Drótos László – Visky Ákos László

Rákóczi-archívum

Mintaalkalmazás egy webarchívum más digitális gyűjteményekkel való összekapcsolására

Előzmények

A könyvtárak digitális gyűjteményeinek alapját általában digitalizált dokumentumok alkotják. Ezeket egészíthetik ki a már eleve számítógépes formában vásárolt, begyűjtött vagy egy repozitóriumba feltöltött, illetve távoli hozzáféréssel előfizetett elektronikus könyvek, hangoskönyvek, folyóiratok, tananyagok és egyéb műfajú fájlok, adatbázisok. A digitális dokumentumok világa azonban ennél a körnél jóval nagyobb, gondoljunk csak a honlapokra, blogokra, wikikre, fórumokra, közösségi oldalakra, a kép-, hang- és videómegosztó platformokra stb., melyeken milliárdos nagyságrendben vannak a könyvtárak érdeklődési és gyűjtőkörébe tartozó tartalmak. Ezért volt szerencsés, hogy a Közgyűjteményi Digitalizálási Stratégia részeként a megyei és egyetemi könyvtárak számára 2019-ben meghirdetett KDS-K pályázatban a digitalizálási célok mellett megjelent a webarchiválás támogatása is.

Az Országos Széchényi Könyvtárban (OSZK) 2017 elején indultak el egy nemzeti szintű webarchívum előkészítő munkálatai az Országos Könyvtári Rendszer (OKR) projekt részeként. 2019 végéig 12 tematikus részgyűjteménybe rendezve mintegy 25 ezer magyar webhelyről készült néhány alkalommal mentés. Külön gyűjtjük az időszaki kiadványok oldalait, ezekből már több mint 4600-at archiválunk rendszeresen. Ezeken kívül időnként fontosabb eseményekhez kapcsolódó weblapokat is lementünk, valamint eddig két alkalommal futtattunk a magyar

webtér egy viszonylag nagy részére, körülbelül 250 ezer szerverre kiterjedő aratást. A letöltött tartalom jogi okokból egy zárt archívumba kerül, és csak kutatási célokra, illetve a könyvtári hálózaton belül lesz majd hozzáférhető. Van viszont két kisebb, demonstrációs célokat szolgáló nyilvános gyűjtemény: egy az OSZK saját online szolgáltatásainak mentéseiből, egy pedig más intézmények és magánszemélyek webhelyeiből, melyekre a tulajdonosaik engedélyt adtak.

A KDS-K pályázat keretében az OSZK webarchiváló munkacsoportja az alábbi célokat tűzte ki maga elé:

- hosszú távú együttműködés megalapozása a pályázatban nyertes könyvtárakkal a webarchívum gyarapítása tekintetében és egyéb munkafázisokban (pl.: minőségellenőrzés, metaadatolás stb.);
- módszertani segítségnyújtás írásos anyagok és előadások formájában a partnerintézményeknek az online tartalmak archiválásával kapcsolatosan;
- mintaalkalmazás készítése szakmai ismeretterjesztéshez és a közoktatásban való felhasználáshoz, amellyel illusztrálható, hogy a webarchívum anyaga hogyan tudja kiegészíteni a könyvtárak hagyományos digitális gyűjteményeit;
- elektronikus tananyag összeállítása a középiskolás korosztály számára az intézményi és a személyes webarchiválásról, a digitális kultúránk megőrzésének fontosságáról és lehetséges módszereiről

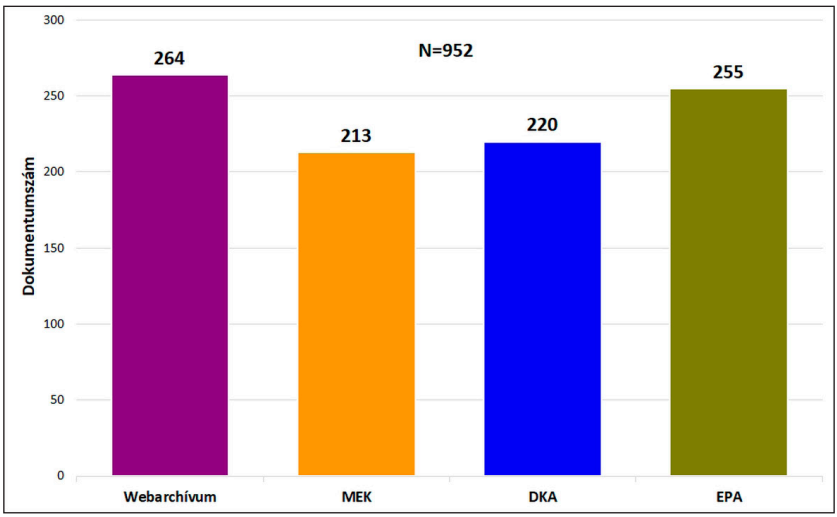
Rákóczi-archívum

A fenti célok közül a harmadikat egy eseményalapú archívum kialakításával terveztük megvalósítani. Az elmúlt években több ilyen gyűjtést is végeztünk már, például a 2018-as országgyűlési, a 2019-es önkormányzati és európai parlamenti választások, továbbá a 2018. évi téli olimpia idején mentettük a hírportálok megfelelő rovatait és a releváns honlapokat. 2020-ban pedig a koronavírus-járvány, a nyári olimpia és a trianoni békediktátum 100. évfordulója kapcsán folytatunk speciális aratásokat. A KDS-K keretében létrehozott mintaalkalmazás apropóját a II. Rákóczi Ferenc-emlékév adta, melyet 2019 júliusától, Rákóczi erdélyi fejedelemmé választására emlékezve (1704. július 8.) hirdetett meg az országgyűlés. Az emlékév eseményei egészen 2020. szeptember 17-éig tartanak, mert ekkor lesz Rákóczi Erdélyország és Magyarország vezérlő fejedelmévé választásának 315. évfordulója.

