Az OSZK webarchiváló kísérleti (pilot) projektjének eredményei és egy üzemszerűen működő magyar webarchívum terve

DRÓTOS László – MOLDOVÁN István

Az alábbi tanulmány a „404 Not Found – Ki őrzi meg az internetet?" című szakmai műhelynapon 2018. november 15-én, az OSZK-ban elhangzott két előadás (Drótos László: Az OSZK webarchiváló pilot projekt eredményei és Moldován István: Egy üzemszerű magyar webarchívum terve és megvalósításának feltételei) szerkesztett változata.

Terra incognita

A webarchiválás ismeretlen terület a hazai memóriaintézményekben, miközben a világ számos országában már tíz-húsz éve folyik ez a tevékenység, sőt a legrégibb és legnagyobb ilyen projekt, az Internet Archive[1] (IA) nem sokkal a World Wide Web technológia népszerűvé válása után, 1996-ban indult, és gyűjteménye 2018 végén több mint *345 milliárd* weboldalból állt. Az Országos Széchényi Könyvtárban (OSZK) az E-könyvtári Szolgáltatások Osztályon közel két évtizede foglalkozunk a digitálisan születő dokumentumok megőrzésével: kezdetben a könyv jellegű művekkel, majd az időszaki kiadványokkal és a képekkel, sőt egy-két honlapot, virtuális kiállítást is tükrözünk és szolgáltatunk a Magyar Elektronikus Könyvtár (MEK) szerverén. 2006-ban megfogalmaztuk egy Magyar Internet Archívum (MIA) koncepcióját, majd ezt követően egy kis webarchívumot is kialakítottunk a HTTrack programot használva, az OSZK Infrastruktúra Szolgáltatások Osztályon pedig *Kovács Péter* informatikus a tömeges webarchiválásra alkalmas Heritrix szoftvert tesztelte. Mégis, amikor 2017 tavaszán elkezdtünk komolyabban foglalkozni ezzel a témával, hogy egy pilot (kísérleti) projekt keretében megteremtsük a nemzeti szintű webarchívum feltételeit, akkor szembesültünk azzal, hogy milyen sokféle tudásra van ehhez szükség. A technikai és jogi kérdésektől kezdve a válogatási és gyűjteményépítési szempontokon át a szolgáltatásig és az egyéb célú hasznosításig igencsak kiterjedt irodalma és széles körű gyakorlata van ennek a szakterületnek.

Ugyancsak kiterjedt már az a *magyar webtér* is, amelyet legalább részlegesen meg kellene őrizni a jövő számára, és amit így lehet definiálni: a magyarországi domén *(.hu*) alá bejegyzett címeken levő webhelyek, valamint a külföldi doméneken magyar természetes vagy jogi személyek által létrehozott webhelyek összessége a jelenben, továbbá minden olyan egyéb weboldal az élő weben, amelyet magyar célközönségnek szánnak vagy magyar vonatkozású.

Hogy mekkora lehet ez a halmaz, arra még közelítő becslést sem könnyű adni. Az Internet Szolgáltatók Tanácsának (ISZT) 2018. december 1-jei statisztikája[2] szerint a .hu országdomén alatt regisztrált nevek száma közel 739 ezer. De ennek jelentős része vagy már, vagy még nem használt, vagy csak névvariáns, és amelyiken van valamilyen szolgáltatás, az sem biztos, hogy webszerver, hanem mondjuk csak levelezésre használják az adott címet. Ugyanakkor egy olyan név alatt, mint például az *oszk.hu* vagy a *blog.hu* több, akár sok száz vagy sok ezer aldomén is lehet, amelyeket már nem tart nyilván az ISZT, és amelyeken önálló webhelyek vannak. Abból a szempontból szerencsések vagyunk, hogy a „hu" rövidítés nem olyan értelmes angol szó, mint például a „me" (Montenegró) vagy a „tv" (Tuvalu), és nálunk nem lehet ingyenesen regisztrálni (ellentétben mondjuk a Tokelau-szigetekkel, ahol emiatt több mint 18 millió.tk végű név van bejegyezve, miközben a három polinéziai atoll összlakossága csak 1500 fő). Ezért a .hu alatt levő webhelyek alapvetően magyaroknak szánt tartalmat szolgáltatnak, nem veszik meg tömegesen a domén neveinket a külföldi cégek azért, hogy utána a saját szerverükre irányítsák át őket. Viszont ugyancsak meghatározhatatlan mennyiség a más országdomének (pl. .ro, .sk), vagy az európai uniós (.eu), vagy a három, illetve több betűs (pl. .com, .info) általános legfelső szintű tartományok alatt működő magyar vagy magyar vonatkozású webszerverek száma. A webkettes, közösségi platformokon pedig már nem is beszélhetünk olyan értelemben webhelyekről, mint a honlap, a blog, vagy a hírportál. A Facebook, Instagram, YouTube, Twitter és társai esetében egy-egy gigászi webhelyen belül vannak elszórva a magyar aloldalak, csatornák, üzenőfalak.

A *magyar webtartalom* még ennél is nagyobb halmaz, ha hozzávesszük azokat a magyar webtérben korábban létezett digitális tartalmakat is, amelyek ugyan az élő weben már nem érhetők el, de valahol (pl. egy webarchívumban vagy egy tárolóeszközön) megőrződtek. Magyarországon tudomásunk szerint először a PetaByte Kft. csinált nagyobb méretű webaratást. 2013–2015/2016 között a HBONE-t használó, kb. ötszáz felsőoktatási és tudományos intézmény honlapját, valamint nagyobb hazai hírportálok anyagát mentették le kutatási célokra, és ez az anyag még mindig megvan. Az amerikai Internet Archive pedig több mint egymilliárd, a .hu doménről lementett weboldalt őriz 1996-ig visszamenőleg. Hogy más országok – elsősorban a szomszédos Szlovákia, Csehország, Ausztria, Szlovénia és Horvátország – nemzeti webarchívumaiban mennyi lehet a magyar vonatkozású tartalom, azt nem tudjuk felmérni, mint ahogy azt sem, hogy a régi szervereken és a különféle tárolókon hány régi honlap anyaga hever, amelyeket még be lehetne gyűjteni, meg lehetne menteni a digitális enyészettől. Természetesen mindent nem lehet és nem is érdemes megőrizni, mindenképpen meg kell határozni egy elsődleges gyűjtőkört, tudva azt, hogy a World Wide Web egy akkora és annyira komplex hipermédia dokumentum, hogy még egy erősen leszűkített gyűjtőkörnél sem lehet sem teljességre, sem precíz lehatárolásra törekedni.

A *magyar webarchívum* fogalmát a következőképpen határoztuk meg: a magyar webtartalom nyilvános, illetve korlátozottan nyilvános, azon belül pedig kiemelten a kulturális, a tudományos, az oktatási és a közéleti jellegű részeinek ismétlődő mentéseiből álló gyűjteménye, hosszú távú megőrzési, kutatási, oktatási, hivatkozhatósági, bizonyíthatósági, helyreállíthatósági és egyéb célokra. (A „korlátozottan nyilvános" jelző olyan weboldalakat jelent, amelyek bárki számára hozzáférhetők ugyan, de előbb el kell fogadni a felhasználási feltételeket vagy regisztrálni kell.)

A projekt

A nemzeti webarchívum megteremtésének első, teljes és csökkentett verziókban is részletesen kidolgozott koncepciója (ha eltekintünk a 2006-os MIA javaslattól) 2016 őszén „*Az egész életen át tartó tanulást megalapozó fejlesztések*" nevű *EFOP-3.7.2-VEKOP-16* programhoz készült el. Ez egy hároméves terv volt, és bár a pályázat nem kapott támogatást, de kétéves, kísérleti projektre átdolgozott változata 2017-ben bekerült az OSZK Országos Könyvtári Rendszer (OKR) kiépítését célzó nagy fejlesztési programjába, így megkezdődhetett az előkészítő munka.

