Drótos László

Mi a MIA?

Javaslat egy Magyar Internet Archívum létrehozására*

Hogy miért fontos a digitális kultúra megőrzése, azt – remélhetőleg – már nem nagyon kell magyarázni. Látjuk, érezzük, hogy mennyire meghatározó lett egy évtized alatt nálunk is az internet, mennyi mindenhez már csak ott férhetünk hozzá; és látjuk, érezzük, hogy rohamosan növekszik az online tartalom – és rohamosan pusztul is. Nemcsak egy-egy láncszem (link) törik el itt-ott, és vezet egy „404-es hibaoldalhoz", hanem egész webhelyek – oldalak ezrei – tűnnek el nyomtalanul egyik napról a másikra, vagy válnak fokozatosan a digitális entrópia áldozatává. Becslések szerint egy weblap átlagéletkora nagyjából egy házilégy egy hónapos élettartamával egyezik meg. És miközben országos felháborodást vált ki, ha a rossz tárolási körülmények miatt meg kell semmisíteni a Nemzeti Könyvtárban őrzött kisnyomtatványok egy részét, vagy csőtörés miatt eláznak régi folyóiratok, a digitális „nyomtatványok" megmentését és megőrzését senki nem érzi feladatának.

Felelősség

Az UNESCO Közgyűlésének 32. ülésszakán, 2003. október 17-én elfogadott „Charta a digitális örökség védelméről" (*http://www.unesco.hu/index.php?type=node&id=508*) világosan megfogalmazza az ezzel kapcsolatos feladatokat, és külön kiemeli, hogy a válogatott archiválás esetén a „born digital"[1] anyagoknak prioritást kell adni. Ez a dokumentum meghatározza a felelősöket is: „A digitális örökség megőrzése a kormányok, alkotók, kiadók, releváns iparágak és az örökségvédelmi intézmények kitartó erőfeszítéseit igényli."

A nemzeti könyvtáraknak természetesen kiemelt felelősségük van, bár nagy kérdés, hogy a hálózaton levő rengetegféle digitális objektumból mi tekinthető a gyűjtőkörük részének? Ezen a téren nagyon eltérő a nemzetközi gyakorlat: van, ahol csak a *kiadvány*nak minősíthető dokumentumokat gyűjtik ezek a könyvtárak (azokat is önkéntes letét vagy köteles példány alapon), van, ahol viszont az egész nemzeti webteret learatják időről időre, és elteszik valami tartósabb tárolóra. Magyarországon 1994-ben egy-két felsőoktatási könyvtárban kezdődött meg a digitális dokumentumok gyűjtése és archiválása, a *Magyar Elektronikus Könyvtár* (*MEK*, *http://mek.oszk.hu*) nevű kezdeményezéssel. A MEK-projektet 1999-ben az *Országos Széchényi Könyvtár* vette át, és azóta egy néhány fős önálló osztály is kialakult mögötte. A MEK a monografikus jellegű, nyilvánosan szolgáltatható, alapvetően szöveges digitális hungarikumok gyűjtését vállalta fel, vagyis lényegében erősen válogatott könyvgyűjtemény – beleértve most már az MP3 hangoskönyveket is. Az osztály munkatársai 2003-ban elindították az *Elektronikus Periodika Archívum és Adatbázis* (*EPA*, *http://epa.oszk.hu*) nevű szolgáltatást is, amelynek adatbázis része teljességre törekedve tartja nyilván a magyar vonatkozású, online vagy offline formában létező elektronikus időszaki kiadványokat; archívum része pedig elsősorban folyóiratokat és hírleveleket ment le, őriz és szolgáltat. A MEK nyitólapjáról elérhető „kiállítóteremben" (*http://mek.oszk.hu/html/kiallitas.html*) teljes honlapok archiválására is akad néhány példa (természetesen készítőik beleegyezésével, sőt kifejezett kérésére): „Erdélyi és csángó költészet", „Váli Dezső oeuvre", „Vészi Endre honlapja", „Lénárd Sándor honlapja". Van tehát már némi gyakorlatunk abban, hogy hogyan lehetne a magyar webnek legalább a legértékesebb részét elmenteni és használhatóvá tenni mind a jelen, mind a jövő számára. Ez a tapasztalat arra már elegendő, hogy lássuk, milyen nagy és komplex ez a feladat, mennyi válogatási, feldolgozási, műszaki és jogi problémát vet fel, és hogy egyetlen intézmény vagy intézménytípus önmagában nem tudja meg-

* A Networkshop 2006 (Miskolc, 2006. április 19–21.) konferencián elhangzott előadás szerkesztett változata. Az előadás videofelvétele megnézhető: *http://vod.niif.hu/index.php?lg=hu&mn=archive&eid=42&sm=listevent &secid=74*

oldani egy *Magyar Internet Archívum* (rövidítsük egyelőre *MIA*-ként) létrehozását és fenntartását. Ezért az OSZK MEK osztálya javasolja egy konzorcium alakítását, amelyben a közgyűjtemények mellett informatikai intézmények és cégek is társulnak a MIA megvalósítása érdekében.