Az internetes sajtóban az első híradások már a javaslat parlament elé kerülését megelőzően megjelentek, a döntés megszületése óta pedig folyamatosan tudósítanak a különböző hazai és határon túli portálok az emlékév keretében tartott rendezvényekről, kiállításokról, konferenciákról, kiadványokról, emléktúrákról. A Rákóczi-archívumban ezeket a híreket próbáljuk megőrizni, de a gyűjtőkört kiterjesztettük a fejedelemmel, annak családjával, a kuruc korral és a szabadságharccal foglalkozó honlapokra, blogokra, hang- és videóanyagokra is. Így aztán

az online újságcikkek mellett az archívumba bekerültek (és bekerülnek még 2020 őszéig) intézmények és szervezetek weblapjai, magánemberek blogbejegyzései, Wikipédia-szócikkek, képgalériák, elektronikus folyóiratokban megjelent tanulmányok, sulinetes oktatóanyagok, YouTube-videók, sőt maga az emlékévről folytatott parlamenti vita jegyzőkönyve is az Országgyűlés honlapjáról.

Mivel szerettük volna azt is bemutatni, hogy hogyan integrálható az archivált tartalom egyéb digitális anyagokkal, ezért a webarchívumot kiegészítettük az OSZK E-könyvtári Szolgáltatások Osztály kezelésében levő három további gyűjtemény állományából a Rákóczihoz és korához kötődő dokumentumokkal. A Magyar Elektronikus Könyvtárból (MEK) könyveket és könyvrészleteket, az Elektronikus Periodika Archívum és Adatbázisból (EPA) folyóirat- és újságcikkeket, a Digitális Képarchívumból (DKA) pedig képeket válogattak össze az ezeket a szolgáltatásokat gondozó munkatársaink. Ezek között az OSZK-ban vagy máshol digitalizált, internetről gyűjtött, illetve szerzők vagy kiadók által beküldött eredeti digitális dokumentumok egyaránt vannak. A MEK egyébként több tucat könyvvel bővült a Rákóczi-archívum miatt, melyek a Google Books és az Internet Archive szervereiről kerültek begyűjtésre. Összességében már közel ezer tételt tartalmaz ez a különgyűjtemény, melyek nagyjából azonos arányban származnak a négy digitális szolgáltatásunkból.



A Rákóczi-archívum összetétele gyűjtemények szerint

Számunkra is tanulságos volt megtapasztalni, hogy a webarchiválási technológia milyen jól kiegészíti az eddigi tevékenységeinket, mert már nemcsak egyedi dokumentumokat tudunk az internetről lementeni és szolgáltatni, hanem sok-sok fájlból álló, komplex weblapokat vagy egész webhelyeket is.

Metaadatok

A webarchívumhoz még 2018-ban kidolgoztunk egy adatstruktúrát, figyelembe véve az amerikai könyvtári szervezet, az Online Computer Library Center (OCLC) által létrehozott Web Archiving Metadata Working Group ez év februári ajánlását, ami elsősorban a Dublin Core adatkészleten alapul, de MARC 21 és MODS megfeleltetéseket is tartalmaz. A főként bibliográfiai információk leírására szolgáló mezőket kiegészítettük technikai és adminisztratív adatmezőkkel is, így több mint százféle adatot tudunk rögzíteni a lementett webhelyekről, valamint az azokból kialakított részgyűjteményekről is. Ezek az XML-formátumú metaadat rekordok a nyilvános webarchívumban megtekinthetők, és az adatséma, valamint a kitöltési útmutató is publikus. Az OSZK-ban bevezetésre tervezett RDA (Resource Description and Access) katalogizálási szabványhoz való hozzáigazításon is elkezdtünk dolgozni tavaly az RDA-HU munkacsoport segítségével.

A FEJEDELEM ZÁSZLAJA ALATT

Azonosító: R19-100126
Típus: Weboldal
Az OSZK-ban archivált dokumentum:
http://webarchivum.oszk.hu/pywb/*?url=https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-i-resz/
Az OSZK-ban archivált dokumentum:
http://webarchivum.oszk.hu/owb/wayback/*/https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-i-resz/
Az OSZK-ban archivált dokumentum:
http://webarchivum.oszk.hu/pywb/*?url=https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-ii-resz/
Az OSZK-ban archivált dokumentum:
http://webarchivum.oszk.hu/owb/wayback/*/https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-ii-resz/
Az Internet Archive mentései:
http://web.archive.org/web/*/https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-i-resz/
Az Internet Archive mentései:
http://web.archive.org/web/*/https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-ii-resz/
A dokumentum eredeti forrása:
https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-i-resz/
A dokumentum eredeti forrása:
https://felvidek.ma/2019/05/a-fejedelem-zaszlaja-alatt-ii-resz/
Főcím: A fejedelem zászlaja alatt
Alcím/egyéb cím: Egy csallóközi család kuruc kori hagyományai
Tartalmazó kiadvány címe: Felvidék.ma
Létrehozó személy: Somogyi Mátyás
Kiadó neve: Szövetség a Közös Célokért
Kiadó honlapja: http://szakc.sk/
Kategória: Szabadságharc, kuruc kor
Tárgyszó: katonatiszt
Tárgyszó: lovasság
Tárgyszó: kuruc
Tárgyszó: Rákóczi-szabadságharc
Személynév: Somogyi Ferenc (derghi)
Nyelv kód: hun
Archiválás dátuma: 2019-10-08
Hozzáférés státusza: nyilvánosan is szolgáltatható
Hozzáférhetőség kezdő dátuma: 2020-01-16