Az új koncepció 1.0-ás változata 2017. április elején készült el, majd az OSZK-ban és az EMMI Közgyűjteményi Főosztályán folytatott egyeztetések, valamint a Kormányzati Informatikai Fejlesztési Ügy-

nökség (KFÜ), a Magyar Tudományos Akadémia KIK és a Nemzeti Audiovizuális Archívum (NAVA) szakembereinek véleményezése után az április 17-i dátumú 3.0-ás verzió került elfogadásra. A megfogalmazott célok a következők voltak: „A fejlesztés célja egy leendő magyar internet archívum (webarchívum) koncepciójának, szervezeti kereteinek és informatikai infrastruktúrájának megteremtése, valamint egy tesztarchívum felállítása és ennek segítségével tapasztalatok gyűjtése egy majdani nagy kapacitású, üzemszerűen működő rendszerhez. Végső célként egy olyan rendszert kívánunk majd létrehozni, amely az interneten nyilvánosan megjelenő magyar és magyar vonatkozású kulturális örökség hosszú távú megőrzésének feladata mellett képes kiszolgálni az oktatás, a tudományos kutatás, az állami szervek, az üzleti szféra és az egyes internethasználók igényeit is. Az archívum megvalósulásával a most csak jelen időben létező magyar internetnek múltja is lesz, és olyan lehetőségek nyílnak meg a mai és a jövőbeli felhasználói számára, amelyek jelenleg nem, vagy csak nehézkesen valósíthatók meg (pl. megszűnt weboldalak megtalálása, régi vagy jelenkori, de feltört, illetve véletlenül törölt webhelyek helyreállítása, weboldalak időbeli változásának elemzése és vizualizálása, stabil hivatkozhatóság, idődimenziót is tartalmazó szöveg- és adatbányászati alkalmazások futtatása, internettörténeti kutatások, hiteles másolatok szolgáltatása, igény szerinti webarchiválás megrendelésre). A rendszert a külföldi legjobb gyakorlatok tapasztalatainak felhasználásával alakítjuk ki, együttműködésben a hazai közgyűjteményi és informatikai szféra egyes szereplőivel, és egyaránt alkalmas lesz nagy tömegű webaratásra, egyes webhelyek vagy weboldalak szelektív lementésére, valamint online tartalmak önkéntes beadására is. Az ehhez szükséges infrastruktúra megteremtése mellett kialakítjuk a webarchiválás és az archivált anyag metaadatolásának módszertanát; megfogalmazzuk az archívum és az eredeti tartalomgazdák, valamint az archívum és annak felhasználói közötti jogviszonyt szabályozó szerződésmintákat; továbbá ajánlások, tanulmányok, előadások és tanfolyamok segítségével ismertetjük meg az érintett szereplőkkel a webarchiválással és a webarchívum-használattal kapcsolatos ismereteket."

A célok kitűzésével párhuzamosan zajlott a projektet megvalósító munkacsoport megszervezése is. Témafelelősnek *Drótos Lászlót,* a MEK munkatársát jelölték ki, akinek közvetlen felettese és a témában helyettesítője jelen cikk másik szerzője, az E-könyvtári Szolgáltatások Osztály vezetője, *Moldován István*. A számítástechnikai feladatokon két informatikus dolgozik. Egyikük a már említett Kovács Péter, aki elsősorban a tömeges aratásokat végzi, míg a rendszergazdai és kisebb programozási feladatok elvégzésére *Vitéz Gábor* kapott megbízást, aki a MEK esetében már sok éve látja el ezt a munkakört. A projekt költségkeretének terhére két főállású munkatársat sikerült felvenni az E-könyvtári Szolgáltatások Osztályra: webkönyvtárosi munkakörbe *Németh Márton*t, aki a digitális könyvtári terület egyik elismert hazai szakembere, jelentős nemzetközi tapasztalatokkal és kapcsolatokkal; webkurátornak pedig *Visky Ákos László*t, aki már korábban is dolgozott archivátorként.

A munkacsoport tagjai közti kommunikáció részben online csatornákon zajlik, a napi szintű Skype megbeszélések mellett egy MIAadm-l nevű zárt levelezőlistán. A közösen szerkesztett dokumentumok és az egyéb fájlok tárolása az OSZK Redmine projektkezelő rendszerében, valamint megosztott Google Drive mappákban történik.

A belső kommunikáció megszervezése mellett már a kezdetektől fontosnak tartottuk, hogy a szakmát és más érintett köröket is tájékoztassuk, ezért 2017. május elején egy ideiglenesnek szánt projekt honlapot[3] indítottunk, amit egy-két hetente frissítünk. Megtalálható rajta a projekt rövid ismertetése; van egy hírrovat az elért eredményekről, a megjelent cikkekről és a megtartott előadásokról; megnézhetők a témában magyar nyelven elérhető publikációk, valamint egy beágyazott hírcsatornán át megjelennek a webarchiválással foglalkozó külföldi kollégák legújabb Twitter-bejegyzései is. Ezek mellett linkek mutatnak a projekt keretében elkészült minden nyilvános szolgáltatásra (demó gyűjtemény, wiki, szakirodalmi bibliográfia, javaslattevő űrlap). Az érdeklődők a mia@mek.oszk.hu e-mail címen vehetik fel a kapcsolatot a munkacsoport tagjaival. Természetesen az OSZK központi honlapján is elérhetők a legfontosabb információk és a publikációk, előadások. Akiket pedig szakmailag is érdekel ez a tevékenység, azok feliratkozhatnak a 2017 májusában indított MIA-l levelezőlistára[4], melyen keresztül nagyjából havonta részletesebb, a technikai kérdésekbe és a problémákba is belemenő tájékoztatást kapnak az aktuális helyzetről, valamint a fontosabb nemzetközi híreket is megosztjuk velük. A listának több mint 30 tagja van, főként hazai közgyűjteményekben dolgozó szakemberek.

Infrastruktúra

A kísérleti projekt idejére a Kormányzati Informatikai Fejlesztési Ügynökség biztosít egy Linux-alapú szervert a nagyméretű aratások memória- és tárhelyigényének tesztelésére. Ennek a gépnek a jelenlegi paraméterei: 16 CPU mag, 64 GB RAM, 100 GB rendszerlemez, 20 TB tárhely, 10 Gbps sebességű hálózati csatlakozás. A megőrzendőnek ítélt mentések során keletkezett fájlok és a hozzájuk tartozó index-állományok összmérete 2018 végére meghaladta a 10 TB-ot. Emellett 2017. június végétől az OSZK-n belül egy kisebb teljesítményű virtuális szerver szolgál az új szoftverek kipróbálására, egyes webhelyek arathatóságának tesztelésére, a nyilvános demóba bekerült honlapok időszakos mentésére, valamint magára a nyilvános szolgáltatásra. Ennek technikai adatai: 4 CPU mag, 16 GB RAM, 1,5 TB tárhely. A tárhely a projekt során többször bővítésre került, jelenleg kb. az 50%-a szabad.

Más nemzeti könyvtárakban működő webarchívumokhoz hasonlóan mi is alapvetően olyan *nyílt forráskódú szoftverek*ből tervezzük felépíteni a rendszert, amiket az Internet Archive és az International Internet Preservation Consortium (IIPC) nevű nemzetközi szervezet is támogat. Ezeket fogjuk kiegészíteni saját fejlesztésű modulokkal vagy önálló alkalmazásokkal, melyek speciális igényeket elégítenek ki, illetve integrálják a webarchívumot az OKR keretében bevezetésre kerülő új könyvtári rendszerrel.

Webhelyek tömeges aratásához a legfontosabb eszköz az Internet Archive által 2003 óta fejlesztett és használt *Heritrix* crawler vagy robot. Bár a Heritrix eredetileg alapvetően még a „hagyományos", 1.0-ás webre lett kitalálva, de később sokat fejlesztettek rajta olyan irányban, hogy a 2.0-ás, JavaScriptekben gazdag, dinamikus megoldásokat használó weboldalakat is ki tudja elemezni és le tudja menteni. Ennek ellenére a tapasztalataink szerint a 3.2-es Heritrix is sokat hibázik, elég gyakoriak a hiányosan vagy hibásan lementett oldalak. A problémák egy része speciális konfigurációs beállításokkal orvosolható (persze tömeges aratásoknál erre nincs emberi kapacitás), más része viszont az eredeti webhelynél használt technológia módosítását igényelné, vagy egy más típusú eszközt (pl. Brozzler vagy Webrecorder) kell használni ilyen esetekben.

