Drótos László – Németh Márton

Az OSZK-ban folyó kísérleti webarchiválási projekt első évének tapasztalatai*

Az Országos Széchényi Könyvtárban az OKR (Országos Könyvtári Rendszer)[1] kifejlesztése keretében 2017−2018 között zajlik egy kísérleti projekt azzal céllal, hogy Magyarországon is megteremtsük a nyilvános webhelyek tömeges archiválásának és hosszú távú megőrzésének feltételeit, elsősorban az ehhez a munkához szükséges informatikai infrastruktúrát és szakértelmet. Ezen a téren több mint 20 éves lemaradást kell ledolgoznunk, mert például az amerikai nonprofit szervezet, az Internet Archive (IA) már 1996 óta foglalkozik ezzel, és azóta példáját számos országban követték, létrehoztak nemzeti, kormányzati vagy intézményi webarchívumokat, gyakran könyvtári, levéltári irányítással vagy közreműködéssel. Az OSZK-ban a 2000-es évek közepén merült fel egy magyar internet archívum (MIA) ötlete, de az ezt előkészítő munka feltételei csak 2017 tavaszán kezdtek megteremtődni. Az egri Networkshop első napján rendezett műhelymunka vitaindító előadásában a 2018 áprilisáig eltelt egy év fejleményeiről számoltunk be, s ezeket az eredményeket és tapasztalatokat foglaljuk össze ebben a cikkben.

Tárgyszavak: *weblap; digitális dokumentum; digitális archívum*

Munkacsoport

Az egyik első fontos lépés az erre a feladatra dedikált munkaerő felvétele és egy munkacsoport megalakítása volt. A webkönyvtárosi és a webadminisztrátori munkakört *Németh Márton*, illetve *Visky Ákos László* tölti be, az informatikai feladatokat pedig *Kovács Péter* és külsős rendszergazdaként *Vitéz Gábor* látja el részmunkaidőben, *Drótos László* témafelelős irányítása mellett. A projekt könyvtárszakmai helye a *Moldován István* vezette *E-könyvtári Szolgáltatások Osztályon* van.

Hardver

A projekt idejére a *KIFÜ* (*Kormányzati Informatikai Fejlesztési Ügynökség*) biztosít egy szervert 128 GB-nyi memóriával és 20 TB háttértárral. A különböző szoftverek tesztelése és a nyilvános demó archívum építése egy kisebb teljesítményű OSZK-s szerveren folyik. A 2018 utáni, remélhetőleg már üzemszerűen működő webarchívum céljára beszerzés alatt van egy nagy kapacitású hardver infrastruktúra.

Szoftver

A külföldi könyvtárak és egyéb intézmények különböző módon építik a webarchívumaikat. Van, ahol házon belül fejlesztenek egy rendszert, mások valamilyen előfizetéses megoldást használnak (pl. az IA Archive-It[2] nevű szolgáltatását), esetleg megvásárolnak egy kész szoftvert, de gyakori az is, hogy ingyenes, többségében nyílt forráskódú programokkal dolgoznak. Utóbbiak fejlesztését a webarchiválással foglalkozó szervezeteket tömörítő IIPC (International Internet Preservation Consortium)[3] ösztönzi és koordinálja. Mi is ilyen szoftverek megtanulásával és tesztelésével töltöttük az elmúlt év jó részét.

Elsősorban az Internet Archive által régóta fejlesztett és használt Heritrix[4] aratószoftvert és OpenWayback[5] megjelenítőt próbálgatjuk. Előbbi egy komplex, sokféleképpen paraméterezhető, jelentős méretű archiválási feladatokra is alkalmas programrendszer, mely a legtöbb nagy archívum eszközkészletében megtalálható. Hátránya, hogy

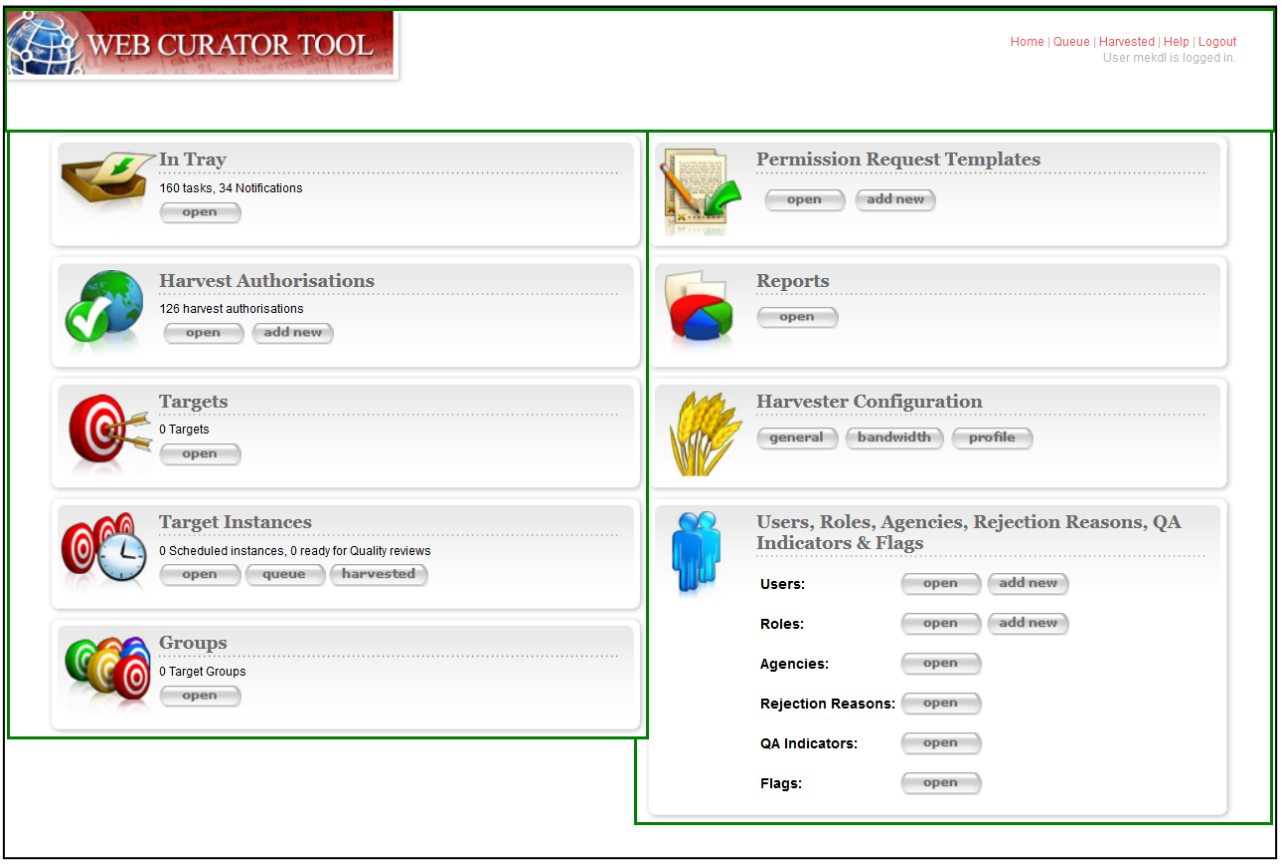
* A Networkshop 2018 konferencia első napján tartott műhelymunka vitaindító előadásának szerkesztett változata