Két részből álló archivált újságcikk metaadat rekordja és oldalképe

Mivel a demó archívumban használt adatszerkezet teljes webhelyekre lett kitalálva, a Rákóczi-gyűjtemény számára viszont nagyrészt egyedi weboldalakat mentettünk le, továbbá a másik három gyűjteményből is át kellett venni a metaadatokat, és erre a munkafázisra csupán néhány hetünk volt, ezért egy egyszerűsített leírás mellett döntöttünk. Csak olyan bibliográfiai adatokat vettünk fel vagy át, amelyek egyaránt értelmezhetők a könyvek, könyvrészletek, cikkek, képek, videók, weboldalak és egyéb online műfajok esetében. Ezek a következők:

a dokumentum fő- és alcíme, az azt tartalmazó kiadvány összefoglaló címe, az eredeti fájl(ok) származási helye, a szerzők és közreműködők neve, a kiadó és annak honlapja, a dokumentum műfaja és tematikus kategóriája, a Köztauruszból vett tárgyszavak, a földrajzi és személynevek, valamint ezek névtér-azonosítói, és végül a nyelvkód. Ezeket az adatmezőket még néhány adminisztratív információval egészítettük ki, például a feldolgozó neve, az archívumba kerülés dátuma, a hozzáférhetőség státusza, az OSZK-ban és az Internet Archive-ban levő mentett verziók URL-je, a címlapképet vagy oldalképet tartalmazó fájl neve. Utóbbiak közül nem mindegyik jelenik meg a honlapon, mert egy részük csak nyilvántartási célokat szolgál.

A metaadatok melletti kép a MEK-es könyvek és könyvrészletek esetében a címlapot ábrázolja, az EPA-ból származó cikkeknél és tanulmányoknál a periodika valamelyik számának a borítóját, a DKA-ban levő képi dokumentumoknál pedig magának a képnek a kicsinyített verzióját jelenítjük meg. A weboldalakról is készült egy nagyobb (1280 pont széles) és egy kisebb (300 pont széles és maximum 600 pont magas) kép a Firefox képernyőkép-készítő funkciójával vagy a Nimbus Screen Capture nevű böngésző-kiegészítővel. Azoknál az eseteknél, amelyeknél (még) nem kaptunk engedélyt a nyilvános szolgáltatásra, csak ez a kis bélyegkép tekinthető meg.

Webarchiválás

Az archiválásra kiválasztott oldalról először a fent említett képet készítjük el, mert így lehet a legpontosabban dokumentálni, hogy hogyan néz ki az a jelenleg használatos böngészőkben. Ez a kép a későbbi minőségellenőrzéshez is hasznos lehet, mert az eredeti forrás nem biztos, hogy a jövőben is (ugyanabban a formában) elérhető lesz.

A Rákóczi-archívum az első olyan részgyűjteményünk, amelyben sok egyedi weboldal van, amiket egyesével mentettünk le, hogy önálló „konténerekbe" (úgynevezett WARC-formátumú fájlokba) kerüljenek az egyes tételek. Erre az engedélyeztetés miatt volt szükség, mert csak így tudjuk nyilvánosan szolgáltatni azokat a mentéseket, amelyekre sikerült szerződést kötni a tartalomgazdákkal. Ezért ennél a kis projektnél nem alkalmazhattuk a tömeges aratásra használt technikát, hanem egy PC-n, Windows alatt futó szoftverekkel készültek a mentések. Ez a munka nekünk is újdonság volt, és sok tanulsággal szolgált, mert például YouTube-videókat vagy Wikipédia-oldalakat korábban még nem archiváltunk. A használt szoftverek mindegyike alkalmas egy kisebbfajta intézményi vagy személyes webarchívum létrehozására, de mindegyiknek van előnye és a hátránya, így érdemes többet is kipróbálni.

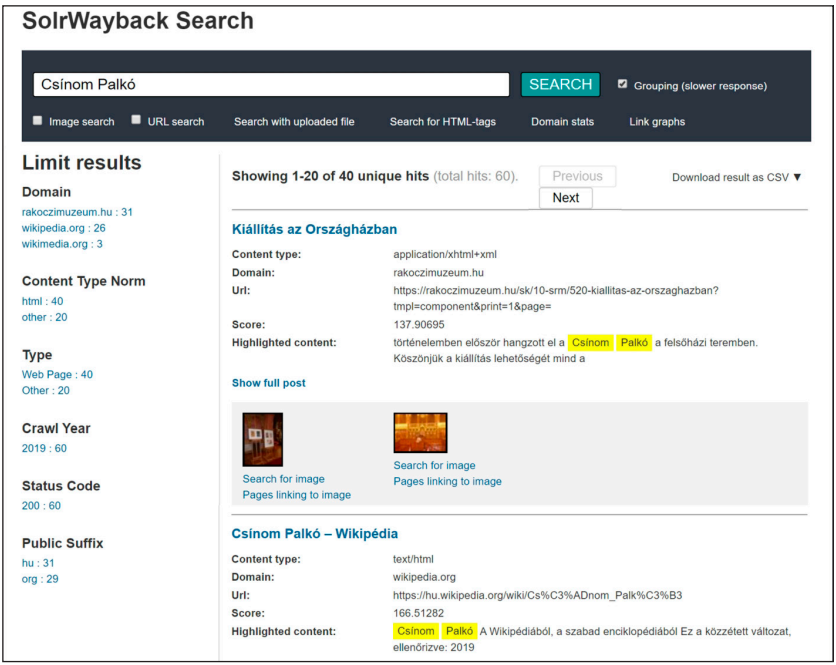
A legtöbb mentés a *PyWb* (Python nyelven írt Wayback) programmal készült. Ehhez fel kell telepíteni a Python környezetet és csak parancsmódban lehet használni. Bár, ahogy a neve is mutatja, a PyWb alapvetően a lementett weboldalak

visszajátszására szolgál, de van egy *record* üzemmódja is. Ha ezt bekapcsoljuk, akkor minden oldalt elment egy szabványos WARC-fájlba, amit megnyitunk a böngészőnkben. Tehát nekünk kell végigkattintgatnunk azokat a linkeket, amelyeket meg akarunk őrizni az archívumban. Ez elég időigényes, főleg azért, mert a program amúgy is elég lassan dolgozik. Viszont általában elég jó minőségben tudja lementeni még a bonyolult hírportálokat és a videók többségét is, amiket szintén el kell indítanunk ahhoz, hogy letöltse őket.

