Összefoglaló a webarchiválásról

A szerző, a *University of South Florida* tanára, két részben foglalta össze azokat az ismereteket, amelyeket a webarchiválás módszertanát oktató leendő egyetemi kurzusához gyűjtött a szakirodalom és a meglevő archívumok áttekintése során. A jelen cikk a téma irodalmának rövid, de tartalmas összefoglalója, mely kitér a webes tartalmak archiválásának minden fázisára: a kiválasztástól és a begyűjtéstől kezdve, az információszervezésen és tároláson át, a leírásig és a hozzáférés biztosításáig. A második publikáció a webarchívumok funkcionalitását elemzi, és a szerző tervez egy olyan kutatást is, amelyben az archiválással napi szinten foglalkozó szakemberekkel készít interjúkat az eddigi tapasztalataikról.

Bevezetés

1996-ban, vagyis a *World Wide Web* elterjedése után, az *Internet Archive* (*IA*) és néhány nemzeti könyvtár elkezdett foglalkozni ennek az új médiumnak az elmentésével és hosszú távú megőrzésével. A 2001-ben indult *International Web Archiving Workshop* (*IWAW*) volt az első olyan fórum, ahol lehetőség nyílt a tapasztalatcserére és ötletek megosztására. A következő fontos lépést az *International Internet Preservation Consortium* (*IIPC*) 2003-as megalakulása jelentette, mely szervezet jelentősen elősegítette a nemzetközi együttműködést, a szabványosítást és a nyílt kódú szoftvereszközök fejlesztését.

Mivel az emberi kultúra egyre nagyobb része a weben keletkezik illetve jelenik meg, ezért mind több könyvtár és egyéb közgyűjtemény szembesül a webarchiválás feladatával és a vele járó kihívásokkal. Ugyanakkor az ehhez a munkához szükséges készségeket még alig néhány helyen oktatják szervezett formában. Az USA-ban 2010 őszén a 32 legfontosabb könyvtár- és információtudományi tanszékből mindössze egy (*University of Michigan*) hirdetett meg egy féléves, kifejezetten a webarchiválással foglalkozó kurzust. Ezenkívül egy-két további egyetem tananyagában lehetett még ilyen irányú ismereteket találni más tantárgyak (pl. „A web tartalomelemzése”, „Digitális objektumok kezelése”) keretében.

A szerző a továbbiakban áttekinti a web archiválásának jelenlegi gyakorlatát, részterületek szerinti bontásban. A hosszú távú megőrzés kimaradt ebből az összegzésből, mert bár kétségtelenül az is fontos elem, de az archivált weboldalak ilyen szempontból már nem különböznek az egyéb digitális objektumoktól, vagyis megőrzésük nem igényel olyan speciális szakértelmet, amit nem lehetne más – például digitális könyvtári – kurzusokon megtanulni.

Értékelés és válogatás

Minden webarchívum rákényszerül, hogy egy vagy több szempont szerint megválogassa a begyűjtendő anyagok körét. Még az Internet Archive is, amely megpróbálja a teljes webet megőrizni, valójában csak a felszíni, robotokkal bejárható webhelyeket gyűjti be, és azokat sem teljes mélységig. A kiválasztási kritériumok sokfélék lehetnek: gyakori szűkítési szempont a domén, illetve aldomén neve (pl. *.gov* vagy *.nasa.gov*), de vannak témakörökre vagy eseményekre (pl. választások, konfliktusok) specializálódott archívumok, és van, amikor a médiatípus (pl. videók) vagy a műfaj (pl. blogok) jelent válogatási szempontot. Mindegyikre, illetve ezek különböző kombinációira léteznek már példák a világban. Sok európai ország menti a teljes nemzeti webteret, vagy akár a más domének alatt levő, de nemzeti nyelvű vagy témájú oldalakat is. Az amerikai *Kongresszusi Könyvtár* többek között a 2001. szeptember 11-i események, illetve az iraki háború internetes lenyomatait mentette el. A *Francia Nemzeti Könyvtár* e-naplókból készített egy válogatást. Az Internet Archive sok más mellett szoftverekből, valamint videojátékokról készült felvételekből alakított ki részgyűjteményeket. A

Preserving Virtual Worlds projekt az online virtuális világok megőrzésére specializálódott.

Az objektív szempontok szerinti válogatás elvileg jól automatizálható. Nem nehéz betanítani az aratást végző szoftvert, hogy fájltípus vagy doménnév szerint szűrje meg a lementendő tartalmat. Az sem bonyolult feladat, hogy a program felismerje az elektronikus újságokat és a blogokat, vagy hogy meg tudja különböztetni a blogbejegyzéseket a kommentektől. Az értékes tartalmú vagy népszerű weboldalak automatikus beazonosítása is elég jól megoldható a rájuk hivatkozó linkek, illetve látogatóik/nézőik száma vagy a felhasználói értékelések alapján. A *Cseh Nemzeti Könyvtár* a WebAnalyzer nevű alkalmazással elemezteti a weblapokat, ami egy előre definiálható szempontrendszer alapján pontozza őket. A határértéket meghaladó pontszámú oldalakat a cseh nemzeti web részének tekintik, és begyűjtetik az aratást végző robottal.

