Beszámoló a webarchiválási pilot projekt során elvégzett könyvtárszakmai feladatokról és a jövőbeli tervekről

Összeállította:
Drótos László témafelelős
<drotos.laszlo@oszk.hu>

Utolsó módosítás: 2017. október 27.

Eddigi eredmények

Tervezés:

A projekttel kapcsolatos előmunkálatok 2017 tavaszán kezdődtek. Elkészült egy 2018 végéig szóló koncepció, melyet az illetékesekkel való egyeztetés után 2017 április 17-én elfogadottnak nyilvánítottunk. Ezen szöveges dokumentum alapján egy részmunkákra lebontott feladattábla is összeállításra került, amiben már a felelősök, az érintettek és a tervezett határidők is szerepelnek, valamint az egyes feladatok aktuális állása is nyomon követhető. Megfogalmazásra került továbbá a webarchívum gyűjtőköre, amely egy ideiglenes szöveg, amit a tapasztalatok alapján az üzemszerű működés megkezdése előtt még szükség esetén módosítani kell.

Munkatársak:

A projekt témafelelőse az OSZK részéről Drótos László (helyettesítője Moldován István), az informatikus pedig Kovács Péter (részmunkaidőben). A projekt idejére sikerült két főállású munkatársat felvenni az OSZK E-könyvtári Szolgáltatások Osztályára: webkönyvtárosi munkakörbe Németh Mártont (2017. május 15-től), webarchívum adminisztrátori munkatársként pedig Visky Ákos Lászlót (2017. július 3-tól). Az archívummal kapcsolatos rendszergazdai és kisebb programozási feladatok elvégzésére pedig a Vitéz Bt. részéről Vitéz Gábor kap megbízást havi 40 órában (2017. október 1-től).

Szoftverek:

A legtöbb nagy webarchívum két fő komponense, a Heritrix crawler és az Open Wayback megjelenítő mellett további szoftverekkel is kísérleteztünk, melyek egy része beépíthető lesz a leendő üzemszerű rendszerbe, illetve annak egyes munkafolyamatait tudja majd segíteni. A legfontosabb ezek közül az új-zélandi fejlesztésű, open source Web Curator Tool, amely egyfajta keretrendszer a Heritrixhez: nyilvántarthatók vele a szelektív archiválásra kiválasztott webhelyek, a tartalomgazdáknak küldött engedélykérő levelek és a kapott engedélyek, ütemezhetők és paraméterezhetők a mentések, és megadhatók a hozzáférési jogosultságok. A WCT-t több nemzeti könyvtár is használja, de a fejlesztése az utóbbi években lelassult, és a legnagyobb problémája, hogy a Heritrix régi, elavult verzióját használja. Biztató fejlemény viszont, hogy a 2017 júliusában kiadott WCT v1.6.3 már a jelenlegi Heritrix 3.x verziót támogatja, igaz, még csak kísérleti módban (ezt még nem tudtuk tesztelni).

Csináltunk próbamentéseket a szintén a Heritrixre épülő és Windows ill. Linux alatt is működő WAIL szoftverrel, amely a tömeges és automatizált aratások helyett az egyéni archiválást segíti és az új verziója a Chrome böngészőt használva olyan Javascriptben gazdag (pl. webkettes) oldalak mentésére is képes, amelyekkel a Heritrix nem igazán boldogul. Ehhez

kapcsolódik a Chrome-ba önállóan is beépíthető a WARCreate, amellyel egy weboldal a Heritrix által is használt szabványos WARC fájlba menthető. Szintén WARC fájlokat állít elő az ingyenes online szolgáltatásként vagy pedig Linux alá telepíthető szoftverként egyaránt használható Webrecorder, amivel egy böngészési folyamat rögzíthető és ami különösen az olyan, emberi közreműködést is igénylő tartalmak mentésére jó, mint például egy Facebook vagy Instagram oldal, vagy egy Twitter csatorna. Az így mentett tartalom megjelenítésére egy Webrecorder Player nevű eszköz is tartozik hozzá, amellyel Windows alatt is megnézhetők a WARC fájlok. Végeztünk továbbá mentéseket a szintén népszerű és nagyon felhasználóbarát felülettel rendelkező webhely-letöltő program, a HTTrack Windows-os változatával. Ez ugyan fájlrendszerbe ment, de létezik egy HTTrack => ARC konverter (ezt még nem tudtuk kipróbálni).

