Webarchiválás a University of Victoria könyvtárában

Ma már közhelynek számít, hogy az elektronikus források mennyire tünékenyek, és miközben a webes tartalmak mennyisége és fontossága rohamosan nő, ezek megőrizhetősége erősen kérdéses. A brit *Telegraph* újságírója 2014-es cikkében jogosan fogalmazott így: „ha nem vigyázunk, a történészek többet fognak tudni a múlt század kezdetéről, mint a mostaniról".

A kanadai *University of Victoria* könyvtára két éve kezdett el webhelyek gyűjtésével és megőrzésével foglalkozni. A szerző az első tapasztalatokon túl összefoglalja a webarchiválás aktuális műszaki, jogi és egyéb nehézségeit, megoldatlan kérdéseit is.

Archiválási technikák

A webhelyek begyűjtésére háromféle lehetőség van. Az első az úgynevezett „tranzakciós" módszer, aminek az az alapötlete, hogy minden kérést, amit a webszerver a felhasználók számára teljesít, egyben egy erre kijelölt archívumba is beküld. Ez a működés megvalósítható például az ingyenes SiteStory kiegészítő (mementoweb.github.io/SiteStory) feltelepítésével, ami természetesen a webszolgáltató együttműködését igényli. A második megoldásnak, a „szerveroldali" archiválásnak szintén előfeltétele az eredeti szolgáltató bevonása, mivel ilyenkor közvetlenül az eredeti gépről másolják át a megőrzendő tartalmat (pl. adatbázist) az archiváló rendszerbe, kikerülve a HTTP protokoll használatát. A francia nemzeti könyvtár által fejlesztett DeepArc (deeparc.sourceforge.net) programmal például jól archiválható XML fájlokká alakíthatók a relációs adatbázisok. Messze a leggyakoribb megoldás jelenleg a harmadik: a „kliensoldali" mentés, melyet távolról való aratásnak is neveznek. Ilyenkor webes robotok HTTP kéréseken keresztül gyűjtik be a weboldalakat, néhány, kezdőpontként szolgáló URL címről kiindulva és az ezeken az oldalakon talált linkeket végigjárva. A robotok viselkedése rugalmasan paraméterezhető, beállítható például az aratás mélysége, vagyis hogy hány szintig kövessék a linkeket.

Az Internet Archive

A legismertebb webarchiválási projekt, amely a kliensoldali megoldást használja, az 1996-ban indult Internet Archive (IA) nevű nonprofit vállalkozás. A könyvtárak csak évekkel később ismerték fel a web megőrzésének fontosságát és kezdtek el saját archívumokat építeni, de ezek mindmáig eltörpülnek az IA 2 petabájtos állományához képest, amely havonta kb. 150 terabájttal nő. A *Bibliothèque nationale de France* 2012-ben 370 terabájtot (18 milliárd oldalt) archivált, a *British Library* 2013-ban az első aratáskor 30 terabájt adatot gyűjtött az *.uk* domén alól, a *Library of Congress* 250 terabájtnál tartott 2011-ben, a *Library and Archives Canada* állománya pedig csak 7 terabájt volt. Az Internet Archive nemcsak mint szolgáltatás meghatározó ezen a területen, hanem több fontos technológiai újítás is köszönhető neki. Mindenekelőtt a népszerű Java-alapú „aratógép", a Heritrix, amely ARC vagy WARC formátumban tárolja el a begyűjtött digitális objektumokat: a szöveges HTML-fájlokhoz hozzácsomagolja a kapcsolódó kép-, média-, JavaScript-, CSS- stb. állományokat és az aratásra vonatkozó metaadatokat is. Ezek az archivált objektumok azután a szintén nyílt forráskódú Wayback Machine visszakereső rendszer segítségével hívhatók elő. A W/ARC fájlok leindexelhetők és teljes szöveggel is visszakereshetők a Lucene keresőmotorra épülő NutchWAX, Solr, vagy Elasticsearch programokkal.

Az intézmények kisebb léptékű archiválási igényeinek kielégítésére több szoftverszolgáltatás (Software as a service = SaaS) is született az elmúlt években. Elsőként – még 2005-ben – az IA indította el az Archive-it nevű előfizetéses szolgáltatását, amit már több közgyűjtemény is igénybe vesz. A *California Digital Library Web Archiving Service* (WAS) rendszere szintén egy ilyen SaaS eszköz,

mely a Heritrix, Wayback, NutchWAX hármasra épül; ezt használja például a Stanford és a Berkeley egyetem is. Az *Internet Memory Foundation* által kínált ArchivetheNet felhasználói között pedig ott találjuk a British Library-t is.

