

Drótos László
OSZK, E-könyvtári Szolgáltatások Osztály

Webes tartalmak digitális megőrzése

A „Born Digital – Digitális tartalom, digitális szolgáltatás” című K2 műhelynapon
2018. október 10-én elhangzott előadás szerkesztett változata

A már eleve digitálisan születő tartalom gyűjtése és hosszú távú megőrzése komoly kihívás a memóriaintézményeknek. Ha ezt a feladatot nem tudják felvállalni, akkor vagy nagy fehér foltok maradnak az utókorra a 21. század első felének kulturális, tudományos, közéleti és személyes történéseiből, vagy csak a non-profit és az üzleti világ szereplői fogják elvégezni ezt a munkát¹. Természetesen ez is nagyon hasznos, de ezeknél a szervezeteknél és cégeknél nem valószínű, hogy évtizedeken vagy akár évszázadokon keresztül megmaradnak, és hogy egyenlő hozzáférést tudnak/akarnak adni mindenkinek a megőrzött tartalomhoz.

Amíg csak egyedi dokumentumokról van szó (pl. könyvek, folyóirat- és egyéb lapszámok, képek, videók), addig a közgyűjteményeknek azok a gyarapítási, feldolgozási és szolgáltatási munkafolyamatait, amelyeket a hagyományos és a digitalizált dokumentumokra kidolgoztak, nagyjából megfeleltethetők a *born digital* típusúakra is, nehézséget „csak” ezek nagy száma, megtalálhatósága, igen vegyes minősége, sokféle formátuma és gyakran tisztázatlan státusza² jelent. De az interneten még ezek a lehatároltnak és (legalább ideiglenesen) lezártak tekinthető dokumentumok sem elkülönülten jelennek meg, hanem be vannak ágyazva egy webes környezetbe: kapcsolódhatnak hozzájuk egyéb tartalmak (pl. kiegészítő multimédia anyagok, linkekkel hivatkozott további dokumentumok, olvasói/nézői vélemények és értékelések), melyeket szintén érdemes volna megőrizni, mert az eredeti kontextus nélkül a digitális közegben született és publikált dokumentumok értelme és értéke is megváltozik.

A fent említett, a könyvtárak számára ismerős dokumentumtípusok mellett ott vannak még az olyan internetes műfajok, mint a honlap, a hírportál, a wiki, a blog, a közösségi média, a fórum, a chat, az elektronikus levél és hírlevél, a videokonferencia, a vlog, a podcast, a sugárzott hang és videó, a 3D kép, az adatbázis, a digitális tananyag, az interaktív térkép, az online játék, a virtuális világ szimuláció, a webes műalkotás, az internetes mém, a linkgyűjtemény, és így tovább – amelyekről még azt sem tudjuk, hogy kinek a feladata lenne ezek legjavának megőrzése és milyen módon. De nemcsak a jövő felé van/lenne ilyen kötelességünk, mert a jelenben is igen komoly probléma az, hogy a sajtóban, a tudományos publikációkban és a tananyagokban egyre gyakrabban hivatkozott online források vagy eltűnnek, vagy elvándorolnak, vagy megváltozik a tartalmuk, így pár év, sőt akár már pár hónap múlva a linkek többsége elavul.

Szerencsére számos közgyűjtemény van szerte a világon, amely a saját állománya digitalizálása mellett foglalkozik a digitálisan keletkező és terjedő tartalom valamely részével is. Csak nemzeti szintű webarchívum projektből mintegy 40 indult 1996 óta, és külföldön az sem ritka már, hogy egyetemi, tudományos vagy közkönyvtárak építenek kisebb-nagyobb gyűjteményeket lementett webhelyekből és egyéb online tartalmakból, akár önállóan, akár másokkal együttműködve. Egyes levéltárak, audiovizuális archívumok és kortárs művészeti múzeumok is beszálltak ebbe a tevékenységbe és mentik az érdeklődési körükbe tartozó

¹ Lásd pl. az amerikai non-profit szervezet, az Internet Archive (<http://archive.org>) állományát, amely 339 milliárd weboldalt, 19 millió könyvet, 4.5 millió videót, 4.7 millió hangfelvételt, 3.2 millió képet és 290 ezer szoftvert tartalmaz. (A könyv-, videó-, hang- és kép-gyűjteményekben vegyesen vannak digitalizált és digitálisan született művek.)