bonyolultsága miatt komoly informatikai szakértelmet igényel és még így is valószínű, hogy csak többszöri próbálkozás után sikerül optimálisan beállítani a paramétereket. További gond, hogy mivel a web 1.0-ás időszakában született, a mai dinamikusan generált, interaktív, programkód- és médiagazdag weboldalakkal kevésbé boldogul. Az OpenWayback az IA több mint 332 milliárd archivált weboldalát megjelenítő Wayback Machine[6] szoftverének *open source* változata, mely a Heritrix vagy más letöltőprogramok által létrehozott WARC (Web ARChive) fájlokat teszi böngészhetővé. (A WARC egy nemzetközi szabvány, egyfajta konténerállomány, amelybe a webszerverekről lekért, tetszőleges típusú fájlok beletehetők.[7])

Az OpenWayback lényegében egy „időgép”, úgy böngészhető vele egy webarchívum tartalma, mintha az élő weben lépegetnénk.

A Heritrix és Wayback páros köré többféle keretrendszert is kifejlesztettek, melyek megkönnyítik ezek használatát. Az egyik legjobban kidolgozott ilyen eszköz a National Library of New Zealand és a British Library 2006-ban elkezdett közös fejlesztésének eredményeként létrejött Web Curator Tool (WCT)[8]

Ezzel egy felhasználóbarát felületen lehet nyilvántartani az archiválásra kiválasztott webhelyeket, adminisztrálni az engedélykérési folyamatot, elindítani és felügyelni az aratásokat, és ellenőrizni a mentések minőségét, feltárni a hibákat vagy hiányokat *(1. ábra)*. Sajnos a WCT fejlesztése az utóbbi években szünetelt, így nem működik rendesen együtt a Heritrix új verzióival, s emiatt sokszor elakadtak a tesztmentéseink. Jó hír viszont, hogy várhatóan 2018 őszén megjelenik egy javított kiadás belőle, ami már megoldja ezeket a problémákat.

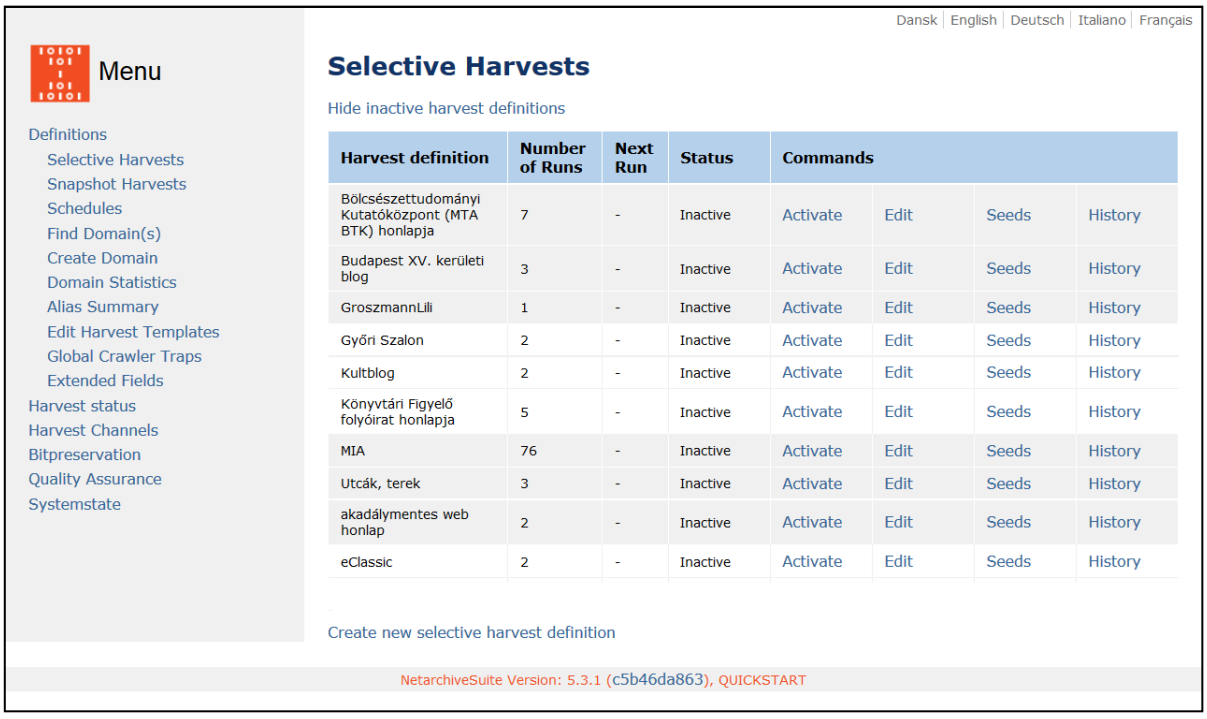


1. ábra **A Web Curator Tool adminisztrátori felülete**

A másik hasonló rendszer, melyet szintén több nemzeti könyvtár is használ, a 2005-től dán programozók által fejlesztett NetarchiveSuite (NAS)[9]. Ezzel is egy böngészőben megjelenő felületen keresztül indíthatunk aratásokat és vezérelhetjük a Heritrix működését *(2. ábra)*. A NAS nem tartalmaz olyan funkciót, amivel nyilvántarthatók az archiválásra és a mentett webhely szolgáltatására vonatkozó engedélyek, a metaadatkezelő része pedig kevésbé fejlett a WCT-hez képest. Viszont feltölthetünk többféle konfigurációs fájlt különböző típusú aratásokhoz, majd hozzárendelhetjük őket az egyes webhelyekhez. Van továbbá egy modulja a WARC állományok esetleges meghibásodásának ellenőrzésére, ami a hosszú távú megőrzés miatt hasznos. A NAS tesztelése során leginkább a rendszer beüzemelése és az általa használt *port*-ok tűzfalon való átengedése jelentett gondot, illetve az, hogy mivel a Wayback-et ún. *proxy* üzemmódban használja, ezért a mentések visszanézéséhez a böngészőben is be kell ezt a módot állítani.

