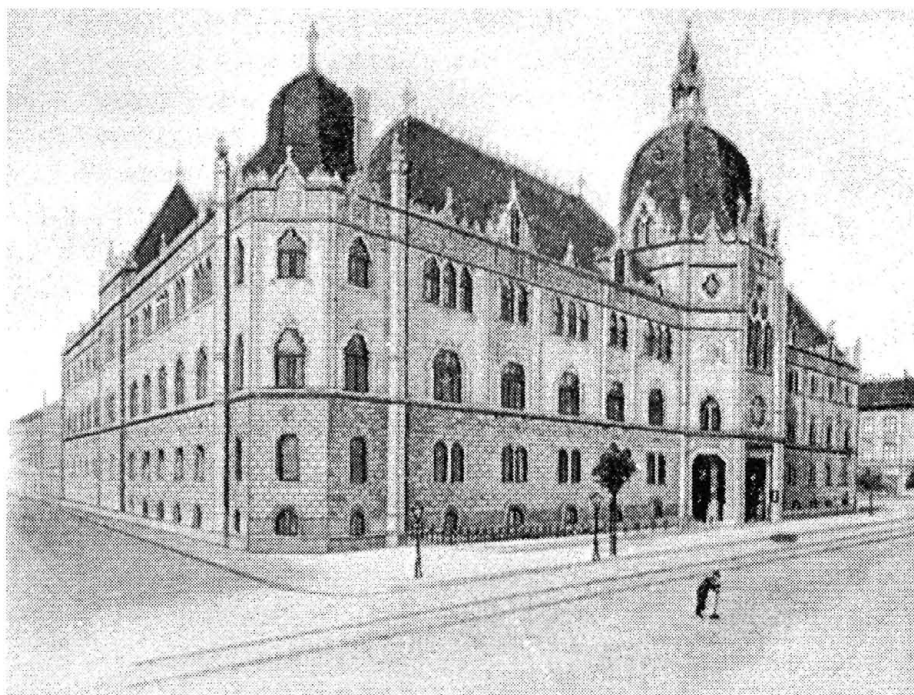
Először 1986-ban tették nyilvánossá a gyűjteményt optikai lézer lemezen. A kezelőrendszer egy DOS-os adatbázis kezelő volt. Később konvertálták az állományt egy XML alapú rendszerbe, amely részben már elérhetővé vált az Interneten, Web-en keresztül is.

A szkennelés után TIFF állományokat hoznak létre, amelyeket igény szerint konvertálnak JPG vagy kisméretű (thumbnail) formátumokra. A leíró adatok indexelését egy helyi Altavista szerverrel végzik.

A digitalizálandó dokumentumokat egy egyetemi bizottság válogatja ki megfelelő előtanulmányok alapján. További digitalizálási terveik között mikrofilm állományok elektronikus konvertálása szerepel. (URL: <http://www.konbib.nl/home-fe.html> – Digital Collections URL: <http://www.konbib.nl/kb/menu/col-dige.html>)

Más, a fentiekhez hasonló projektekről is kaptunk információkat. A koppenhágai Királyi Könyvtár ugyancsak foglalkozik középkori kéziratok digitalizálásával. Egy kiállítás kapcsán („Living words & luminous pictures. Medieval Book Culture in Denmark.”) most 12 középkori kézirat digitális változatát tették online elérhetővé. (URL: <http://www.kb.dk/elib/mss/mdr/index-en.htm>)

Egy érdekes, az előzőektől eltérő digitalizációs projektről tartott beszámoló a Norvég Nemzeti Könyvtár munkatársa, Svein Arne Brygfjeld is. A Norvég Műsor-sugárzó Vállalat (Norwegian Broadcasting Corporation) a nemzeti könyvtárral közös projekt keretében úgy döntött, hogy digitalizálják hangszalagokon meglévő