Egy tematikus vagy egy eseményhez kötődő válogatás esetében viszont szükség van az emberi ítélőképességre is. Mivel a „kézi" válogatás időigényes és költséges, ezért inkább csak a kisebb archívumokra jellemző. Takarékosságból egyes projekteknél elfogadják a felhasználók/tartalomgazdák által ajánlott URL címeket is, vagy felhasználják a már meglévő tematikus webkatalógusok címlistáit, illetve az adott terület szakértőinek segítségét kérik a fontos helyek beazonosításához. A folyamatot úgy is lehet gyorsítani, ha a válogatás nem weboldalak, hanem webhelyek vagy akár webhelycsoportok szintjén történik, és legfeljebb csak kizárnak ezekből egyes részeket, amelyek jelentősen más témájúak.

A válogatási szempontok tovább szűkíthetők értékalapú elemzéssel. A *National Taiwan University* például csak olyan webes forrásokat gyűjt, amelyek történeti, kulturális, társadalmi, oktatási vagy tudományos értékük miatt fontosak. A spamszűrés szintén egyfajta módszer az értékes és értéktelen tartalom elkülönítésére. A letöltött weblapokból való reprezentatív mintavétellel is lehet szűkíteni az archiválandó anyag mennyiségét. A francia könyvtárosok a mintavételezési stratégiát a kiinduló címlista és szűrőrendszer összeállításánál alkalmazzák: egy olyan archívumot akarnak létrehozni, amely a francia társadalom és kultúra sokszínűségét tükrözi, függetlenül a lementett tartalom értékétől vagy népszerűségétől. Ezért a gyűjtőkörbe egyaránt belefér a „legjobb" (pl. a szépirodalom vagy a szakirodalom), illetve a „legrosszabb" (pl. a reklám vagy akár a pornográfia), és a legnagyobbaktól a legkisebbekig minden webhelynek esélye van az archívumba való bekerülésre.

Begyűjtés

A webes tartalmak gyűjtésének többféle formája lehetséges, az archívum méretétől, az archívum és a webhelygazdák közötti kapcsolattól, valamint a megőrzendő anyag jellegétől függően. A könyvtárak és a levéltárak bevett gyarapodási forrásai az állami szervektől érkező dokumentumok, a könyvadományok és a kiadóktól kapott kötelespéldányok. Ezek a webarchiválásnál is lehetséges állománybővítési módok. Például a *U.S. National Archives and Records Administration* (*NARA*) megkérte mindegyik szövetségi minisztériumot, hogy adjanak be egy pillanatfelvételt a honlapjaikról *Clinton* elnök hivatali idejének lejártakor.

Az adatbázis-alapú, dinamikusan generált webhelyek nem másolhatók le egyszerűen és hosszú távú megőrizhetőségük is kérdéses. Ennek a problémának az egyik lehetséges, viszonylag egyszerű megoldása az, ha az adatbázis tartalmát valamilyen nyílt formátumra (pl. XML-re) konvertálják egy olyan eszközzel, mint amilyen a DeepArc.

Csak a webarchívumokra jellemző sajátos „szerzeményezési" módszer az aratás. Ennek az a lényege, hogy egy induló címlista alapján szoftverrobotok (ún. *crawler*ek járják be a weboldalakat, és miután letöltötték azok tartalmát, követik a bennük található hiperlinkeket, amelyek további oldalakra vezetik őket. A robotok viselkedését és a letöltendő fájlok körét szűrőkkel lehet szabályozni. Arra is van példa (*Arizona State Library*), hogy egy eredetileg beadásra tervezett archívumot aratásra állítottak át, mert a tartalomgazdák nem depozitáltak megbízhatóan. Bizonyos forrásokat a robotok nem tudnak rendesen begyűjteni (pl. térinformatikai GIS adatállományok, dinamikus webtartalmak, sugárzott média). A NARA 2004-ben összeállított egy útmutatót azokra a speciális esetekre, amelyeknél az automatikus módszerek nem használhatók.

A ismételt aratásoknál begyűjtött változatlan tartalmú oldalak fölöslegesen fogyasztják az erőforrásokat, így ezeket érdemes kiszűrni. Szerencsére az olyan szoftverek, mint amilyen a Heritrix, már elég intelligensek ahhoz, hogy felismerjék a duplumokat és ne töltsék le, illetve ne tárolják el ezeket. A nagy és gyorsan változó webhelyek periodikus mentése során egy további probléma is fellép: a crawler akár több napig is dolgozik, mire lement

egy nagy méretű site-ot, ám eközben annak tartalma folyamatosan frissül. Vagyis valójában egy olyan website kerül megőrzésre, amely ebben a formájában sosem létezett, mert az egyes oldalairól különböző időpontokban történt a pillanatfelvétel.