² Például: mi tekinthető kiadványnak? mi esik a köteleespéldány szabályozás alá? mennyiben más, mint a nyomtatott kiadása? ki az illetékes jogtulajdonos? milyen feltételekkel szolgáltatható?

szegmensét az internetnek. Magyarországon eddig csak az egyedi dokumentumok archiválása volt „üzemszerű”, bár az sem tömeges méretekben. Az 1994-ben indult, majd 1999-től az OSZK-ba került MEK³ a digitális könyvek megőrzését és szolgáltatását vállalta fel, a 2004-től létező EPA⁴ az elektronikus periodikákkal foglalkozik, a 2007-ben alapított DKA⁵ pedig a képi dokumentumokra koncentrál. Bár mindhárom gyűjteményben vannak digitalizált anyagok is, gyarapodásuk másik fontos forrása az internet. 2006-ban elkészült az OSZK-ban a MIA⁶, vagyis egy leendő Magyar Internet Archívum terve is, amely a webhelyekre és más online műfajokra terjedne ki, de ennek a megvalósítása csak 2017-ben kezdődhetett el, az Országos Könyvtári Rendszer⁷ kiépítését szolgáló projekt részeként. Az elsődleges feladat a könyvtári szempontból legfontosabb médium, a web megőrzése lenne. Egy fenntartható és közgyűjteményi együttműködés keretében működtethető nemzeti webarchívum technikai, szakmai és jogi feltételeit igyekszünk megteremteni az ez év végéig tartó előkészítő fázisban.

A webnek nevezett digitális univerzum – a fizikailag létező világegyetemhez hasonlóan – egyetlen pontból, a CERN szerverén 1990 decemberében létrehozott HTML fájlból⁸ terjedt ki egy határtalan, folyamatosan születő és pusztuló világhálónak, amelyben bár vannak lokális struktúrák: fájlok, weblapok, webhelyek, webhely-csoportok, de a linkek révén minden mindennel kapcsolatban van, így az egész web egyetlen óriási hipermédia dokumentum. Természetesen ahhoz, hogy könyvtári szempontból valamit kezdeni lehessen vele, muszáj valahogy szegmentálni, s valamilyen gyűjtőkört és várható felhasználást megfogalmazni.

A jelenlegi fő célkitűzésünk ez: A magyar webtérben nyilvánosan elérhető – kiemelten a kulturális, a tudományos, az oktatási és a közéleti jellegű – digitális tartalmak rendszeres mentése és hosszú távú megőrzése kutatási, oktatási, hivatkozhatósági, bizonyíthatósági, helyreállíthatósági és egyéb célokra.

A „magyar webtér” alatt pedig a következőt értjük: A magyarországi domén (.hu) alá bejegyzett címeiken lévő webhelyek, valamint a külföldi doméneken magyar természetes vagy jogi személyek által létrehozott webhelyek összessége a jelenben; továbbá minden olyan egyéb weboldal az élő weben, amely magyar vonatkozású, ill. magyar célközönségnek szól.

Ennél bővebb a „magyar webtartalom” fogalma, ami a magyar webtérben létező vagy valaha létezett digitális tartalmak összessége, beleértve tehát azokat is, amelyek már az élő weben nem elérhetők. Mivel az első hazai webszerver 25 éve, 1993-ban indult el a BME-n⁹ és ez alatt a negyedszázad alatt weboldalak milliói tűntek el a magyar webtérből, ezért fontos lenne a még valahol (pl. az Internet Archive-ban, a szomszédos országok webarchívumaiban, a lekapcsolt szerverek winchesterein, a fiókokban elfekvő optikai lemezekben) fellelhető régi magyar webtartalom begyűjtése is.

