



Archiválás a nagyméretű digitalizálás korában (Fehér könyv)

A Könyvtári és Információs Erőforrások Tanácsa (Council on Library and Information Resources = CLIR) megbízásából Oya Y. Rieger, a Cornell Egyetem könyvtárának munkatársa készítette ezt az összefoglaló tanulmányt. Azt a problémát vizsgálja, hogy vajon a digitalizált anyagok minősége és fontos metaadatokkal való ellátása miként valósul meg, most, amikor könyvek millióit digitalizálják olyan hatalmas projektekben (large-scale digitization initiatives = LSDI), mint a *Google Book Search* vagy a *Microsoft Live Search Book*. Az elkészítés során széleskörűen – kérdőívek módszerével is – tájékozódott a könyvtáros közösségben, valamint a szakirodalomban. Ezek a projektek vegyes érzelmeket keltettek egyes kutatókban és könyvtárosokban, akik sokat tettek a nyomtatott gyűjtemények jó minőségű digitális szurrogátumaiért. Hasonlóan kérdések vetődnek fel a nagy digitális gyűjtemények és elérhetőségük hosszú távú fennmaradásával kapcsolatban. A CLIR célja a tanulmány elkészítésével az volt, hogy stimulálja a kérdések megvitatását és az együttműködési szándékot.

A bevezetést (1. fejezet) követően a tanulmány további 5 fejezetre és egy függelékre tagolódik.

A másodikban megismerkedhetünk négy nagyméretű digitalizálási projekttel és digitalizálási stratégiákkal. A már említett két kereskedelmi vállalkozás mellett két nem profit célú projektet mutat be a szerző: az OCA-t (*Open Content Alliance*) és a *Million Book Project*-et.

Előljáróban tárgyalja azokat a motivációs tényezőket, amelyek a projektekben részt vevő számos könyvtárat, illetve üzleti és nem profit vállalkozásokat az együttműködésre készítette.

A könyvtárakat három fő érv készíti a részvételre: az első alapvető küldetésükhöz kapcsolódik, fejleszteni a hozzáférést gyűjteményeikhez, mégpedig világszerte. Míg korábban a nagyobb egyetemi könyvtárak az Egyesült Államokban 5–10 000

könyvcímet voltak képesek digitalizálni évente, a *University of Michigan–Google LSDI* ma 30 000 kötetet hetente, így a könyvtár teljes minősített gyűjteményét öt év alatt feldolgozzák. A második szempont az állomány megőrzése. A kiadók gyakran nem tartanak meg példányokat kifutott kiadványaikból, így az archiválás a hosszú távú megőrzést szolgálja. A harmadik motivációs cél a könyvtár saját kutatás-fejlesztési munkájának támogatása, akár gyakorlat megszerzése nagy digitális állományok kezelésében, akár például új keresőeszközök fejlesztése.

A vállalkozások céljai eltérőek. A *Google Book Search* program teljes könyveket digitalizál, mind szabad használatúakat (public domain), mind jogvédetteket. Ez utóbbiakból csak kis részleteket dolgoznak fel. A végső eredmény egy kereshető index lesz. 2007 augusztusában 27 könyvtár vett részt a munkában. A Google ezen kívül több mint 10 000 kiadóval működik együtt.

A 2005-ben indult *Microsoft-projekt* először az OCA-val, majd további könyvtárakkal szabad felhasználású, 1923 előtt kiadott könyvek teljes szövegű adatbázisát fejleszti.

A két projektben eltérő a részt vevő könyvtárak helyzete, mások a saját, illetve a közös anyagok további használatának feltételei. Kiegészítésként a Microsoft kiadókkal is végez munkát a *Live Search Books Publishers Program* keretében. A cég saját interfészt készített a szövegekben végzett kereséshez.