Hogy kell-e vagy szoktak-e engedélyt kérni az archiváláshoz, az is több tényező függvénye: a gyűjtemény nagysága, az archivált anyag jellege, a működtető szervezet típusa és a hatályos jogi környezet egyaránt befolyásolja ezt a dolgot. Új-Zélandon például, ahol a kötelespéldány szabályozás a webes forrásokra is kiterjed, az erre feljogosított könyvtárnak nem szükséges engedélyeket beszereznie az országban készült tartalmak lementéséhez. Az olyan nemzeti levéltárak, mint amilyen a NARA vagy a *UK National Archives*, szabadon archiválhatják a közintézmények anyagait. A kisebb archívumoknál gyakoribb, hogy előzetesen engedélyt kérnek a copyright-tulajdonosoktól, mert az igazán nagy volumenű projekteknél ez gyakorlatilag megvalósíthatatlan. Utóbbiak (pl. az Internet Archive) inkább az *opt-out* megoldást választják, vagyis a robotjaik egyrészt engedelmeskednek a tartalomszolgáltatók által beállítható robotkizárási előírásoknak, másrészt a jogtulajdonosoknak utólagosan is lehetőségük van kérni az anyagaik törlését. A copyright törvény 2006-os módosítása megengedte a Francia Nemzeti Könyvtárnak, hogy figyelmen kívül hagyja a robotokat kizáró fájlban talált szabályokat, de a gyakorlatban csak a kisméretű, fókuszált aratásoknál szokták néha figyelmen kívül hagyni őket, mert ezeknél könnyebb kezelni az esetleges következményeket. A *Library of Congress* a blogok és a híroldalak mentésekor igyekszik megszerezni a tulajdonosok engedélyét, de más típusú webhelyeknél csak egy értesítést küld ki arról, hogy a könyvtár archiválja a site tartalmát.

Szervezés és tárolás

A webarchívumoknak meg kellene őrizni az archivált anyagok hitelességét és integritását. Hogy ezt milyen fokon és módon oldják meg, az az archívumok jellegétől és céljaitól függ. Vannak esetek, amikor elegendő csak a szellemi tartalom megőrzése, máskor (pl. egy bíróság által is elfogadható bizonyítékhoz) az eredeti szerkezetet és kontextust is meg kell tartani. Minden archiválásra kiválasztott site-hoz tartozik egy külső struktúra, vagyis hogy hol helyezkedik el más webhelyekhez viszonyítva: honnan és milyen módon hivatkoznak rá, és ő milyen kifelé mutató linkeken át kapcsolódik más helyekhez. És tartozik hozzá egy belső struktúra is, amelyet a részegységei és weboldalai közötti belső linkek határoznak meg. Hasonlóképpen beszélhetünk külső és belső szerkezetről az egyes weboldalak szintjén is, hiszen ezeknél is vannak kívülről rájuk és róluk kifelé mutató hiperlinkek, valamint van egy saját struktúrájuk: a szövegek, képek és egyéb elemek elrendezése az oldalon. Az ismétlődő archiválás során ezek mellett egy történeti kontextus is keletkezik, ami azt mutatja, hogy hogyan változott egy webhely vagy weblap az időben.

A lementett tartalmak archívumba szervezésére háromféle módszer terjedt el eddig: helyi fájlrendszer, webszerű elrendezés és nem webszerű elrendezés. Ezek mindegyike képes az intellektuális tartalom megőrzésére, de a szerkezetet és a kontextust eltérő mértékben tudják csak megtartani. A lokális fájlarchívumnál a linkeket át kell konvertálni relatív URI címekre, amelyek a helyi rendszerbe mentett fájlokra mutatnak, azért, hogy a felhasználók navigálni tudjanak az oldalak között. Egy webszerű archívumban a weblapok és a hozzájuk tartozó metaadatok konténerfájlokba kerülnek, és megtartják az eredeti URI azonosítóikat valamint linkjeiket. Utóbbiakat persze automatikusan át kell irányítani olyankor, amikor egy felhasználó követni próbálja őket, hogy továbbra is az archívumban tudjon maradni, és ne vigyék ki őt az élő webre. Ez a megoldás őrzi meg leginkább az eredeti állapotot. A harmadik, nem webszerű tárolási módszernél kiveszik a dokumentumokat a hipertext környezetükből és vagy katalógusszerűen kereshető adatbázisba teszik, vagy egyszerűen PDF fájlokká konvertálják őket.