A webarchívumot előkészítő projekthez két új munkatársat vettünk fel az E-könyvtári Szolgáltatások Osztályra, akik két részmunkaidős informatikussal és jelen cikk szerzőjével, mint témafelelőssel alkotnak egy munkacsoportot. Egyelőre két ideiglenes szerveren folynak a tesztek. Egy nagyobb teljesítményű (128 GB memória, 20+4 TB tárhely) gépet a KIFÜ¹⁰

³ Magyar Elektronikus Könyvtár: <http://mek.oszk.hu>

⁴ Elektronikus Periodika Archívum és Adatbázis <http://epa.oszk.hu>

⁵ Digitális Képtárhely: <http://dka.oszk.hu>

⁶ Drótos László: Mi a MIA? – Javaslat egy Magyar Internet Archívum létrehozására
<http://mek.oszk.hu/html/irattar/eloadas/2006/mia.htm>

⁷ OKR-projekt: <http://www.oszk.hu/okr-projekt>

⁸ CERN – Home of the first website: <http://info.cern.ch>

⁹ BME Irányítástechnika és Informatika Tanszék: http://www.fsz.bme.hu/www/other_h.html

¹⁰ Kormányzati Informatikai Fejlesztési Ügynökség: <http://kifu.gov.hu>

biztosít, amelyen az egyszerre sok száz vagy sok ezer webhelyre kiterjedő, több napos aratások futnak, és van az OSZK-ban egy kisebb szerver a szoftvertesztek, az egyedi próbamentések céljára és a nyilvános demó gyűjtemény szolgáltatásához. A tervek szerint 2019-ben egy ennél lényegesen komolyabb infrastruktúra áll majd rendelkezésre az üzemszerű működéshez, ennek beszerzése folyamatban van.

Weboldalak és webhelyek letöltésére többféle szoftver és szolgáltatás létezik, köztük sok ingyenes. A Windows alatt is használhatók (pl. ScrapBook X¹¹, Web ScrapBook¹², WARCreate¹³, WAIL¹⁴, Webrecorder¹⁵) inkább a magáncélú és kis volumenű archiválásra szolgálnak, de például a nagyon felhasználóbarát és még magyar felülettel is rendelkező HTTrack¹⁶ programot mind a mai napig használják az 1996-ban indult ausztrál nemzeti webarchívumot, a PANDORA-t¹⁷ építő könyvtárakban is. Ezeknek a szoftvereknek egy része képes az Internet Archive-nál kidolgozott és 2009-ben ISO 28500 néven szabványosított WARC¹⁸ formátumba menteni, ami tulajdonképpen egy fájlkonténer: minden, amit a webszerver küld, beleértve a weboldal összes elemét és a technikai metaadatokat is, egyetlen .warc kiterjesztésű állományba kerül, amit azután még tömörítenek is általában.

Az Internet Archive emellett még két fontos szoftvert is kifejlesztett, melyeket szintén sok webarchívumnál használnak: a Heritrix¹⁹ nevű aratógépet és a Wayback²⁰ megjelenítőt, amivel a Heritrix robotjával begyűjtött és WARC-ba mentett webtartalom úgy böngészhető, mintha az élő weben navigálnánk. Mivel ezek parancsokkal és konfigurációs fájlokkal vezérelhető programok, ezért az évek során barátságosabb kezelőfelületek is készültek hozzájuk, s ezek plusz funkciókat is tartalmaznak (pl. metaadatok bevitelének lehetősége, az ismétlődő aratások ütemezése, a szolgáltatási engedélyek nyilvántartása, a mentett anyag minőségellenőrzése, részgyűjtemények kialakítása). Ilyen keretrendszer a már említett, amerikai fejlesztésű WAIL, valamint az új-zélandi Web Curator Tool²¹ és a dán NetarchiveSuite²². Szintén dán könyvtári fejlesztés a WARC-ban tárolt weboldalak megjelenítése mellett teljes szövegű keresőt és statisztikai, ill. vizualizációs funkciókat is tartalmazó SolrWayback²³, aminek a tesztelésébe mi is bekapcsolódtunk. Továbbá egy saját kereső prototípusát is elkészítettük SolrMIA²⁴ néven, mellyel a teljes szövegű találati listák tovább szűkíthetők a metaadatok közt tárolt főtéma, téma, altéma, műfaj és típus szerint; a listában szereplő fájlok alatt pedig megjelenik az eredeti webhelyek neve. (Ezeket az egységesített „főcímekeket” szintén az általunk XML-ben rögzített metaadatok közül veszi át a program.) Az eddig említettek mellett még egy olyan archiváló szoftver van, amit elkezdtünk tesztelni és valószínűleg szintén használni fogunk majd az üzemszerűen működő rendszernél

