Webarchiválás a webkettes világban

Archiválási módszerek Ausztráliában

A *National Library of Australia* vezető szerepet játszik az ausztrál web begyűjtésében és megőrzésében 1996, a *PANDORA* archívum (*pandora.nla.gov.au*) létrehozása óta. Emellett léteznek más, szűkebb körű projektek is, mint például a tasmániai *Our Digital Island* (*odi.statelibrary.tas.gov.au*), vagy a kontinens Northern Territory nevű részén működő *Territory Stories* (*territorystories.nt.gov.au*). A nemzeti könyvtár jelenleg már háromféle módon archivál: a PANDORA gyűjteménybe szelektíven válogat online forrásokat, továbbá az Internet „Archive" segítségével a teljes *.au* domént learatja, valamint elkezdte használni az „Archive-It" szolgáltatást is. Elmondható tehát, hogy az ausztrál online tartalom jelentős részét sikerül így megmenteni a jövő számára. De a technológiai változások miatt a könyvtárnak folyamatosan alkalmazkodnia kell: fejleszteni az archiváló eszközeit, bővíteni a gyűjtött tartalmak körét és újabb partnerekkel szövetkezni, hogy eredményesen tudja folytatni ezt a fontos munkát.

A nemzetközileg is elismert PANDORA projektben jelenleg kilenc intézmény vesz részt: a nemzeti könyvtár és az egyes állami könyvtárak, az *AIATSIS* (Ausztrália őslakosságának kutatóintézete), az *NFSA* (nemzeti film- és hangarchívum), valamint az *Australian War Memorial* (háborús emlékhely és múzeum). 2008 júliusában az archívum 19 307 katalogizált tételt tartalmazott – összesen 53 112 080 fájlt, amelyek 2,2 terabyte tárhelyet foglaltak el. A begyűjtött anyagban többek között elektronikus folyóiratok, kormányzati kiadványok, valamint fontos tudományos és kulturális site-ok találhatók. Az egyes tételek jellege nagyon változó: az egyetlen PDF dokumentumtól a több ezer állományból álló komplett webhelyekig terjed, de ezek mellett archiválnak blogokat, podcastokat és videókat is. A válogatás, a begyűjtés és a hosszú távú archiválás kiforrott elvek mentén zajlik, az egész folyamatot a saját fejlesztésű PANDAS rendszerrel menedzselik, és ez szabályozza az archivált tartalomhoz való hozzáférést is. Minden digitális objektum stabil, feloldó rendszert is tartalmazó azonosítót kap. Együttműködéseket alakítottak ki indexelő és kivonatoló szolgálatokkal, amelyek a PANDORA-ban archivált publikációkat dolgozzák fel – ezek a dokumentumok is állandó URI-t kapnak, hogy hosszú távon is hivatkozhatók és előhívhatók legyenek.

Célzott és szelektív archiválással csak a nemzeti webtér egy viszonylag kis szeletét: a hosszú távon is jelentős kulturális vagy kutatási értékkel bíró tartalmat lehet megőrizni. A National Library of Australia tisztában van ennek a módszernek a korlátaival, ezért 2005 óta együttműködik az *Internet Archive* (*archive.org*) szervezettel, mely évente egyszer a robotjával bejárja az *.au* domén alá tartozó webszervereket. Egy-egy ilyen aratás nagyjából egy hónapig tart, és az így begyűjtött anyag mennyisége mellett eltörpül a PANDORA gyűjteménye: 2007-ben ez alatt az egy hónap alatt 18 TB-nyi digitális állomány gyűlt össze, miközben a PANDORA-ban 11 év alatt 2 TB-ot sikerült archiválni. A robot 2008-as futtatásakor mintegy egy milliárd fájl begyűjtésére számítottak. A *Heritrix* (*crawler.archive.org*) szoftverrel zajló aratás, bár igen kiterjedt, de messze nem teljes, hiszen egyrészt évente csak egyszer történik (és a közbülső idő alatt sok tartalom jelenik meg és tűnik is el), továbbá a robot engedelmeskedik a *robots.txt* fájlokban előírt tiltásoknak, és végül – bár a Heritrix nagyon sok mindent tud – vannak webhelyek, amelyeket nehéz vagy lehetetlen bejárni vele. Mindezen korlátok ellenére így is olyan hatalmas mennyiségű az összeszedett tartalom, hogy az mindenféle minőségellenőrzést reménytelenné tesz. Míg a PANDORA-ban megvan rá a lehetőség, hogy minden tételnél azonosítsák, és lehetőség szerint kijavítsák a letöltéskor keletkezett hibákat, a teljes webtér aratásakor ez lehetetlen. Egy másik különbség, hogy míg a PANDORA esetében a tartalomszolgáltatóktól engedélyt kérnek az archiválásra és az archivált verzió szolgáltatására, itt ez megvalósíthatatlan lenne. És mivel az ausztrál

copyright törvény szerint az online publikációk nem tartoznak a kötelespéldány-beszolgáltatási körbe, ezért a Heritrix segítségével készült archívum nem lehet nyilvános. Ez nem jelenti azt, hogy az anyag egyáltalán nem hasznosul, kutatók ugyanis dolgoznak rajta, csak a nagyközönség nem férhet hozzá jelenleg.