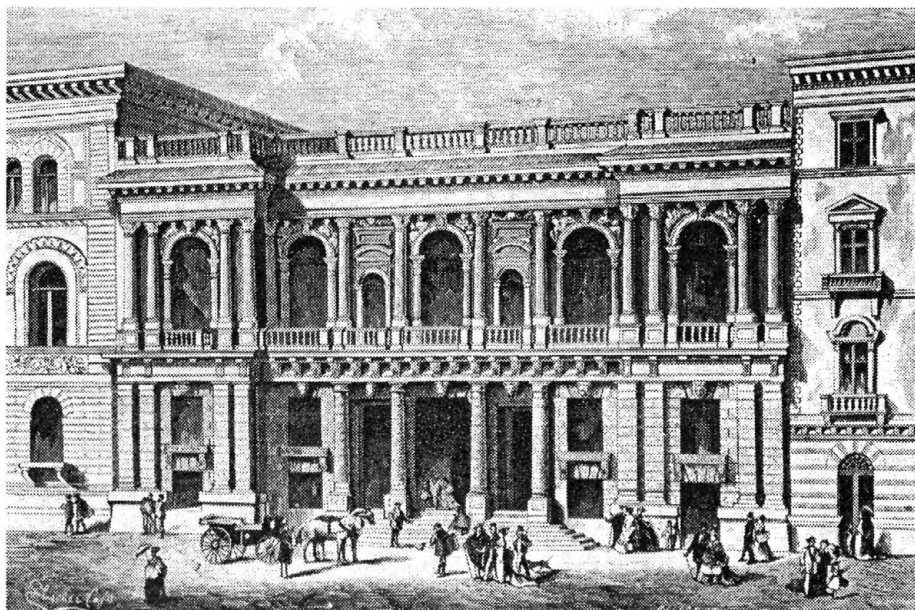


hangarchívumukat. A projekt kb. 6 éve tart, folyamatosan szállítják a hangszalagokat a könyvtár 1000 km-re lévő laboratóriumába, ahol 3 UNIX szerveren digitalizálják őket folyamatosan. A hanganyagokat alacsony hangminőségben nyilvánosan hozzáférhetővé teszik az Interneten november 1-jével induló szolgáltatásukban, magasabb hangminőségben azonban csak a kutatóknak áll rendelkezésére a hangarchívum. A digitális könyvtárban az Interneten mindenki számára elérhető lesz viszont Norvégia rádióműsorként közvetített 20. századi hangos történelme. (URL: <http://www.nb.no/dra/> – sajnos csak norvég felülete van.)

* * *

A fentiekén kívül számos érdekes témában hangzottak még el előadások és rövidebb délutáni bemutatók, amelyekről a konferencia honlapján található előadás-kivonatokból és az ott található URL-ek nyomán tájékozódhat az érdeklődő.

Valószínűleg nem az első és nem is az utolsó olyan konferencia volt ez, amely a könyvtárak világát, a könyvtárosokat is érintve az általános digitalizációs folyamatról szól, amelynek révén a múlt és a jelen kulturális értékei egyre nagyobb mértékben elektronikus hordozókra, CD-ROM-okra vagy az Internetre kerülnek. A konferencia érdekességét inkább az adta, hogy az Internet angolszász, amerikai túlsúlya közepette az ezen a területen folyó európai tevékenységekre és elért eredményekre igyekezett felhívni a figyelmet, lehetőséget adva a terület szakembereinek tapasztalatcserére, ismerkedésre. Valószínű, hogy sok egyéb digitalizációs



projekt, tevékenység nem került bemutatásra a konferencián, de erről a területről nehezen is alakítható ki egy átfogó összkép.

Mindazonáltal erre tesz kísérletet az IFLA/UNESCO felmérése (Survey on Digitisation and Preversation), amelynek keretében igyekeznek összegyűjteni a könyvtárak és egyéb kulturális intézmények digitalizációs tevékenységeit. A felmérést a közelmúltban végezték, eredményét nemsokára teljes egészében közzéteszik. Az információk gyűjtését folyamatosan végzik és teszik majd nyilvánossá az UNESCO Memory of the World projektje keretében. (URL: <http://www.unesco.org/webworld/mdm/>)

Röviddel a konferencia után csak néhány gyors, vázlatos megállapítást, benyomást igyekszem még megfogalmazni a fentiek alapján:

A konferencia főként a nagyobb könyvtári digitalizációs projektekre koncentrált. Jelentőségükhöz képest némiképp elhanyagolta azokat a tevékenységeket, amelyek nem ebben a szférában történnek, mindazonáltal a kulturális örökség jelentős digitalizációját végzik. Gondolok itt olyan jelentős magánkezdeményezésekre, mint pl. a Krén Emil által kezdeményezett magyar festészeti adatbázis vagy számos CD-ROM kiadó tevékenysége. A CD-ROM-okon megjelenő digitalizált kulturális örökség különös fontosságú, mivel jelentős mennyiségű információ jelenik meg így, azonban a nagyon hardver- és szoftverfüggő alkalmazások jövőbeni alkalmazhatósága erősen megkérdőjelezhető.