Amivel a PyWb nem boldogul, azt meg lehet próbálni a *Webrecorder* nevű, hasonló elven működő, és Windows alá is telepíthető, vagy ingyenes online szolgáltatásként igénybe vehető eszközzel is. Ha nem akarunk kattintgatni, akkor pedig a WAIL-rendszert lehet használni, amiben kétféle robotfunkció is van: az egyik a Chrome böngésző motorján keresztül tölti le a megadott weboldalt és egy szintig követi a benne levő linkeket is, a másik pedig a tömeges aratásoknál is használt *Heritrix crawlert* futtatja, de ez is a kezdőlaptól számítva legfeljebb csak három szintig megy lefelé a linkeken, így teljes webhelyek letöltésére nem alkalmas. Utóbbi célra a magyar nyelvű felülettel is rendelkező *HTTrack* ajánlható. Ezzel is készült néhány próbamentés a Rákóczi-archívum számára, de mert ez nem WARC-formátumban tárolja az anyagokat, ezért ezeket a mentéseket még át kellett konvertálni.

Az archivált cikk első része a PyWb-megjelenítőben

A WARC-fájlok visszanézésére a már említett PyWb szolgál, de a metaadatok közé felvettük az Internet Archive Wayback Machine szolgáltatásához hasonló Open Wayback (OWB) megjelenítő felületre mutató linket is, bár ezzel kevésbé jók a tapasztalataink. Belinkeltük továbbá az Internet Archive saját mentéseit, ha pedig az amerikai archívumban nem volt még mentés az adott weboldalról, akkor a *Save page now* szolgáltatásuk segítségével készíttettünk egyet. A weboldalak szövegében való keresésre a SolrWayback szoftvert használjuk, amivel a részletes technikai metaadatok is megnézhetők minden egyes tételnél, a *Toolbar* panel bekapcsolásával pedig különböző statisztikák és grafikonok is generálhatók.



Teljes szövegű keresés találati listája a SolrWayback programban

Problémák

Mivel a Rákóczi-archívumban sok az egyedi weboldal (pl.: újsághír, blogbejegyzés, Wikipédia-szócikk), ezért olyan újfajta problémákkal is találkoztunk, amelyekkel a korábbi, teljes webhelyekre kiterjedő aratásoknál nem szembesültünk. Az első ilyen kérdés mindjárt az, hogy meddig terjed egy webes dokumentum? Például egy hírportálon megjelent cikkhez hozzátartoznak-e a mellette megjelenő reklámok, az ajánlott további hasonló vagy éppen egész más témájú, de aktuálisan népszerű hírek, a külön megnyitható képgaléria, a beágyazott videók, a más szerverekről belinkelt PDF- és egyéb fájlok, az olvasói vélemények? A Wikipédia esetében külön mentsünk és írjunk le minden egyes szócikket, vagy

tekintsük egy könyvtári egységnek mondjuk a *https://hu.wikipedia.org/wiki/Rákóczi-szabadságharc* oldalt az abban hivatkozott – például az egyes hadjáratokat, csatákat, hadvezéreket részletesen leíró – szócikkekkel együtt? Hozzávegyük a Wikimédia Commons tematikus médiagyűjteményét is, amelyből az illusztrációk be vannak ágyazva a lexikon szöveges részébe? És az egyes nyelvi verziókat, melyek néha csak fordítások, máskor viszont teljesen újraírt szócikkek? Archiváljuk-e a szócikkek alakulását dokumentáló *laptörténet* és *vitalap* aloldalakat is?

A problémák másik részét az archiváláskor keletkező technikai nehézségek adják. Volt olyan weblap, amelynél már a teljes oldalt kitakaró cookie figyelmeztetésen vagy CAPTCHA teszten sem tudott túljutni az archiváló szoftver. Más esetekben az ékezetes fájlnevek vagy a csak JavaScript kódok lefuttatásával keletkező URL-címek akadályozták meg a programot a linkek követésében és a fájlok letöltésében. A biztonsági problémák miatt az internetről fokozatosan eltűnő Flash-formátum már a mentéseknél is gondot okoz: a böngészőn keresztül való archiváláskor külön engedélyezni kell ezeknek a fájloknak a letöltését és futtatását. (Az viszont kérdéses, hogy néhány év múlva egyáltalán lesz-e még olyan szoftver, amivel ezekről a Flash-alapú weboldalakról készült mentések visszanézhetők lesznek.) Viszonylag gyakori az is, hogy a webszervereken levő robots.txt fájlokban olyan útvonalak vannak – sokszor nem szándékosan – elzárva a robotok elől, amelyekre a Google keresőjének nincs szüksége, de archiválási szempontból fontosak: például a külalakot meghatározó CSS-fájlokat, vagy a navigációhoz használt programkódokat tartalmazó alkönyvtárak.

