A PubMed Central archívuma és a visszamenőleges szkennelés projektje

A *PubMed Central* (*PMC*: http://www.pubmedcentral.gov) az USA *Nemzeti Orvostudományi Könyvtár*ának (*National Library of Medicine = NLM*) élettudományi folyóiratokat tartalmazó archívuma. A 2000-ben az elektronikus folyóiratok letéti helyeként létrehozott adatbázist a *Nemzeti Biotechnológiai Információs Központ* (*National Center for Biotechnology Information*) munkatársai kezdték építeni és tartják karban. A PubMed Central ingyenesen és minden korlátozás nélkül elérhető. A kiadók önkéntes alapon csatlakozhatnak, de bizonyos szabványoknak eleget kell tenniük.

2002-ben indult a visszamenőleges szkennelési projekt azokkal a folyóiratokkal, amelyek legújabb számaiból a tartalomjegyzéket a PMC megkapta. A digitalizálási program finanszírozását az NLM vállalta. Meg kellett találni a szkennelést végző vállalkozást, el kellett kezdeni a folyóiratok visszamenőleges gyűjtését, ki kellett dolgozni a minőségi követelményeket (*Conversion System-Design Document = CSDD*), meg kellett tervezni a dokumentumok leírásához szükséges XML elemeket stb. A kezdeményezésben részt vállaló kiadóknak két szerződést kellett aláírniuk: az egyik a legfrissebb számok tartalomjegyzékének a PMC-be történő eljuttatására vonatkozott, a másik a visszamenőleges szkennelésre. Ez utóbbi a biztosítéka annak, hogy az NLM-be eljuttatják a szkennelt folyóirat egy teljes, eldobható számát. Mivel a szkenneléskor a folyóiratokat szét kell szedni, az NLM nem tudja visszaküldeni őket.

A projekt egyik első résztvevője az *Amerikai Mikrobiológiai Egyesület* volt, ennek köszönhetően folyóirataik teljes egészükben bekerültek az adatbázisba, például a *Journal of Bacteriology* 1916-tól, vagy a *Bulletin of Medical Library Association* indulásától, 1911-től. 2004-ben az NLM együttműködési szerződést kötött az Egyesült Királyságban működő *Welcome Trust*tal és a *Közös Információs Rendszerek Bizottságá*val (*Joint Information Systems Committee = JISC*) a feldolgozandó és szolgáltatandó folyóiratok körének bővítésére. Ennek eredményeképpen számos fontos folyóiratot digitalizáltak és tettek a PMC-n ingyenesen elérhetővé.

A digitalizálási projekt fő célja a teljes eredeti folyóirat digitalizálása és kereshetővé tétele volt. Ehhez képkategóriákat kellett meghatározni a folyóiratban található különböző tartalmú oldalakhoz, mint a tartalomjegyzék, a borító, az adminisztratív anyagok (felhívás rendezvényekre, szerzőknek szóló útmutató stb.), hirdetések (ha vannak), cikk (a CSDD-előírások szerint).

Azokhoz a cikkekhez, amelyek bibliográfiai adatai még nem voltak meg a PubMed/Medline adatbázisban, az NLM munkatársai készítették el az XML rekordokat. Minden bibliográfiai tételhez egy fájlcsomag tartozik a következőkkel:

- minden oldalról 600 dpi felbontású, fekete-fehér TIFF formátumú fájl;
- optikai karakterfelismerővel készült szövegfájl (ASCII, nem szerkesztett) a kereséshez és a hivatkozások összekötéséhez;
- 300 dpi felbontású színes vagy szürke árnyalatokban megjelenő képek TIFF formátumban;
- az NLM-ben készült pdf fájl.

A munka legnehezebb és legköltségesebb része az XML rekordok összeállítása, mivel még egy olyan egyszerű adat, mint a cikk típusának (szerkesztőségi, könyvszemle, olvasói levél stb.) pontos meghatározása is lényegesen befolyásolja a cikk részeinek jelölését és megjelenítését. A bibliográfiai leírásban szereplő mezők (szerző, cím, lábjegyzetek stb.) jelöléséhez a CSDD is tartalmaz előírásokat.