Egy további megőrzési módszer az *Archive-It*, amit az Internet Archive tesz lehetővé a saját szerverén. Az első és ez ideig egyetlen ausztrál szervezet, amely ezt igénybe vette, a nemzeti könyvtár *Asian Collections* nevű különgyűjteménye (*nla.gov.au/asian/asianwebarchive.html*). Itt arra használják ezt a szolgáltatást, hogy az Ausztrálián kívüli webszerverekről archiválják a gyűjtőkörbe tartozó társadalmi és politikai események digitális dokumentumait, melyeket várhatóan egyetlen regionális intézmény sem fog megőrizni (pl. egyes ázsiai országokban zajló parlamenti választások és zavargások hírei, illetve ottani kormányzati és egyházi oldalak). Azért választották ezt a külső hoszton levő megoldást, mert gyors és egyszerű lehetőségnek tűnt egy webarchívum kialakítására, ami így nem igényel saját műszaki hátteret, számítástechnikai szakértelmet és sok élőmunkát. Hamar kiderült, hogy ez csak részben igaz, mert az eredetileg elképzeltnél jóval több időt vesz igénybe a megfelelő webhelyek kiválasztása és a gyűjtemény gondozása. Hátrány az is, hogy miután összeállították a robot számára a kiinduló URL-ek listáját, már nincs mód kézzel belenyúlni a folyamatba, törölni vagy javítani a hibás vagy hiányzó tartalmakat, így ezek a sikertelen letöltések is benne maradnak a gyűjteményben és megjelennek a felhasználók előtt. További probléma, hogy ha megszakad az éves előfizetés megújítása, akkor az Internet Archive beolvasztja a gyűjteményt a saját nagy archívumába, és többé már nem érhető el önálló egységként. Mindezen hátrányok ellenére a könyvtár tervei között továbbra is szerepel ennek a szolgáltatásnak a használata, a saját archiválás mellett.

Fájlok begyűjtése

A PANDORA indulásakor a letöltő szoftver még csak az egyszerű HTML állományokkal boldogult, már a frame-es szerkezetű weblapokkal is gondjai voltak. Azóta ráadásul megjelentek a JavaScript, applet, CSS, Flash és más egyéb webes technikák és formátumok, melyek mindegyike újabb és újabb fejtörést okoz az archiválással foglalkozó szakembereknek. A formátumok közül különösen a multimédia-tartalmak okoznak problémát ilyen szempontból. A RealPlayer videóktól a podcast hangfelvételekig nemcsak a tárolási formátumok komplexitása jelent nehézséget, hanem azok szolgáltatási módja is.

Az ausztrál nemzeti könyvtár eddigi legnagyobb webarchiválási vállalkozása a 2007-es választások anyagának összegyűjtése volt. Mindent igyekeztek lementeni, beleértve az egyes pártok, lobbicsoportok és jelöltek honlapjait, blogjait, videóit és az internetes média vonatkozó oldalait. Összesen 350 webhelyet mentettek le, sokat közülük többször is, a változó tartalom miatt. Az igazi gondot a videók okozták; nem is annyira maguk a fájlok, hanem ahogy beágyazták és sugározták őket. A webmesterek különféle módokon próbálják minél kényelmesebbé tenni felhasználóiknak a mozgóképek megtekintését, ezért archiváláskor is eltérő megoldásokat kellett használni az egyes site-oknál. Azoknál az egyszerűbb eseteknél, amikor egy weblapon csak egy film volt, ingyenes videoletöltő szoftverek segítségével mentették le őket egyenként (mivel az „aratógépek" rendszerint nem gyűjtik be automatikusan a videókat), és konvertáló programokkal alakították át az *.flv* típusú fájlokat valamilyen elterjedtebb, (pl. *.mpeg*) formátumra. Ahol több videó volt egy lapra belinkelve, ott inkább meghagyták az eredeti flash formátumot és egy FVL-lejátszót tettek bele a lementett weboldalakba, így a felhasználók ugyanolyan könnyen meg tudják nézni ezeket a felvételeket, mint az eredeti szerveren. Amikor az ausztráliai választási kampány YouTube oldalának lementésére került sor (*nla.gov.au/nla.arc-76644*), amely több mint 700 videóból állt, szakértői segítséget kellett kérni a helyi informatikusoktól, akiknek végül sikerült kinyerni a videók URI azonosítóit, letölteni őket és elvégezni a szükséges változtatásokat az archivált weblapokon. Ezek nem egyszerű, hanem hosszadalmas, komoly szakértelmet kívánó munkák, amelyekre szükség van, ha azt szeretnénk, hogy az archívumban is lejátszhatók legyenek a videók.

