Drótos László

Az internet archiválása mint könyvtári feladat*

A nyilvános internetről minden nap tömeges méretekben letörölt vagy máshová költöző dokumentumok és egyéb információforrások egyre nagyobb problémát jelentenek a tudományos publikációkban és a tananyagokban való hivatkozhatóság szempontjából, de az átlagos internetező is állandóan belefut az eltűnt weboldalakat jelző 404-es hibákba. A világháló alapvetően egy jelen idejű médium, de legalább egy részét érdemes lenne megőrizni és kutathatóvá tenni a jövő generációi számára. Ez a cikk arra a kérdésre keresi a választ, hogy ki, mit, hogyan, mivel és miért mentsen az internetről, és hol van itt a könyvtárak és a könyvtárosok feladata és felelőssége? Bemutat néhány hasznos eszközt és szolgáltatást, majd röviden ismerteti a nemzetközi helyzetet és az OSZK-ban 2017 tavaszán elindult kísérleti webarchiválási projektet.

Tárgyszavak: *internet; archiválás, OSZK; honlaptérkép*

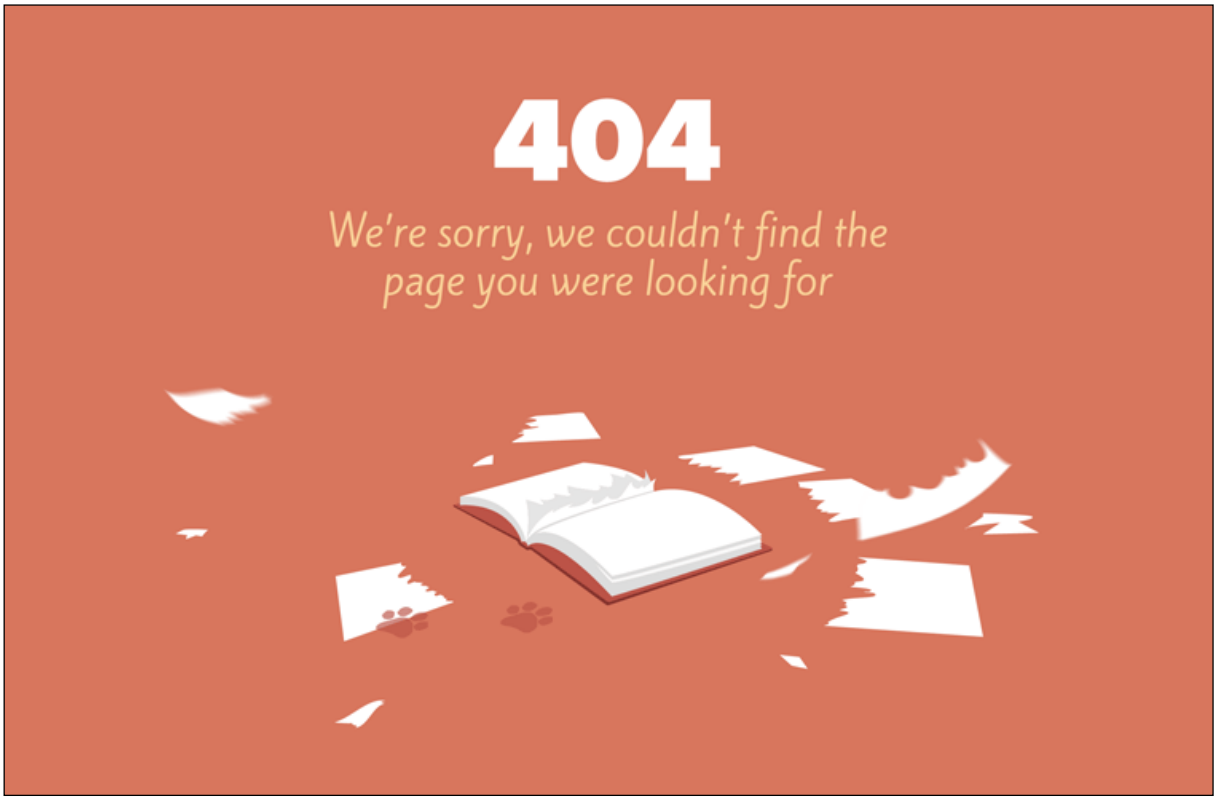
Bevezetés

Képzeljünk el egy könyvtárat, amelyben valakik módszeresen tépdesik ki a könyvek oldalait, vagy esetleg más lapokat tesznek a helyükre. De nemcsak oldalakat tüntetnek el, hanem egész könyveket, könyvespolcokat, sőt olvasótermeket, ahogy például a *Microsoft* tette 2011-ben, amikor bezárta a fénykorában 120 millió regisztrált taggal rendelkező blogszolgáltatását, a *Windows Live Spaces*-t, vagy a *Yahoo*! 2009-ben a több millió honlapból álló *GeoCities*-t, vagy a Google 2016 novemberében a 100 millió, földrajzi helyhez kötött fotót tartalmazó *Panoramio*-t. Nem kellene szólni a könyvtárosoknak, hogy csináljanak már valamit?

Kérdés persze, hogy az internet világkönyvtár-e? Nemcsak egy olyan jelen idejű kommunikációs eszköz, mint a telefon vagy a rádió? (A magnó feltalálása előtt persze.) Ha csupán a tudományos publikációkban egyre nagyobb számban megjelenő URL címekre gondolunk, vagy arra, hogy az oktatásban milyen fontosak