A University of Victoria webarchiváló projektje

2013-ban a University of Victoria könyvtárosai – egy konzorciumi együttműködés keretében – úgy döntöttek, hogy előfizetnek az Archive-it szolgáltatásra és elkezdenek egy saját archívumot építeni a gyűjtőkörükbe tartozó weboldalakból. Korábban nem foglalkoztak ezzel a kérdéssel, így nem volt egy kialakult víziójuk arról, hogy milyen legyen ez az archívum és hogy hogyan épüljön be a digitális könyvtári stratégiába. Ezért aztán 8 hónap is eltelt, mire kísérletezni kezdtek néhány kiválasztott webhely lementésével, és csak 2014 elején indították el a komolyabb gyűjteményépítést. Jelenleg ilyen részhalmazaik vannak: az egyetem 50 éves fennállását ünneplő események dokumentumai, tematikus válogatások a könyvtár egyes különgyűjteményeihez kapcsolódóan (pl. anarchizmus, transzneműség, környezetvédő szervezetek), helyi hírek, önkormányzati honlapok, digitális bölcsészettel foglalkozó webhelyek. A lementett fájlokat az IA tárolja és a könyvtárosok egy webes felületen át tudják az archívumot menedzselni.

Kihívások, problémák

A könyvtárak és a levéltárak eltérő típusú dokumentumokkal és megőrzési logikával dolgoznak. Előbbiek jellemzően önálló objektumokkal (könyvek, folyóiratok) foglalkoznak, melyeket valamilyen közös jellemző – például a témakörük – alapján helyeznek el a polcokon. Utóbbiak viszont nagyrészt elsődleges forrásokat kezelnek, melyeket nagyon más elvek szerint rendeznek fondokba, például származási hely szerint, vagy időrendben, vagy az eredeti sorrend megőrzésével. A webhelyek azonban nem követik ezt a hagyományos kettős felosztást, mert bár lehetnek rajtuk önálló „publikációk" (pl. cikkek, jelentések, hírlevelek), de sokszor ezek elrendezése, egymáshoz való kapcsolódása is fontos és megőrzendő. Valójában tehát egy dokumentumhálót kell(ene) archiválni, így ahhoz, hogy ezt sikeresen meg lehessen tenni, az egyetemi könyvtárosoknak és levéltárosoknak együtt kell működniük, átlépve a megszokott határaikon. A University of Victoria webarchiválási tevékenységét egy Archive-it munkacsoport felügyeli, melynek informatikus és gyűjteményépítő szakembereken kívül az egyetem oktatási és kutatási információforrásaiért felelős könyvtáros, a gyűjteményszervezési részleg vezetője, a különgyűjtemények igazgatója és az egyetemi levéltáros is tagja.

Tisztázandó a felelősség kérdése is. A könyvtárak hagyományosan olyan anyagokat őriznek, amelyekért fizettek. Az elektronikus tartalmak megőrzésére is vannak már megoldások, mint például a Portico, a LOCKSS és a Scholar's Portal, ám ezek is többnyire az előfizetéses forrásokra koncentrálnak. Ugyan a nyílt hozzáférésű folyóiratok hosszú távú archiválására is indult egy ígéretes kezdeményezés: a Public Knowledge Project, mely az Open Journal System típusú szoftverekkel készülő kiadványokat tárolja el a LOCKSS program rendszerében, de az egyéb műfajú *open access* publikációk – különösen azok, amelyeket kisebb kiadók jelentetnek meg – megmaradása erősen kérdéses, egyszerűen azért, mert egyetlen könyvtár sem érzi magát felelősnek ezekért. Míg az országos vagy állami szintű közgyűjtemények kötelesek bizonyos fajta webhelyeket archiválni, addig az egyetemi könyvtáraknak és levéltáraknak nincsenek ilyen kötelezettségeik. A webarchiváláshoz szükséges idő- és munkaerő-ráfordítás helyett inkább a fizetős tartalmak beszerzésére és menedzselésére megy el az energia ezekben az intézményekben.