¹¹ <http://mekosztaly.oszk.hu/mediawiki/index.php/ScrapBook>

¹² http://mekosztaly.oszk.hu/mediawiki/index.php/Web_ScrapBook

¹³ <http://mekosztaly.oszk.hu/mediawiki/index.php/WARCreate>

¹⁴ <http://mekosztaly.oszk.hu/mediawiki/index.php/WAIL>

¹⁵ <http://mekosztaly.oszk.hu/mediawiki/index.php/Webrecorder>

¹⁶ <http://mekosztaly.oszk.hu/mediawiki/index.php/HTTrack>

¹⁷ [http://mekosztaly.oszk.hu/mediawiki/index.php/PANDORA_\(ausztr%C3%A1l\)](http://mekosztaly.oszk.hu/mediawiki/index.php/PANDORA_(ausztr%C3%A1l))

¹⁸ <http://mekosztaly.oszk.hu/mediawiki/index.php/WARC>

¹⁹ <http://mekosztaly.oszk.hu/mediawiki/index.php/Heritrix>

²⁰ <http://mekosztaly.oszk.hu/mediawiki/index.php/Wayback>

²¹ <http://mekosztaly.oszk.hu/mediawiki/index.php/WCT>

²² <http://mekosztaly.oszk.hu/mediawiki/index.php/NetarchiveSuite>

²³ <http://mekosztaly.oszk.hu/mediawiki/index.php/SolrWayback>

²⁴ <http://webadmin.oszk.hu/solrmia/>

is: a Brozzler²⁵. A böngésző (*browser*) és a keresőrobot (*crawler*) szavakból összerakott név arra utal, hogy a Heritrix, vagy például a Google által is használt, a weboldalakba ágyazott linkeket követő szoftverrobot ki lett egészítve egy böngészőmodullal (mégpedig a Chrome motorjával), így jobb minőségben lehet vele menteni a modern, dinamikus generált weboldalakat, mint az eredetileg még az 1.0-ás webhez készült Heritrix-szel.

A webhelyek archiválása számítástechnikailag egy meglehetősen bonyolult feladat. Részben a weben használt sokféle formátum, műszaki és design megoldás, program- és parancsnyelv, szerverbeállítás stb. miatt, részben pedig azért, mert a weboldalakat emberek számára fejlesztik, Ezért gyakran olyan interaktív funkciókat és vizuális megoldásokat tartalmaznak, amelyek egy ember számára kézenfekvőek, vagy legalábbis könnyen megtanulhatók, ám egy értelem és érzékszervek nélküli szoftverrobot nem veszi ezeket észre vagy nem tudja őket végrehajtani (pl. továbbgörgetni egy oldalt, vagy leokézni egy figyelmeztető ablakot). A problémák másik része pedig abból származik, hogy a lementett tartalom nem úgy jelenik meg az archívumban, mint az élő honlapon, mert például a külalakot meghatározó stílusfájlok egy olyan mappában vannak, ahonnan ki vannak tiltva a robotok, vagy mert a helyes megjelenítéshez és a webhelyen belüli navigációhoz olyan programok futnak az eredeti webszerveren, amelyek nem menthetők le, ill. nem működőképesek az archívumot üzemeltető gépen. Azért, hogy legalább képként megőrizzük pontosan azt a látványt, ahogyan egy honlap az adott időszakban elterjedt böngészőkben megjelent, az aratásokkal egy időben a webhelyek kezdőoldaláról PNG képfájlokat is készítünk. A web hosszú távú megőrzését nagyban segítené, ha a fogyatékkal élők számára bevezetett akadálymentes felületekhez hasonlóan robotbarát²⁶ és archívumbarát²⁷ megoldásokat is beépítenének a webfejlesztők és webmesterek a szolgáltatásaikba.