Kipróbáltuk még a Windows-os gépekre is feltelepíthető WAIL[10] szoftvert, aminek az új változata a Heritrix mellett egy, a Chrome böngészőmotorját használó archiváló eszközt is tartalmaz. Ezzel sokkal jobb minőségben menthetők az olyan komplex weblapok, mint például a hírportálok vagy a Facebook oldalak, továbbá van egy modulja Twitter üzenetek mentésére is. A WAIL elsősorban személyes archiváláshoz hasznos, tömeges aratásokat nem lehet vele csinálni. Némileg hasonló hozzá az ingyenes online szolgáltatásként is igénybe vehető Webrecorder[11]. Ez szintén egy böngészőn keresztül ment, de nem egy előre megadott mélységig járja be az archiválásra kiválasztott webhelyet, hanem csak azokat az oldalakat menti el, amelyeket megnézünk egy böngészési folyamat során, tehát ahogy a neve is utal rá, egy videomagnóhoz hasonlóan felveszi (menti) mindazt, amit megnézünk. Az így keletkezett WARC fájlok a webrecorder.io felhőtárhelyén megőrződnek és bármikor visszanézhetők online, vagy pedig letölthetjük őket a saját gépünkre és ott a Webrecorder Player[12] segítségével jeleníthetők meg a bennük levő weboldalak (természetesen nem feltétlenül abban a sorrendben, ahogy felvettük őket). Egyedi weblapok egykattintásos mentésére szolgál a WARCreate[13] nevű beépülő böngészőmodul is, ami szintén szabványos WARC fájlba ment.

A WARC állományokban levő szöveges és képi tartalom kereshetővé tételére többféle megoldást is kifejlesztettek az évek során. Mi a NAS projekthez kapcsolódó SolrWayback[14] rendszert választottuk, ami a keresés mellett olyan különleges



2. ábra **A NetarchiveSuite adminisztrátori felülete**

funkciókat is tud, mint a link-gráfok rajzolása vagy a képfájlokban levő koordináták alapján a képek térképre vetítése. E mellett elkezdtünk egy saját keresőt is fejleszteni SolrMIA néven, mely szintén az Apache Solr platformra épül, támogatja a magyar szavak automatikus szótövezését, így a ragozott vagy képzett alakok is megtalálhatók vele; a találati listában megjelenik a webhely neve és az eredeti URL címe, valamint a keresett szavak szövegkörnyezete; a találati halmaz pedig nemcsak doménnévre, fájlformátumra, vagy mentési évre szűkíthető, hanem a webhelyek besorolására általunk használt témakörökre (pl. „Képzőművészet") és műfajokra (pl. „Elektronikus periodika") is. Mindkét kereső kipróbálható a nyilvános demó archívumban[15].

Még két, Windows alatt is használható, sokféle funkciót és beállítási lehetőséget kínáló ingyenes webarchiváló programot érdemes megemlíteni. Az egyik a HTTrack[16], amit például az ausztrál PANDORA archívumot építő könyvtárakban is használnak. Ez nem WARC konténerekbe, hanem fájlrendszerbe ment, így a lementett oldalak visszanézéséhez nem kell Wayback vagy Webrecorder Player, elég egy hagyományos böngésző. További előny számunkra, hogy magyarított felület is van hozzá, akárcsak a másik általunk kipróbált szoftverhez, a ScrapBook X nevű böngészőkiegészítőhöz. A ScrapBook X[17] is egy könnyen megtanulható és egyszerűen használható eszköz weboldalak vagy webhelyek letöltésére, majd ezekből gyűjtemények kialakítására, sőt beépített teljes szövegű keresője és szerkesztő felülete is van, amivel a mentett oldalakhoz jegyzeteket fűzhetünk, módosíthatjuk őket, kiemelhetjük vagy törölhetjük egyes részeiket stb. Ezzel sem lehet WARC fájlokat készíteni, de van hozzá egy további kiegészítő, ami egyéb archív formátumokra tudja konvertálni a letöltött anyagot. Mivel a Firefox új, Quantum verziójával nem kompatibilis, ezért a Firefox ESR változatát kell telepíteni hozzá. A programozója 2017-ben egy új szoftver fejlesztésébe kezdett Web ScrapBook néven, aminél már nincs ez a korlát, de ez még elég kezdetleges állapotban volt 2018 tavaszán, amikor mi teszteltük.

A weboldalak nemcsak WARC konténerekbe, illetve önálló HTML fájlokba menthetők, hanem képként is megőrizhetők. Természetesen ez esetben a szöveg kereshetőségéről, vagy a linkekre való kattintásról és az egyéb interaktív funkciókról le kell mondanunk, cserébe viszont az adott weboldalnak az adott időszakban népszerű böngészőben megjelent külalakját tudjuk így eltenni, ami szintén fontos lehet. Ezért aztán gyakori, hogy a webarchívumok ilyen *screenshot*-okat is tartalmaznak. Teljes méretű oldalképek készítésére mi a Grab Them All[18] és a Nimbus Screen Capture[19] böngészőkiegészítőket próbáltuk ki. Előbbinek nagy előnye, hogy egy szövegfájlban bármennyi URL cím megadható, melyeket sorban betölt a Firefox-ba, majd lefényképezi őket és elmenti PNG vagy JPG formátumban. Sajnos egy idő után ennél is belefutottunk a Quantum inkompatibilitási problémába.

Metaadatok

Egy könyvtári webarchívumnál jogos elvárás, hogy a teljes szövegű keresés mellett metaadatok alapján is kereshető, illetve böngészhető legyen. Az archiválási munkafolyamat és a hosszú távú megőrizhetőség pedig azt is megkívánja, hogy a bibliográfiai leírások mellett adminisztrációs és technikai metaadatokat is rögzítsünk. Ezek egy része automatikusan is előállítható (pl. a fájlokban található metaadatokat kigyűjtő Metadata Extraction Tool[20] segítségével), más részük viszont emberi intelligenciát és adatrögzítést igényel. Egy nemzeti archívum esetében viszont olyan tömegű digitális dokumentumról van szó, hogy nagyon meg kell gondolni, milyen szinten és milyen részletességgel készítünk emberi munkával metaadatokat. A digitális kulturális örökség szempontjából fontos, kiválogatott webhelyeket (pl. elektronikus folyóiratokat) nyilván érdemes részletesebben leírni, sőt akár ezek önálló alegységeiről is felvenni néhány adatot, míg a nagy tömegű aratások során begyűjtött sok tízezer honlapról, blogról és egyéb online forrásról elégséges lehet csak részgyűjtemény szintű leírásokat készíteni.

Az amerikai könyvtári szervezet, az OCLC 2016 elején életre hívott egy Web Archiving Metadata Working Group[21] nevű munkacsoportot azzal a céllal, hogy felmérje a jelenlegi helyzetet és a felhasználói elvárásokat a webarchívumok metaadataival kapcsolatosan, majd megfogalmazzon egy ajánlást az archivált webhelyeket leíró, elsősorban bibliográfiai jellegű adatokra. Mi is ezt a Dublin Core-alapú ajánlást vettük figyelembe a saját adatszerkezetünk kialakításánál, melyet kiegészítettünk olyan adminisztrációs és technikai adatmezőkkel, amiket az eddigi tapasztalataink alapján érdemes nyilvántartani. Például: mennyire sürgős az adott webhely archiválása?, milyen szoftverrel készült a mentés?, hol vannak a mentés során keletkezett naplófájlok és WARC állományok?,

ellenőrizte-e valaki a lementett anyagot és milyen hibákat talált?