A hibák harmadik csoportja a visszanézéskor lép fel. Hiába van benne a WARC-konténerben a weboldalt alkotó összes fájl, ha azok egy részét a megjelenítő szoftver valamiért nem tudja megtalálni vagy megmutatni. Bár sokszor még azt sem könnyű eldönteni, hogy tényleg le lett-e mentve minden szükséges fájl és valóban csak a megjelenítő korlátaiba ütköztünk. A Wikipédiánál futottunk bele abba a problémába, hogy ott egy úgynevezett *srcset* attribútummal adják meg a szócikkekbe ágyazott képek méretét, ami azt eredményezi, hogy a webszerver a felhasználó képernyőfelbontásához optimalizálva küldi át a képeket. Vagyis ha böngészőn keresztül mentünk, akkor csak adott méretű képfájlok kerülnek az archívumba, amelyek egy más felbontású monitoron nem jelennek meg. További tipikus problémát jelentenek a JavaScript-, Java- vagy Flash-alapú képnézegetők, hang- és videólejátszók. Ezek a megjelenítő felületen sokszor nem indulnak el, még ha maguk a médiafájlok le is lettek töltve.

Engedélyeztetés

A technikai problémák mellett a jogi korlátozások is nehezítik a webarchiválást. A webhelyek szerzői jogi szempontból védett és nem védett alkotásokat is tartalmazhatnak, jellegüket tekintve gyűjteményes műnek minősülnek, felhasználásuk tekintetében – a tulajdonjog mellett – ez a szabályozás az iránymutató.

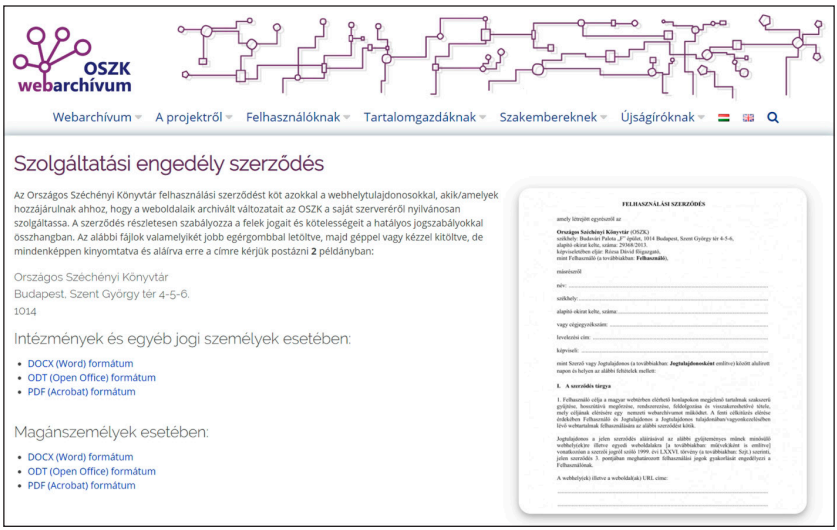
Bár egy szerzői jogilag védett alkotásról a másolatkészítés is engedélyköteles, a szabályozás szerint a közgyűjtemények szabad felhasználás keretében, nem haszonszerzés céljából végezhetnek többszörözést a védett művekről is, ami vonatkozik a webarchiválásra is. Azonban ez a szabály csak a megőrzésre és a helyben vagy zárt hálózaton történő szolgáltatásra ad lehetőséget, a nyilvános közzétételre nem. Utóbbi a jogtulajdonos engedélyével történhet, de ehhez a szerzői jogi törvény szerint kétoldalú szerződés kell. Mivel a webhelyek szerzői jogilag gyűjteményes művek, ezért nem kell külön-külön szerződni minden jogtulajdonossal. Tulajdonjogi szempontból viszont lehet több tulajdonos, akikkel külön kell megállapodni.

A Rákóczi-archívum jogvédett webes tartalmai esetében ugyanazt az engedélyeztetési eljárást alkalmaztuk, mint amit korábban a demonstrációs célból létrehozott gyűjtemény esetében kialakítottunk. Annyi eltéréssel, hogy most nemcsak teljes webhelyekre kértünk szolgáltatási engedélyt, hanem gyakran csak egy-egy weblapra, mivel főleg cikkek és blogbejegyzések kerültek ebbe a gyűjteménybe; továbbá áttértünk a határozatlan idejű szerződésekre. A többi műnél nem kellett az engedélyekkel bajlódni, mivel a MEK, EPA és DKA állományából válogatott dokumentumok már rendelkeznek ezekkel, vagy eleve szabad felhasználásúak – ahogy az ilyen minősítésű webhelyek esetében sem kértünk felhasználási engedélyt.

Persze ahhoz, hogy valakivel szerződni tudjunk, ismerni kell a kilétét és valahogy fel kell venni vele a kapcsolatot. De ahogyan a nyomtatott kiadványoknál is problémás néha a jogtulajdonos és az elérhetősége kiderítése, nincs ez másként a webhelyek esetében sem. Egy hagyományos honlapon még csak-csak találunk impresszumot vagy legalább egy kapcsolati e-mail-címet, netán jogvédettségre vagy CC licencre utaló kitételt, a blogoknál vagy a közösségi oldalaknál ezek gyakran hiányoznak. Jobb esetben van mód online üzenetküldésre, rosszabb esetben csak egy bejegyzés kommentelésére van lehetőség, ami lássuk be, nem a legjobb formája a hivatalos kapcsolatfelvételnek. Némi szerencsével, kerülő utakon juthatunk információhoz, ha például egy Facebook-fiók nevéből vagy adatlapjáról kiderül a tulajdonos kiléte és alternatív forrásból találunk hozzá kapcsolati adatot is.