A továbbiakban nézzük át, hogy miből áll, hol tart máshol ez a munka, és hogy mi – eddigi tapasztalataink alapján – mit tartunk reálisan járható útnak itt és most.

Technikai vonatkozások

Az „internetarchívum" vagy „webarchívum" kifejezés többféle dolgot is jelent a szak- és köznyelvben. A továbbiakban komplett webhelyek, szolgáltatások időszakosan ismétlődő mentésével létrejövő és a hosszú távú megőrzésre szánt másolatot értem ezen. Nem tekintem idetartozónak a MEK-hez vagy az EPA-hoz hasonló, egyedi dokumentumok vagy egyedi kiadványok mentéseit, illetve a webes keresőrendszerekhez indexelés céljára begyűjtött, ideiglenesen archivált állományokat.

A teljes szolgáltatások mentése *kétféle* archívumot eredményezhet.

1. Fájlrendszerbe való mentés
Ennél a megoldásnál valamilyen letöltő program segítségével (HTTP vagy FTP protokollon át) egy másolat készül egy adott webhelyről, amennyire csak lehet megőrizve annak eredeti arculatát és szerkezetét. A mentés során a belső hivatkozásokat relatív útvonalakra kell konvertálni, így a másolat ugyanúgy navigálható maradhat, mint az eredeti. Természetesen a szolgáltatói oldalon futó szkriptek nem biztos, hogy működnek az archív szerveren is. Ha adatbázis van az eredeti szolgáltatás mögött, akkor a teljes szoftver- és operációsrendszer-környezetet meg kell teremteni az archiváló gépen a funkcionalitás megőrzéséhez, ami költséges és időigényes feladat. A csak statikus HTML dokumentumokból álló honlapok könnyen és jó minőségben elmenthetők így, de ezeknél is szükség van egy fájlmenedzsment rendszer kialakítására, a rohamosan szaporodó állományok nyilván- és karbantartása érdekében. A tárhelykímélés céljából hasznos duplikátumszűrés is komoly probléma, ez szintén a belső ugrópontok átalakítását igényli. Ez a technika rosszul skálázható, tömeges és gyakori mentésre nem alkalmas, de a felhasználók számára jól böngészhető, valódi „webmúzeum" érzését kelti.

2. Adatstruktúrába való mentés
Ez esetben egy harvester (szüretelő) vagy crawler (portyázó) robotot alkalmaznak, amely akár egy teljes felső szintű domén tartalmát is le tudja tölteni. A begyűjtött anyag egységes szerkezetű (pl. XML-re konvertált, metaadatokkal ellátott és tömörített) archív állományokba kerül, majd adatbázist és indexeket készítenek hozzájuk. Az archivált anyag egyes részeinek elérése vagy *URI* (*Uniform Resource Identifier = egységes forrásazonosító*) alapján, vagy teljes szövegű kereséssel történhet. A kikeresett weblapról a további navigálás csak nagyon korlátozottan lehetséges az archívumon belül, vagy a belső ugrópontok az eredeti forrásra visznek tovább (ha az még létezik). Egy ilyen archívum felépítése komolyabb műszaki feladat, viszont jól skálázható, hatalmas mennyiség gyűjthető be, szinte teljesen automatizáltan. Azoknak a felhasználóknak jó, akik böngészés helyett szeretnének célzottan keresni egy nagy archívumban.

Az ismertetetteken kívül másféle felosztás is lehetséges. Besorolhatjuk például az internetarchívumokat:
- médiatípus szerint: web-, newsgroup-, sugárzott multimédia-gyűjtemények;
- válogatási szempontok szerint: teljes nemzeti webteret vagy egyéb nagyobb domént gyűjtő, vagy csak minőségi webhelyeket, vagy nagyobb témaköröket, illetve csupán néhány témát vagy egy-egy eseményt gyűjtő archívumok;
- a gyűjtés iránya szerint: pulltípusú lementés, illetve pushtípusú önkéntes vagy kötelező feltöltés.

Ezeknek különböző variációi is elképzelhetők, és vannak is már működő példák.

Egy webarchívum több szoftverkomponensből áll össze, *kulcsrakész* rendszerek nem léteznek ezen a téren; a már működő szolgáltatásokat részben saját fejlesztésű, részben kész – többnyire nyílt és ingyenes – elemekből rakták össze. Az alapszintű működéshez a letöltő, illetve szüretelő programok mellé kell egy adatbázis-kezelő a metaadatok tárolására, egy teljes szövegű indexelő és kereső, valamint egy szolgáltatási felület. Az egész munkafolyamat kézben tartására és a minőségbiztosítás céljára pedig ki kell alakítani egy menedzsment keretrendszert.