2017 nyarától 2018 októberéig többféle aratást is végeztünk a Heritrix programmal.²⁸ Csináltunk úgynevezett szelektív archiválásokat: könyvtárak, levéltárak, múzeumok, egyetemek, kutatóintézetek és önkormányzatok honlapjait, valamint irodalmi témájú webhelyeket és az EPA-ban „távoli”-ként nyilvántartott időszaki kiadványokat mentettük le 1-3 alkalommal. Néhány hétig folyamatosan mentettük azokat a weboldalakat, amelyek a 2018-as téli olimpiával, illetve az országgyűlési választásokkal foglalkoztak. A téma-, műfaj-, illetve esemény-alapú gyűjtések mellett végül egy országos méretűnek tekinthető aratást is lefuttattunk nagyjából egy hét alatt, amely 291 ezer, a .hu alá bejegyzett doménre terjedt ki. A másfél év alatt összegyűjtött, tömörítve mintegy 10 terabájtnyi anyag elsősorban tesztelési célokat szolgál, hogy felmérjük a magyar webtér nyilvános részének megőrzéséhez és az archívumra építhető szolgáltatásokhoz szükséges infrastruktúra igényt.

De hogy minél előbb legyen egy nyilvánosan használható szolgáltatása is a projektnek, egyedi engedélyeket kértünk a lementett webhelyek egy részének tulajdonosaitól és 2018 januárjában megjelentettünk egy kis demó gyűjteményt²⁹, amely mintegy 120 honlapból, blogból és időszaki kiadványból áll, s a korábban említett két teljes szövegű keresőt is beépítettük. (1. ábra) Minden webhely esetében megnézhető az általunk lementett néhány *memento*³⁰, az első mentéskor készült oldalkép, a kifelé mutató linkekből rajzolt gráf, az Internet Archive által mentett anyag, az eredeti honlap, valamint a részletes metaadatok. (2. ábra) Az adatszerkezet

²⁵ <http://mekosztaly.oszk.hu/mediawiki/index.php/Brozzler>

²⁶ http://mekosztaly.oszk.hu/mediawiki/index.php/Crawler-friendly_website

²⁷ http://mekosztaly.oszk.hu/mediawiki/index.php/Archive-friendly_website

²⁸ Általában csak a kezdőoldaltól számított két-három szint mélységig ment le a robot és videofájlokat többnyire nem töltöttük le.

²⁹ <http://mekosztaly.oszk.hu/mia/demo/>

³⁰ <http://mekosztaly.oszk.hu/mediawiki/index.php/Memento>

kialakításánál az amerikai könyvtári szervezet, az OCLC egyik munkacsoportjának³¹ ajánlását vettük alapul, és ezt az elsősorban bibliográfiai adatmezőkből álló struktúrát bővítettük ki olyan – főként adminisztratív és technikai jellegű – mezőkkel és almezőkkel, amelyekre szükségünk volt ahhoz, hogy az egyes munkafolyamatok során keletkező valamennyi információt rögzíteni tudjunk. Így összesen több mint százféle adatot tudunk eltárolni egy webhellyel kapcsolatban, és emellett készítettünk egy valamivel egyszerűbb adatszerkezetet a webarchívumot alkotó egyes részgyűjtemények leírásához is.