Bizonyos esetekben azonban hiába fordulunk hozzájárulásért a weboldal tulajdonosához, hiszen az oldal és a tartalom tulajdonosa nem mindig esik egybe. Ugyanis az internet technológiája a különböző tartalmak összekapcsolására épül, aminek alapvetően háromféle formája lehet: hivatkozás, beágyazás és tényleges átmásolás. Bár tartalmilag és archiválási szempontból egy egységnek számíthat az adott webhely, és a felhasználók számára sem okoz problémát, hogy ezek a különböző forrásból származó tartalmak nem különülnek el egymástól élesen – hiszen egy felületen látszanak –, jogi szempontból ezek külön esetek lehetnek. Ilyenkor a többi jogtulajdonos kilétét is ki kell deríteni, és velük is fel kell venni a

kapcsolatot. Paradox módon előfordulhat olyan eset is, hogy az archivált weblap egyik részére van szolgáltatási engedély, míg más részére, például a beágyazott videóra nincs. Ilyenkor csak az engedélyezett rész látszódhat nyilvánosan, aminek technikai előfeltétele, hogy a résztartalmak külön kerüljenek archiválásra, s emiatt a jogosultság tisztázása meg kellene hogy előzze az archiválás műveletét. Ez persze olyan elvi kérdéseket vethet fel, mint például, hogy az archív példány mennyiben tekinthető az eredeti kompilált tartalom másolatának? Vagy mi legyen akkor, ha a kiegészítő tartalomra van engedély, de arra nincs, amibe be volt ágyazva?



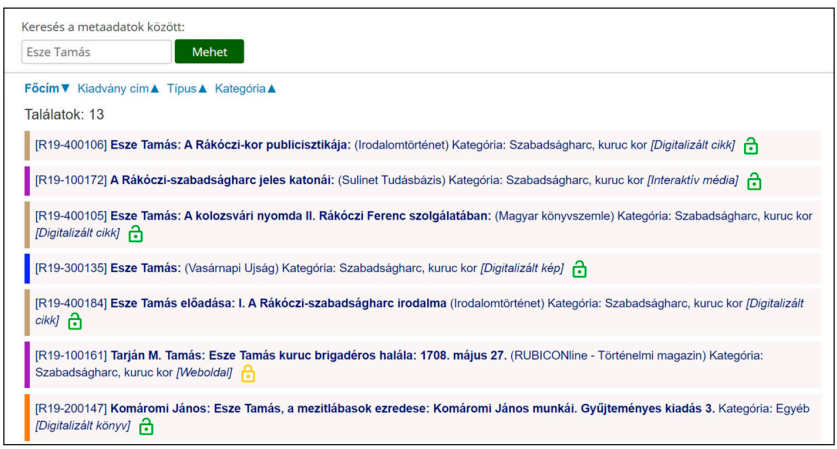
Letölthető felhasználási szerződés a webarchívum honlapján

Engedélyeztetéskor elektronikus úton próbáljuk felvenni a kapcsolatot a tartalomgazdákkal. A tájékoztató levélben röviden írunk a webarchiválásról, annak technikai hátteréről, a jogi tudnivalókról, valamint a nyilvános szolgáltatáshoz szükséges szerződést is mellékeljük. Sajnos leveleink többsége reakció nélkül marad, körülbelül harmadrészükre kapunk választ, de ez nem kirívó, mert a külföldi webarchívumok munkatársai is ilyen arányról számolnak be. Arról nem rendelkezünk információval, hogy a válasz elutasítása, feledékenység, esetleg a levél céltévesztése miatt maradt-e el. Egy bizonyos idő után szoktunk emlékeztetőket küldeni, és az ezekre adott reakciók azért sejtetik, hogy hangsúlyos a feledékenység a válaszadás elmaradásában. A válaszok túlnyomó többsége pozitív, kategorikus elutasítás ritkán fordul elő elvi okból, sokkal inkább jellemző az, hogy a jogi tisztázatlanság áll a visszautasítás hátterében. Sok egyéb visszajelzést is kapunk, és többen maguktól ajánlják archiválásra a webhelyeiket.

A legtöbb probléma abból adódik, hogy a szerződést csak papíron lehet megkötni, eredeti példányokkal, mert a szerzői jogi törvény csak ezt a formát ismeri

a felhasználási engedélyre.* Ezt sokak kifogásolják, életszerűtlennek tartják a mai világban, és olyanra is volt példa, hogy bár elektronikus úton elküldték nekünk a kitöltött szerződést, de papíron már nem, ami természetesen így nem érvényes. A Rákóczi-archívum esetében az is gondot okoz, hogy mivel az egész emlékév alatt bővítjük a gyűjteményt, ezért az új tartalmak miatt folyamatosan kell engedélyeket kiküldeni, így például egy hírportálnál könnyen lehet, hogy több szerződést kell kötnünk. Igyekszünk feleslegesen nem terhelni a partnereinket, ezért nem egyesével kérünk engedélyt, hanem bizonyos időközönként összegyűjtve több cikkre. Sajnos így is előfordult, hogy míg az első engedélykérésünkre kimondottan pozitív reagálás érkezett, az újabb megkeresésre már elmaradt a válasz.

Az emberi erőforrások szűkössége is gondot okoz. A nyilvános szolgáltatás kezdete óta még nem sikerült minden szolgáltatni kívánt webhely ügyében elküldeni a szerződést, se a demó, se a Rákóczi-archívum esetében. Míg az archiválási folyamatok nagy része automatizálható, a szerződéskötések adminisztrálása és a szükséges levelezés lebonyolítása aránytalanul sok időt igényel az egyéb feladatok mellett. Ez az oka annak, hogy a Rákóczi-gyűjteményben az archivált weboldalak közel fele még(?) nem hozzáférhető. A böngészhető listákban látszik, hogy milyen státuszban van az engedélyeztetési folyamat: a zöld lakat jelzi a szabad hozzáférést; a sárga azt, ha az még nem zárult le; és a piros jelezné az elutasítást, de eddig ilyen szerencsére még nincs.