A választások miatt a nemzeti könyvtár azt is feladatul kapta, hogy mentse el az előző kormányzat online anyagait. Erre már amúgy is számítottak a PANDORA archiválói, tanulva a korábbi kormányváltások tapasztalataiból, és még a választás időpontja előtt lementették minden minisztérium honlapját, amikor az még élő és karbantartott volt. Az előrelátás nem volt haszontalan, mert ezúttal is sok kormányzati weboldalt és online dokumentumot vettek le a nyilvános szolgáltatásból, különö-

sen azoknál a szervezeti egységeknél, amelyeknek megváltozott a feladatköre.

Gyűjtési irányok

Azzal, hogy a nemzeti könyvtár begyűjti az *.au* domén alá eső szerverek tartalmát, és emellett szelektíven is archiválja a PANDORA rendszerben a fontosabb webhelyeket és dokumentumokat, elmondható, hogy meg tudja menteni az ausztrál internetes tartalom jelentős részét. De hogy pontosan mekkora ez a rész, azt nem lehet megállapítani. Azzal tisztában vannak, hogy mindenképpen nagy hiányok maradnak. Például nem archiválják átfogóan azokat az ausztrál site-okat, amelyek nem az *.au* domén alatt vannak (de remélhetőleg az Internet Archive azért ezek többségét megőrzi). Nem gyűjtik viszont azt a – bizonyos szempontból a hagyományos elektronikus publikációknál is fontosabb – kreatív tartalmat, amit a magánemberek produkálnak a video-, foto- és művészeti webhelyeken, a blogokban, a virtuális világokban és a közösségi helyeken. Vannak ugyan próbálkozások ezeknek a begyűjtésére is, de csak kis, célzott projektek. (Ilyen pl. az egyik, nemrég indult kezdeményezés, amely az ausztráliai táncokkal kapcsolatos anyagot szedi össze a különböző weboldalakról és videomegosztó helyekről.) Bár a nemzeti könyvtár megegyezett a Flickr-rel, és engedélyt kapott a MySpace-től és a YouTube-tól is az archiválásra, de eddig még nagyon kevés anyagot mentettek le ezekről a helyekről. Az olyan forrásokról, mint például a virtuális világok (*Second Life* és társai) és a közösségi hálózatok (*Facebook, Bebo* stb.) pedig egyáltalán nincsen másolatuk. A fő ok, amiért nem mentenek le valamit, vagy jogi: olyan copyright és személyiségi jogi előírások vannak, amelyek nem engedik az archiválást; vagy pedig az adott forrás természete olyan, ami miatt nem tekinthető a nyilvános internet részének.

A könyvtárosok egyénileg is segíthetik a digitális kulturális örökség fennmaradását, például úgy, hogy törekednek arra, hogy a könyvtáruk, illetve az anyaintézményük weboldalain megjelentetett tartalom meg legyen őrizve. A kormányzati és az akadémiai szektorban a publikáció a nyomtatottól egyre inkább az online irányba tolódik. A tapasztalatok azt mutatják, hogy nemcsak hosszú, hanem rövid távon sem lehet bízni abban, hogy ami megjelenik egy honlapon, az elérhető is marad. Az egyetemeket már kötelezték arra, hogy szellemi produktumaikat repozitóriumokban helyezzék el, így ezen a módon a digitális publikációk hozzáférhetők lesznek a jövőben is. Hasonlóképpen elvárható lenne, hogy a kormányzat által fenntartott site-okon megjelenő kiadványok is elérhetők maradjanak, de ez egyáltalán nincs így. Vagyis, ha egy online publikáció fontos egy könyvtár gyűjteménye, illetve olvasói számára, akkor a könyvtárosoknak érdemes tenni valamit azért, hogy az biztonságosan megőrződjön valahol hosszú távon is.

Jövőbeli trendek

Az interneten mindig újabb és újabb technológiák jelennek meg, és a web archiválásával foglalkozók mindig újabb hiányokat fedeznek fel a begyűjtött anyagban a ténylegesen létező online tartalomhoz képest. Ezekkel folyamatosan foglalkozni kell; a webarchiválás sosem lesz teljes körűen kidolgozott és bejáratott állománygyarapítási folyamat. Állandóan fejleszteni kell az archiválási technikát, és felfedezni, majd összegyűjteni az újfajta tartalmakat, mert arra nem várhatunk, hogy ezek majd maguktól jönnek be hozzánk. A web túl dinamikus, a technológiája túl változékony, a tartalomelőállítók száma túlságosan nagy ahhoz, hogy valaha is egy olyan jól skálázható letéti rendszert lehetne kialakítani, mint amilyet a nyomtatott anyagokhoz létrehoztak a közgyűjtemények.