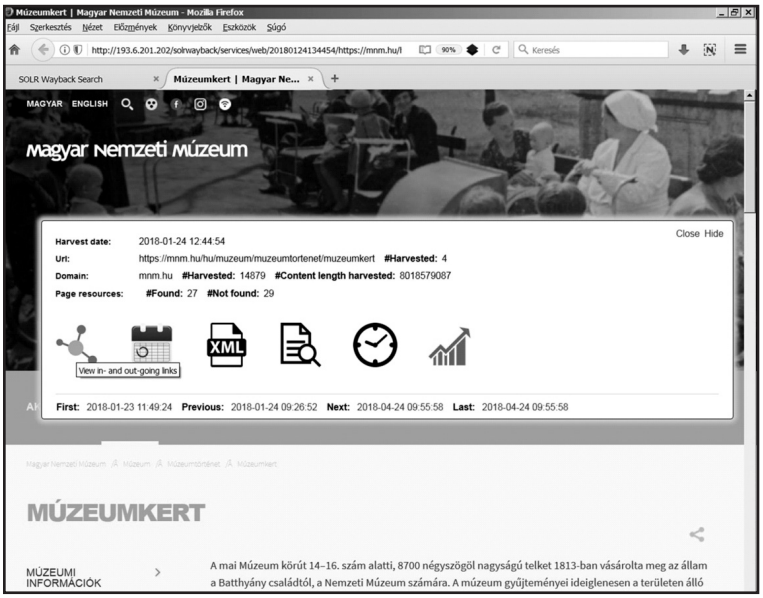
A Heritrix ún. WARC formátumú konténerekben tárolja a robot által talált linkekről letöltött fájlokat. Ezek megjelenítéséhez az Internet Archive által készített *OpenWayback* (OWB) szoftvert használjuk mindkét szerveren. Az OWB időgép-szerűen teszi lehetővé a lementett webtartalom böngészését egy URL cím megadása után. Először naptári nézetben mutatja meg az adott URL-ről készült mentések dátumait, majd egy időpontot kiválasztva megjeleníti a fájlt, és ha vannak benne linkek, akkor azokra kattintva ugyanúgy böngészhető az archívum, mint az élő web – mindaddig, amíg egy olyan linkre nem kattint a felhasználó, ami már nincs benne az archívumban.

Mivel a Heritix és az OWB vezérlése komolyabb informatikai ismereteket igényel és időnként a szerverhez való közvetlen hozzáférésre is szükség van, ezért az évek során több olyan keretrendszert is kifejlesztettek, amelyekkel felhasználóbarát felületen indíthatók és állíthatók le az aratások, valamint indexelhetők és nézhetők vissza a lementett fájlok. Ezek a szoftverek általában még további funkciókat is tartalmaznak, amelyek az archiválási folyamat néhány egyéb munkafázisát is támogatják. Az egyik ezek közül a 2006-tól létező *Web Curator Tool* (WCT), mely eredetileg az új-zélandi és a brit nemzeti könyvtár közös projektje volt és több más nemzeti webarchívumnál is elkezdték használni, de később a British Library kiszállt, az új-zélandiak pedig nem fejlesztették tovább. Emiatt az általunk 2017 júniusában először kipróbált WCT még nem támogatta a Heritrix 3-as verzióit. Mivel időközben a Holland Nemzeti Könyvtár (Koninklijke Bibliotheek) is bekapcsolódott a szoftverfejlesztésbe, 2018 novemberére elkészült a WCT 2.0-ás változata, amit jelenleg mi is tesztelünk. Ez már teljesen integrálva lett a 3-as Heritrix-szel és a tervek szerint idővel más crawlert is támogatni fog. A WCT az aratások paraméterezésén és ütemezésén kívül számos egyéb funkciót is tartalmaz, például metaadatolhatók és csoportokba sorolhatók a szelektív archiválásra kiválasztott webhelyek, kiküldhetők vele engedélykérő levelek, majd nyilvántarthatók a kapott engedélyek, többféle módon segíti a lementett tartalom ellenőrzését, illetve a megismételt aratásoknál a minőség javítását, a hiányok pótlását stb. Kifejezetten csoportmunkára készült, így akár egy intézményen belül, akár más intézményekből többen is használhatják, mert szigorúan elkülöníti a jogosultságokat, különböző szintű felhasználó csoportokat lehet benne definiálni.

Megnéztük továbbá a *NetarchiveSuite* (NAS) és a *Web Archiving Integration Layer* (WAIL) keretrendszereket, valamint a *Brozzler* (browser + crawler) nevű, böngészővel kombinált robottal is végeztünk néhány próbát, melyet 2016-ban mutatott be az Internet Archive. A tesztelt webarchiváló és megjelenítő

programok közül említést érdemelnek még a következők: a *WARCreate* nevű Chrome böngésző-kiegészítő, amivel WARC formátumba menthetők egyedi weboldalak; az ingyenes online szolgáltatásként is használható *Webrecorder*, mellyel egy böngészési folyamat rögzíthető; a Windows alatt futó *Webrecorder Player*, amivel bármilyen WARC fájl megjeleníthető; valamint a nagyon felhasználóbarát webhely-letöltő program, a *HTTrack*, ami WARC helyett fájlrendszerbe ment.

Az OpenWaybacknek is megvannak a maga korlátai (pl. csak URL cím szerint lehet keresni vele, az egyes mentett weboldalakról pedig nagyon kevés információt ad), és ezért szintén alternatív megoldásokat kezdtek el fejleszteni néhány helyen. A legígéretesebb ezek közül a dán *SolrWayback*, ami teljes szövegű- és képkereső funkciót is tartalmaz, az archív weblapokon kinyitható Toolbar panelen pedig olyan funkciókat találunk, mint az adott doménről kifelé mutató, illetve az arra kívülről hivatkozó linkekből rajzolt gráf, az oldal mentéseinek időbeli eloszlása naptárszerű nézetben, és a mentésekből generált bélyegképek, a weboldalt alkotó elemek letöltési idejének eltérései, valamint az illető doménről aratott anyag mennyiségének évenkénti változása az archívumban (*1. ábra*).



1. ábra
A SolrWayback megjelenítő eszköztára egy archivált weboldalon

A SolrWayback keresőjét beépítettük a nyilvános gyűjteményünkbe, és mivel voltak különböző problémáink vele, felvettük a kapcsolatot a dán fejlesztővel, aki készségesen segített és jó munkamegosztás alakult ki, például mi tesztelhettük elsőként az új funkciókat. 2018 tavaszán egy saját keresőt is elkezdtünk fejleszteni *SolrMIA* néven, ami szintén a Solr indexre épül, de a találatok az OpenWaybackben jelennek meg. További különbségek még, hogy ez használja a magyar nyelvhez készült szótövező algoritmust, szűrőfeltételként felajánlja a webhelyek metaadatai között rögzített témakört, műfajt és típust is, továbbá az egyes találatok alá kiírja a webhely egységesített nevét.

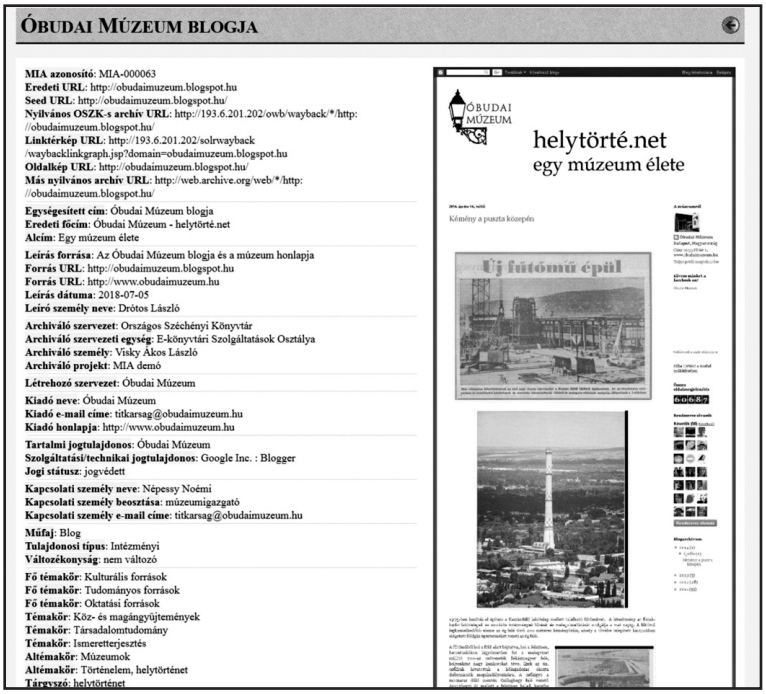
Az archívumban való böngészéshez, illetve az archivált verzió és az eredeti weboldal közötti eltérések összehasonlításához, valamint a mentés idején használatos böngészőkben való megjelenés megőrzéséhez fontos, hogy oldalképek is készüljenek legalább a kezdőoldalról és esetleg néhány fontosabb aloldalról. Erre a *Grab Them All* (GTA) nevű Firefox kiegészítőt, a *Nimbus Screen Capture* Chrome-ba beépülő változatát, valamint a linuxos szervereken a *Puppeteer* nevű programmal vezérelt Chromiumot (a Chrome böngésző motorját) használjuk.