A minőségbiztosításhoz az NLM-ben egy olyan összetett rendszert dolgoztak ki, amely lehetővé teszi a szkennelt oldal megjelenítését és hibajelentés készítését. A kiadványok 5%-ából véletlensze-

rűen összeállítanak egy csomagot úgy, hogy abban minden fájltípusból legyen. Az ellenőrzést ezen a csomagon manuálisan végzik, összevetve az eredeti folyóiratoldalt és a szkennelt képet. Az ellenőrzés a cikk teljességére, az XML adatok pontosságára, a képek élességére, a színfelbontásra és az OCR teljességére terjed ki. A hibák számától függően az ellenőrzés után a csomag *Elfogadott* vagy *Visszautasított* státuszba kerül. Egy csomag csak egy folyóiratcímet tartalmazhat, terjedelme általában 3000 oldal. Az elfogadási kritérium minden kategóriában 99–100%. A végleges döntést az ellenőrzés második szintjén a minőség-ellenőrzés utáni vizsgálat eredményeként hozzák meg. Ezután dolgozzák fel az egyes csomagokat a webes megjelenéshez. Mielőtt a folyóiratot „élővé" tennék, a digitalizált változatot jóváhagyásra elküldik a kiadónak.

A visszamenőleges állomány építése mellett a kiadók folyamatosan küldik a legújabb számok tartalmát, és teszik a megjelenéstől számított 6–24 hónapon keresztül ingyenesen elérhetővé. A teljes szövegű tartalom előállításához több kiadó saját dokumentumtípus-meghatározást használ, míg mások az NLM által kidolgozott definíciókat.

A PubMed Central kiadói statisztikája szerint – nem meglepő módon – a leggyakrabban a legújabb számokból töltik le az oldalakat. A statisztika az egyes címekhez a folyóiratonkénti megoszlást is tartalmazza.

/FISHEL, Martha–MYERS, Carol J.: The PubMed Central Archive and the back issues scanning project. = Journal of Interlibrary Loan, Document Delivery & Electronic Reserve, 17. köt. 3. sz. 2007. p. 109–116./

(Viszocsekné Péteri Éva)

ENRICH

A kulturális örökségre vonatkozó információforrások európai hálózata. EU projekt 2007. december−2009. november közötti időtartammal

Az *ENRICH* (gazdagítás) címmel indított ún. célzott projektet az *Európai Unió* eContentPlus, a digitális formában rendelkezésre álló információk nemzetközi hasznosításának továbbfejlesztését támogató programja keretében finanszírozza. A projektben Magyarországot a *BME Országos Műszaki Információs Központ és Könyvtár* (*OMIKK*) képviseli, és további egyetemi könyvtárak bekapcsolódására is számítani lehet.
A projektet 2007. december 3-án, a Prágában tartott nemzetközi értekezlettel indították útjára a részt vevő országok képviselői.

A projekt célja, hogy Európa különböző kulturális intézményeiben fellelhető kézirat- és ősnyomtatvány vagyon digitalizált formában elérhető részéhez egységes és hatékony hozzáférést biztosítson anélkül, hogy a felhasználónak foglalkoznia kellene az egyes rendszerek sajátosságaiból adódó különbségekkel. Más szóval, a projekt egy közösen használható virtuális gyűjteményt kíván létrehozni, egyrészt a kutatók, másrészt a kulturális kérdések, tudománytörténet, irodalomtörténet stb. iránt érdeklődők széles köre számára. Ez konkrétan azt jelenti, hogy a projekt az európai nemzeti könyvtárak eddig digitalizált kézirat/ősnyomtatvány/régi illetve ritka könyvállományának mintegy 85%-át egységesen és közvetlenül hozzáférhetővé kívánja tenni az interneten keresztül. Ezt a szétszórtan már rendelkezésre álló digitális gyűjteményt a jövőben további értékes anyaggal egészítik ki a részt vevő országok egyetemi és egyéb könyvtárai. A konzorcium végeredményben mintegy 5 millió digitalizált oldal tartalmához kíván hozzáférést biztosítani.

A projekt a prágai *Nemzeti Könyvtár* által kialakított „Manuscriptorum" digitális könyvtár eddigi tapasztalataira és anyagára épül, mely a http://www.manuscriptorium.eu honlapon érhető el. Ez jelenleg 46 cseh és külföldi gyűjtemény digitalizált változatához biztosít hozzáférést, és 15 éves fejlesztői munka eredményeként jött létre, melyet a Cseh Köztársaság nemzeti könyvtára és az *AIP Beroun Ltd*. cég együttműködve hajtott végre. Ez jelenleg a leggazdagabb digitalizált kézirat-gyűjtemény Európában, mely már 1 millió oldalnyi digitalizált anyagot tesz hozzáférhetővé, és biztonságos digitális archívummal rendelkezik. A digitalizálással kapcsolatos munkát a cseh állam támogatta. A felhasználók, akiknek kb. 50%-a a Cseh Köztársaságon kívülről származik, cseh és angol nyelven kereshetnek. A rendszer egy változata a középiskolák oktatási tevékenységét is segíti. A kezdeményezés létrejöttében nagy szerepe volt az *UNESCO* „Világme-