A jogi helyzet sem egyszerű egy webarchívumnál. A University of Victoria illetékesei még dolgoznak az irányelveken, ezért a gyűjtemény nagy része jelenleg nem érhető el nyilvánosan. Alapvetően kétféle megközelítés létezik: az *opt-in* és az *opt-out*. Előbbinél az archiváló intézmény felveszi a kapcsolatot a tartalomszolgáltatókkal és engedélyt kér tőlük. Az utóbbi esetben viszont minden érdekes tartalmat begyűjt (tiszteletben tartva persze a robots.txt fájlban megadott tiltásokat) és a tartalomgazdák utólag külön kérésekkel töröltethetnek az archívumból bizonyos részeket. A British Library 2013 előtt az opt-in megoldást alkalmazta, de csak 24 százalékban kaptak választ az engedélykérésekre, ami azt jelentette, hogy így az Egyesült Királyság webhelyeinek csupán egy százalékát lehetett volna archiválni. Ez a helyzet 2013 áprilisában változott meg, amikor a brit parlament elfogadta a nem nyomtatott információforrások kötelespéldány-szabályozását. Az Internet Archive az opt-out elvet követi: a tartalomszolgáltató kérésére megszüntetik a nyilvános elérés lehetőségét

a Wayback Machine felületén, de a központi nyilvántartásból nem törlik a tartalmat. A Library of Congress egy közbülső megoldásra törekszik: bizonyos site-ok esetében külön engedélyt kérnek, a többinél pedig valamilyen módon megpróbálják értesíteni a szolgáltatót az aratásról és az archiválásról. A University of Victoria a *fair use* elvből indul ki, feltételezve, hogy a szabadon hozzáférhető webes tartalmakat szolgáltatók hallgatólagosan beleegyeztek azok leindexelésébe és tudományos célú lementésébe.

Technikai kérdések

A web nem csupán méretében növekszik rohamosan, hanem egyre komplexebbé is válik. Már nem csak a kilencvenes és a korai kétezres évekre jellemző összelinkelt statikus HTML fájlok alkotják, amelyeket könnyű volt egy Heritrix-szerű programmal learatni. Elterjedt az XML, a JavaScript, a JSON és az AJAX a weboldalakon, s mögöttük mindenféle adatbázisokban van a tartalom. A *New York Times* online kiadásának 2014. augusztus 28-i nyitólapját az Archive-it szolgáltatással lementve, kiderült, hogy az 235 URL-t tartalmazott, melyek 61 különböző szerverről származtak (a szöveges részek mellett 85 képfájl és 35 JavaScript alkotta az oldalt).

A modern, kliens oldalról nehezen archiválható webhelyekre jó példa a *Colonial Despatches* nevű gyűjtemény, mely a *Brit Gyarmatügyi Hivatal* és a Vancouver Island valamint British Columbia területén élő telepesek közötti korabeli levelezést dolgozza fel. A szolgáltatás mögött egy nagy eXist adatbázis van, TEI jelölésű XML dokumentumokkal. Bár ránézésre a kezelőfelület egyszerű (lásd: bcgenesis.uvic.ca/docsByDate.htm), valójában egy bonyolult megoldás van a háttérben: a felhasználó által kiválasztott dokumentumot a TEI/XML-ből egy XSLT fájl alapján XHTML-re konvertálja a rendszer és AJAX technológiával, JavaScriptet és CSS-t használva illeszti be ugyanabba a weboldalba. Az AJAX és a hozzá hasonló technikák elterjedése előtt nem volt arra lehetőség, hogy dinamikusan változtassák az éppen nézett oldal tartalmát, hanem többnyire ilyenkor egy másik weboldal jelent meg (URL címében a keresési paraméterekkel) és így ezt a linket követve a Heritrix robotja be tudta gyűjteni a találatként megjelenő dokumentumot.

Mivel a felhasználó tevékenysége függvényében változó, illetve a csak űrlapos kereséssel hozzáférhető tartalmú, adatbázis-alapú webhelyek aránya növekszik, mind nagyobb kihívást jelent ezek robotokkal való learathatósága és Wayback Machine-szerű megjeleníthetősége. A HTML5 elterjedésével a weboldalak egyre jobban hasonlítanak majd a mobil applikációkra, a web a statikus HTML dokumentumleíró nyelvről a JavaScriptes programnyelvre vált át. Ezt a változást az archiváló technikáknak is követniük kell. Az Archive-it szolgáltatásban 2014 júniusában megjelent egy Umbra nevű eszköz a Heritrix kiegészítőjeként, amely a Heritrix robotjától kapott URL címeken megpróbálja utánozni a valódi böngészőprogramok és a valódi felhasználók viselkedését: lefuttatja a kliens oldali scripteket, szimulálja az egérkurzor mozgását és a kattintásokat az egyes oldalelemek felett, lefelé görgeti az oldalt, hátha ilyenkor további tartalmak is letöltődnek a szerverről (ahogy pl. a Facebook hírfolyamánál is történik). A University of Victoria webarchiválóinak egyelőre vegyesek a tapasztalatai az Umbra-val. A Facebook és a Twitter esetében jól működik, mert ezekre lett optimalizálva, de például a fent említett Colonial Despatches honlap AJAX-os böngészőfelületével nem boldogult (talán mert a szokásos <a> tag helyet a címkét használták az oldal készítői az egérkattintáskor végrehajtandó Javascriptes utasításokhoz).