Bár a WCT és a NAS is képes bizonyos metaadatok nyilvántartására, de ezen a téren egyik sem elég rugalmas és fejlett (a WCT-ben nem is bővíthetők szabadon az adatmezők), így egyéb megoldásokat is számításba vettünk. Készítettünk néhány tesztrekordot az OSZK Tudástárak alprojektjéhez készült adatrögzítő programmal, valamint a KOHA nevű ingyenes könyvtári rendszerrel is tervezünk még ilyen próbákat MARC21 alapon. Továbbá definiáltunk egy XSD fájlt[22], amelynek alapján például az XML Notepad[23] szerkesztővel adatbázisok és kompromisszumok nélkül is tudunk XML formátumban adatokat rögzíteni akár egyes webhelyekről vagy webhelyrészekről, akár komplett részgyűjteményekről. A terveink szerint ezeknek az adatoknak egy része az OSZK katalógusába is bekerül majd.

Aratások

Az első néhány héten terhelési teszteket végeztünk, hogy lássuk, milyen memória- és tárhelyigényei vannak a Heritrixnek attól függően, hogy hány URL címen és milyen beállításokkal indítjuk el, majd 2017. április végén lefutott az első komolyabb aratás: az EPA (Elektronikus Periodika Archívum és Adatbázis) által nyilvántartott, de az állományában nem archivált kb. 2 ezer magyar időszaki kiadvány weboldalait próbáltuk meg lementeni. Ezt a gyűjtést július elején megismételtük, kihagyva az URL listából a problémás (pl. megszűnt vagy az aratórobot által nem bejárható) címeket, és csak az időközben megváltozott oldalakat tárolva el újból. Így 1456 webhelyről mintegy 13 millió fájlt töltöttünk le 1,3 terabájt összméretben 17 nap alatt. Ezzel párhuzamosan elkezdtük a *Könyvtári Intézet* által gondozott *Nyilvános Könyvtárak Jegyzéké*ben található könyvtári honlapok mentését is, amit a magyar levéltárak, múzeumok és galériák weboldalainak archiválása követett. Az elmúlt év során további gyűjteményeket is csináltunk, így például az egyetemek, a kutatóintézetek és az önkormányzatok honlapjait mentettük (utóbbiak listáját a magyar Wikipédiából kaptuk meg) – összességében több mint 5 ezer *site*-ot. Jelenleg elsősorban az irodalom és művészet témájában válogatunk megőrzésre érdemes webhelyeket. A tematikus aratások mellett két eseményalapú archiválást is csináltunk: a 2018-as téli olimpiával, valamint az országgyűlési választásokkal kapcsolatos online forrásokat gyűjtöttük néhány hétig. Külön mentettük az OSZK saját webes tartalmait, például a honlapot, a blogot, a virtuális kiállításokat és a Facebook oldalt. Terveink közt szerepel a *.hu* alatt található nyilvános magyar webhelyekről évente egy-két reprezentatív jellegű (tehát nem teljes mélységű és nem minden fájltípusra kiterjedő) aratás is, de ennek a feltételei még nem adottak.

A kísérleti projekt keretében eddig lementett – még eléggé ideiglenes és töredékes – anyag egy nem publikus tárhelyen van, mivel a nemzeti könyvtárnak még nincs törvényi felhatalmazása ennek szolgáltatására, de remélhetőleg a jövőben majd helyben vagy zárt hálózaton, ellenőrzött körülmények között hozzáférhető lesz az archívum az olvasók és kutatók számára. Addig is egy kis, valamivel több mint száz webhelyből álló gyűjtemény nyilvánosan is megnézhető a mekosztaly.oszk.hu/mia/demo címen *(3., 4. és 5. ábra)*. Az itt látható oldalak tulajdonosaitól egyedi engedélyeket kértünk arra, hogy legalább a kísérleti projekt végéig szolgáltathassuk az általuk közzétett tartalmak mentéseit. Ennek az intézményi és személyes honlapokból és blogokból álló, valamint néhány elektronikus periodikát is tartalmazó szolgáltatásnak az a célja, hogy demonstrálja a webarchiválás és a webarchívumban való keresés technológiájának jelenlegi lehetőségeit és korlátait. Az OSZK-ban található mentés(ek)re mutató link mellett minden webhely esetében megnézhető a letöltéskor készült képernyőfotó, az adott domain linktérképe, az Internet Archive-ban levő többi mentés, valamint természetesen az élő weboldal is megnyitható egy másik ablakban. Az eredeti és az archív példányok összehasonlításával láthatóvá válnak az archiválás során keletkezett hibák és hiányok.

Problémák

A webhelyek szelektív aratása során jó néhány tipikus hibaforrásba ütköztünk bele. Általánosan érvényes, hogy a Heritrix program által indított robotok a mélyweb tartalmával nem tudnak mit kezdeni. Nem képesek belemenni adatbázisokba, s nem tudják jól lementeni a közösségi média tartalmait sem. Az online adatbázisok világa gyaníthatóan még hosszú ideig kívül fog esni a webarchiválás hatókörén. A közösségi hálózatok tartalmainak mentésére pedig talán – a még elég kezdetleges állapotú – böngésző-emuláló programok lesznek majd képesek, amelyek voltaképpen a weben szörföző ember viselkedését utánozzák egy parancsvezérelt böngészőn keresztül.

OSZK WEBARATÁS – DEMÓ ARCHÍVUM

KERESÉS A TELJES SZÖVEGBEN...

KÖNYVTÁR | LEVÉLTÁR | MÚZEUM | MŰVÉSZET | KUTATÁS | OKTATÁS | ÖNKORMÁNYZAT | TÖRTÉNELEM | KÖNYV | E-PERIODIKA | BLOG | SZEMÉLYES

Ez a kis gyűjtemény az OSZK-ban zajló kísérleti webaratás projekt keretében készül azokból a mentésekből, amelyeknél az eredeti honlaptulajdonos engedélyt adott a Nemzeti Könyvtár számára az archivált példány(ok) nyilvános szolgáltatására – egyelőre 2018 végéig. (További felajánlásokat, javaslatokat örömmel veszünk ezen az űrlapon.) A célja az, hogy demonstráljuk vele a jelenlegi webarchiválási technológia lehetőségeit és korlátait. Bár a demóhoz kiválasztott webhelyek az automatikus módszerekkel viszonylag jól archiválhatók közül kerültek ki, még így is előfordulnak hibák és hiányok a lementett példányokban. Ezen problémák egy része a site-ok robot-barát és archívum-barát kialakításával orvosolható.