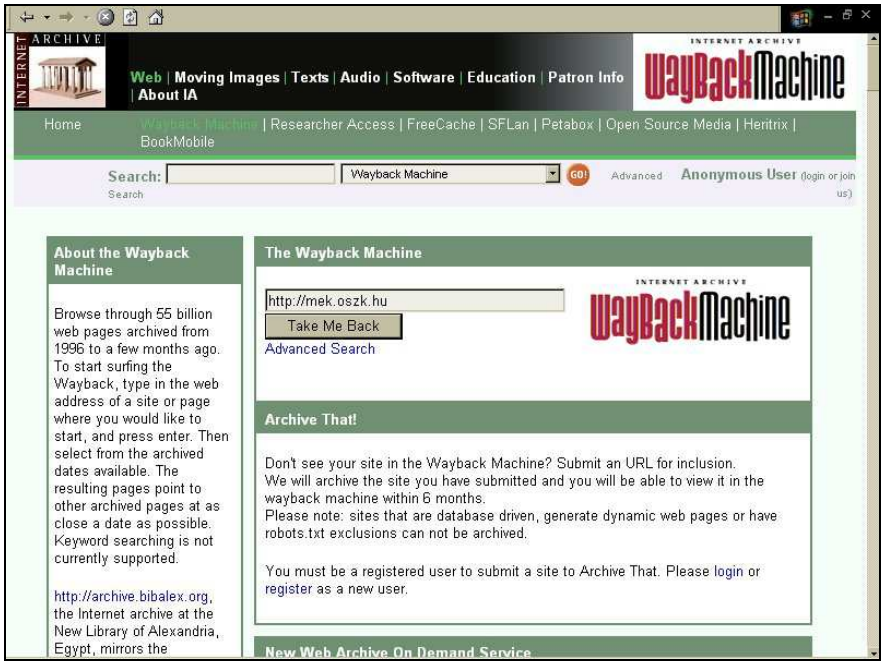
Az archiválás során rengeteg technikai problémával szembesül az üzemeltető: az internet – és azon belül a web önmagában is – nagyon bonyolult, gyorsan változó, nehezen megőrizhető médi-

um. Hosszú távú és mindenre kiterjedő megoldás (a papírra nyomtatáson kívül) még sehol sem létezik. Nagyon nagy szükség volna egy olyan W3C ajánlásra, amely – az online szolgáltatások akadálymentesítéséhez hasonlóan – javaslatokat tartalmazna arra, hogyan kell egy webhelyet úgy kialakítani, hogy az könnyen és hosszú távon is archiválható legyen. Megoldás lehet, hogy az archiváló robotok számára egy alternatív verziót „exportál" a honlap üzemeltetője (pl. statikus HTML lapok formájában); a nyilvános szolgáltatást nem kell átalakítani emiatt.

Külföldi példák

A világban egy-két tucat ilyen nyilvános szolgáltatás, illetve projekt létezik. Ezek jellegükben és méretükben is erősen különböznek. Jellemző, hogy sok köztük a pilot jellegű próbálkozás, amelyek egy előre lehatárolt, néhány éves időszakra terjednek ki, kevés a már üzemszerűen működő, valóban nagy archívum. Ugyancsak jellemző mostanában az archívumok integrálódása, a közös szabványokra, technológiákra és a munkamegosztásra való törekvés.

A legrégibb és leghíresebb kezdeményezés természetesen az *Internet Archive* (*IA*, http://www.archive.org) nevű nonprofit szervezet San Franciscóban, amely a webarchiváláson mint alapcélon túllépve, a *könyvtár* fogalmát a legszélesebb értelemre kiterjesztve, a digitális objektumok *Alexandriai Könyvtárává* szeretne válni. Legismertebb szolgáltatásuk az *Alexa Internet* cég (jelenleg az *Amazon.com* tulajdona) által összegyűjtött anyagra épülő *Wayback Machine* (*1. ábra*). Ez 2006 februárjában mintegy 55 milliárd weboldalt tudott előkeresni URL alapján. Most folyik a *Nutch* nevű teljes szövegű kereső beépítése. A szüretelés 1996-ban indult, és elvileg a teljes nyilvános webre kiterjed, de a népszerűbb honlapokat gyakrabban begyűjtik (az átlaggyakoriság 2 hónap, az átlagnövekedés havi 20 terabájt), az anyag hat hónap késéssel válik nyilvánossá, az archívum egyes részei azonban csak kutatóknak érhetők el. Tipikusan a második csoportba tartozó rendszer: a mintegy 1 petabájtnyi anyag megfelelő részének előkeresése az archív fájlokból meglehetősen hosszú válaszidőket eredményez, sok a hiányzó objektum, úgyhogy a pontos URL ismerete és némi szerencse is kell ahhoz, hogy hiánytalanul megtaláljunk egy régi weboldalt. Az IA együttműködő partnerei között van az amerikai nemzeti könyvtár is, és egyik alapító tagja a 2003 nyarán létrejött *IIPC*-nek (*International Internet Preservation Consortium, http://netpreserve.org*), melyet a *Bibliothèque Nationale de France* vezet. A jelenleg 12 tagú IIPC az internetarchiválás módszertanának kidolgozását koordinálja (*2. ábra*).