A hozzáférés szintjét jelző ikonok a metaadat-kereső találati listájában

* A szerzői jogról szóló 1999. évi LXXVI. törvény 45. § (1) bekezdése írja elő a kötelező írásbeliséget, az pedig a Ptk-ból következik, hogy e-mail útján nem lehet írásbeli szerződést kötni.

Szolgáltatás

A Rákóczi-archívumhoz 2019 utolsó negyedében egy önálló webhelyet készítettünk, amely egyben prototípusként szolgált az OSZK webarchívumának új honlapjához. Korábban ugyanis csak egy ideiglenesnek szánt, kézzel szerkesztett HTML-fájlban tettük közzé a projekt híreit, és innen volt elérhető a nyilvános demó gyűjtemény, a szakirodalmi bibliográfia, a wiki, az éves workshop oldala és az ajánló űrlap, amivel bárki javasolhat megőrzésre érdemes magyar és magyar vonatkozású webhelyet. Az új honlap már WordPress-alapú, és ezzel a tartalomkezelővel készült a Rákóczi-gyűjtemény felülete is, amihez még különböző kereső- és böngészőfunkciókat fejlesztettünk. Az archívum anyaga kilistázható tematikus kategóriák, dokumentumtípusok, illetve gyűjtemények szerint. Keresni lehet egyszerre az összes metaadatban, vagy több mező kombinálásával, a találati listák pedig négyféle szempont szerint rendezhetők. A teljes szövegű kereső jogi és technikai okok miatt csak a weboldalakra és azok közül is csak a nyilvánosan szolgáltathatókra terjed ki.

A honlap hatnyelvű, a magyar mellett készült angol, német, francia, lengyel és szlovák verzió is. Ezt a sokféle változatot részben az indokolja, hogy Rákóczi és a szabadságharc hatása európai jelentőségű volt, és bár az archívumban levő dokumentumok többsége magyar nyelvű, mégis hasznos lehet ezekre is felhívni a külföldi kutatók figyelmét. A másik ok pedig az, hogy úgy gondoljuk, más országok könyvtárai számára is érdekes egy ilyen szolgáltatás annak demonstrálására, hogy hogyan lehet archivált weboldalakból és más digitális vagy digitalizált dokumentumokból egy tartalomszolgáltatást kialakítani.

Együttműködés

A KDS-K pályázaton nyertes könyvtárak vezetőit megkerestük azzal a kéréssel, hogy jelöljenek ki egy-egy kapcsolattartót, akivel tudunk egyeztetni az együttműködés keretében megvalósítható feladatokról. Szerencsére az intézmények többsége számára nem teljesen ismeretlen ez a terület, mert munkatársaik közül néhányan már részt vettek vagy az OSZK-ban évente megrendezett *404 Not Found – Ki őrzi meg az internetet?* című workshopok, vagy a Könyvtári Intézet által szervezett *Az internet archiválása mint közgyűjteményi feladat* című tanfolyamok valamelyikén. A kapcsolattartókkal és a közreműködő kollégákkal 2020 február végén és március elején videobeszélgetések formájában tekintettük át a május végéig elvégezhető munkát.

Elsősorban az archiválásra érdemes webhelyek válogatásában kértük a segítségüket, hiszen a helyi kollégák jobban ismerik az adott régióban fontos online információforrásokat, illetve a saját honlappal, bloggal, Facebook- vagy Instagram-oldallal rendelkező intézményeket, közszereplőket, művészeket vagy akár olyan magánembereket, akik szélesebb kör számára is érdekes tartalmakat tesznek közzé az interneten. A címek nyilvántartására egy megosztott táblázatot hoztunk

létre, melyben minden könyvtár egy külön munkalapon rögzítheti a tágabb régiójába tartozó webhelyek nevét és URL-címét, illetve esetleg az engedélyeztetés ügyében illetékes elérhetőségét is, ha az nem deríthető ki könnyen. A táblázatba a webarchívumban használt, illetve tervezett tematikus kategóriák szerint lehet bevinni az adatokat (pl.: helytörténet-helyismeret, irodalom, művészet, kutatás, oktatás, művelődés, vallás, média, sport), és van egy kiemelten fontos műfaji kategória is az elektronikus periodikáknak. Hogy mely webhelyeket tartunk már nyilván és archiválunk időszakosan, azt egy úgynevezett *seed-kereső* segítségével lehet ellenőrizni, ahol elég csak egy jellemző részletet beírni az URL-ből és kapunk egy listát azokról a webhelyekről, amelyeknek a címében ez a betűcsoport előfordul, és benne vannak az OSZK webarchívumában. A táblázat kitöltésére vonatkozó tudnivalókat írásban is elküldtük a MIA-L levelezőlistára, melynek a partnerkönyvtárak kapcsolattartóin kívül bárki tagja lehet, aki érdeklődik az internetes tartalmak hosszú távú megőrzése iránt.

Ha jut rá idő, akkor szeretnénk bevonni a megyei és városi könyvtárakban dolgozó kollégákat az engedélyeztetés folyamatába, valamint legalább kísérleti jelleggel a minőségellenőrzésbe és a metaadatolásba, hogy a magyar könyvtárosoknak is legyen gyakorlati tapasztalatuk ezen a szakterületen. Továbbá igény esetén szívesen tartunk kihelyezett előadásokat és bemutatókat vagy akár tanfolyamokat is. Reméljük, hogy a most kiépülő szakmai kapcsolatok a KDS-K projekt határidejének lejárta után is megmaradnak, és valamilyen formában továbbra is együtt tudunk majd dolgozni ezekkel a könyvtárakkal.