Tesztgyűjtemény

Az első néhány héten terhelési teszteket végeztünk a KIFÜ-s szerveren, 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 Elektronikus Periodika Archívum és Adatbázis (EPA) által nyilvántartott, de az állományában nem archivált kb. kétezer magyar időszaki kiadvány weboldalait próbáltuk meg lementeni. Ezt a gyűjtést júliusban megismételtük, kihagyva az URL listából a problémás (pl. megszűnt vagy az aratórobot által nem bejárható) címeket, és csak az időközben megváltozott oldalakat tároltuk el újból. Ezzel párhuzamosan elkezdtük a Könyvtári Intézet által gondozott *Nyilvános Könyvtárak Jegyzéké*ben található könyvtári honlapok és egyéb könyvtári témájú oldalak mentését is, amit a magyar levéltárak, múzeumok és galériák, egyetemek és kutatóintézetek webhelyeinek archiválása követett. 2018 elején a kormányzati szervek és az önkormányzatok honlapjait is lementettük (utóbbiak listáját a magyar Wikipédiától kaptuk meg), valamint egy még korántsem teljes körű mentést csináltunk 2018 júliusában az irodalmi témájú webhelyekről. Ezeket a válogatott címlistákat 2018 őszén különböző források alapján további URL-ekkel egészítettük ki, majd elkezdtük az újraaratásukat. 2018. december végéig több mint hétezer webhelyet archiváltunk kettő, négy alkalommal, jellemzően három szint mélységben. A tematikus gyűjtések mellett esemény-alapú archiválásokat is csináltunk: a 2018-as téli olimpiával, valamint az országgyűlési választásokkal kapcsolatos online forrásokat gyűjtöttük néhány hétig, illetve a Múzeumok Éjszakája alkalmából játék témájú webhelyeket mentettünk. 2017 júliusában külön mentettük az OSZK saját webes tartalmait, pl. a honlapot, a blogot, a virtuális kiállításokat és a Facebook oldalt, 2018 novemberében pedig a megszűnő régi Corvina honlapot.

A KIFÜ szerveren végzett tesztaratások anyaga nem igazán alkalmas még házon belül való használatra sem (sok a hiány és a hiba benne, nincs teljes egészében leindexelve, nincsenek webhely szintű metaadatok stb.). Viszont szerettünk volna mielőbb egy nyilvános gyűjteményt is létrehozni, amivel a szakemberek és a nagyközönség számára is demonstrálni lehet a webarchiválás lehetőségeit és korlátait, ezért a 2017-ben learatott anyagból ősszel kiválogattunk több mint 150 jobban sikerült mentést és ezek tulajdonosaitól engedélyt kértünk a projekt végéig a nyilvános szolgáltatásra. Szerencsére egészen sokan pozitívan válaszoltak, így az ő webhelyeiket 2018 folyamán néhány alkalommal lementettük az OSZK-s szerverre is és létrehoztunk belőlük egy demó gyűjteményt, ami 2018. február 14-től érhető el a projekt



2. ábra

Egy lementett blogról készített metaadatrekord és oldalkép részlete

honlapjáról. Ez a lista még később is bővült néhány tétellel, így végül 118 webhely található benne. Nagy részük intézményi honlap vagy blog, de van néhány személyes oldal és időszaki kiadvány is köztük. Az OSZK-s szerveren található mentésekre mutató link mellett minden webhely esetében megnézhető a letöltéskor készült oldalkép, az adott domén linktérképe, az amerikai 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. Ezeket a mentéseket alaposan ellenőriztük, a minőségi problémákat pedig rögzítettük az XML metaadat rekordokban, melyek HTML-re konvertálva szintén megnézhetők, a kicsinyített oldalképekkel együtt (*2. ábra*). Márciusban a szolgáltatófelületbe beépítettük a SolrWayback és a SolrMIA keresőket is, amelyekkel a teljes állományban lehet szavakra és képekre keresni.

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 azt is megkívánja, hogy a bibliográfiai leírások mellett adminisztrációs és technikai metaadatokat is rögzítsünk. A metaadatolás történhet gyűjtemény, részgyűjtemény, webhely, webhelyrész és fájl szintjén. (Utóbbi természetesen csak a fájlokból automatikusan kinyerhető vagy generálható adatokból, hiszen több százmillió fájl esetében más megoldás nem jöhet szóba.) Mi egyelőre a részgyűjtemény- és a webhely-szintű metaadatolással foglalkoztunk. Az amerikai Online Computer Library Center (OCLC) Web Archiving Metadata Working Group nevű munkacsoportjának 2018 februárjában közzétett, Dublin Core-alapú ajánlását figyelembe véve dolgoztuk ki áprilisban a saját adatstruktúránkat, amelynek egyes mezőit – ahol lehetett – megfeleltettük a MARC21-nek is. Több mint százféle adatot definiáltunk mindkét szinten, melyeket három csoportba soroltunk: leíró, adminisztrációs és technikai adatok körére. Egy webarchívumnál a nagy számosság miatt nem várható el, hogy minden webhelyről azonos részletességű leírások készüljenek, mert ez a tevékenység nagyon élőmunka-igényes és a legszűkebb keresztmetszet mindig a rendelkezésre álló munkaerő. Ezért a webhelyek esetében csak három kötelező mezőt határoztunk meg: azonosító, név, URL, minden további adat kitöltése opcionális, illetve reményeink szerint részben majd automatizálható lesz. Az adatok tárolása egyelőre XML fájlokban történik, melyek betölthetők lesznek majd az OSZK leendő új digitális gyűjteménykezelő rendszerének adatbázisába.

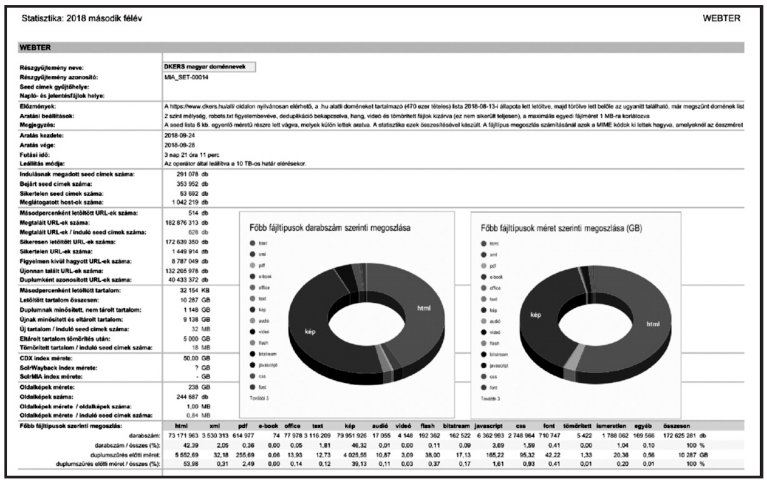
Webtér archiválás

A válogatott aratásokkal a nemzeti webtér csak egy kis töredéke őrizhető meg, igaz, az nagyobb mélységben, jobb minőségben és gyakoribb mentésekkel. A szelektív archiválás mellett sok országban végeznek a teljes országdoménre kiterjedő aratásokat is, általában évente egyszer-kétszer, hogy legalább egy reprezentatív „pillanatképet” készítsenek az adott országban működő webszerverek minél nagyobb halmazáról. Az első ilyen tesztet 2018 szeptemberében végeztük, amikor is több mint 291 ezer .hu végű névből álló címlistán indítottuk el a robotot, majd négy napi futás után az eredetileg meghatározott 10 TB összeméret elérésekor leállítottuk. A deduplikáció (vagyis az azonos fájlok kiszűrése) és a tömörítés után a ténylegesen eltárolt tartalom kb. 5 TB lett. A robotot úgy állítottuk be, hogy a kezdőoldaltól max. 2 szint mélységig kövesse a linkeket, a hang, a videó és a tömörített állományokat ne mentse le. A Heritrix négy nap alatt összesen kb. egymillió szervert járt be és több mint 172 millió fájlt töltött le, amelyeknek 42%-a HTML, 46%-a pedig kép. A 291 ezer kiinduló URL címről több mint 244 ezer esetben sikerült oldalképet is csinálni, ezek összesen 238 GB-ot foglalnak el (*3. ábra*).