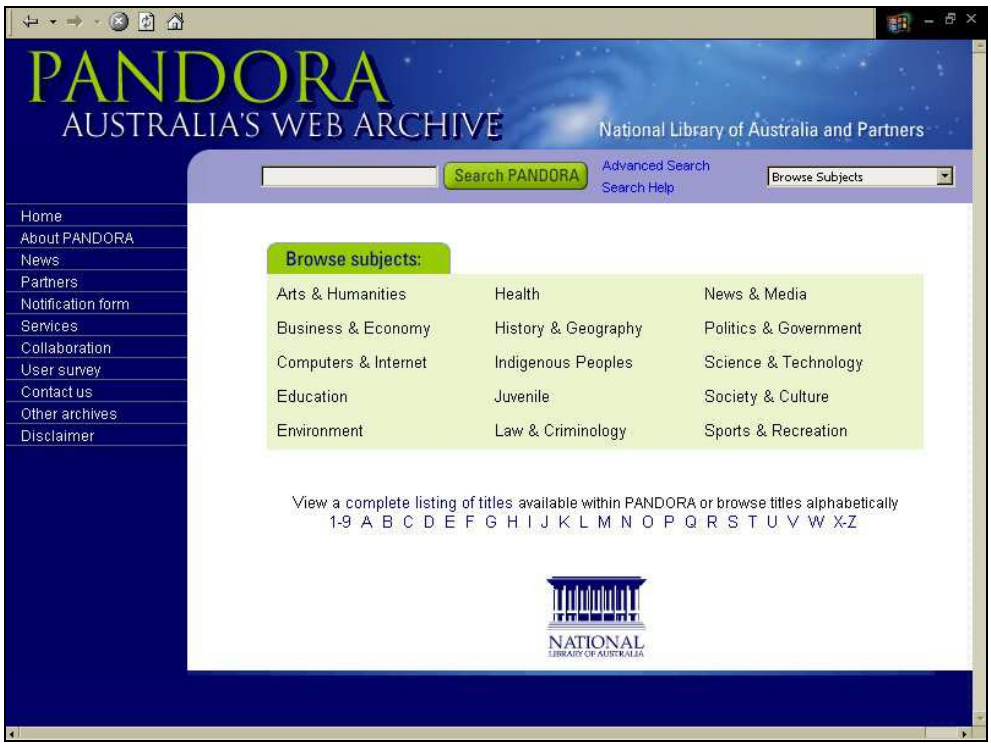


1. ábra **Az Internet Archive által működtetett „időgép", a Wayback Machine**

2. ábra **Az International Internet Preservation Consortium honlapja**

A skandináv országok nemzeti könyvtárai 2000 szeptemberében indították a *Nordic Web Archive* (*NWA*, *http://nwa.nb.no*) nevű projektjüket, amely 2002 júniusában zárult, és a *Nordunet2* informatikai programból finanszírozták. Az egyes országok webtereinek kísérleti jellegű archiválása mellett több nyílt forráskódú eszközt is kifejlesztettek a webarchiválás céljára, amelyeket átadtak más országoknak is (pl. Csehország, Észtország, Litvánia). Az NWA az Internet Archive-hoz hasonló második típusú technológia: előbb a *NEDLIB* harvesterrel kísérleteztek, majd áttértek az Internet Archive által is használt *Heritrix* programra. A begyűjtött objektumok automatikusan készülő metaadatokkal együtt XML fájlokba kerülnek, ezeket indexelik, és egy *WERA* (*WEb aRchive Access*) nevű felületen lehet bennük keresni, URL cím vagy teljes szöveg alapján. Svédországban a *Royal Library* már 1996-ban foglalkozni kezdett a webarchiválással. Az első próbálkozás 1997-ben történt, ezt 2003-ig tíz alkalommal ismételték meg, és így 185 millió fájl (több mint 5,5 terabájt) gyűlt össze. Jelenleg már évi 2-3 alkalommal aratják le a svéd szervereket. 2003 óta nyilvános az archívum, a Wayback Machine-hoz hasonló hozzáférést tesz lehetővé. A finn webtér első archiválása 11,7 millió fájlt eredményezett 2002-ben (kb. 500 gigabájt), ez a szám 2003 októberére 15 millióra nőtt. Az északi országok 2003-ban szintén csatlakoztak az IIPC-hez, „apportként” felajánlva az NWA-hoz kifejlesztett szoftvereket, és az elmúlt években szerzett tapasztalataikat.