A projekt kezdete óta folyamatosan igyekszünk minden lényeges információt megosztani szakmai és szélesebb körökben is, mert a magyar internet megőrzése olyan méretű feladat, amit nem tud megoldani egyetlen intézmény és benne néhány ezzel foglalkozó munkatárs. Fontos lenne, hogy minél többen ismerjék meg ennek a szakterületnek az alapjait és kapcsolódjanak be a munkába, akár úgy, hogy megőrzésre érdemes, de kevésbé ismert magyar webhelyeket ajánlanak az erre szolgáló úrlapon³², vagy archívumbaráttá alakítják át a honlapjukat, vagy segítenek a mentések minőségellenőrzésében és metaadatolásában, de akár úgy is, hogy helyi webarchívumokat hoznak létre. Az ismeretterjesztést szolgálja a projekt ideiglenes honlapja³³, a jelenleg már 30 fős MIA-L levelezőcsoport³⁴, a közel 600 szócikket tartalmazó MIA wiki³⁵, a több mint 450 tételes és többféle formátumban is elérhető szakbibliográfia³⁶, az elmúlt két évben publikált jó néhány cikk és megtartott előadás (ezek szintén megtalálhatóak a honlapon), a Könyvtári Intézet szervezésében tervezett továbbképzési tanfolyam és e-learning tananyag, valamint a 2018. november 15-én már második alkalommal megrendezésre kerülő „404 Not Found – Ki őrzi meg az internetet?” című félnapos workshop.

Ami a további terveket illeti: Újabb tematikus gyűjtéseket csinálunk majd és mellettük újraaratjuk az eddigieket is, figyelembe véve a korábbi ellenőrzések során talált problémákat, valamint legalább részgyűjtemény szinten leírjuk az összes eddigi mentést. A metaadatok egy részét már lehetőleg automatikus megoldásokkal állítjuk elő. Bővítjük a .hu domén alatt levő webhelyek listáját az eddig lementett weboldalakban levő linkekből kinyerhető további aldomén címekkel, és félévente lefuttatunk ezekre is egy-egy nagy aratást. Statisztikai funkciókat építünk be és kialakítunk egy raktári rendszert a WARC fájlok, az oldalképek és az egyéb segédállományok számára. Elkészítjük az üzemszerű működéshez és az Országos Könyvtári Rendszerhez való illesztéshez szükséges informatikai és munkafolyamat terveket. Belső útmutatókat, szabályzatokat írunk, segítjük a tartalomgazdákkal kötendő szerződés, valamint a webarchiválást szabályozó törvénytervezet szövegének megfogalmazását. Részt veszünk az internet megőrzésével foglalkozó intézményekből álló szervezet, az International Internet Preservation Consortium³⁷ munkájában, főként az oktatási munkacsoport keretében.³⁸ És tovább szorgalmazzuk a hazai együttműködést is a közgyűjtemények között a digitálisan születő, a papír-alapú világnál sokkal veszélyeztetettebb és tűnékenyebb kultúránk megőrzése érdekében.

³¹ http://mekosztaly.oszk.hu/mediawiki/index.php/OCLC_WAM

³² <https://goo.gl/forms/Y1qIIxcM7APPi443>

³³ <http://mekosztaly.oszk.hu/mia/>

³⁴ <http://mekosztaly.oszk.hu/cgi-bin/mailman/listinfo/mia-l>

³⁵ <http://mekosztaly.oszk.hu/miawiki>

³⁶ <http://mekosztaly.oszk.hu/mia/doc/webarchivalas-irodalom.html>

³⁷ <http://mekosztaly.oszk.hu/mediawiki/index.php/IIPC>

³⁸ A 2003-ban alapított IIPC-nek kb. 45 országból vannak tagjai és 2018-ban csatlakozott hozzá magyar részről OSZK is.

Ajánlott irodalom:

Dancs Szabolcs: Webarchiválási politikák

In: Könyv, könyvtár, könyvtáros, 2011. (20. évf.), 10. sz.

Drótos László: Az internet archiválása mint könyvtári feladat

In: Tudományos és Műszaki Tájékoztatás, 2017. (64. évf.), 7-8. sz.