A nemzeti webtér archiválása olyan méretű és bonyolultságú feladat, hogy ezt a legtöbb országban elosztott módon, több intézmény együttműködésével végzik. Nálunk az első időszakban egy olyan kooperáció képzelhető el, hogy az archiválást és a szerződések megkötését az OSZK végzi, a hazai közgyűjtemények pedig a többi munkafázisba segítenek be. Az archivált webtartalmak közül a számukra fontosakat a könyvtárak a saját digitális szolgáltatásaikba is beépíthetik, vagy úgy, hogy a nyilvános archívumból belinkelik azokat, vagy pedig úgy, hogy az OKP projektben tervezett OSZK-pontokon, vagyis a könyvtárakba kihelyezett terminálokon keresztül a nem publikus gyűjteményhez is hozzáférést adnak a felhasználóknak.

Webcímek

KDS-K: Pályázat a Közgyűjteményi Digitalizálási Stratégia végrehajtásához szükséges könyvtári digitalizálás támogatására: http://www.oszk.hu/kds-k/palyazat_2019 (2020.03.20.)

Rákóczi-archívum: https://rakoczi2019.webarchivum.oszk.hu (2020.03.20.)

A webarchiválás projekt régi honlapja: http://mekosztaly.oszk.hu/mia (2020.03.20.)

A webarchívum új honlapja: https://webarchivum.oszk.hu (2020.03.20.)

Nyilvános Demó archívum: https://webarchivum.oszk.hu/demo-kezdolap/ (2020.03.20.)

MIA Wiki: http://mekosztaly.oszk.hu/miawiki (2020.03.20.)

Metaadatséma és útmutató: http://mekosztaly.oszk.hu/mia/xml/ (2020.03.20.)

Seedkereső: http://webadmin.oszk.hu/seed-kereso (2020.03.20.)

Publikus ajánló űrlap: https://goo.gl/forms/Y1qIIxcM7APPiq443 (2020.03.20.)

MIA-L levelezőcsoport: http://mekosztaly.oszk.hu/cgi-bin/mailman/listinfo/mia-l (2020.03.20.)

Információs e-mail-cím: mia@mek.oszk.hu

Irodalom

Drótos László: *Webes tartalmak digitális megőrzése.* = Könyv, Könyvtár, Könyvtáros, 27. évf. 2018. 10. sz. 11-17. p. https://epa.oszk.hu/01300/01367/00307/pdf/EPA01367_3K_2018_10_011-017.pdf (2020.03.20.)

Drótos László: *A webarchívum és a KDS kapcsolata.* = *Könyvtárak kincsei digitális formában – a magyar könyvtárak digitalizálási stratégiája* konferencia. OSZK, Budapest, 2019. 04.17.
Az előadás prezentációja az alábbi oldalon érhető el: https://webarchivum.oszk.hu/Webarchivum_KDS/ (2020.03.20.)

Drótos László: *Az OSZK webarchívumának újdonságai.* = *„404 Not Found – Ki őrzi meg az internetet?" workshop.* OSZK, Budapest, 2019.11.14.
Az előadás prezentációja az alábbi oldalon érhető el: https://webarchivum.oszk.hu/Drotos_Laszlo_Az_OSZK_webarchivumanak_ujdonsagai/ (2020.03.20.)

Az előadásról készült videófelvétel az alábbi oldalon érhető el:
http://videotorium.hu/hu/recordings/35066 (2020.03.20.)

Drótos László – Moldován István: *Az OSZK webarchiváló kísérleti (pilot) projektjének eredményei és egy üzemszerűen működő magyar webarchívum terve.* = Könyvtári Figyelő, 29. (65.) évf. 2019. 1. sz. 38-51. p. https://epa.oszk.hu/00100/00143/00355/pdf/EPA00143_konyvtari_figyelo_2019_01_038-051.pdf (2020.03.20.)

Drótos László – Moldován István: *Ki őrzi meg a helyi webet? Helyismereti vonatkozású internetes tartalmak archiválása webaratással.* = MKE Helyismereti Könyvtárosok Szervezete XX. Országos Konferenciája, Győr, 2018.07.26.
Az előadás prezentációja az alábbi oldalon érhető el: https://webarchivum.oszk.hu/Ki_orzi_meg_a_helyi_webet_MKE_2018/ (2020.03.20.)

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 (2020.03.20.)

Halász Annamária: *A webarchiválás jogi feltételrendszerének biztosítása. = 404 Not Found – Ki őrzi meg az internetet? workshop.* OSZK, Budapest, 2018.11.15.
https://videotorium.hu/hu/recordings/28736/ (2020.03.20.)

Ilácsa Szabina: *Webhelyek metaadatolási problémái. = 404 Not Found – Ki őrzi meg az internetet? workshop.* OSZK, Budapest, 2019.11.14 .
Az előadás prezentációja az alábbi oldalon érhető el: https://webarchivum.oszk.hu/Ilacsa_Szabina_Webhelyek_metaadatolasi_problemai/ (2020.03.20.)
Az előadásról készült videófelvétel az alábbi oldalon érhető el:
https://videotorium.hu/hu/recordings/35078/07-ilacsa-szabina-webhelyek-metaadatolasi-problemai (2020.03.20.)

Kokas Károly: *Szegedikum a webarchívumban. A helyi érdekű webarchiválás lehetőségei az OSZK webarchiválási programja keretében. = 404 Not Found – Ki őrzi meg az internetet? workshop.* OSZK, Budapest, 2019.11.14.
Az előadásról készült videófelvétel az alábbi oldalon érhető el: https://videotorium.hu/hu/recordings/35069/ (2020.03.20.)

A cikkben szereplő képernyőfotók a Rákóczi-archívum honlapjáról készültek.