A *National Library of Australia* által irányított konzorcium *PANDORA* (*Preserving and Accessing Networked Documentary Resources of Australia*) projektjének (*http://pandora.nla.gov.au*) kezdetei 1996-ig nyúlnak vissza, és az első típusú archívumok közé tartozik *(3. ábra)*. A jól szervezett, munkamegosztáson alapuló rendszerben válogatott honlapokat mentenek le, amelyek 15 nagyobb témacsoport szerint böngészhetők, és teljes szöveggel is kereshetők. Saját fejlesztésű *PANDAS* (*PANDORA Digital Archiving System*) nevű rendszerük a félautomatikus munkafolyamat minden fázisát támogatja: az archiválásra kiválasztott URL-ek nyilvántartása, a mentés időzítése és indítása, a mentett anyag minőség-ellenőrzése és hibajavítása, metaadatok hozzárendelése, előkészítés a nyilvános szolgáltatásra, a hozzáférési korlátozások, statisztikák és jelentések összeállítása. Az első verzióban, 2001 júniusában elkészült PANDAS-t azóta kétszer is továbbfejlesztették, a harmadik változat megjelenését ez év első felére ígérik. 2006 januárjában mintegy 11 ezer honlap vagy dokumentum mentése volt az archívumban



3. ábra **Az Ausztrál Nemzeti Könyvtár PANDORA archívuma**

(a különböző időpontokban történt ismételt mentéseket is beleszámítva összesen 21,5 ezer tétel), ez összesen 29 millió fájlt és kb. egy terabájtot jelentett. A válogatás szempontja: ausztrál témájú vagy ausztrál szerzőjű, társadalmi, politikai, kulturális, vallási, tudományos vagy gazdasági témájú, hosszú távon is kutatásra érdemes anyagok, melyeknek az archiválására a készítőjük engedélyt ad. A szolgáltatás fontos része, hogy minden dokumentumot stabil URI-val látnak el, így azokra megbízhatóan lehet hivatkozni akkor is, amikor az eredeti helyükről már eltűntek. Az archivált anyagok katalógustételei a nemzeti bibliográfiai adatbázisba (*Kinetica*) is belekerülnek. Az ausztrál nemzeti könyvtár szintén az IIPC tagja, és létrehoztak egy tematikus honlapot is *PADI* (*Preserving Access to Digital Information, http://www.nla.gov.au/ padi*) néven a nemzetközi tapasztalatok összegyűjtése céljából.

Az Egyesült Királyság nemzeti könyvtárának első webarchiválási kísérlete 2001-ben volt, ekkor 100 brit történelmi és kulturális honlapot mentettek le, de a gyűjtemény nem vált nyilvánossá. Az angol *Web Archiving Consortium* (*http://www.webarchive.org.uk*) 2004 júniusában alakult hat nagy intézmény (köztük a *British Library*, a *National Archives*, a *Joint Information Systems Committee of the Higher and Further Education Councils*) összefogásával. Egy kétéves projektet indítottak, amelynek során mintegy hatezer webhely archiválását tervezik (2006 elején ebből kb. ezer már elérhető) a fontosabb tudományos, oktatási, kulturális és közéleti honlapok közül a szolgáltatók engedélyével. A tagok felosztották egymás közt a begyűjtendő anyagokat, hagyományos gyűjtőkörük és szakértelmük alapján. Az ausztrál PANDAS rendszert vették át és fejlesztik tovább saját igényeiknek megfelelően, letöltő programnak pedig a *HTTrack*et használják. 2003-ban a British Library is csatlakozott az IIPC-hez.

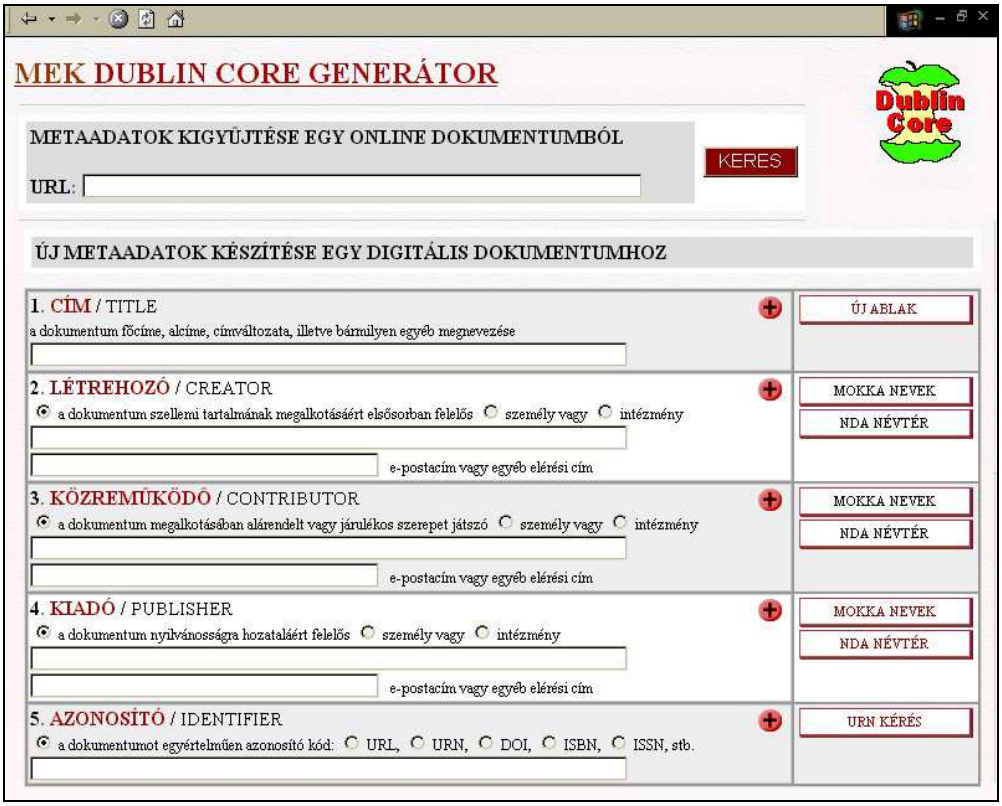
A *Library of Congress* 2000-ben indította a *Minerva* (*Mapping the Internet the Electronic Resources Virtual Archive*, *http://www.loc.gov/minerva*) projektet, első lépésben 35 webhely HTTrackkel való mentésével. Az Internet Archive-val és egyetemi intézményekkel együttműködve a 2000-es elnökválasztási kampány alatt már 200 honlapról készítettek napi mentéseket. Azóta több mint 35 ezer honlapot mentettek le, általában valamilyen eseményhez kapcsolódókat: pl. a 2001. szeptember 11-i terrortámadás, a 2002-es téli olimpia, az iraki háború. 2002 második felében 1,3 terabájtnyi anyagot gyűjtöttek össze. Az archívumnak csak egy része nyilvános, nagy hangsúlyt fektetnek a copyrightszabályok betartására. A metaadatok leírása a saját fejlesztésű, MARC-elemeket is tartalmazó, XML-alapú, *MODS* (*Metadata Object De-*

