A Könyvtári Figyelő 2020/1 számában megjelent cikk kézirata

noWebarchive

A „404 Not Found – Ki őrzi meg az internetet? - 2019” konferencián elhangzott előadás alapján

Holl András

Vajon maradandót alkotunk-e? Mennyi marad meg, mennyi „megy át” abból, amit létrehoztunk? Mi az, amit majd felhasználnak mások, mi az ami nem kell senkinek? Ott rejlik-e a hamu mélyén a csillagfényű gyémánt?[1]

A könyvtáros megközelítése: amelyik gondolat, információ könyvesült, az érték, megőrzendő. Persze ez nem teljesen igaz, de jó közelítés. Vannak könyvek, amelyeket sokan olvasnak, és el is felejtenek (megérdemelten), vannak, amelyeket senki sem olvas. Néha csak hosszú nyomozással lehet kideríteni, hogyan vállt korszakalkotóvá egy mű, amit állítólag „senki sem olvasott”.[2] Legyen a könyv megőrzendő vagy sem, a megőrzés többé-kevésbé megoldható: a nyomtatott könyvek készítése akkora befektetést igényel, ami a létrehozott könyvek mennyiségét valamennyire korlátozza, és ezzel a megőrzést elvégezhetővé teszi.

Meg kell-e őrizniük a könyvtáraknak a kisnyomtatványokat? Ha nem mindegyiket, hol a határ? Annyit állíthatunk, hogy mindent nem lehet, és nem is érdemes megőrizni. Gondoljuk tovább a problémát: megőrzendő-e, és megőrizhető-e minden hangfelvétel, fénykép, film (videó)? Minden tárgy (mindenből egy), amit az emberiség készített? És ha nem, MIVEL nem, hogy válogassuk ki a fontosakat vagy a jellemzőeket? Jó az, ha a szemétdombokról szedjük össze a kólás dobozokat (amiből sok maradt, az fontos, vagy legalábbis jellemző)? A régészek számára a régi korok anyagi kultúrája gyakran a szeméten – szemétgödrökön, pöcegödrökön, kutak mélyén felhalmozódott iszaprétegeken – keresztül deríthető fel. A FORTEPAN gyűjtemény alapjait lomtalanításokból gyűjtötték össze – az eredményt egy nemrég zárult a kiállításon láthattuk a Nemzeti Galériában.[3]

Nem odázhatjuk tovább, hogy a tárgyra térjünk: a web archiválásáról szeretnénk megosztani néhány gondolatunkat. Mint talán a hosszúra sikerült, elkalandozó bevezetésből is kiderült, fontosnak tartjuk az emberiség kulturális lenyomatának megőrzését, de nem tartjuk lehetségesnek, sem helyesnek, hogy mindent megőrizzünk, ami tárgyiasult (bájtosult). Azt sem, hogy csupán az elterjedtség, a népszerűség legyen a szelekciós kritérium. A World Wide Web sokoldalúbbnak és hasznosabbnak bizonyult, mint amire tervezték. Egyre több információ került fel, kerül fel a webre – ami jó, a baj csak az, hogy másutt nem marad: egyre inkább a weben, és webből leszünk. Ez az új médium illékony, az értékek megőrzéséről gondoskodni kell. De mivel a webre könnyű (olcsó) információkat felvinni, a szelekció szükségessége is felmerül. Még fontosabb, hogy ezt a szelekciót véleményünk szerint nem szabad az archiválóra (és a véletlenre) bízni. Az információ létrehozójának kell eldöntenie, mi az, amit hosszú távon megőrzendőnek tekint, és e szerint kell eljárnia. A digitális média is lehetőséget ad a biztonságos megőrzésre – csupán használni kell a megfelelő technológiákat (megőrizhető struktúrák és formátumok, egyedi azonosítók, licencek, redundáns archívumok, csak néhány fontosabbat említve).

---

1 Cyprian Norwid „A Kulisszák mögött” c. verséből kölcsönözve, Radó György fordítása alapján.

2 Owen Gingerich: The Book Nobody Read. Kopernikusz „De revolutionibus...” c. könyvének fogadtatásáról, Koestler állításának cáfolatáról.

3 Minden múlt a múltam #huszadikszázad #privátfotó #Fortepan. Magyar Nemzeti Galéria, 2019. április 16. – szeptember 29. https://mng.hu/kiallitasok/minden-mult-a-multam-huszadikszazad-privatfoto-fortepan/

1.) Információk könnyen archiválható formában

A web munkaeszköznek indult, tájékoztató eszközzé vállt (kirakat, hirdetőtábla), végül az információk egyre nagyobb mértékben felköltöztek a webre: ott jelennek meg, gyakran más médiában nem is léteznek. Pedig a web sok szempontból alkalmatlan az információk megőrzésére: illékony, változó (az új információ felülírja a régit), elosztott (a hiperhivatkozások más oldalakra vezetnek, amelyeknek az archiválását szintén meg kellene oldani) és technológia-érzékeny. A weboldalak emlékezethiánya a W3C koncepciókban meg is jelenik (Jacobs és Walsh, 2004). A kérdés részletesebb kifejtését Herbert Van de Sompelnél (2009) találhatjuk.