Ez a hatalmas anyag további elemzéseket igényel. Szeretnénk például kigyűjteni belőle az aldoménekre mutató linkeket, hiszen a kiinduló címlista csak a központi honlapokra mutatott, így ha a domén tulajdonosa aldoméneket is létrehozott, ahol önálló webhelyek vannak, azokat most 2 szintig sem járta be a robot. Ki kellene továbbá gyűjteni, majd törölni a listából azokat a doméneket, amelyekről csak valamilyen hibaüzenet jött, vagy a letöltött tartalom mérete nem ér el egy határértéket, mert azok vagy üresek, vagy ki vannak tiltva róluk a robotok, vagy robottal nem járhatók be (mert például bejelentkezést igényelnek). A következő domén szintű aratásnál pedig finomhangolni kell a konfigurációs paramétereken, hogy minél több értelmes/értékes fájlt mentsünk le és minél kevesebb „szemetet”.

Kapcsolatépítés

Mivel más országokban már tíz, húsz éves tapasztalatokkal rendelkeznek ezen a szakterületen, sőt van,



3. ábra
Az első domén-szintű aratásról készült statisztika

ahol már egy második generációs webarchívumot működtetnek, számunkra különösen fontos a náluk felgyűlt gyakorlati és elméleti tudás mielőbbi elsajátítása. Ezért is kezdeményeztük 2017 nyarán, hogy az OSZK csatlakozzon az internet archiválásával foglalkozó intézményeket és cégeket tömörítő szervezethez, az International Internet Preservation Consortium (IIPC) munkájához, ami az október 6-án elküldött kérelmet követően 2018. január elsejével meg is történt.[5] Az IIPC 2013-ban jött létre a francia nemzeti könyvtár kezdeményezésére 12 alapító intézménnyel, jelenleg már kb. 45 országból vannak tagjai. Szeptember 18-án videokonferencia formájában lehetőségünk nyílt beszélni az IIPC végrehajtó bizottságának tagjával, *Jefferson Bailey*-vel, aki egyben az Internet Archive egyik vezetője is. Bár a szervezet 2018 novemberében tartott új-zélandi közgyűlésén és konferenciáján a nagy távolság és a magas költségek miatt nem tudtunk részt venni, de újonnan csatlakozott tagként lehetőségünk volt arra, hogy egy öt perces videó prezentációval bemutatkozzunk a rendezvényen. Az IIPC oktatási munkacsoportjának tevékenységébe is bekapcsolódtunk 2018 elején. A kommunikáció online konferenciabeszélgetések, valamint megosztott dokumentumok formájában történik. A munkacsoport az internetes tartalmak megőrzéséhez szükséges ismeretek oktatásának módszertanát dolgozza ki. Ennek érdekében felmérték a tagországokban eddig elkészült tananyagokat, a jó gyakorlatokat, és kérdőíves kutatást végeztek az igényekről, elvárásokról, melynek előzetes eredményeiről mi is beszámoltunk előadások és cikkek formájában.

2017 őszén Németh Márton elvégezte a dániai Aarhus egyetemén működő NetLab kutatócsoport angol nyelvű távoktatási kurzusát, ahol öt témakörben folyt a képzés. Október 24-én Bécsben, majd november 9-én Pozsonyban voltunk egynapos tanulmányúton, ezt követően 2017 decemberében a szlovák webarchívumot fejlesztő Tempest cég képviselői az OSZK-ban is bemutatták a rendszerüket. 2018 őszén pozsonyi konferenciákon többször is volt alkalmunk a szlovák és a cseh kollégákkal beszélni, akiket például a SolrWayback-kel folytatott kísérleteink érdekeltek. A SolrWayback és a NetLab kurzus miatt a dániai szakemberekkel is kapcsolatba kerültünk, és a terület egyik elismert dán kutatójának, Niels Brüggernek a COST Action nevű EU-s programra beadott Transnational Researcher Use of Web ARChives (TRUeWARC) című projekttervét – több más országgal együtt – mi is támogattuk.

Az Internet Archive-val, amely az egész világra kiterjedően gyűjti a weboldalakat 1996 óta és ezért a legnagyobb magyar archív webanyaggal rendelkezik, 2018 júliusában kezdeményeztük a kapcsolatfelvételt. A tőlük kapott tájékoztatás szerint a gyűjteményükben több mint egymilliárd .hu végű URL cím van. Bár a Wayback Machine felületén nincs mód országdomén szerinti szűrésre, de egy API-n keresztül lehetőség van erre, ahogyan ez a német nemzeti könyvtár által előfizetett szolgáltatás esetében már kipróbálható.[6] Egy ilyen keresőfelületet majd mi szeretnénk beépíteni az archívumunk honlapjába, amivel a régi magyar web is kereshetővé válik.

A magyar internet legalább részleges megőrzése olyan nagyságrendű és bonyolultságú feladat, amit

egyetlen intézmény néhány munkatárssal egészen bizonyosan nem tud fenntartható módon megoldani. A nemzeti könyvtárnak együttműködéseket kell kialakítania nemcsak a közgyűjteményi szféra egyes intézményeivel (pl. a válogatás, engedélyeztetés, minőségellenőrzés és metaadatolás élőmunka-igényes feladatainak megosztása érdekében), hanem a nagyobb tartalomszolgáltató cégekkel is (pl. hogy robot- és archívumbaráttá tegyék a szolgáltatásaikat, illetve hogy küldjék be az archívumba az automatikusan nem aratható nyilvános tartalmaikat); továbbá a képzéssel és továbbképzéssel foglalkozó egyetemi tanszékeket is segítenünk kell abban, hogy ez a terület is bekerüljön az oktatásba. Ezen a téren is megtettük a kezdeti lépéseket: felvettük a kapcsolatokat az érintett intézmények egy részével, megbeszéléseket folytattunk és bemutatókat tartottunk, és bár még sokan idegennek érzik ezt a szakterületet, reményeink szerint a webes tartalmak megőrzésével való foglalkozás néhány év múlva ugyanúgy beépül a közgyűjtemények tevékenységébe, mint például a digitalizálás vagy az intézményi repozitóriumok.

Egy webarchívumnak maguk az internethasználók is potenciális partnerei. Bevonhatók lennének a mentések ellenőrzésébe, a metaadatok gazdagításába, de leggyakrabban a megőrzésre érdemes webhelyek összegyűjtésében szoktak tőlük segítséget kérni a külföldi archívumok, különösen olyankor, amikor egy eseménnyel kapcsolatos fontos online források URL címeit kell néhány hét vagy nap alatt összeszedni. Mi is létrehoztuk egy Google-űrlapot[7] 2018 márciusában, melyen keresztül bárki javasolhat kevésbé ismert, de archiválásra érdemes magyar webhelyeket. Az űrlap kitöltése előtt egy saját fejlesztésű egyszerű keresővel ellenőrizni lehet, hogy az adott URL nem szerepel-e már a nyilvántartásunkban.

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. A projekt honlap és a MIA-l lista mellett 2017 tavaszán egy wikit[8] is elkezdtünk összeállítani az internet archiválással kapcsolatos ismeretekről. A MIA wikiben már közel 600 szócikk található, ilyen kategóriákba sorolva: fogalmak, formátumok, fórumok, hasznosítás, irodalom, projektek, rendezvények, szabványok, szervezetek, szoftverek, szolgáltatások, és a szócikkekből több ezer link mutat kifelé az elsődleges információforrásokra.

2017 előtt magyar nyelven 15 lényegesebb publikáció jelent meg a webarchiválásról, ezek többsége külföldi cikkek referátuma vagy szakirodalmi szemle volt. Szakmai konferenciákon sem jelent meg ez a téma korábban nálunk, eltekintve a MIA létrehozását sürgető néhány régi előadásunktól. A projekt ideje alatt már egy tucat cikket és tanulmányt publikáltunk a hazai szakfolyóiratokban (Tudományos és Műszaki Tájékoztatás, Könyv, Könyvtár, Könyvtáros, Könyvtári Figyelő, Digitális Bölcsészet) és hazai, illetve külföldi konferencia-kiadványokban, továbbá legalább ugyanennyi előadást tartottunk különböző rendezvényeken, melyek között volt néhány angol nyelvű is: Debrecenben a 2017-es CogInfoCom és Rigában a 2018-as BOBCATSSS konferenciákon, valamint Pozsonyban az Egyetemi Könyvtár rendezvényein. Jelentkeztünk a 2019-es BOBCATSSS-re is, valamint a zágrábi IIPC konferenciára, melynek a szervezésében is részt veszünk. A tavalyi Networkshop konferencián az előadás mellett egy hosszabb műhelymegbeszélésre is sor került, és az idei évi rendezvényen is tervezünk egy közel három órás tutoriált (irányított megbeszélést) a nulladik napon.