A piros nyilakra linkelt mentések a Heritrix szoftverrel készültek 2017 decemberével kezdődően, nem teljes mélységben, a videók és más nagy méretű állományok letöltésének kizárásával, és az eredeti webhelyen levő robots.txt fájlban levő tiltások tiszteletben tartásával. A megjelenítés az Open Wayback szoftverrel történik, melyben a mentés dátumára kell kattintani az archív példány megtekintéséhez. A mentett változatra és az első mentés időpontjában készül képernyőfotóra mutató nyíl mellett van egy-egy link az amerikai Internet Archive-ban található mentésekre, valamint az eredeti "élő" honlapra is (amennyiben az még létezik). A nyilakra kattintva a weblapok új böngészőlapon nyílnak meg, így könnyen összehasonlíthatók a mentett verziók és az eredeti oldalak. A sárga gomb pedig az adott doménről kifelé mutató, illetve a rá kívülről hivatkozó linkek gráfját rajzolja ki - az archívumban levő mentések alapján (mivel a demó archívum még kicsi, ezért ez utóbbi, *ingoing* típusú linkek száma nagyon kevés).

A fejlesztés alatt levő SolrMIA keresőnk mellett kísérleti jelleggel a Solr Wayback Search nevű felület is kipróbálható, amellyel a mentett webhelyek teljes szövegében lehet keresni és a találatok szűkíthetők doménnevek, fájltípusok és a mentés éve szerint. A találati listában a weboldal vagy fájl (nagy betűvel kiemelt) címére kattintva jutunk el az archivált verzióra, az *Url:* sorban levő link pedig az eredeti honlapra/fájlra visz. A *Show full post* felirat alatt megnézhetők az adott találat részletes adatai. A keresés magyar nyelvű automatikus szótövezéssel történik, ami azt jelenti, hogy a ragozott és képzett alakokat is megtalálja a program. A leggyakoribb kötőszavak és névmások viszont ki lettek zárva az indexelésből, így ezeket figyelmen kívül hagyja a kereső. Az *Image search* és azon belül a *Geo search* opciókkal képekre és földrajzi nevekre lehet rákeresni, de ez a keresés csak az alkönyvtár- és fájlnevekben történik, így kevés találatot ad.

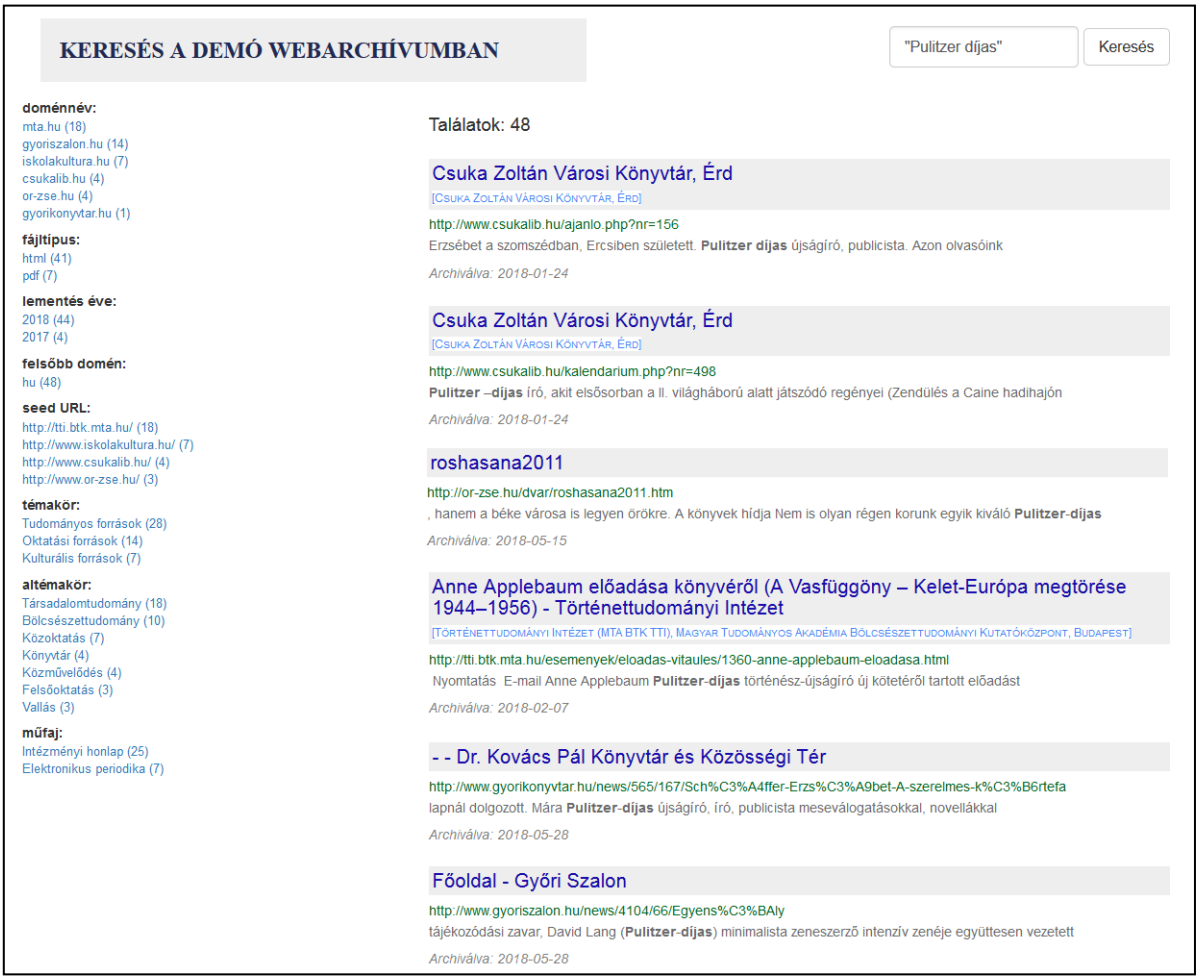
KÖNYVTÁRI HONLAPOK

A webhely neve	URL címe	OSZK mentés	Képernyőfotó	Linktérkép	Internet Archive	Eredeti oldal
Berzsenyi Dániel Városi Könyvtár, Marcali	www.marcalikonyvtar.hu					
Csuka Zoltán Városi Könyvtár, Érd	www.csukalib.hu					
Evangélikus Országos Könyvtár	konyvtar.lutheran.hu					
Esztergomi Főszékesegyházi Könyvtár	www.bibliotheca.hu					
EX LIBRIS Könyvtár	exlibriskonyvtar.hu					
Fejér György Városi Könyvtár, Keszthely	www.fgyvk.hu					
Helischer József Városi Könyvtár, Esztergom	www.vkesztergom.hu					