Az előbbiek miatt szükséges a web archiválása. De miért nem öntjük könnyen archiválható formába (és tesszük el archívumokba) a megőrzendő információkat? A Könyvtári Figyelő olvasóinak talán nem meglepő, ha archiválásra alkalmas formátumként a könyvet ajánljuk. Nem kell feltétlenül kinyomtatni, bár az sem árt, ha néhány tucat nyomtatott példány eljut a könyvtárak polcaira. A digitális könyv kompakt, jól archiválható objektum, megfelelő, hosszú távú megőrzésre alkalmas formátumok léteznek (pl. PDF/A), nyilvántartásuk a meglévő információs rendszerekben megoldott. Készítsünk évkönyveket, almanachokat, katalógusokat – öntsük ilyen formában azokat az információkat, amire az utókornak is szüksége lehet (humán olvasóközönséget feltételezve)!

2.) Adatbázis a weboldal és az almanach mögött

Nem csupán az archiválás miatt előnyös a weboldalakon megjelenő információkat adatbázisban tárolni. A tartalommenedzselő rendszerek (Content Management System, CMS) használata, a weboldalak információinak szakrendszerekből (pl. személyzeti adatbázis a munkatársak weboldal esetében) való átvétele lehetővé teszi, hogy ugyanabból az operatív adatbázisból készüljön el a tájékoztatást szolgáló naprakész weboldal és az archiválást szolgáló intézményi évkönyv, almanach.

Az MTA KIK kiállításainak esetében alkalmaztuk már azt a technikát, hogy a weboldal kis adatbázis segítségével a kiállítás témakörébe tartozó, a repozitóriumban tárolt dokumentumokat mutatott be. A kiállításvezető digitális változata is elérhető volt a weblapról – és a repozitóriumban archiválásra került.[4] Ha a weboldal elvész, nem baj (a kiállítás is bezárt valamikor). Az eseményről megőrzendő információk a repozitóriumban megtalálhatóak: a kiállított művek digitális másolatai, a könyvtár éves jelentése (benne a kiállítások felsorolásával), a kiállításvezető (az eredeti dokumentumok pedig természetesen megtalálhatóak Kézirattár gyűjteményében).

3.) Adatbázisok archiválása

Amennyiben az adatainkat adatbázisokban tartjuk, azoknak az archiválását kell megoldani. Fentebb már említettünk egy lehetőséget: almanachok, évkönyvek készítése. Az adatbázisok gyakorta bonyolultabbak (és unalmasabbak) annál, hogy könyv formában érdemes vagy egyáltalán kivitelezhető lenne az archiválásuk. A relációs adatbázisok archiválásával foglalkozik az E-ARK projekt, a Nemzeti Levéltár részvételével (Lux, 2018).

Az adatbázisok is dinamikusak: kérdés, hogy mikor, melyik állapotot érdemes archiválni? A tudományos kutatásban esetenként azt a megoldást alkalmazzák, hogy amikor a kutatáshoz szükséges információkat egy dinamikus adatbázisból lekérdezik, a lekérdezés eredményét DOI azonosítóval ellátva archiválják. Amikor az adatbázist nem használták, a változások naplózása

4 Hazádnak rendületlenül – 180 éves a Szózat http://vorosmarty.mtak.hu/

hibakeresési szempontból fontos lehet, de az egyes verziók, az adatbázis minden változatának archiválására nincs szükség.

4.) Repozitóriumok archiválása

A repozitórium nem feltétlenül ideális tájékoztató eszköz – de aggregálható és archiválható. A hosszú távon megőrzendő dokumentumoknak a repozitóriumban van a helye, és a megőrzendő információk legfontosabb, legnagyobb érdeklődésre számot tartó részét pedig repozitóriumban archiválható dokumentum formába kell önteni (legyen az könyv, film, kép, stb.). Bár a repozitóriumok maguk is webes szolgáltatások, véleményünk szerint érdemes a repozitóriumban őrzött dokumentumok egy részét hagyományos weboldalakba szervezve bemutatni, továbbá a repozitóriumokban való keresésre a repozitórium szoftvere által nyújtott lehetőségek meghagyásával az olvasóknak más lehetőségeket biztosítani és ajánlani: ilyenek a közös keresők (discovery systems) vagy az aggregátorok.

A repozitórium az a webes szolgáltatás, ami maga is hosszú távú megőrzésre szolgál. Egyszer azonban a repozitórium (a szoftver, de akár az üzemeltető intézmény vagy részleg) is megszűnik. Hogyan lehet a repozitórium életénél hosszabb távra is gondoskodni az információk megőrzéséről?