2017. október 13-án az OSZK-ban műhelynapot szerveztünk „404 Not Found – Ki őrzi meg az internetet?" címmel, melyet 2018. november 15-én megismételtünk. Az előadások után kerekasztal-beszélgetésre is sor került a közgyűjteményi szakemberekkel az együttműködés lehetséges formáiról. Mindkét alkalommal 70–80 résztvevője volt a műhelynapnak, és a visszajelzések alapján sikeresnek minősíthetjük ezeket a félnapos rendezvényeket.

2017 nyarán a Könyvtári Intézettel közösen elkezdtünk szervezni egy 30 órás továbbképző tanfolyamot „*Az internet archiválása, mint közgyűjteményi feladat*" címmel olyan könyvtári, múzeumi, levéltári szakembereknek, akik akár saját intézményi archívumot szeretnének kialakítani, akár bedolgoznának az OSZK-ban tervezett nemzeti webarchívum munkálataiba. A tanfolyam végleges terve októberre készült el és a következő évben lezajlott akkreditációs eljáráson is sikeresen túljutott. Erről 2018 novemberében kaptuk meg az értesítést, így a továbbképzés első alkalommal 2019. április-májusban, majd pedig szeptemberben kerül meghirdetésre. A 4 napos tanfolyam tervezése mellett 2018. január végén felmerült egy „blended learning" típusú (vagyis távokta-

tási jellegű, de emellett néha személyes jelenlétet is igénylő) tananyag összeállításának ötlete, amely az OSZK leendő e-learning rendszerén keresztül lesz elérhető. A 25×45 perces oktatóanyaggal szeptemberre készültünk el és a szöveges részek mellett videókat, prezentációkat és képernyőfotókat is tartalmaz. A távoktató keretrendszerbe való betöltése várhatóan 2019 elején kezdődik el.

A gyakorlati tevékenység mellett a projekt kezdete óta folyamatosan figyeljük a webarchiválással, illetve általában az internetes tartalmak megőrzésével kapcsolatos tudományos kommunikációt is. Gyűjtjük és olvassuk a nemzetközi szakirodalmat, melyből összeállítottunk egy már több mint 450 tételes bibliográfiát,[9] ami különféle formátumokban böngészhető és tölthető le a projekt honlapjáról.

Jogi környezet[10]

A webarchiválás üzemszerű működtetésének jogi feltételei is vannak. Számos országban ezt a kötelespéldány szabályozás keretében oldották meg, különböző módokon. Jogi szabályozás vonatkozik a nemzeti szintű webarchiválásra pl. Németországban, Észtországban, Izlandon, az Egyesület Királyságban.

Magyarországon évek óta napirenden van a kötelespéldány rendelet módosítása[11], a legfrissebb információk alapján 2019 első negyedévében már várható ennek hatályba lépésbe. Kézenfekvőnek tűnt, hogy ezt kellene kiegészíteni a tartalomszolgáltatók és a webarchívumot működtető nemzeti könyvtár jogait és kötelességeit definiáló részekkel, számos külföldi példa is ezt támasztotta alá. Ennek elősegítése érdekében 2017 májusában összefoglalókat készítettünk néhány nemzeti webarchívum (dán, cseh, észt, brit, ausztrál) gyűjtőköréről, jogi és szervezeti kereteiről, majd az észt kötelespéldány törvény angol nyelvű fordításából kiemeltük és kommentáltuk azokat a részeket, amelyeket a magyar jogalkotók számára is megfontolandónak ítéltünk[12].

A 2019-ben meginduló Közgyűjteményi Digitalizálási Stratégia (KDS) keretében 2018 nyarán egy külön bizottság kezdett foglalkozni a stratégia jogi szabályozásának feltételeivel. Ennek a munkának a során irányult a figyelem a webarchiválás szabályozására is, mivel a KDS terveibe belekerült a webarchiválás tevékenysége is a digitalizálás mellé. A bizottság arra az álláspontra jutott, hogy a kötelespéldány-szolgáltatás szabályozása miniszteri rendeletként nem biztosít elegendő jogi fedezetet a webarchiválásra, ezért azt külön törvényben kellene szabályozni. A bizottság kérésére kidolgoztunk egy koncepció-tervezetet a törvénytervezet számára[13], amelyben meghatároztuk a webarchiválás alapvető fogalmait, kereteit, kiterjedését, a fenntarthatóság feltételeit, továbbá költségvetési tervet is készítettünk a folyamatos fenntarthatóság biztosítására. A törvénytervezetet az ősz folyamán már nem a KDS jogi bizottsága, hanem az EMMI Könyvtári és Levéltári Főosztály kezdte el gondozni. 2018 decemberében egyeztettünk törvénytervezet szövegéről az említett főosztállyal és az EMMI Közigazgatási Államtitkárság osztályvezetőjével. A legfrissebb információk alapján a tervezet közigazgatási egyeztetésre bocsátását megelőző egyeztetésen a minisztériumi és közigazgatási szervek abban állapodtak meg, nem szükséges a törvényi szintű szabályozás, megfelelő lesz a kormányrendelet megalkotása. A jogszabály előkészítése ennek megfelelően folytatódik 2019 folyamán. A jogszabály hatályba lépése 2020. január elsejétől várható.

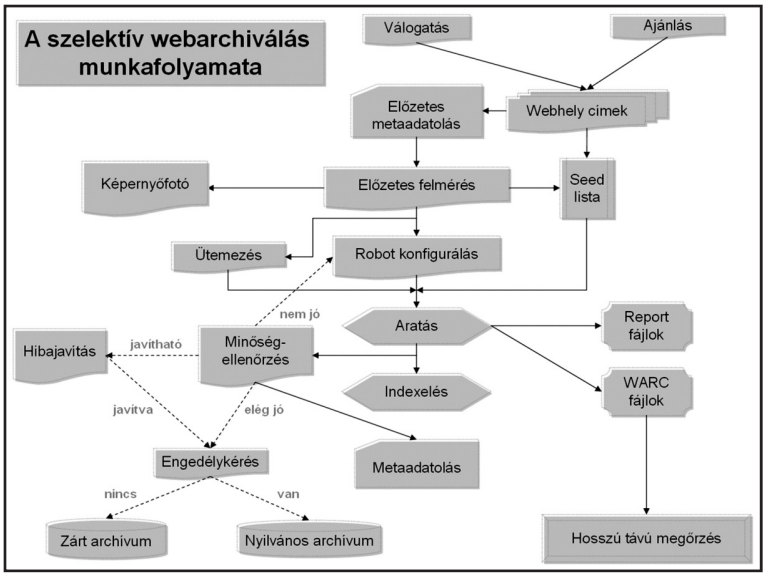
A jogi szabályozás egyik fontos feladata a nemzeti domain címek begyűjtése a teljes körű webaratás érdekében. Ezt több országban a nemzeti könyvtár és a helyi domain nyilvántartó szervezetek közötti megállapodás biztosítja. Ennek érdekében több lépést tettünk az OSZK-ban, hogy az Internet Szolgáltatók Tanácsával[14] hasonló együttműködési megállapodást kössünk, sajnos eredménytelenül. A jogi bizottság is úgy vélte, erre ez a szervezet jogszabállyal nem kötelezhető, ezért alternatív megoldást javasoltak. A minisztérium azt az eljárást fogalmazta bele végül a törvénytervezetbe, miszerint a tartalomgazdáknak kötelező lesz egy, az OSZK által később kidolgozandó űrlapon bejelenteni az induló honlapjaikat, a regisztráció, a nyilvántartás és a webaratás érdekében.

Üzemszerű működésű webarchívum terve

Az OKR projekt[15] keretében elindított kísérleti webarchiválás tervezett határideje 2018. december 31. volt. A konferencián elhangzott előadás idején azzal az ütemezéssel számoltunk, hogy a kísérleti szakasz 2018. decemberi lezárása után 2019 januárjától az üzemszerű webarchiválással folytatjuk a munkát. Egyéb tényezők miatt az OKR projekt lezárása 2018. december 31-éről elhalasztódott 2019. június 30-ra. A webarchiválási kormányrendelet életbe lépése előtt még nincs törvényi kötelezettsége a nemzeti könyvtárnak a webarchiválásra és a tartalomgazdák bejelentési kötelezettsége sem lépett még hatályba.