Az archívumban való böngészéshez, illetve az archivált verzió és az eredeti weboldal közötti eltérések összehasonlításához, valamint a mentés idején használatos böngészőkben való megjelenés megőrzéséhez fontos, hogy mind az élő, mind pedig az archivált webhelyekről (legalább a kezdőoldalról és esetleg néhány fontosabb aloldalról) képernyőfotók készüljenek. Erre a célra a Grab Them All nevű Firefox kiegészítőt teszteltük, mellyel jó minőségű, a teljes weboldalt mutató PNG vagy JPG képek készíthetők egy megadott URL címlista alapján (továbbá parancsmódban és szerveren is futtatható).

Összességében az a tapasztalatunk, hogy a tesztelt ingyenes programok különböző részfeladatokra alkalmasak, jól ki tudják egészíteni a Heritrix aratószoftvert, de szinte mindegyiknek vannak olyan hibái, hiányosságai, amelyekre figyelni kell a használatuk során, vagy amelyeket remélhetőleg a későbbi verziókban kijavítanak majd a fejlesztőik.

Tesztaratások:

2017 tavaszától októberig folyamatosan zajlottak terhelési és egyéb tesztek, melyekhez a kiinduló (seed) URL címek egyrészt egy 2011-es, 400 ezer darab .hu alatti domain nevet tartalmazó listáról lettek véletlenszerűen kiválasztva, másrészt mi is összeállítottunk különböző címlistákat: az EPA-ban (Elektronikus Periodika Archívum és Adatbázis) „távoli”-ként regisztrált több mint 2 ezer online időszaki kiadvány, a 2016 decemberi Nyilvános Könyvtárak Jegyzékében szereplő kb. 500 könyvtári site, továbbá kb. 600 múzeum és galéria honlapja, valamint több mint 100 levéltári témájú webhely került lementésre egy-két alkalommal. Ezek mellett az OSZK saját online felületeiről (honlap, Facebook oldal, virtuális kiállítások) is csináltunk próbamentéseket. Az archivált verziókat szúrópróbaszerűen – az Open Waybackkel megjelenítve – összevetettük az eredeti honlappal, valamint sok esetben az Internet Archive-ban található legutolsó mentéssel is, és kigyűjtöttük a legjobb és a legproblémásabb eseteket. Előbbiek jelentik a leendő nyilvános demó szolgáltatás alapját, melyekhez az engedélykérő levelek kiküldését már elkezdtük. Utóbbiaknál pedig igyekszünk majd kideríteni a hibák okait, és vagy a Heritrix beállításainak módosításával, vagy más aratószoftverrel, vagy pedig az eredeti honlap gazdájának bevonásával próbáljuk meg lehetőség szerint orvosolni az archiválásnál fellépett problémákat.

Ismeretszerzés:

Az internetes tartalmak archiválásáról rengeteg ismeret összegyűlt az elmúlt két évtizedben, mióta az Internet Archive és egyes közgyűjtemények foglalkoznak ezzel feladattal. A külföldi publikációk feldolgozását, átolvasását folyamatosan végezzük és egy válogatott bibliográfiát is készítettünk belőlük. Egy fontos szakkönyvet, Adrian Brown „Archiving websites : a practical guide for information management professionals” című művét pedig kérésünkre megvásárolta a Könyvtártudományi Szakkönyvtár. Bár ez egy 2006-os kiadás, de olyan alapelveket tartalmaz, amelyek azóta is érvényesek. A Szakkönyvtár emellett beszerezte a WARC

fájlformátumot definiáló ISO 28500:2009 szabványt, valamint az ISO/TR 14873:2013 számú Technical Report-ot, amely a webarchívumok statisztikai, fogalmi és minőségi jellemzőinek meghatározásához hasznos. Sok értékes ismerettel szolgál továbbá a dán Aarhus University által a NetLab projekt keretében meghirdetett online kurzus a webarchiválásról és a webarchívumokról, amelyhez Németh Márton a kevés külföldi résztvevő egyikeként tudott csatlakozni.