Az OCA-projektet az *Internet Archive* és a *Yahoo* szervezte nem profit és kormányzati, valamint kulturális és technológiai intézmények részvételével. Célja szabad hozzáférésű digitális gyűjtemények létrehozása és elérhetővé tétele az *Internet Archive* és a *The Open Library* keretében. Az OCA valódi digitális könyvtárat kíván létrehozni több nyelven, multimédiás dokumentumokat is beleértve. Ezek a tartalmak kereshetők is a főbb kereső-

gépekkel. Az állományok megtalálhatók az Internet Archive-ban a Microsoftnál és a *Library of Alexandria*-ban, a jövőben más repozitóriumokban is. Jelenleg szabadon használható anyagokra összpontosítanak, de tárgyalnak kiadókkal is. A projektet a Microsoft és az Adobe is támogatja.

A Million Book Projectet a *Carnegie Mellon University School of Computer Science* és az *University Libraries* szervezik; magja a digitális könyvtárakra vonatkozó kutatási program. Az elmúlt hat évben az USA *Nemzeti Tudományos Alapjának* támogatásával és nemzetközi kapcsolatokkal (26 intézmény) több mint 1,4 millió könyvet szkenneltek Kínában, Indiában és Egyiptomban. Partnerük volt még az Internet Archive is. A projekt 2007-ben befejeződött, de a résztvevők folytatni kívánják a munkát. A részt vevő országok saját anyagaikat maguk kívánják tárolni, a hozzáféréstől még nincs megegyezés.

A harmadik fejezet a digitalizálás és az archiválás fontosabb adminisztratív és technikai kérdéseit tekinti át. Ez a leírás ugyan nem lehet teljes, azonban érinti, pontosabban felteszi a legfontosabb kérdéseket.

A négy témakör: kiválasztás digitalizálásra és maga a digitalizálás; a tartalom előállítása, ezen belül a kép minősége; az archiválási metaadatok, ezek leíró és szerkezeti típusai, a minőség ellenőrzése; a műszaki infrastruktúra, végül a szervezési infrastruktúra.

A kiválasztás sok kérdést vet fel. Tisztázni kell azonban, hogy a digitalizálás fogalma nem azonos az archiválást célzó (digitális) átalakítással. Az első kérdés mindjárt az, hogy minden digitalizált anyagot meg kell-e őrizni hosszú távon? Közismertek az olyan adatok, hogy a gyűjtemények 10–20%-a adja a cirkulált tételek 80–90%-át. Nyilván, a korlátozott forrásokat célszerű a használati gyakoriság szerint felhasználni. De vajon a digitalizálás, és ezért a könnyebb hozzáférés, nem módosítja-e a használat gyakoriságát? Bár az említett nagy projektekben ritkán szerepelnek értékes, ritka vagy speciális gyűjteményrészek, és a könyvdigitalizálás módszerei is finomodtak az elmúlt időszakban, mégis vizsgálni kell, hogy fenyegeti-e sérülés a kiválasztott tételeket.

Mint a digitalizálásnál mindig, most is gondosan kell vizsgálni a jogvédelemhez kapcsolódó kérdéseket.

A digitális archiválás számos döntést és előkészületet feltételez. Szabványos elvek és eljárasmódok alapozzák meg a sikeres programokat. A futó LSDI-k legtöbbje nem felel meg ezeknek a kritériumoknak. A tanulmány táblázatokban foglalja össze egyrészt azokat a folyamatokat, amelyeket végig kell járni a digitális archiválás során, másrészt az elemzett projektekben elfogadott műszaki specifikációkat. Emellett itt is bemutatja a legfontosabb vitatott kérdéseket. Az például egy évtized óta elfogadott szakmai megállapodás, hogy archiválási célra a veszteségmentesen tömörített TIFF formátumot használjuk. A TIFF-nek számos előnye van, de 1992 óta nem változott. A nagyméretű digitalizálás programjaiban azonban sokan fordulnak a JPEG2000 felé, amely ISO szabvánnyá vált, és sok előnyös műszaki tulajdonsága van, amellett, hogy a metaadatok beépíthetők a fájlba. Lassú terjedésének az a fő oka, hogy a böngészők nem támogatják megfelelő módon.