A bemutatott, ismertetett projektek jelentős része régi, a szerzői jog által már régen nem védett könyvekkel, kéziratokkal foglalkozott. Ezek kulturális örökségünk fontos részei, ennek ellenére igazából valószínűleg csak egy szűk körű, kutatói közösség számára hasznosíthatóak, jelentenek igazán értéket. A XX. századi írott és nem-írott kulturális termékek digitális hozzáférhetővé tétele a szerzői jogi problémák miatt a technikai lehetőségektől jócskán elmarad. Néhány konferencia-

résztevő elmondta, hogy digitalizációs projektjeik során – a régi kéziratok, könyvek kétségtelen kulturális értéke mellett – az anyagok kiválasztásában nem elhanyagolható szempont volt a szerzői jogi problémák kikerülése.

Az ismertett projekt jelentős része „képalapú” digitalizálást választott, még ott is, ahol a jelentős szöveges anyag karakteres felismerést igényelt volna. Néhány olyan esetben alkalmaztak csak karakteres felismerést, ahol kevés szöveg mélyebb, tudományos-nyelvészeti elemzését valósították meg.

A digitalizált állományok mellett minden esetben a hatékony visszakeresést megkönnyítő mondatok voltak találhatóak. Ezeknél előszeretettel alkalmaznak valamilyen SGML alapú formátum-leírást. A metaadatok majd minden esetben önálló alkalmazásban voltak lekérdezhetőek. Csak a további fejlesztések kapcsán valósították meg – vagy éppen tervezik – ezek kapcsolatát a hagyományos könyvtári katalógusokkal.

Az SGML alapú formátumok elterjedt alkalmazása volt tapasztalható, bár többnyire a képi állományokat leíró metaadatok esetében. Magukat a digitalizált szövegeket csak néhány esetben dolgozták fel SGML technológiával, többnyire kisebb szövegállományoknál, amikor azokkal további tudományos szintű kutatásokat végeztek.

A projektek többsége ugyan már jónéhány évre tekintett vissza, és szép eredményeket tudott felmutatni, de többségük CD-ROM-on került nyilvánosságra. Kevés eredménye nyilvánosan elérhető az Interneten, de jónéhány projekt még csak most készül, korlátozottan, a világhálóra. Különösképpen egy áttekintő amerikai előadás fényében volt érzékelhető az a különbség, hogy míg az ismertett tengerentúli projektek túlnyomó részben az Interneten jelentek meg, addig az európaiak csak „második lépésként” célozták meg az online szolgáltatást.

Végezetül hadd jegyezzem meg érdekességgéppen, hogy a konferencia címe, témája ugyan kimondottan a digitalizáció volt, de a Magyar Elektronikus Könyvtár alapvető céljához hasonló, a meglévő digitalizált állományokat összegyűjtő könyvtári tevékenységről, projektről nemigen hallottam külföldi kollégáktól. Megnyugodva tapasztaltam, hogy kulturális örökségünk digitalizálásában nem vagyunk sokkal elmaradva európai kollégáinktól.

Sőt, eredményeink nagyobb jelenlétre jogosítanak fel a nemzetközi szintéren, mint ami ténylegesen megvalósul. Hátrányként említhető, hogy nálunk kevésbé terjedt el még a tudományos igényű, SMGL alapú szövegdigitalizálás, -kutatás. A középkori dokumentumok digitalizálásának egyik kétségtelen előnye, hogy azok még többnyire az akkori közös európai nyelven, latinul íródtak. A jelenkori dokumentumoknál azonban már jelentkezik Európa nyelvi sokszínűsége.

* * *

Remélem, a fenti tapasztalatok felhasználhatóak lesznek a hazai digitalizációs projektekben, segítenek abban, hogy az Interneten a francia, német, holland és egyéb kulturális dokumentumok mellett a magyar kultúra öröksége is mind szélesebb körben elérhető legyen.

Köszönet a Nemzeti Kulturális Örökség Minisztériuma Könyvtári Osztályának, hogy anyagi támogatásával lehetővé tette részvételemet ezen a konferencián.

Moldován István