Először is néhány alapfogalmat ismertetünk, melyek a webes tartalmak előre tervezett archiválásával kapcsolatosak.

i.) *Sötét archívum*: az adattartalom folyamatos, de a külvilág számára nem látható archiválása egy másik intézmény archívumában.

ii.) *Végrendelet*: írásos megállapodás az archívum megszűnésének esetére, melyben az adatgazda megnevezi az „örököst”, aki megszűnése esetén az anyagot a hozzá tartozó jogokkal együtt örökli, az örökös intézmény pedig vállalja az anyag megőrzését (továbbszolgáltatását) ebben az esetben. A gyakorlatban a sötét archívum megoldással együtt lehet működőképes.

iii.) *Sírkő*: lehetőség szerint a már nem működő szolgáltatások vagy archivált tételek URL-jét át kell irányítani a szolgáltatást megörökölő archívumra. Amennyiben ez nem lehetséges, az URL-en legalább egy „sírkő”-nek meg kell jelennie: ami a 404-es hibajelzés helyett tájékoztat, hogy milyen szolgáltatás (dokumentum) volt korábban elérhető az URL-en, és a digitális objektumot hol kell ezután keresni.

Az alapfogalmak ismertetése után lássuk a repozitóriumok megőrzésének lehetőségeit.

a.) A repozitóriumok megoszthatják egymás között a tartalmakat, növelve a redundanciát. A könyvtárak hosszú távú, megosztott archiválási megoldásaihoz – mint amilyen a LOCKSS[5] (Rosenthal és Reich, 2000) és a CLOCKSS (Kiefer, 2015) hasonló megoldásokat kellene kialakítani. Szisztematikus tartalommegosztásról hazai repozitóriumok között nem tudunk beszámolni, de eseti tartalomcserék, átvételek, közös digitalizálások előfordulnak.

b.) Célszerű a repozitóriumokból olyan formátumú mentéseket készíteni, amelyek alkalmasak arra, hogy viszonylag könnyen más technológiai (pl. repozitórium szoftver) környezetben is újraéleszthetőek legyenek. Az Eprints-et futtató repozitóriumokból például lehet EprintsXML formátumban elmenteni a tételeket, a teljes bináris dokumentumo(ka)t karakteres formába átkódolva és becsomagolva az XML dokumentumba.

5 LOCKSS: Lots Of Copies Keep Stuff Safe

c.) A repozitóriumok archiválása kívülről, a repozitórium aktív közreműködése nélkül is lehetséges, az OAI-PMH segítségével begyűjtve a metaadatokat, majd a teljes dokumantumokat az XML-ben szereplő URL-ekről. Van de Sompel (2016) további lehetőségeket is ismertet.

5.) Proaktiv archiválás

Hiba, ha a webarchiválást végző szervezet(ek)re hárítjuk az archiválás minden terhét. Mikor szükséges egy weboldalt archiválni? Mit lehet az archivált kópiával kezdeni? A minimum, hogy megfelelő, géppel olvasható formában megadjuk a weboldalak frissítésének időpontját. De még jobb, ha az archiválandó weboldal jelezni tudja a webarchiválónak: most archiválj engem! Fontos, hogy a weboldalak megfelelő licencekkel és útmutatásokkal legyen ellátva: archiválható-e, nyilvánossá tehető-e az archivált kópia? A célra léteznek megfelelő, de legalábbis szükség szerint továbbfejleszthető technológiai alapok. Azt sem tartanánk rossz ötletnek, ha a weboldalak készítői valamilyen módon az oldal metainformációi között jeleznék, mennyire tartják archiválásra érdemesnek az adott oldalt? Amikor mégis a webarchiválás a megoldás az információk megőrzésére, akkor a weboldal tulajdonosának együtt kell működnie az archiválóval, segítenie kell az archiválást!

Irodalom

Gingerich, O.: The Book Nobody Read. Penguin, 2005.

Kiefer, R.S.: Digital preservation of scholarly content, focusing on the example of the CLOCKSS Archive. Insights 28 (1): 91-96. DOI: 10.1629/uksg.215

Lux, Z.: Implementation of New Technologies to Ensure the Sustainability of Digital Content. CDA2018, Bratislava, 2018. ISBN 978-80-89303-67-0. 95. o.

Rosenthal, D.S.H. - Reich, V.: Permanent Web Publishing. Proceedings of the FREENIX track. 2000 USENIX Annual Technical Conference. 2000.

Van de Sompel, H. – Nelson, M.L. - Sanderson, R. - Balakireva, L.L. - Ainsworth, S. - Shankar, H.: Memento: Time Travel for the Web, arXiv 0911.1112v2, 2009.

Van de Sompel, H. - Rosenthal, D.S.H. - Nelson, M.L.: Web infrastructure to Support e-Journal Preservation (and More), arXiv 1605.06154, 2016.