Kritika illeti azt a gyakorlatot, hogy ma is a kilencvenes évek technikájával hajtjuk végre a digitalizálást. Példaként említi a rögzített értékű felbontást és bitmélységet, anélkül, hogy a digitalizáló eszköz minőségi paramétereit tesztelték volna. Vizsgálatok mutatják, hogy azonos felbontással két különböző szkennelvel végzett digitalizálás különböző képeket mutathat. Nem szabad eltekinteni a be rendezések gondos tesztelésétől.

Külön alfejezetek foglalkoznak az archiválási metaadatokkal. Ezek létrehozási stratégiáival a *PREMIS (PREservation Metadata: Implementation Strategies)* dokumentum foglalkozik. Felhasználható még a digitális képekre vonatkozó, 2006-ban kiadott *ANSI/NISO Z39.87 (Technical Metadata for Digital Still Images)* szabvány is, de a munkákat megkönnyítő műszaki metaadatok szerepe még nem jól dokumentált.

A futó programokban a keresést és dokumentumkezelést elősegítő leíró metaadatok a helyi OPAC-ok bibliográfiai rekordjaiból származnak, míg a strukturális metaadatokból csak kevés található. Törekvés látható arra, hogy globálisan egységes állandó azonosítókkal lássák el a digitalizált könyveket.

A minőségirányítási alfejezet áttekinti az egész munkafolyamat során teendő ellenőrzési és szabályozási lépéseket, és azt veti fel az egyik legfontosabb kérdésként, hogy az ilyen nagy tömegű digitalizálási munkában lehetséges-e teljes körű ellenőrzés, mert ez nagyon idő- és költségigényes fel-

adat. Bár bizonyos technikai elemek automatikusan is ellenőrizhetők, mégis emberi vizsgálatra marad a paraméterek többsége. A vizsgált projektek eltérő módszereket követnek, és például a Google-kezdemenyezés kapott már kritikákat a digitalizálás minősége miatt.

A műszaki infrastruktúrával foglalkozó rész számos sérülékeny elemet említ meg, és kiemeli a tároláshoz kapcsolódó kérdéseket. A technikai fejlődés miatti avulás is sok problémát okoz, mégis a könyvtárak számára – a felmérések szerint – a hatalmas tömegű anyag kezelése jelenti a nagyobb gondot.

A tanulmány ismerteti a *Getty Research Institute* nemrégiben készült felmérését, amely szintén arra a következtetésre jutott, hogy az ilyen programokban részt vevő könyvtáraknak újra kell gondolniuk az infrastruktúra- és adattárolási modelljeiket.

Egyedül a technológia azonban nem tudja megoldani az archiválási problémákat, az intézményi politikák, stratégiák és finanszírozási modellek is fontosak. Bár egy évtized óta folytatnak vitákat különböző fórumok a digitális archiválás mikéntjéről, alig maroknyi könyvtárnak van olyan programja, amely támogatná a nagy tömegű archiválást. A tanulmány felsorolja az alapvető szervezési követelményeket, és röviden bemutat néhány szabványt és ajánlást, mint az *Open Archival Information System (OAIS)*, a *Trustworthy Repositories Audit & Certification (TRAC)*, a *Digital Repository Audit Method Based on Risk Assessment (DRAMORA)*, és a *Defining Digital Preservation*.

A negyedik fejezetben a tanulmány megkísérli előrejelezni az LSDI-k hatását a nyomtatott gyűjteményekre.

Mivel sok könyvtár súlyos helygondokkal küzd, valószínűleg nyomásnak lesz kitéve, hogy a digitalizálás után a nyomtatott eredeti példányokat vegye ki az állományból. Azonban a szakirodalom arra is rámutat, hogy a digitális állományok hatására még növekedhet is az igény a nyomtatott változatokra. Az olvasók közül sokan maradnának inkább ezek mellett, különösen a testesebb könyvek esetében. Történtek már kísérletek a „nyomtatás igény esetén” típusú megoldásra, persze csak a szerzői jogszabályok betartásával.