Leírás és metaadatok

A nagy webarchívumok gyakran megelégszenek az automatikusan generálható adatokkal: a lementés pillanatát jelző időbélyeg, a webszervertől kapott státuskód (pl. 404 = nem található, 303 = átirányítás), a fájlméret, az URI, a MIME típus (pl. text/html), a HTML fejlécben levő metaadatok stb. A *Greek Web Archive* rendszere például a weblapokban talált kulcsszavak és az ugrópontok szövege alapján osztályozza és rendezi klaszterekbe az archivált oldalakat. A kisebb léptékű projektek megtehetik, hogy manuális módszerekkel állítanak elő bizonyos metaadatokat. A University of California kampányszövegeket gyűjtő archívumánál például Dublin Core adatmezőket, Library of Congress tárgyszavazást és saját besorolási állo-

mányokat használnak a katalogizáláshoz. A *Digital Archive for Chinese Studies* sinológusokat kért fel a leíró metaadatok elkészítéséhez. A *National Taiwan University Web Archives* fejlesztői háromszintű osztályozási rendszert és speciális katalogizálási szabályokat dolgoztak ki a webes tartalmakhoz. Más rendszereknél a felhasználók is címkézhetik, kommentálhatják és értékelhetik az archivált anyagokat. A Library of Congress MODS rekordokat készít azokból az adatokból, amelyeket az archiválandó oldalakat javaslók szolgáltatnak, majd ezeket a rekordokat a katalogizálók még kiegészítik és pontosítják.

Gyakori megoldás, hogy előbb a nagyobb egységeket (pl. a webhelyeket) metaadatolják, majd ha van rá ember, akkor weblapszinten is elvégzik a leírást. Fájlszintű katalogizálásra (pl. az oldalakon található minden egyes kép önálló leírására) ritkán van példa, de bizonyos automatikusan generálható metaadatokat (pl. formátum, méret, módosítási dátum) ezen a szinten is elő lehet állítani. Minél kisebb egységet választunk, annál pontosabb leírások készíthetők, és természetesen annál több metaadatrekord fog keletkezni. A *Harvard University* webarchívumánál csak egyetlen, az online katalógusban is visszakereshető MARC rekordot készítenek a könyvtárosok az egyes részhalmazokról, amelyek rendszerint több webhelyből állnak. A Library of Congress hasonlóképpen, részgyűjteményenként katalogizálja az archivált anyagát, de emellett minden website-hoz saját MODS rekord is készül – utóbbiak azonban csak az archívumon belül kereshetők, az OPAC-ban nem jelennek meg. Az ausztrál PANDORA esetében a leírási szint egyaránt lehet a teljes webhely vagy annak valamilyen kisebb egysége.

Hozzáférés és használat

Hogy az archivált tartalomhoz ki és hogyan férhet hozzá, azt elsősorban az adott országban érvényes jogi szabályozás határozza meg. Új-Zélandon nemcsak a publikus weboldalak archiválását engedi meg a kötelespéldány-törvény, hanem az archívum nyilvános szolgáltatását is. Az Egyesült Államokban a Library of Congress csak a bibliográfiai leírásokat teszi teljes körűen visszakereshetővé, nyilvános hozzáférést csak azokhoz a webhelyekhez tesz lehetővé, amelyek tulajdonosai erre engedélyt adtak. Sok webarchívum zárt vagy csupán helyben használható – ilyen például a francia, a finn, a dán, a norvég, a szlovén, a svájci és az osztrák. Más esetekben csak csökkentett funkcionalitással vagy pedig késleltetéssel engedik a nyilvános hozzáférést. A Harvard University Library WAX rendszerénél például legalább 3 hónap a késleltetés, az *IA Wayback Machine* szolgáltatásánál pedig 6-12 hónap a várakozási idő azért, hogy ne jelentsenek konkurenciát az eredeti, „élő" webhelyeknek.

A keresési lehetőségeket az alkalmazott technológia és a metaadatok részletessége határozza meg. A Library of Congress és a National Library of New Zealand archívuma – a *subject headings* szerinti osztályozásnak köszönhetően – authoritylisták segítségével böngészhető. Ezzel szemben a Wayback Machine csak URL cím alapján tud megtalálni egy oldalt. A *NutchWax* keresőgépet használó rendszerek teljes szövegű keresést is biztosítanak. Vannak érdekes vizualizációs kísérletek is: az Egyesült Királyság archívumához adatbányász módszerekkel címkefelhőket készítettek, illetve egy 3D-ben animált falon lehet megnézni az egyes weblapok alakulását az időben. Japán kutatók pedig diavetítés és grafikon segítségével kísérelték meg bemutatni azt, hogy egy URL cím mögött hogyan változik a tartalom.

/NIU, Jinfang: An Overview of Web Archiving. = D-Lib Magazine, 18. köt. 3–4. sz. 2012./

(Drótos László)