Ugyanezekkel a problémákkal küzdenek a nagy keresőgépeket üzemeltető cégek is, melyek nem tudják leindexelni a dinamikusan generálódó tartalmakat. A Google meg is fogalmazott bizonyos ajánlásokat a webmestereknek, amelyeket betartva azok bejárhatóbbá tehetik a webhelyeiket a robotoknak (*support.google.com/webmasters/answer/35769*). Mivel a SEO, vagyis a keresőoptimalizálás sok tartalomszolgáltatónak – az üzleti vállalkozásoknak különösen – fontos szempont, ezért sokan tesznek is ennek érdekében, valamit például XML honlaptérképet készítenek. A webarchiválással foglalkozó könyvtárak hasonlóképpen megpróbálhatnák meggyőzni a számukra fontos szolgáltatásokat, hogy lehetőleg már a webhely tervezésekor gondoljanak a megőrzési szempontokra is. A University of Victoria archiválói el is kezdtek egy ilyen egyeztetést a helyi *Humanities Computing and Media Centre* informatikusaival, hogy építsenek be a dinamikus webhelyeikbe egy olyan funkciót, amellyel sima HTML fájlokra konvertálható a tartalom az Archive-it aratógépe számára.

Persze felmerül a kérdés, hogy mit is akarunk valójában archiválni? A Colonial Despatches esetében például, még ha tökéletesen sikerülne is lementeni a weben megjeleníthető tartalmat és reprodukálni a felhasználói felület teljes funkcionalitá-

sát és interaktivitását, a valódi értéket a háttérben működő eXist adatbázisban levő, hatalmas élőmunkával készült TEI-kódolt XML dokumentumok jelentik, amelyekhez nem fér hozzá az Archive-it, így ezek ezzel a módszerrel nem őrizhetők meg. Ezért arról is elindult a tárgyalás a fejlesztőkkel, hogy hogyan lehetne a teljes rendszert egy virtuális gépre tükrözni az archívumban.

Hozzáférés és hosszú távú megőrzés

Az Archive-it szolgáltatással készült WARC fájlokat az Internet Archive tárolja az Egyesült Államokban. A kanadai szervezetek számára ez már csak azért is problematikus, mert az amerikai DCMA törvényben megfogalmazott *notice-and-takedown* rendelkezés jelentősen különbözik a Kanadában érvényes *notice-and-notice* rendszertől. Így előfordulhat az a helyzet, hogy az Internet Archive kénytelen eltávolítani egy olyan tartalmat, amit Kanadában nem lenne kötelező, ha ott nyújtana be ilyen kérelmet valamelyik tartalomgazda.

Az Internet Archive által tárolt WARC fájlok meg is sérülhetnek. Az IA saját óriási archívuma esetében bizonyos szintű adatvesztés megengedhető, hiszen az egész gyűjtemény eleve csak egy statisztikai mintavételnek tekinthető a teljes élő webből. Viszont egy közgyűjteménynek már komoly gondot jelenthet az adatvesztés egy általa létrehozott és gondozott webarchívumban. Az Archive-it szerencsére megengedi azt, hogy az előfizetői a WARC fájljaikat letöltsék és helyben is kezeljék őket. A University of Victoria szakemberei azt tervezik, hogy kérnek majd ilyen másolatokat és az Archivematica nevű, nyílt forráskódú, a digitális megőrzést támogató szoftverrel feldolgozva a COPPUL Private LOCKSS Network rendszerében tárolják el őket. Az Archivematica – sok más formátum mellett – képes a WARC fájlok befogadására is, ezt követően pedig el lehet vele végezni rajtuk az ISO-OAIS archiválási modellnek megfelelő műveleteket, majd BagIt típusú, hierarchikus csomagokat (Archival Information Packages) lehet készíteni belőlük. Ezek a becsomagolt állományok feltölthetők lesznek majd a COPPUL-PLN hálózat tárhelyére, ahol, ha valamelyik fájl megsérülne, az elosztott rendszerben tárolt további példányokból kijavítható. Az Archivematica és a LOCKSS összekapcsolása persze nem triviális feladat, de érdemes megoldani azért, hogy helyben lehessen menedzselni az archivált webes tartalmakat a hosszú távú megőrzés és hozzáférés érdekében.

/DAVIS, Corey: Archiving the Web: A Case Study from the University of Victoria. = The Code4Lib Journal, 26 sz. 2014-10-21
http://journal.code4lib.org/articles/10015/

(Drótos László)