Az ötödik fejezet ajánlásokat fogalmaz meg az archiválási stratégiákhoz, amelyeknek az az első

rendű célja, hogy segítse az LSDI-ben érdekeltet közötti megbeszéléseket.

A 13 ajánlás öt téma köré csoportosítható:

1. A digitalizálás, mint archiválási módszer. Újra kell gondolni az archív képekre vonatkozó másfél évtizede született digitalizálási követelményeket. Új módszerek, új műszaki megoldások, új olvasói igények születtek azóta. Hajlékony minőségbiztosítási programot kell kidolgozni. Kompromisszumot kell keresni az archiválás és a hozzáférés követelményei között.
2. Tartós hozzáférés. Javítani kell a hozzáférést a digitalizált tartalomhoz, mind a hagyományos módon, mind a legújabb technikák használatával. A hosszú távú használat feltétele az is, hogy világosan értsük a szerződéses korlátozások hatását az archiválási felelőségekre, arra, hogy mit tehetünk és mit nem azokkal a digitalizált anyagokkal, amelyek a kereskedelmi partnerekkel közösen készültek.
3. Az archiválás menedzsmentje. Újra kell gondolni a nyomtatott állomány tárolását, különösen a ritkán használt vagy redundáns tételekét. Meggondolandók a földrajzilag megosztott tárolás lehetőségei. Támogatni kell a digitális mesterpéldányok regiszterének használatát.
4. Digitális archiválási stratégiák. Körvonalazni szükséges egy archiválási akciótervet a nagyméretű digitalizálási kezdeményezések számára. Miután minden könyvtári anyagot nem lehet digitalizálni, ki kell dolgozni a digitális archiválás szintjeit kijelölő elveket.
5. A kutatókönyvtárak stratégiái. A könyvtáraknak szemmel kell tartani és egymás között meg kell osztani a költséginformációkat. Újra kell gondolniuk könyvtári prioritásaikat és stratégiáikat. Gyorsabban kell válaszolniuk a kihívásokra, és el kell mozdulniuk egy mozgékonyabb és nyílt tervezési modell felé. Újra kell gondolniuk állományfejlesztési elveiket.

Végül a hatodik fejezet következtetései döntően a kulturális intézmények együttműködésének szükségességéről szólnak. Egyetlen könyvtár sem képes a Google vagy a Microsoft léptékű vállalkozás végrehajtására, azonban másokkal együtt megfelelhetnek egy ilyen típusú projekt kihívásainak. Az erők egyesítésének van helye, ha ez költséghatékonyabb és minden együttműködő partner nyere a közös akcióban.

A kulturális intézmények azonban a tudományos örökség kezelői, és erről nem mondhatnak le akkor sem, ha a kereskedelmi érdekek számukra is nyújtanak bizonyos előnyöket. Alapvető szempontnak

kell maradnia a tartós hozzáférés fenntartásának és a költséghatékonyságnak.

A kutató-könyvtáraknak módosítani kell szerepüket és programjaikat azért, hogy kielégíthessék a 21. század olvasóinak igényeit.

A tanulmány jellegénél fogva széles olvasóközönségnek szól a digitalizálásban részt vevő intézmények munkatársaitól a gyűjteményeket használó kutatókon át a támogatást nyújtó alapítványok vagy kormányzati szervek vezetőiig. Az anyag

komoly értéke, hogy gondos és terjedelmes lábjegyzeteiben sok-sok hivatkozás található, amelyek a további tájékozódást megkönnyítik az érdeklődő olvasónak.

/RIEGER, Oya Y.: Preservation in the age of large-scale digitization. A white paper. =

<http://www.clir.org/activities/details/lodi.pdf/>

(Horváth Péter)