scription Schema) segítségével történik, az adatok a könyvtár katalógusába is bekerülnek. A Library of Congress a vezetője az *NDIIPP* (*National Digital Information Infrastructure and Preservation Program, http://www.digitalpreservation.gov*) együttműködésnek, amely az Egyesült Államokban folyó digitális archiválási tevékenységet koordinálja. Természetesen tagjai a nemzetközi IIPC-nek is.

A *Cseh Nemzeti Könyvtár* egy egyetemi partnerrel együttműködve 2000-ben kezdett egy kétéves pilot projektbe, és hozta létre *WebArchive* (*http://www.webarchiv.cz*) nevű szolgáltatását. A teljes nemzeti webtér begyűjtését célozták meg a NEDLIB harvesterrel. 2002-ben megismételték a szüretelést, és elindult a begyűjtött anyag integrálása a könyvtár online szolgáltatási felületébe, valamint a Cseh Nemzeti Bibliográfiába. A projekt mellékter-mékeként URN szervert, Dublin Core és MD5 checksum[2] generátort is beüzemeltek.

Javaslatok

Mint a fenti példákból is látszik, a nemzeti könyvtárak minden országban vezető vagy legalábbis kezdeményező szerepet játszottak az internet-archiválás elindításában. Az is látszik azonban, hogy szinte mindenhol partnereket kerestek maguknak ehhez a munkához: elsősorban informatikai intézményeket, egyetemi tanszékeket és/vagy cégeket. Nálunk is ez tűnik a leginkább járható útnak, ezért kellene egy konzorciumot létrehozni mindazoknak a szervezeteknek, amelyek érdekeltek, érintettek ebben a kérdésben. A technikai feltételek lényegében nálunk is adottak a feladat elvégzéséhez. Vannak nagy sebességű vonalaink a letöltéshez, a terabájtos tárolók is egyre elterjedtebbek, van URN-szerverünk (*http://nbn.urn.hu*), rövidesen elkészül a magyar DC-generátor (*http://mek.oszk.hu/dc – 4. ábra*), van országos meta-adatgyűjtő rendszerünk (*http://www.nda.hu*) és saját fejlesztésű keresőnk (*http://keres.sztaki.hu*). A szükséges további szoftverek részben szabadon hozzáférhetők, részben megkaphatók az IIPC-től, ha csatlakozunk hozzá. Ami hiányzik, az egyrészt az információs és kulturális kormányzati akarat és költségvetési támogatás, másrészt a szükséges jogi környezet – ezeket a konzorciumnak ki kell lobbiznia. Kell továbbá egy reálisan megvalósítható közös vízió a középtávon elérni kívánt célról vagy célokról.



4. ábra **A MEK DC metaadat-generátorának részlete**

A kezdéshez célszerű lenne egy nagyon szűk körű előkészítő csoportot alakítani, amely javaslatot tesz a lehetséges konzorciumi tagokra, és elkészíti a projekt stratégiai tervét. A konzorcium megalakulása és a célkitűzések elfogadása után pedig egy 2-3 éves pilot projektet kellene beindítani, melynek céljai: a kérdéskör áttekintése, a külföldi eredmények és a nemzetközi szabványok/trendek megismerése, egy első rendszerterv elkészítése, gyakorlati tesztek lefolytatása – röviden: a majdani üzemszerű működéshez szükséges elméleti ismeretek és gyakorlati tapasztalatok megszerzése, valamint a szervezeti háttér kiépítése. Ennek az időszaknak nem elsődleges célja nyilvános szolgáltatás(ok) indítása, de teszt/demó szinten a lehetséges szolgáltatási módokkal is foglalkozni kell.