A fentiek alapján ezért a webarchiválás 2019-ben is egyfajta projekt státusban folytatódik, és lépésről

lépésre haladva próbáljuk kiépíteni a hosszú távú tevékenység elemeit, az egyes munkafolyamatokat (*4. ábra*) támogató informatikai rendszert.



4. ábra
A szelektív webarchiválás munkafolyamata

Szervezeti feltételek

A webarchiválásnak, mint folyamatos könyvtári tevékenységnek helyet kell kapnia a nemzeti könyvtár szervezetében, melynek logikus helye az E-könyvtári Szolgáltatások Osztály, ahol a pilot program szakmai vezetője is dolgozik. Az osztály 1999-ben jött létre és a digitális, online tartalmak gyűjtésével, megőrzésével és szolgáltatásával foglalkozik, tevékenysége folyamatosan bővül: kezdetektől a Magyar Elektronikus Könyvtárban[16] (MEK) digitális könyv jellegű kiadványokkal, majd 2004-től az Elektronikus Periodika Adatbázis és Archívum[17] (EPA) keretében elektronikus időszaki kiadványokkal, 2007-től pedig a Digitális Képarchívumban[18] (DKA) digitális képi dokumentumokkal. Az osztályon fogadják az OSZK Digitális Könyvtárába[19] (DK) kötelespéldány jelleggel beérkező, főként kereskedelmi e-könyveket is. Ezek a gyűjtemények számos online kiadványt tartalmaznak, sok közülük az eredeti helyén már régen eltűnt, csak a könyvtár őrzi, szolgáltatja őket. Ezekben a gyűjteményekben alapvetően dokumentum szintű gyűjtés és megőrzés folyik, az adott dokumentumokat a maguk teljességében őrizzük, archiváljuk és szolgáltatjuk, s látjuk el egyenként önálló metaadatokkal.

A webarchiválás ezt a tevékenységet terjeszti ki a honlapok szintjére. Itt viszont technológiai akadályok miatt sok esetben a webhelyeket nem lehet, illetve nem tudjuk a maguk teljességében gyűjteni (általában csak maximum 3 szint mélységig), valamint részletes metaadatokkal csak a nyilvánosan is szolgáltatható honlapokat tudjuk ellátni.

2019 januárjától a projekt folytatásában sikerült továbbra is biztosítani a webarchiválással foglalkozó két kolléga határozott idejű foglalkoztatását, valamint a külsős informatikus megbízását. Az OSZK kidolgozott egy új szervezeti működési szabályzat tervezetet, melyet eljuttatott a fenntartó minisztériumnak, melyben ezen az osztályon rögzíti a webarchiválás tevékenységét, átnevezve azt E-könyvtári Szolgáltatások és Webarchiválás Osztálynak. Jelenleg várjuk az EMMI jóváhagyását a szervezeti változásokat tartalmazó új SZMSZ tervezetre.

Infrastruktúra feltételek

A projekt alatt, a tesztmérések alapján a KIFÜ-vel közösen kidolgoztunk egy specifikációt az üzemszerű webarchiválás infrastruktúrája számára. A webarchiváláshoz szükséges szerverek és tárhely részét képezik az OKR projekt nagy infrastrukturális beszerzésének. A tervezett eszközök a szükséges

közbeszerzési eljárás lefolytatása után 2018 decemberében megérkeztek az OSZK-ba. Pontosabban ezek egy része, mivel a leendő nagy infrastruktúra a biztonságos üzemeltetés, megőrzés érdekében két telephelyen fog párhuzamosan működni (egy része az OSZK-ban, másik része a KIFÜ gondozásában). A beérkezett eszközök üzembeállítása 2019 nyarán várható. Ezért 2019 januárjától még a pilot projektben használt infrastruktúra szolgálja ki a webarchiválási tevékenységet. Az átmeneti időre azonban tárhely bővítést igényeltünk a KIFÜ-től a náluk működő nagyobb virtuális szerverhez, amelyet január hó folyamán meg is kaptunk 4 TB méretben.

Metaadatok kezelése

A teszt során kidolgoztuk a webhelyek és a részgyűjtemények katalogizálásához szükséges részletes metaadat struktúrát, amely mind a bibliográfiai, mind a technikai, mind pedig az adminisztratív metaadatokat tartalmazza. Az adatrögzítéshez több eszközt is teszteltünk, de végül nem találtunk megfelelő alkalmazást, ezért döntöttünk amellett, hogy a készülő metaadatokat ideiglenesen kézi szerkesztéssel, XML fájlokban hozzuk létre és tároljuk. Az XML fájlokban tárolt metaadatok a későbbiekben beépülhetnek a nagy projekt kapcsán létrejövő OKP (Országos Könyvtári Platform) katalogizáló rendszerébe. Ennek jelenlegi céldátuma 2021. november 25-e, de bízunk benne, hogy a folyamatos fejlesztés során a webarchiválás metaadatolása már korábban is beépülhet a leendő rendszerbe. Ugyancsak később tudjuk majd összekapcsolni az elkészült és elkészülő metaadatokat az OKP keretében elinduló Nemzeti Névtérrel is.

A korlátozott munkaerő kapacitás miatt részletes metaadatokkal csak a nyilvános szolgáltatási engedéllyel rendelkező archivált webhelyeket tervezzük ellátni.

Gyarapítás külső forrásokból

A webarchívumot nemcsak az OSZK által aratott tartalmakkal tervezzük bővíteni, hanem más forrásokból, máshol meglévő tartalmakkal is. Mivel a hazai üzemszerű webarchiválás elindulása legkorábban 2020-tól várható, és bár részlegesen a projekt keretében már 2017-től archiválunk magyar webhelyeket, fontos lenne a korábbi éveket, a több mint 25 éves magyar web[20] tartalmának még fellelhető részét is archiválni. Ennek érdekében felvettük a kapcsolatot az Internet Archive-val, mely, mint már szó volt róla, 1996 óta az egész világra kiterjedően gyűjti a weboldalakat és egyelőre a legnagyobb magyar archív webanyaggal rendelkezik. Az IA-tól tájékoztatást kértünk az általuk tárolt magyar tartalom átvételének lehetőségeiről. 2018 novemberében kaptunk is egy technikai ajánlatot[21], amelyet 2019 januárjában a pénzügyi árajánlat követett. Az ajánlatukban különböző lehetőségeket adtak meg a náluk lévő magyar anyag leválogatására, elkülönített szolgáltatására. Mivel az árajánlat a 2018-as költségvetésterv elkészülte után érkezett, még függőben van, lesz-e rá forrás, és ha igen, akkor melyik ajánlott szolgáltatásra.

Szolgáltatás

A webarchiválás célja a változó webhelyek tartalmának elmentése, megőrzése az utókor számára, de ennek csak akkor van értelme, ha ezek hasznosulnak is valamilyen szolgáltatás keretében. Az archivált honlapok jelentős részéhez szerzői jogi okokból majd csak az OSZK-ban lehet hozzáférni, dedikált számítógépekről, hasonlóan a NAVA rendszeréhez. A belső szolgáltató felület kidolgozása még az elkövetkező feladatok része. A tervek szerint egy közös informatikai megoldás lesz az OSZK-ban a védett, csak helyben olvasható digitális tartalmak olvasására, kutatására, hiszen számos forrásból, jelenleg több adatbázisban is rendelkezünk ilyenekkel (e-könyvek az OSZK DK-ban, jogvédett digitalizált könyvek az ELDORADO-ban, védett folyóiratok a DSpace-ben, NAVA szolgáltatás).

Az archivált webhelyek egy kisebb, válogatott részéhez törekszünk nyilvános szolgáltatási engedélyeket szerezni a MEK, EPA, DKA szolgáltatásokhoz hasonlóan. A korábbi tervek alapján 2018 decemberében elkészítettük és véglegesítettük azt a felhasználási szerződést, amellyel – egyelőre határozott időre – formálisan is jogszerű engedélyt szerzünk az archivált webhelyek egy részének nyilvános szolgáltatására.[22] 2019 januárjától elkezdjük a már szolgáltatott webhelyek tartalomgazdáival a szerződéskötést, az engedélyek megújítását, valamint tovább bővítjük új nyilvános tartalmakkal a webarchívumot.