Ismeretterjesztés:

A fenntarthatóság és a munkamegosztás miatt kiemelten fontosnak tartjuk, hogy a magyar közgyűjteményekben legyenek olyan szakemberek, akik értenek az online tartalmak lementéséhez és hosszú távú megőrzéséhez. Ennek érdekében folyamatosan és egyre több csatornán terjesztjük az eddig megszerzett ismereteket és tapasztalatokat. Legegyszerűbben a projekt ideiglenesnek szánt weblapján tájékozódhatnak az érdeklődők a témában, de már az OSZK központi honlapján is elérhetők a legfontosabb információk és a publikációk, előadások. A szakkifejezésekről, formátumokról, projektekről, szoftverekről és szolgáltatásokról, fórumokról és rendezvényekről elkészült egy folyamatosan karbantartott wiki, mely jelenleg kb. 550 szócikket tartalmaz.

Több bemutatót és előadást is tartottunk az elmúlt fél évben: a Könyvtári Intézetben, a MEK Egyesület közgyűlésén, „A jövő könyvtára felé...” webináriumon, az MKE Vándorgyűlésén, az idei Coginfo.com konferencián (és már jelentkeztünk előadással a 2017 novemberében megrendezésre kerülő MUZEUM@DIGIT, valamint a 2018-as BOBCATSSS konferenciákra is). 2017. okt. 13-án az OSZK-ban egy workshop-ot szerveztünk „404 Not Found – Ki őrzi meg az internetet?” címmel, melyet mind az előadók, mind pedig a résztvevők sikeresnek ítéltek meg. (A rendezvényen kb. 70 fő vett részt.) Ezen célzott alkalmak mellett egyéb eseményeken, szakmai találkozókon, tanórákon is kitérünk mostanában már a webarchiválásra, mint közgyűjteményi feladatra. A téma iránt komolyabban érdeklődők számára májusban elindítottunk egy levelezőlistát, melynek jelenleg 23 tagja van.

Általunk írt tanulmányok és szemlék jelentek meg vagy vannak megjelenés alatt olyan szakfolyóiratokban, mint a TMT, a Könyvtári Figyelő és a Digitális Bölcsészet. Elkezdtünk megszervezni egy 30 órás akkreditált továbbképző tanfolyamot „Az internet archiválása mint közgyűjteményi feladat” címmel. A tanfolyam tervezete elkészült és ha a kb. 100 ezer forintos akkreditációs költségre is sikerül forrást találni, akkor a Könyvtári Intézet 2018 tavaszán vagy őszén már meg tudja ezt hirdetni a közgyűjteményi dolgozóknak.

Kapcsolatépítés:

E-mailben illetve személyesen már jó néhány hazai és külföldi intézménnyel (pl. közgyűjtemények, egyetemi tanszékek, webarchívumok) felvettük a kapcsolatot. A külföldiek közül a holland, a belga, a dán, a szlovén, a szlovák és az osztrák kollégákkal vagyunk kapcsolatban. (Utóbbiakkal okt. 24-én személyesen is találkoztunk Bécsben, nov. 9-én pedig Pozsonyba tervezünk ellátogatni, az ottani egyetemi könyvtár által működtetett nemzeti webarchívumba.) Elindítottuk az OSZK csatlakozási kérelmét az internet archiválással foglalkozó intézményeket tömörítő IIPC (International Internet Preservation Consortium) konzorciumhoz és okt. 13-án meg is érkezett a kedvező válasz, mely szerint az IIPC Steering Committee jóváhagyta a magyar csatlakozást, így már csak a szerződés aláírása és a tagdíj befizetése van hátra. Teljes jogú tagként még több lehetőségünk nyílik majd a kapcsolatépítésre és a nemzetközi munkamegosztásban való részvételre.