Amikor a National Library of Australia nekikezdett a nemzeti web archiválásának, kevés eszköz létezett, és kevés olyan intézmény volt, amelyekkel együtt tudott volna működni, vagy amelyektől tanulni lehetett volna e téren. Ezért saját rendszert és saját szoftvereszközöket találtak ki, s mind a mai napig a házilag fejlesztett PANDAS segítségével menedzselik az archívumot. Ez a rendszer már a harmadik verziójánál tart, és várhatóan ez volt az utolsó fejlesztési fázis, mert a könyvtár a továbbiakban már nem tud önmagában finanszírozni egy ekkora fejlesztést. Ugyanakkor, köszönhetően annak, hogy időközben a webarchiválás a világ más részein is bevett gyakorlattá vált, vannak már partnerek, akikkel meg lehet osztani a feladatok egy részét. Ezen a területen az *IIPC* (*International Internet Preservation Consortium*) nevű nemzetközi konzorcium – melynek más nemzeti könyvtárak és egyéb intézmények mellett az ausztrálok is tagjai – határozza meg a fejlődés irányait, így most már a közösen kifejlesztett eszközök adaptálásával lehet tovább folytatni az ausztrál digitális örökség megőrzését.

/CROOK, Edgar: Web archiving in a Web 2.0 world. = The Electronic Library, 27. köt. 5. sz. 2009. p. 831–836. ***http://www.emeraldinsight.com/10.1108/02640470910998542/***

(Drótos László)

Több mint 10 ezer fotó az interneten a forradalom előtti Oroszországról

Több mint tízezer, fekete-fehér és színes fotó kerül fel az internetre a forradalom előtti Oroszországról a *Runyiversz könyvtárportál*ra.

A honlapra a kor híres fotóművészei – *Alekszandr Grekov*, *Ivan Barsevszkij*, *Karl Bulla*, *Andrej Denyer*, *Makszim Dmitrijev*, *Andrej Karelin*, *William Carrick*, *Szergej Levickij*, *Szergej Prokudin*-Gorszkij és mások – által a XIX. században és a XX. század elején készített felvételeket teszik fel. Eddig 2000 kép került fel a világhálóra, 2010 végére pedig az internetezők 10 ezer fotót tekinthetnek meg. A képeket orosz levéltárak és magángyűjtők anyagaiból válogatták.

A Runyiversz elnöke, *Mihail Baranov* szerint "a képek segítségével a látogatók 'élőben' ismerkedhetnek meg a forradalom előtti Oroszország mindennapjaival". Az orosz fotóművészek gazdag örökségéből különös figyelmet érdemelnek Szergej Prokudin-Gorszkijnak az orosz birodalom nevezetességeiről készült színes felvételei. Az első fényképészműhely az 1840-es években nyílt meg Oroszországban, szinte közvetlenül a fényképezés feltalálása után. A század végére a számuk közel ezerre emelkedett.

A Runyiversz történelmi projektje a 2008-ban létrehozott digitalizált fakszimilekönyvtár. A honlapon ma a XIX. században és a XX. század elején kiadott könyvek, mindenekelőtt orosz történészek és filozófusok művei, enciklopédiák, dokumentumgyűjtemények olvashatók – olyanok, amelyeket kivontak a kulturális forgalomból és közel száz évig nem jelentek meg újra.

/SG.hu Hírlevél, 2010. január 2., *http://www.sg.hu/*

(SzP)

Gyászhír

Tárczy Ferenc – a TMT nyomdai munkájáért felelős kollégánk – 2010. február 16-án, 55 éves korában, tragikus hirtelenséggel elhunyt.

Tárczy Ferenc 1973 óta dolgozott az *OMIKK* nyomdájában szakképzett nyomdászként. 2001-től a *Reprográfiai üzem* vezetőjeként folytatta tevékenységét, 2009-től pedig a *BME OMIKK* gondnoki feladatait is ellátta.

Olyan jó embert és kedves kollégát vesztettünk el, akit mindenki szeretett megbízhatóságáért, becsületességéért, segítőkészségéért, szorgalmáért, pozitív gondolkodásáért. Közvetlen munkatársként Feri fáradhatatlan volt: mindenki kérésére – hivatali beosztástól függetlenül – azonnal rendelkezésre állt, gyorsan és önállóan oldotta meg a feladatokat.

Szerettük vidámságát, humorát és optimizmusát. Családjáról, unokájáról büszkén és sok szeretettel beszélt.
Komoly tervei voltak a munka vonatkozásában, melyek sajnos már nem valósulhatnak meg …
Emlékét kegyelettel és mély szeretettel megőrizzük.

A szerkesztőség.