Drótos László - Kokas Károly: Webarchiválás és a történeti kutatások

In: Digitális Bölcsészet, 2018. (1. évf.), 1. sz.

Drótos László - Németh Márton: Az OSZK-ban folyó kísérleti webarchiválási projekt első évének tapasztalatai

In: Tudományos és Műszaki Tájékoztatás, 2018. (65. évf.), 7-8. sz.

Németh Márton: A webarchiválásról történeti megközelítésben

In: Könyv, könyvtár, könyvtáros, 2018. (27. évf.), 2. sz.

Németh Márton: Nemzetközi körkép a webarchiválás gyakorlatáról

In: Könyvtári Figyelő, 2017. (63. évf.), 4. sz.

OSZK WEBARATÁS – DEMÓ ARCHÍVUM - Mozilla Firefox

Éjlj Szerkesztés Nézet Előzmények Könyvjelzők Eszközök Sűgő

http://mekosztaly.oszk.hu/mia/demo/

OSZK WEBARATÁS - DEM... x +

OSZK WEBARATÁS – DEMÓ ARCHÍVUM

KERESÉS A TELJES SZÖVEGBEN...

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

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

A piros nyilakra linkelt mentések a **Heritrix** szoftverrel készültek 2017 decemberével kezdődően, nem teljes mélységben, a videók és más nagy méretű állományok letöltésének kizárásával, és az eredeti webhelyen levő **robots.txt** fájlban levő tiltások tiszteletben tartásával. A megjelenítés az **Open Wayback** szoftverrel történik, melyben a mentés dátumára kell kattintani az archivált példány megtekintéséhez. A mentett változatra és az első mentés időpontjában készült oldalképre mutató nyíl mellett van egy-egy link az amerikai **Internet Archive**-ban található mentésekre, valamint az eredeti "élő" honlapra is. (Ha szürke a nyíl, akkor már nem él az oldal.) A nyilakra kattintva a weblapok új böngészőn nyílnak meg, így könnyen összehasonlíthatók a mentett verziók és az eredeti oldalak. A sárga gomb pedig az adott doménről kifelé mutató, illetve a rá kívülről hivatkozó linkek gráfját rajzolja ki - az archívumban levő mentések alapján (mivel a demó archívum még kicsi, ezért ez utóbbi, *ingoing* típusú linkek száma nagyon kevés). Az utolsó oszlopban levő barna nyilak az archivált webhelyeket leíró metaadat rekordokra visznek (a böngészőben a Ctrl/U megnyomásával látszik az eredeti XML forráskódjuk).

A fejlesztés alatt levő **SolrMIA** keresőnk mellett kísérleti jelleggel a **Solr Wayback Search** nevű felület is kipróbálható, amellyel szintén a mentett webhelyek teljes szövegében lehet keresni és a találatok szűkíthetők doménnevek, fájltypusok és a mentés éve szerint. A találati listában a weboldal vagy fájl (nagy betűvel kiemelt) címére kattintva jutunk el az archivált verzióra, az *Url*: sorban levő cím pedig az eredeti honlapra/fájltra mutat. A *Show full post* felirat alatt megnézhetők az adott találat részletes adatai. Egy találatra kattintva további információk is megjeleníthetők az adott oldalról vagy doménről, ha a bal felső sarokban levő *Toolbar* eszköztárat lenyitjuk.