További tervek

Nyilvános demó:

Az eddigi tesztek alapján a jól archiválható honlapok közül azokból, amelyekre az eredeti tartalomgazdától engedélyt kapunk, mielőbb szeretnénk egy nyilvánosan is böngészhető demó archívumot készíteni egy egyszerű, ideiglenes szolgáltatófelülettel.

Szoftvertesztek:

A nyár folyamán már megnézett programok időközben megjelent új verzióinak, illetve saját szerverre telepíthető változatainak tesztelése mellett olyan további szoftverek kipróbálását tervezzük, mint a szintén a Heritrixre épülő NetarchiveSuite és az ehhez hozzáfejlesztett BCWeb keretrendszer (utóbbi jelenleg nem open source, de ígéretünk van rá, hogy kipróbálhatjuk), valamint a böngésző-alapú Brozzler és Crawljax nevű crawlerek, továbbá a különböző WARC fájl és CDX indexfájl manipuláló eszközök.

Rendszerépítés:

Az új hardver konfiguráció beszerzésével párhuzamosan, a szoftvertesztek tapasztalataira támaszkodva elkezdjük egy, már tervszerű archiválásra is alkalmas rendszer egyes alkotóelemeinek összerakását: aratószoftver(ek), ütemező modul, engedély-nyilvántartás, metaadat kezelés, képernyőfotók tárolása, teljes szövegű és metaadatkereső, a webarchívumok közös kereshetőségét biztosító Memento protokoll, stb. Továbbá a webarchívum rendszerét és munkafolyamatait integrálnunk kell majd az OSZK szintén beszerzés alatt levő új könyvtári rendszerének egyes moduljaival: metaadatolás, jogkezelés, közös kereső (discovery eszköz), hosszú távú megőrzés.

Szelektív archiválás:

További tematikus és eseményalapú URL címgyűjteményeket állítunk össze (összességében egy-két éven belül néhány ezer címet, végső célként kb. 15 ezret) és ezeket a webhelyeket észszerű mélységig (3-10 szint, a tartalom fontosságától és mennyiségétől függően) észszerű gyakorisággal (a legnagyobb híroldalakat naponta, a többi webhelyet heti vagy 1-2 havi rendszerességgel, a változékonyságuktól függően) elkezdjük menteni, kizárva az aratásból a nagyméretű (pl. videó ill. csomagolt) fájlokat és korlátozva az egy webhelyről letöltött tartalom mennyiségét is. A fontos magyar honlapok válogatásába és a mentések metaadatolásába egyes partnerintézmények munkatársait is szeretnénk bevonni. A szelektíven mentett webhelyek egy része (amelyeknél erre engedélyt kapunk) nyilvánosan is elérhető lesz, a teljes anyag pedig dedikált gépekről (pl. OSZK- és NAVA-pontok) lesz hozzáférhető. Egy-egy webhely minden korábbi mentését szolgáltatni szeretnénk, tehát ennél a gyűjteménynél mindent diszken kell tárolni a gyors eléréshez, és a metaadatok keresése mellett a teljes szövegben való keresést is biztosítanunk kell.

Domain szintű aratás:

Reprezentatív „pillanatfelvételek” készítése céljából megkíséreljük a .hu alatt bejegyzett, jelenleg már több mint 700 ezer domain mentését évi két alkalommal, legfeljebb 3 szint mélységig és erős szűrési feltételeket beállítva a maximális egyedi fájlméretre, az egy site-ról mentendő maximális összméretre, az egy site-ról letöltött objektumok maximális számára, az egy site-on eltöltött maximális időre, az egyes kérések közötti várakozási időre, a teljes aratásra fordított maximális időre, a teljes letöltött tartalom maximális összméretére és a kizárt fájlformátumokra, továbbá figyelembe véve a webszervereken levő robots.txt fájlokban megfogalmazott tiltásokat. Ha sikerül az internetes biztonsággal és a gyermekek védelmével