3. ábra **Részlet a nyilvános demó archívumból**

De ha a Heritrix robotja le is arat egy oldalt, még akkor sincs semmi garancia arra, hogy a begyűjtött tartalom megjelenítése az OpenWayback programban hibátlan lesz. Nagyon sok múlik az adott oldal technológiáján, az alkalmazott webprogramozási elemeken. Általános probléma például, hogy a különféle nyelvi verziókkal rendelkező oldalaknál olyan megoldást használnak a nyelvváltásra, amit az aratórobot nem tud lekövetni. Egy másik jellemző hiba, amikor szétesik az oldal grafikai elrendezése, mert a honlap külalakját meghatározó technológia miatt a szükséges elemek nem menthetők le vagy nem jelennek meg a Waybackben. Az is előfordul, hogy az aratórobot automatikusan átirányítódik a webhely akadálymentes változatára, vagy pedig a robots.txt fájlban le van tiltva a honlap elrendezését rögzítő stíluslaphoz való hozzáférés. (Ez adódhat abból is, hogy a webdizájnt olyan szellemi terméknek tekintik az oldal tervezői, amit nem akarnak letölthetővé tenni.) A tartalom ugyan mindkét esetben archiválva van, de a grafikai elrendezés hiányában annak értelmezése általában nagy nehézségekbe ütközik. Persze van olyan eset is, amikor azért nem tudunk aratni egy webhelyet, mert egyszerűen ki vannak tiltva róla a robotok. Ennek egy enyhébb formája, amikor nem a robots.txt fájlba, hanem a weboldalak forráskódjába tesznek olyan utasításokat, melyek azok szöveges tartalmának indexelését vagy a rajtuk található linkek követését tiltják meg. Ilyenkor megjelenik a dizájn és esetleg a kezdőoldal is hiánytalanul, de a honlap egyéb részeire vezető linkek már nem fognak működni az archivált verzióban.

Persze mindenki olyan hozzáférést enged a webes tartalmaihoz, amilyet szeretne. De, aki fontosnak tartja, hogy az általa közzétett tartalom a jövő számára is megőrződjön, az néhány szabály betartásával jelentősen megkönnyítheti ezt, hasonlóan a honlapok – ma már egyre elterjedtebb – akadálymentesítéséhez. Külön tanácsok vonatkoznak



4. ábra **A SolrMIA kereső találati listája, szűkítési feltételekkel**

5. ábra **Egy archivált weboldal az OpenWayback megjelenítőben**

arra, hogy miként legyen robotokkal könnyen bejárható egy webhely (*crawler-friendly website*)[24], illetve miként legyen jó minőségben archiválható és hosszú távon is megőrizhető (*archive-friendly website*)[25]. Egy robotbarát webhely releváns tartalma könnyen és teljesen felderíthető a keresőgépek és a webarchívumok által indított robotokkal, az érdektelen (pl. naplófájlok, segédállományok) vagy lementhetetlen (pl. adatbázisok, webáruházak, naptárak) részei viszont el vannak rejtve előlük. Egyebek mellett ilyen megoldásokkal tehető bejárhatóbbá egy webszerver tartalma:

- honlaptérkép (lehetőleg XML-ben), amely minden lényeges aloldalra elvezeti a robotot;
- a tartalom értékes része nincs túl mélyen a kezdőlapról indulva és linkeken keresztül is elérhető, nemcsak egy keresőűrlapon át;
- szabályos HTML linkek a Javascript-, Flash-, Java-alapú stb. megoldások helyett/mellett, amelyeket a robot is követni tud;
- az azonos tartalomra mutató sokféle belső link, vagy a végtelen körben egymásra hivatkozó linkek kerülése vagy kanonizálása a robotok számára;
- frame-ek, egérkattintásra aktiválódó layerek, dinamikusan generálódó tartalmak elkerülése, vagy legalább statikus és önálló URL címekkel rendelkező alternatívák generálása ezekből a robotok számára;
- jól konfigurált robots.txt, amely beengedi a robotokat, de csak a tényleges tartalmat szolgáltató, illetve számukra optimalizált részekre.

A robotbarát webhelyek kialakítására vonatkozó ajánlások elsősorban a keresőoptimalizáláshoz íródtak, de nagy részük az archiválási célból indított robotok esetében is hasznos, viszont az utóbbiak esetében még más szempontok is fellépnek. S itt érkezünk el az archívumbarát webhely fogalmához: azon felül, hogy az ajánlásnak megfelelő webhelyek könnyen bejárhatók robotokkal, a lementett tartalom jó minőségben archiválható is. Ez azt jelenti, hogy az archív változat tartalmában, megjelenésében és funkcionalitásában kellően hű mása az eredetinek. Az ehhez szükséges legfontosabb követelmények a következők:

- logikus site-struktúra, amelynek a felépítése az URL címekben is tükröződik, mert így könnyebb kiválasztani az archiválásra érdemes részeket és utólag ellenőrizni az eredményt;
- valid HTML és CSS kód, ami lehetővé teszi a helyes megjelenítést a szabványokat követő böngészőkben a jövőben is;
- ékezetek és egyes speciális karakterek kerülése az alkönyvtárak és a fájlok neveiben;
- lehetőleg nyílt fájlformátumok használata, melyek hosszú távon is megjeleníthetők maradnak;
- nincs a webhelyen olyan speciális formátumú tartalom, amihez külön megjelenítőt vagy böngészőkiegészítőt kell telepíteni;
- a hang- és a videotartalom nem sugárzott (stream) módon van beágyazva, hanem letölthető fájlok formájában (is);
- a robots.txt fájlban nincs letiltva a külalakot szabályozó (pl. CSS) fájlok letöltése;
- nem tartalmaz olyan szerver oldalon futó scripteket, programokat, vagy adatbázist, amelyek nélkül a website használhatatlan;
- a webszerver nem használ olyan session vagy persistent típusú cookie-kat, amelyek alapvetően befolyásolják a megjelenő tartalmat;
- részletes beágyazott metaadatok vannak a weboldalak fejlécében és az egyéb dokumentumokban (pl. képek, PDF fájlok), melyek megkönnyítik a begyűjtött digitális objektumok beazonosítását és automatikus metaadatolását;
- a készítés vagy az utolsó módosítás dátumának feltüntetése a weboldalakon és a dokumentumokban, hogy az archivált változat használója meg tudja állapítani, mikor készültek (ne csak azt lássa, hogy mikor lettek archiválva);
- a webhely jogi közleményében kitér az archiválásra is (pl. „archiválható, de csak fél év után szolgáltatható és csak könyvtáron belül”), vagy egy Creative Commons licenccel szabályozza a felhasználást az archivált példány esetében is.

Az Archive Ready szolgáltatás, illetve alkalmazási felület (API)[26] segítségével bárki saját maga is ellenőrizheti, hogy a honlapja eleget tesz-e a fontosabb ajánlásoknak. Fontos lenne a jövőben, hogy a magyar interneten is minél szélesebb körben teret nyerjenek ezek az ajánlások, mint ahogy az akadálymentesség terén ez már elég szépen megvalósult.

Hasznosítás

Ha jó minőségben sikerül aratni online tartalmakat és archívumot építeni belőlük, logikus kérdésként vetődik fel, hogy mire lehet jó egy ilyen gyűjtemény? A kísérleti projekt szerves részét alkotja annak felmérése is, hogy miképpen lehet hasznosítani a webarchívumokban begyűjtött információkincset.