KÖNYVTÁRI HONLAPOK

Azonosító	A webhely neve	URL címe	OSZK mentés	Oldalkép	Linktérkép	IA mentés	Eredeti	Metaadat
MIA-000034	Berzsényi Dániel Városi Könyvtár, Marcali	www.marcalikonyvtar.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000029	Csuka Zoltán Városi Könyvtár, Érd	www.csukalib.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000106	Evangélikus Országos Könyvtár	konyvtar.lutheran.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000006	Esztergomi Főszékesegyházi Könyvtár	www.bibliotheca.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000007	EX LIBRIS Könyvtár	exlibriskonyvtar.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000032	Fejér György Városi Könyvtár, Keszthely	www.fgyvk.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000030	Helischer József Városi Könyvtár, Esztergom	www.vkesztergom.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000035	Jókai Mór Városi Könyvtár, Pápa	www.jmvk.papa.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000026	Keresztély Gyula Városi Könyvtár, Bátaszék	www.kgyvk.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000129	Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér, Győr	www.gyorikonyvtar.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000125	MaNDA DB	mandadb.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000033	Martonosi Pál Városi Könyvtár, Kiskunhalas	www.vkhalas.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000014	Nagy Gáspár Városi Könyvtár, Budakeszi	ngvk.hu	🔴	🔵	🟡	🟣	🟢	🟠
MIA-000031	Széchenyi István Egyetem Egyetemi Könyvtár, Győr	lib.sze.hu	🔴	🔵	🟡	🟣	🟢	🟠

1. ábra: A nyilvános demó webarchívum részlete

DR. KOVÁCS PÁL MEGYEI KÖNYVTÁR ÉS KÖZÖSSÉGI TÉR, GYŐR HONLAPJA

MIA azonosító: MIA-000129
Eredeti URL: <http://www.gyorikonnyvtar.hu>
Seed URL: <http://www.gyorikonnyvtar.hu/>
Nyilvános OSZK-s archív URL: http://193.6.201.202/owb/wayback/*/
<http://www.gyorikonnyvtar.hu>
Linktérkép URL: <http://193.6.201.202/solrwayback/waybacklinkgraph.jsp?domain=gyorikonnyvtar.hu>
Oldalkép URL: <http://www.gyorikonnyvtar.hu/>
Más nyilvános archív URL: http://web.archive.org/web/*/
<http://www.gyorikonnyvtar.hu/>
Egységesített cím: Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér, Győr honlapja
Eredeti főcím: Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér
Leírás forrása: A Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér honlapja
Forrás URL: <http://www.gyorikonnyvtar.hu>
Leírás dátuma: 2018-07-03
Leíró személy neve: Drótos László

Archiváló szervezet: Országos Széchényi Könyvtár
Archiváló szervezeti egység: E-könyvtári Szolgáltatások Osztálya
Archiváló személy: Visky Ákos László
Archiváló projekt: MIA demó

Létrehozó szervezet: Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér
Fotós: Szabó Béla

Kiadó neve: Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér
Kiadó e-mail címe: info@gyorikonnyvtar.hu
Kiadó honlapja: <http://www.gyorikonnyvtar.hu>

Tartalmi jogtulajdonos: Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér
Jogi státusz: jogvédett

Kapcsolati személy neve: Dr. Horváth Sándor Domonkos
Kapcsolati személy beosztása: igazgató
Kapcsolati személy e-mail címe: hsd@gyorikonnyvtar.hu
Kapcsolati személy telefonszáma: +36-96-516-671

Műfaj: Általános honlap
Tulajdonosi típus: Intézményi
Változékonyság: hetente változó

Fő témakör: Kulturális források
Fő témakör: Oktatási források
Témakör: Köz- és magángyűjtemények
Témakör: Közművelődés
Altémakör: Könyvtárak
Tárgyszó: megyei könyvtár
Tárgyszó: művelődési ház
Tárgyszó: közösségi tér
Földrajzi név: Győr
Földrajzi név azonosító: 3052009

Rövid leírás: A győri Kisfaludy Károly Megyei Könyvtár és a Gálgozci Erzsébet Városi Könyvtár összefonásával 2013-ban létrejött Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér 2017. júliusában megújult központi honlapja.

Kapcsolódó részgyűjtemény azonosítója: MIA_SET-00001
Kapcsolódó archivált webhely MIA azonosítója: MIA-000130
Kapcsolódó élő webhely neve: A Dr. Kovács Pál Könyvtár és Közösségi Tér blogja



2. ábra: Egy archivált honlap „katalóguscédulájának” részlete