A pilot projekt idejére a konzorciumi tagok munkacsoportokat állítanának fel, amelyek félévente jelentésekben számolnak be tevékenységükről, a projekt első fázisának lezárásaként pedig egy összefoglaló tanulmányt készítenek az általuk vizsgált témáról.

A javasolt munkacsoportok

A válogatással és lehatárolással foglalkozó munkacsoport

Feladata: Megvizsgálni az archiválandó anyag kiválasztásának, illetve lehatárolásának szempontjait, mind az egyedi, mind a generális begyűjtés céljára. Az egyedi webhelyek mentéséhez meg kell határozni a válogatás kiindulópontjait (pl. ugrópontgyűjtemények), a válogatás tartalmi, minőségi szempontjait (pl. magyar intézmények által fenntartott vagy magyar tartalommal rendelkező, kulturális, tudományos, közéleti stb. webszolgáltatások), és javaslatot kell tenni a válogatás felelősére. A generális aratáshoz meg kell határozni a magyar webtér kiterjedését (a *.hu* domén és a rajta kívül eső, magyar tartalmat szolgáltató szerverek), és létre kell hozni az együttműködést a doménszolgáltatókkal a magyar webtérbe tartozó szerverek naprakész nyilvántartásához. Meg kell határozni továbbá, hogy a begyűjtés milyen mélységben és milyen típusú objektumokra terjedjen ki.

A begyűjtés és tárolás technikai kérdéseivel foglalkozó munkacsoport

Feladata: Áttekinteni az egyedi honlapok mentésének, valamint a robotokkal való aratás technológiájának állását, tesztelni és véleményezni az ezen a téren rendelkezésre álló szoftvereket, kezdeményezni ezek honosítását, illetve a hiányzó vagy túl drága komponensek hazai kifejlesztését. Kidolgozni a begyűjtött digitális objektumok tárolásának technikáját, az egyre inkább szabványosodó nemzetközi gyakorlatnak megfelelően. Felbecsülni a szükséges tárolási kapacitást és annak növekedési ütemét. Ajánlást kidolgozni a jól begyűjthető és jól archiválható webhelyek kialakítására, illetve a problémás helyek tartalmának ilyen célra alkalmas exportálására, és ennek figyelembevételére ösztönözni a nagyobb intézményi tartalomszolgáltatókat.

A metaadatok kérdéseivel foglalkozó munkacsoport

Feladata: Áttekinteni az internetarchívumok metaadat-használatának nemzetközi gyakorlatát. Javaslatot kidolgozni az egyedi honlapok mentéseinek metaadataira: lehetőleg maga a tartalomgazda lássa el Dublin Core-leírással a webhelyen levő nagyobb tartalmi egységek nyitólapjait még archiválás előtt. Ahol ez nem történik meg, ott a feladatot könyvtáraknak kell elvégezniük, felosztva egymás között a szakterületeket. Mind az egyedi webhelyek mentésénél, mind pedig a nagy tömegű automatikus aratásnál ki kell dolgozni a dokumentumból automatikusan kinyerhető, illetve a digitális objektumokról generálható metaadatok előállításának és tárolásának technológiáját. Döntést kell hozni arról, hogy ezekből mi, és milyen módon kerüljön a nemzeti bibliográfiába, illetve a könyvtárak katalógusaiba, valamint az NDA-ba.

A hasznosítás/szolgáltatás kérdéseivel foglalkozó munkacsoport

Feladata: Az archívumba kerülő anyag lehetséges felhasználási formáinak áttekintése. Javaslatokat tesz a nyilvános és nem nyilvános, a nonprofit és az üzleti célú hasznosításra. Felméri a használói igényeket, és piackutatást végez, majd becsléseket tesz a várható forgalomra, illetve a lehetséges bevétel nagyságára. Teszt/demó szinten beüzemel egy vagy több keresőfelületet a pilot fázisban begyűjtött anyagban való kereséshez és böngészéshez. (Ennek még nem kell feltétlenül nyilvánosnak lennie.)

A jogi kérdésekkel foglalkozó munkacsoport

Feladata: Áttekinteni az internetarchiválással kapcsolatban felmerülő jogi vonatkozásokat: a kötelespéldány-törvény kiterjeszthetősége az internetre, a copyright és privacy (személyiségi jogi) kérdések, az archívum tulajdonjoga és hasznosításának joga stb. Szerződéstervezetet dolgoz ki az egyedi webhelyek archiválásához, amely rögzíti az eredeti honlap gazdájának és az archívumnak a jogait, illetve kötelességeit. Jogi nyilatkozatterveze-

tet dolgoz ki az egyedi engedélyekkel, illetve az automatikusan, egyedi engedélyek nélkül begyűjtött anyag státusára és felhasználására vonatkozóan. Törvényjavaslatot készít elő, amely az internet-archiválás kötelezettségének előírása mellett különleges jogokat biztosít a nemzeti könyvtárnak a begyűjtésre és a szolgáltatásra (a NAVA-törvényhez hasonlóan).