Webarchívum honlap

A pilot időszakban létrehoztunk egy honlapot, ahol a projekt eredményeiről, a webarchiválás eszközeiről, módszereiről számoltunk be folyamatosan. Itt található egy több száz szócikket tartalmazó wiki, egy részletes tematikus bibliográfia és az engedélyezett archivált honlapok nyilvános gyűjteménye

is. A hosszú távú szolgáltatás érdekében elkezdtünk megtervezni egy új, jobban strukturált, állandó honlapot a tevékenységnek és a szolgáltatásnak. Itt kap majd helyet az az űrlap is, amelyen keresztül – ha a jogszabály hatályba lép – a tartalomgazdák bejelenthetik az új webhelyeiket. A honlap tervezése még decemberben elkezdődött, januárban folytatódik tovább, a kivitelezők és a kivitelezés szervezésével párhuzamosan.

Oktatás

Ugyancsak ebben a kísérleti szakaszban terveztünk meg egy akkreditált tanfolyamot, amely a Könyvtári Intézet keretei között mutatja be a közgyűjtemények munkatársainak a webarchiválás gyakorlati módszertanát. A tanfolyamot 2018 folyamán a Könyvtári Intézet sikeresen akkreditáltatta és már megjelent a 2019 tavaszi képzési tervükben, áprilisban indul az első tanfolyam[23].

Irodalom

DRÓTOS László: Az internet archiválása mint könyvtári feladat. = Tudományos és Műszaki Tájékoztatás, 64. évf. 2017. 7–8. sz. 361–371. p. http://epa.oszk.hu/03000/03071/00109/pdf/EPA03071_tmt_2017_07_08_361-371.pdf [2019. január 21.]

NÉMETH Márton: 404 Not Found – Ki őrzi meg az internetet? Webarchiválás workshop az Országos Széchényi Könyvtárban. = Tudományos és Műszaki Tájékoztatás, 64. évf. 2017. 11. sz. 577–582. p. http://epa.oszk.hu/03000/03071/00112/pdf/EPA03071_tmt_2017_11_577-582.pdf [2019. január 21.]

NÉMETH Márton: Nemzetközi körkép a webarchiválás gyakorlatáról. = Könyvtári Figyelő, 63. évf. 2017. 4. sz. 575–582. p. http://epa.oszk.hu/00100/00143/00349/pdf/EPA00143_konyvtari_figyelo_2017_04_575-582.pdf [2019. január 21.]

DRÓTOS László – NÉMETH Márton: Hungarian web archiving pilot project in the National Széchényi Library. In: 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (pp. 000209–000212). IEEE. https://doi.org/10.1109/CogInfoCom.2017.8268244

DRÓTOS László – KOKAS Károly: Webarchiválás és a történeti kutatások. = Digitális Bölcsészet, 1. évf. 2018. 1. sz. 35–55. p. https://doi.org/10.31400/dh-hun. [2019. január 12.]

DRÓTOS László – NÉMETH Márton: Az OSZK-ban folyó kísérleti webarchiválási projekt első évének tapasztalatai. = Tudományos és Műszaki Tájékoztatás, 65. évf. 2018. 7–8 sz. 389–400. p. http://tmt.omikk.bme.hu/tmt/article/view/7153/8156

DRÓTOS László –NÉMETH Márton: Web museum, web library, web archive; The responsibility of public collections to preserve digital culture. In: Llelde Petrovska, Baiba Īvāne-Kronberga, Zena Meldere (eds.). The Power of reading: proceedings of the XXVI BOBCATSSS Symposium, Riga, Latvia, January 2018. Riga: The University of Latvia Press. 124–126 p.] http://bobcatsss2018.lu.lv/files/2018/08/BOBCATSSS_2018_TheProceedings.pdf [2019. január 21.]

NÉMETH Márton: A webarchiválásról történeti megközelítésben. = Könyv, Könyvtár, Könyvtáros, 27. évf. 2018. 2. sz. 48–52. p. http://ki2.oszk.hu/3k/2018/06/a-webarchivalasrol-torteneti-megkozelitesben/ [2019. január 21.]

DRÓTOS László: Webes tartalmak digitális megőrzése. = Könyv, Könyvtár, Könyvtáros, 27. évf. 2018. 10. sz. 11–17. p. http://ki2.oszk.hu/3k/category/27-evfolyam/2018-10/ [2019. január 21.]

NÉMETH Márton: A webarchiválás nemzetközi környezete; Mozaikok az IIPC 2018. kongresszusról. = Könyv, Könyvtár, Könyvtáros, 27. évf. 2018. 12. sz. 23–27. p. http://ki2.oszk.hu/3k/category/27-evfolyam/2018-12/ [2019. január 21.]

Jegyzetek

1. Internet Archive Wayback Machine: http://web.archive.org
2. A .hu közdomainek alatt delegált domainek számának alakulása: http://www.nic.hu/statisztika
3. A webarchiváló projekt ideiglenes honlapja: http://mekosztaly.oszk.hu/mia
4. A MIA-l levelezőcsoport weblapja: http://mekosztaly.oszk.hu/cgi-bin/mailman/listinfo/mia-l
5. Az OSZK az IIPC honlapján: http://netpreserve.org/about-us/members/orszagos-szechenyi-konyvtar/
6. German Web Archive Search: http://webservices.archive.org/dnb/
7. Webhely archiválási javaslat: https://goo.gl/forms/Y1qIIxcM7APPiq443
8. MIA Wiki: http://mekosztaly.oszk.hu/miawiki
9. A webarchiválás válogatott bibliográfiája: http://mekosztaly.

oszk.hu/mia/doc/webarchivalas-irodalom.html

10. A webarchiválás jövőbeni fenntarthatóságáról szóló előadás a 2018. novemberi helyzet alapján készült. Mivel az előadás nyomán született cikk két hónappal később íródott, ezért ezt már aktualizáltuk az azóta történt eseményekkel, változásokkal, nagyjából a 2019. januári helyzet alapján.
11. A kötelespéldányszolgáltatás rendelettervezete = Könyvtári Intézet, 2013. március 04. [2019. január 21.] https://ki.oszk.hu/hir/konyvtari-intezet/kotelespeldanyszolgaltatas-rendelettervezete
12. Észt webarchívum kötelespéldány szabályozás: https://docs.google.com/document/d/1XtDlgsawK-qA6tLm5__0RxhoeSu7VJq2e74mUQw1IAE/ [2019. január 21.]
13. Koncepcióterv a webarchiválási törvénytervezethez, 2018. augusztus 08. https://docs.google.com/document/d/1Rk6z3m75fvDYfLMDDZzT8hcDNox245gr1OXLKGE-o3Y/ [2019. január 21.]
14. http://www.iszt.hu/iszt/
15. http://www.oszk.hu/okr-projekt
16. http://mek.oszk.hu
17. http://epa.oszk.hu
18. http://dka.oszk.hu
19. http://oszkdk.oszk.hu
20. 20 éves a magyarországi internet: http://mek.oszk.hu/18700/18732/
21. https://docs.google.com/presentation/d/1R-XOv1XthYJGl5UEuxwv8gqS87YWUR70oBPUA6hkt68/ [2019. január 21.]
22. http://mekosztaly.oszk.hu/mia/demo/
23. https://ki.oszk.hu/tanfolyamok/az-internet-archivalasa-mint-kozgyujtemenyi-feladat [2019. január 21.]

Beérkezett: 2019. január 21.

„Segíts te is megőrizni a magyar webet!" akció

Az idei Internet Fiesta alatt (2018. március 21–28.) az OSZK webarchiváló projektje keretében URL-cím gyűjtő akciót tart az OSZK E-Könyvtári Szolgáltatások Osztálya, melynek keretében a http://mekosztaly.oszk.hu/mi/if.html címre várják a kollégák javaslatait irodalmi vagy művészeti témájú honlapok, blogok vagy egyéb webhelyek hosszú távú megőrzésére a magyar webarchívumban.

(Forrás: Katalist, 2019. március 19. Németh Márton felhívása)

Az olvasás jövője

A European Cooperation in Science and Technology (COST) által támogatott kutatási kezdeményezés, az Evolution of Reading in the Age of Digitisation (E-READ) közel kétszáz európai tudóst fog össze, akik az olvasás, a publikálás, valamint az olvasás- és íráskészség kérdéseit kutatják, és arra vállalkoztak, hogy a digitalizálás olvasásra tett hatásait vizsgálják. 2018 októberében a norvégiai Stavangerben tanácskoztak, hogy összegezzék négyéves kutatási programjuk eredményeit, és közreadtak egy nyilatkozatot az olvasás jövőjéről (Declaration Concerning the Future of Reading), mely elolvasható a http://ereadcost.eu/wp-content/uploads/2019/01/StavangerDeclaration.pdf címen.