Az első nagy témakör ezen a téren az igény szerinti archiválás[27]. Például az Internet Archive fizetős szolgáltatásaként működő Archive-It vállalja, hogy bármely intézmény vagy cég számára meg-

adott paraméterekkel mentést készít az ügyfél saját webhelyeiről vagy más honlapokról és egyéb online forrásokról. A learatott anyag a megrendelő tulajdonába megy át, aki tárolhatja azt az Archive-It szerverein és/vagy a saját eszközein is. A begyűjtött információkkal pedig a törvényes keretek között saját maga rendelkezik.

Sokszor felmerül igényként, hogy a tudományos vagy oktatási jellegű publikációkban való stabil hivatkozhatóság miatt a webes dokumentumoknak állandó URI címük és változatlan tartalmuk legyen akkor is, ha az eredeti dokumentumok URL-je vagy tartalma időközben megváltozik az élő weben, vagy egyszerűen eltűnnek onnan. Felvetődhet továbbá, hogy a weboldalakról készült mentések és képernyőfotók hitelesítéssel legyenek ellátva, melyek akár bizonyítékként is felhasználhatók hivatalos eljárások során. Mindkét esetre a webarchívumok tudnak megfelelő választ adni és ezek a fajta felhasználási formák az üzemszerű archiválás megindulását követően Magyarországon is megjelenhetnek. A jövőben remélhetőleg sikerül partnerséget kialakítani piaci szereplőkkel az ilyen igények kiaknázására.

A webarchívumok hasznosításához kapcsolódó második nagy terület a digitális bölcsészeti, társadalomtudományi, történeti vizsgálódások köre. Mostanában válik egyre nyilvánvalóbbá, hogy milyen sokszínű módokon lehet az archívumokban begyűjtött információhalmazt feldolgozni, értelmezni és újrahasznosítani. A lementett online források a különféle társadalmi jelenségek, mozgások újfajta elemzéseinek szolgálhatnak nyersanyagul. A nagy mennyiségű adat feldolgozása, értelmezése, az abból merített következtetések levonása nagyon izgalmas új kutatási utakat nyit meg például az adatbányászat és a történettudomány találkozásával. A tudományos vizsgálati módszerek lehetnek kvantitatívak, kvalitatívak, illetve vegyes jellegűek is. Nemrégiben indult el az „Internet Histories”[28] című folyóirat az ilyen irányú kutatások bemutatására és ösztönzésére. Az újszerű tudományos projekteket esettanulmányok keretében ismertető, *Niels Brügger* és *Ralph Schroeder* által szerkesztett „The Web as a History”[29] című tanulmánykötetről pedig magyar nyelvű recenzió[30] is készült. A webhistoriográfiai és az archivált online forrásokra épülő egyéb irányú kutatásoknak el kellene már indulniuk nálunk is, és ehhez természetesen szükség lesz a nemzeti könyvtár és az egyetemek, kutatóintézetek közötti együttműködésekre is.

Együttműködés

A webarchiválás sikeressége szempontjából az együttműködés igénye egyéb szinteken is megjelenik. A kísérleti projekt honlapjáról elérhető egy javaslattevő űrlap, amellyel bárki javasolhat értékes magyar webhelyeket archiválásra. Ezt a lehetőséget szélesebb szakmai körben is meghirdettük: hírlevelekben, könyvtáros fórumokon, levelezőlistákon, közösségi médiafelületeken. Szeretnénk intenzívebben nyitni a határon túli magyar könyvtárak és kulturális szervezetek irányába is, hogy segítsenek nekünk a magyar kulturális örökség részét képező, digitálisan születő tartalmak válogatásában.

A közgyűjteményi partnerekkel (könyvtárak, múzeumok, levéltárak) feltétlenül szükségesnek tartjuk a szakterületi, illetve földrajzi jellegű munkamegosztás kialakítását. Egyetlen országban sem képes a nemzeti könyvtár az internet megőrzésének teljes vertikumát felvállalni. Ott működnek igazán jó archívumok, ahol egy egész intézményhálózat szakmai tudása és szolgáltatási képessége áll mögöttük.

Az együttműködések fontos előfeltétele egy olyan képzési háttér kialakítása, melynek révén el tudják sajátítani kollégáink a webarchiváláshoz szükséges készségeket, képességeket. A magyarországi továbbképzés megteremtése érdekében a Könyvtári Intézettel együtt tanfolyamot tervezünk „Az internet archiválása mint közgyűjteményi feladat” címmel. (Jelenleg az akkreditáció folyamata zajlik, reményeink szerint legkésőbb jövő év elején el tudjuk indítani a továbbképzést.) Ez egészülne ki az *Országos Könyvtári Platform* projekt e-learning ágához kapcsolódva egy olyan online tanulási felülettel, ami szintén naprakész tudás elsajátítását teszi lehetővé. Az itthoni terveinket nagyban segíti az IIPC tavaly szerveződött oktatási és képzési munkacsoportjának tevékenysége. Ennek keretében nemzetközi szinten zajlik a tananyagok és kurzusok fejlesztése. A munkacsoport egyik aktív tagja Németh Márton, aki 2017 őszén részt vett a dániai *Aarhusi Egyetem* Netlab kutatócsoportja által szervezett e-learning képzésen is. Ezen a kurzuson alaposan körbejárhattuk a webarchiválás különféle technikai kihívásait, illetve az archívumok tudományos célú hasznosításának kérdéseit is. Az oktatás témájáról bővebben a Networkshop 2018 konferencia előadásaiból szerkesztett – e cikk írásakor még megjelenés alatt álló – kötetben levő tanulmányunkból lehet részletesebben tájékozódni.

Az intézményi együttműködésnek a webarchiválás terén két fő formája lehetséges. Az egyik esetben a partnerek önállóan végeznek archiválást, amihez az OSZK biztosítja a tárhelyet, a learatott anyagok pedig bekerülnek a Magyar Internet Archívumba. A másik mód pedig az, amikor egy intézmény saját szerverén, saját infrastruktúrával épít archívumot. Ebben az esetben is megoldható az archívumok összehangolása az URL-ek lekérdezésének szintjén, az ún. *memento* protokoll[31] segítségével. Így lehetővé válik, hogy ha valaki felad egy keresőkérdést, akkor több archívum anyagából is kapjon találatokat, melyek különböző időpontbeli mentésekre mutatnak. Az üzemszerű webarchiválás magyarországi megindulása után a memento protokoll a nemzetközi együttműködésben is nagy lehetőségeket rejt. Lekérdezhetővé tudjuk tenni a Magyar Internet Archívum anyagát partnereink felé és ezzel együtt mi is hozzáférhetünk a más intézményekben (pl. a szlovák vagy az osztrák nemzeti könyvtárban) őrzött magyar vonatkozású weboldalakhoz. Emellett felmerülhet az Internet Archive által 1996 óta lementett jelentős magyar tartalom egészének vagy részhalmazainak megvásárlása is.