A finanszírozás kérdéseivel foglalkozó munkacsoport

Feladata: Megoldani a projekt finanszírozásának problémáját abban a fázisban, amíg a résztvevők költségvetésébe nem épül be ez a tevékenység, illetve amíg az archívum hasznosításából származó bevétel nem járul hozzá a fenntartáshoz. Ennek érdekében felméri a pilot projekt időszakának várható költségeit, majd szponzorokat és üzleti partnereket keres, célzott támogatásokat és pályázati lehetőségeket kutat fel, és ezek segítségével igyekszik megszerezni a szükséges anyagi forrásokat.

Mivel egy nemzeti internetarchívum beindítása és folyamatos működtetése hatalmas és szerteágazó feladat, mindenképpen a lépcsőzetes, pragmatikus építkezés a célravezető, mert egy maximalista hozzáállással túl sok feltételnek kell megfelelni, és túl sok problémát megoldani, ami valószínűleg zsákutcába vezet. Fontos továbbá a munkák megosztása és az erőforrások koncentrálása, ugyanis csak így érhető el viszonylag rövid időn belül, hogy egy jelentős méretű és hasznos szolgáltatás jöjjön létre.

A projekt hosszú távú fennmaradásához azt is végig kell gondolni, hogy az UNESCO által ránk rótt kötelezettség teljesítésén kívül milyen előnyei vannak egy internetarchívumnak? Nem csak azért hasznos, mert a *404-es hibák* egy részére megoldást ad, ha van egy másolat az illető webhelyről egy másik szerveren. Egy ilyen archívum új – időbeli – dimenziót ad az amúgy jelen idejű internetnek. Mivel egységes szerkezetben, stabil szerveren, metaadatokkal és állandó URI azonosítóval ellátva, több időbeli állapotot rögzítve vannak benne a weblapok, informatikusi szempontból sokkal „jobban viselkedik", mint az eredeti, kaotikus és efemer internet. Egy ilyen gyűjteményre közhasznú és üzleti célú szolgáltatások sora építhető, például:

- tematikus összeállítások készíthetők évfordulókra, eseményekhez;
- részhalmazok képezhetők, és azokhoz speciális keresők rendelhetők médiatípus, témakör, célközönség vagy egyéb szempontok alapján;
- idődimenziót is tartalmazó nyelvi elemzők és egyéb statisztikai programok futtathatók rajta;
- szöveg- és adatbányászati rendszerek, tématérképek építhetők rá;
- szakirodalmi hivatkozásoknál és webes hivatkozásoknál jól használható stabil és rövid URN vagy URL címek rendelhetők nemcsak az egyes webhelyekhez, hanem akár azok minden elemére (pl. fejezetcímekhez, táblázatokhoz, ábrákhoz) külön is.

Egyszóval egy ilyen gyűjtemény hatalmas értéket képvisel, amelyet jól kihasználva a projekt idővel önfenntartóvá válhat.

A legelső dolog azonban, amit azonnal el kellene kezdenünk: a 90-es évek első felében megszületett magyar online szolgáltatások maradványainak összegyűjtése, és egy kis webmúzeum kialakítása belőlük, mielőtt a hazai internet legizgalmasabb korszakának emlékei végképp eltűnnek a digitális nirvánában.

Jegyzetek

[1] Eleve „digitálisan született", hagyományos hordozón nem publikált dokumentumok.

[2] A mentett fájlok integritásának vizsgálatára alkalmas ellenőrző összeg.

Irodalom

DIPPOLD Péter: A hagyományos nemzeti bibliográfia és az Internet: Válaszlehetőségek az új kihívásokra. Doktori disszertáció, Budapest, ELTE BTK, 2005. http://mek.oszk.hu/03500/03557

HAKALA, Juha: Archiving the Web: European experiences. Presentation in CONSAL XII, 20-23 October 2003, Brunei, URN:NBN:fi-fe20031951, http://www.lib.helsinki.fi/tietolinja/0203/webarchive.html

MAGYAR Gábor: Internetarchiválás, illeszkedés az NDA-hoz. NDA-konferencia, 2004. december 14. http://www.nda.hu/resource.aspx?ResourceID=magyarg_intenetarchivalas_041214_V1

MOLDOVÁN István: Archiválás a digitalizáció korszakában. Informatikai és Könyvtári Szövetség, OSZK, Budapest, 2002. szeptember 17. http://mek.oszk.hu/html/irattar/eloadas/2002/iksz-oszk.ppt

Beérkezett: 2006. V. 9-én.

Drótos László
az Országos Széchényi Könyvtár Magyar Elektronikus Könyvtár osztályán főkönyvtáros.
A Magyar Elektronikus Könyvtárért Egyesület Elnökségi tagja.
E-mail: *mekdl@iif.hu*