foglalkozó szervezetektől tiltólistákat kapni, akkor ezekkel előzetesen megszűrjük a teljes .hu listát, hogy csökkentsük a szemét illetve a káros tartalom mennyiségét, tárhelyet és gépidőt takarítva így meg. Egy-egy ilyen domain szintű aratást – ha ez technikailag megoldható – néhány hét, max. egy hónap alatt lenne célszerű lebonyolítani, hogy viszonylag egyidejű pillanatkép készüljön, vagyis ha az archívumban egy weboldalon egy linkre kattintunk, akkor a linkelt oldalon lehetőleg ne egy egészen más tartalom legyen, mint ami eredetileg volt, amiatt, mert a két oldal mentése közt több hónap telt el és közben nagyon megváltozott valamelyik tartalma. Az így lementett anyagból (egyértelmű és részletes jogszabályi felhatalmazás híján) még nem lesz semmilyen szolgáltatás, zárt archívumként őrizzük csak az utókornak, ezért – a statisztikai kielemzés elvégzése után – egyelőre elegendő lesz a szalagos tárolás. Hosszabb távon elképzelhető majd, hogy kutatási célból kisebb-nagyobb halmazokat diszkre kell visszamásolni belőle meghatározott ideig, amíg a kutatás lezajlik.

Statisztikák:

Mind a szelektív, mind pedig a domain szintű aratás során begyűjtött tartalomról különböző statisztikákat és grafikonokat készítő scripteket, programokat kell írni és/vagy beüzemelni, amelyek áttekinthető formában mutatják meg annak összetételét különféle szempontok szerint, valamint ezen jellemzők időbeli változását (pl. hostok, objektumok és linkek száma, tárhelyfoglaltság, fájlméretek és MIME típusok eloszlása).

Ajánlás:

A tapasztalataink és a hasonló külföldi útmutatók alapján egy ajánlást fogunk kidolgozni a robotokkal jól bejárható, jól archiválható és hosszú távon is jól megőrizhető webhelyekre, és összeállítunk egy mintapéldákat tartalmazó gyűjteményt is a típushibákról. Ezekre felhívjuk majd a webmesterek figyelmét (elsősorban a szelektív archiválásra kiválasztott site-ok esetében).

Crowdsourcing:

Létrehozunk egy online űrlapot, amelyen keresztül bárki javasolhat magyar vagy magyar vonatkozású webhelyet megőrzésre. A beérkező javaslatokat ellenőrizzük és amennyiben a gyűjtőkörbe esnek és még nincsenek benne a szelektív archiválás adatbázisában, akkor bevisszük őket és engedélyt kérünk a tartalomgazdáktól.

Szabályozás:

Elkezdjük kidolgozni a webarchiválással és a webarchívummal kapcsolatos belső szabályzatokat (pl. hozzáférés, személyi adatok védelme) és jogszabály-módosítási javaslatokat fogalmazunk meg a jogalkotók számára.

Hasznosítás:

Ösztönözni fogjuk, hogy a kutatóintézetekben és az egyetemi tanszékeken induljanak el olyan kutatások, amelyek a webarchívumokban található tartalom tudományos vagy egyéb célú hasznosítására vonatkoznak (pl. történeti kutatások, adat- és szövegbányászat, vizualizáció).

Retrospektív archiválás:

Tárgyalásokat kezdünk arról, hogy az élő webről már eltűnt, de valahol (pl. az Internet Archive, a Common Crawl és a PetaByte Kft. anyagában, a szomszédos országok archívumaiban, az itt-ott „elfekvő” winchestereken) még megtalálható régebbi magyar webes tartalmakat hogyan lehetne az OSZK állományába átvenni. Ez egyrészt sok szervezést és esetenként (pl.

Internet Archive) sok pénzt is igényel, de ha esetleg minden feltétel adott lesz, akkor jelentős plusz tárhelyet is biztosítani kell majd ehhez a visszamenőlegesen begyűjtött anyaghoz.

Távlati tervezés:

2018 őszén el kell készítenünk egy üzemszerűen működő, mindenféle online tartalom gyűjtésére, biztonságos megőrzésére és szolgáltatására alkalmas magyar internet archívum rendszertervét és fenntarthatósági tervét.