A nemzetközi kapcsolatok építése amúgy is fontos része a projektnek. 2018 januárjában az OSZK is csatlakozott az IIPC konzorciumhoz, melynek már kb. 45 országból vannak tagjai. Személyesen felvettük a kapcsolatot a szlovák, a cseh és az osztrák archívumok képviselőivel. Sikerült részt vennünk az IIPC webarchiválással foglalkozó ülésszakán az IFLA 2017 konferencián. Ez utóbbi egy globális esemény volt, így nyugat-európai, amerikai, ausztrál kollégáknak is sikerült bemutatkoznunk, tapasztalatokat cserélni velük. Az ismerkedésre jó alkalom nyílott a már említett, Aarhusból szervezett online szemináriumon is. Itt a dán kollégák mellett felvettük a kapcsolatot a szintén velünk nagyjából egy időben indult belga projekt munkatársaival. A webarchiváláshoz szükséges technikai háttér meghatározása, a leendő szolgáltatások tervezése kapcsán eredményes együttműködést folytatunk velük. Nagyon fontos külföldi partnerünk a magyar származású *Kees Teszelszky*, aki Hollandia nemzeti könyvtárában az internetarchiválási projektet irányítja. Széles körű beágyazottsága, kapcsolatrendszere, önzetlen támogatása hatalmas segítséget jelent számunkra.

Ismeretterjesztés

A magyar internetarchívumot előkészítő projekt fontos feladatának tartjuk, hogy mind a szakmai körökben, mind pedig a szélesebb nyilvánosságban minél többen értesüljenek róla, hogy elindult egy ilyen irányú tevékenység a nemzeti könyvtárban, és hogy akiket ez érdekel, azok kapcsolódjanak be, vagy a tapasztalatainkat felhasználva kezdjenek el saját – akár magán, akár intézményi – archívumokat építeni. Az ehhez szükséges ismeretterjesztést szolgálja a projekt ideiglenes honlapja[32] és a rajta megjelenő hírek, dokumentumok, Twitter üzenetek; a világban működő webarchívumokat és szervezeteket, a főbb rendezvényeket és fórumokat, az ehhez a munkához hasznos szoftvereket és szolgáltatásokat, a formátumokat és fogalmakat ismertető, már több mint 580 szócikkből álló MIA wiki[33]; a válogatott külföldi és hazai bibliográfiák[34]; és a zárt levelezőcsoportként működő MIA-L lista[35], amelyre várjuk a téma iránt érdeklődők feliratkozását. Hasonló célt szolgálnak a témában publikált cikkek, a konferenciákon és egyéb rendezvényeken tartott előadások, a webkettes csatornákon közzétett hírek, és az első alkalommal 2017 októberében tartott „404 Not Found – Ki őrzi meg az internetet?” című workshop, melyet szeretnénk még legalább néhány évig megismételni az OSZK-ban.

Irodalom

Dancs Szabolcs: Webarchiválási politikák. In: Könyv, könyvtár, könyvtáros, 2011. (20. évf.), 10. sz. pp. 14–20.

Drótos László: Mi a MIA? : Javaslat egy Magyar Internet Archívum létrehozására. In: Tudományos és Műszaki Tájékoztatás, 2006. (53. évf.), 6. sz. pp. 267–274.

Drótos László: Az internet archiválása mint könyvtári feladat. In: Tudományos és Műszaki Tájékoztatás, 2017. (64. évf.), 7–8. sz. pp. 361–371.

Drótos László – Németh Márton: A webarchiválás oktatása. In: Networkshop 2018 konferenciakötet (megjelenés alatt!)

Drótos László – Kokas Károly: Webarchiválás és a történeti kutatások. In: Digitális Bölcsészet (megjelenés alatt!)

Hegyközi Ilona: Hol tart ma a webarchiválás? In: Könyvtári Figyelő, 2014. 4. sz. pp. 527–534.

Kornhoffer Mónika: Internet-archívumok hazánkban és Közép-Európában. In: Felderítő Szemle, 2011. (10. évf.), 3–4. sz. pp. 63–78.

Németh Márton: A webarchiválásról történeti megközelítésben. In: Könyv Könyvtár Könyvtáros, 2018. (27. évf.), 2. sz. pp. 48–52.

Németh Márton: Nemzetközi körkép a webarchiválás gyakorlatáról. In: Könyvtári Figyelő, 2017. (63. évf.), 4. sz. pp. 575–582.

Hivatkozások

1 Az OKR projekt ismertetője az OSZK honlapján: http://www.oszk.hu/okr-projekt

2 http://archive-it.org

3 http://www.netpreserve.org

4 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Heritrix

5 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Wayback

6 https://web.archive.org

7 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/WARC

8 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/WCT

9 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/NetarchiveSuite

10 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/WAIL

11 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Webrecorder

12 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Webrecorder_Player

13 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/WARCreate

14 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/SolrWayback

15 MIA wiki szócikk: http://mekosztaly.oszk.hu/mia/demo/

16 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/HTTrack

17 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/ScrapBook

18 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Grab_Them_All

19 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Nimbus_Screen_Capture

20 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Metadata_Extraction_Tool

21 A Web Archiving Metadata Working Group weboldala: https://www.oclc.org/research/themes/research-collections/wam.html

22 http://mekosztaly.oszk.hu/mia/xml/

23 Az XML Notepad szócikke az angol Wikipédiában: https://en.wikipedia.org/wiki/XML_Notepad

24 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Crawler-friendly_website

25 MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Archive-friendly_website

26 Archive Ready honlap: http://archiveready.com
Archive Ready API: http://archiveready.com/docs/api.html

27 On-demand alapú webarchiváló szolgáltatások felsorolása a MIA wikiben: http://mekosztaly.oszk.hu/mediawiki/index.php/Igény_szerinti_archiválás

[28] https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=rint20

[29] http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf

[30] Németh Márton. A webarchiválásról történeti megközelítésben. Könyv Könyvtár Könyvtáros 27. évf. 2. szám (2018.) pp. 48–52. http://ki2.oszk.hu/3k/2018/06/a-webarchivalasrol-torteneti-megkozelitesben/

[31] MIA wiki szócikk: http://mekosztaly.oszk.hu/mediawiki/index.php/Memento_Project

[32] http://mekosztaly.oszk.hu/mia

[33] http://mekosztaly.oszk.hu/miawiki

[34] http://mekosztaly.oszk.hu/mia/doc/webarchivalas-irodalom.html
http://mekosztaly.oszk.hu/mediawiki/index.php/Kateg%C3%B3ria:IRODALOM

[35] http://mekosztaly.oszk.hu/cgi-bin/mailman/listinfo/mia-l

Beérkezett: 2018. VI. 22-én.

Drótos László
könyvtáros
OSZK – E-könyvtári Szolgáltatások Osztálya.
E-mail: *drotos.laszlo@oszk.hu*

Németh Márton
webkönyvtáros
OSZK – E-könyvtári Szolgáltatások Osztálya.
E-mail: *nemeth.marton@oszk.hu*