mória" című programjának, amiért a cseh Nemzeti Könyvtárat az UNESCO 2005-ben a *Jikji-díjjal* jutalmazta. A Manuscriptorum létrehozatalával kapcsolatos munka tapasztalatait, az abból adódó ismereteket azóta több más ország hasznosította.

Az ENRICH projekt eredményeként a jelenleginél sokkal több adat válik hozzáférhetővé Európa számos részéből. A dokumentumokat leíró ún. metaadatokat a központi adatbázis céljaira a projekt a nemzetközi *OAI* (*nyitott archívum*) protokoll alkalmazásával fogja összegyűjteni. A dokumentumok leírását olyan kapcsolati adatok egészítik ki, melyek a leírást összekötik a részt vevő intézmények adatbázisaiban tárolt képekkel. A szükséges átalakítások elvégzése érdekében a projekt minden részt vevő intézménynél megfelelő számítógépes programokat kíván telepíteni.

Az ENRICH útján kiszolgálni kívánt felhasználói körbe egyrészt maguknak a dokumentumoknak a tulajdonosai, másrészt könyvtárak, múzeumok és archívumok, kutatók és hallgatók, politikusok és általában a kulturális múlt iránt érdeklődők tartoznak. Ez a projekt lehetővé teszi számukra az érdeklődési körükbe vágó olyan dokumentumok keresését és elérését, amelyekhez más módon nehezen férnének hozzá. Emellett a rendszer történelmi dokumentumok teljes szövegének, audio- és videoanyagoknak, illetve számos történelmi térképnek elérését is lehetővé teszi. Az ENRICH konzorcium szoros együttműködést tervez az *TEL*-el (*Európai Könyvtár – The European Library*) és az *Európai Digitális Könyvtár* alkotóelemévé fog válni, amint az megvalósul.

A felhasználók számára olyan eszközök állnak rendelkezésre, melyek lehetővé teszik, hogy létrehozzák saját dokumentumaikat és digitális könyvtáraikat a Manuscriptoriumban. Ez az alkalmazás több nyelven biztosít hozzáférést a Manuscriptoriumhoz, és – éppúgy mint a többnyelvű ontológiák – engedélyezi a keresést egy adott felhasználói nyelven és az adatok visszanyerését a forrás nyelvén.

Az ENRICH konzorcium 18 partnerből áll, és a projektet számos egyéb intézmény támogatja.

A projektet a Cseh Nemzeti Könyvtár (National Library of the Czech Republic) két cseh partnerrel − az AiP Beroun Ltd. céggel és a *Crossczech Prague Inc*. céggel közösen irányítja. Az első két cseh intézmény mellett egyes feladatcsoportok tekintetében vezető szerepet tölt be az *Oxford University Computing Services*, az *Università degli Studi di Firenze – Centro per la comunicazione e l'integrazione dei media*, az *Institute of Mathematics and Informatics* Vilnius-ban, a *SYSTRAN Paris* és a *National Library of Spain*. További fontos technikai partnerek: *Københavns Universitet – Nordisk Foskningsinstitut*, *Biblioteca Nazionale Centrale di Firenze*, *University Library Vilnius*, *University Library Wrocław*, *Stofnun Árna Magnússonar í íslenskum fræðum Reykjavík*-ban, *Computer Science for the Humanities – Cologne University*, *St. Pölten Diocese Archive* (*Monasterium project*, Ausztria), *National and University Library of Iceland*, *Budapesti Műszaki és Gazdaságtudományi Egyetem* és a *Poznań Supercomputing and Networking Centre*.

Az együttműködés iránti érdeklődést kifejezték további nemzeti könyvtárak, nevezetesen Magyarországon, Kazahsztánban, Moldovában, Lengyelországban, Romániában, Szerbiában és Törökországban éppúgy, mint a pozsonyi, bukaresti és heidelbergi könyvtárak. A tagok listája a projekt időtartama alatt remélhetőleg tovább fog bővülni.

(BME OMIKK)