az online források, akkor nem nehéz belátni, hogy a világhálónak legalább egy részére könyvtárként kellene tekinteni és vigyázni. De még az olyan, látszólag kérészéletű műfajok, mint a hirdetési és árverési oldalak, a reklámok vagy a *Facebook* posztok is értékes információkat tartalmazhatnak a jövő számára. Elég csak arra utalni, hogy a régi újságok apróhirdetései, a plakátok, a kézírásos naplók mennyi érdekes részletet árulnak el az akkor élt emberek életéről, melyeket a korabeli tudományos és irodalmi művek nem rögzítettek.

De nemcsak a jövő, illetve a tudomány és az oktatás érdekében kell valamit tenni az online források folyamatos erodálódása ellen: az átlagos internetezőnek is mindennapos élménye a 404-es error, a szerverek által küldött „Not Found” hibaüzenet, amikor egy, már nem létező webcímet próbál megnézni. A 404-es hibaoldalt a webmesterek át tudják tervezni és például egy keresőmezőt, vagy egy honlaptérképet, vagy legalább egy, a főoldalra mutató linket rá szoktak tenni, ennyivel segítve a zsákutcába tévedt felhasználót. És hogy az eltűnt forrás okozta frusztrációt is csökkentsék, sokszor valami vicces képet, animációt, vagy akár egy böngészőben játszható játékot is kiraknak ide. A 404-es hiba ábrázolása önálló kortárs művészeti ággá nőtte ki magát *(1. ábra)*, de nem biztos, hogy ez a legjobb módszer a probléma kezelésére. Sokkal elegánsabb és hatékonyabb megoldás lenne, ha a böngészőkben vagy a webszerverekben lenne egy olyan funkció, amely ilyenkor felajánlja a keresett weboldal korábbi állapotait és a felhasználó eldönthetné, hogy melyiket szeretné

* A szerző azonos címmel, „A jövő könyvtára felé...” webinárium-sorozat keretében a II. Rákóczi Ferenc Megyei és Városi Könyvtárban 2017. június 1-jén tartott előadásának szerkesztett és kibővített változata. A PowerPoint prezentáció letölthető a MEK Irattárából: http://mek.oszk.hu/html/irattar/eloadas/2017/internet_archivalas.ppt



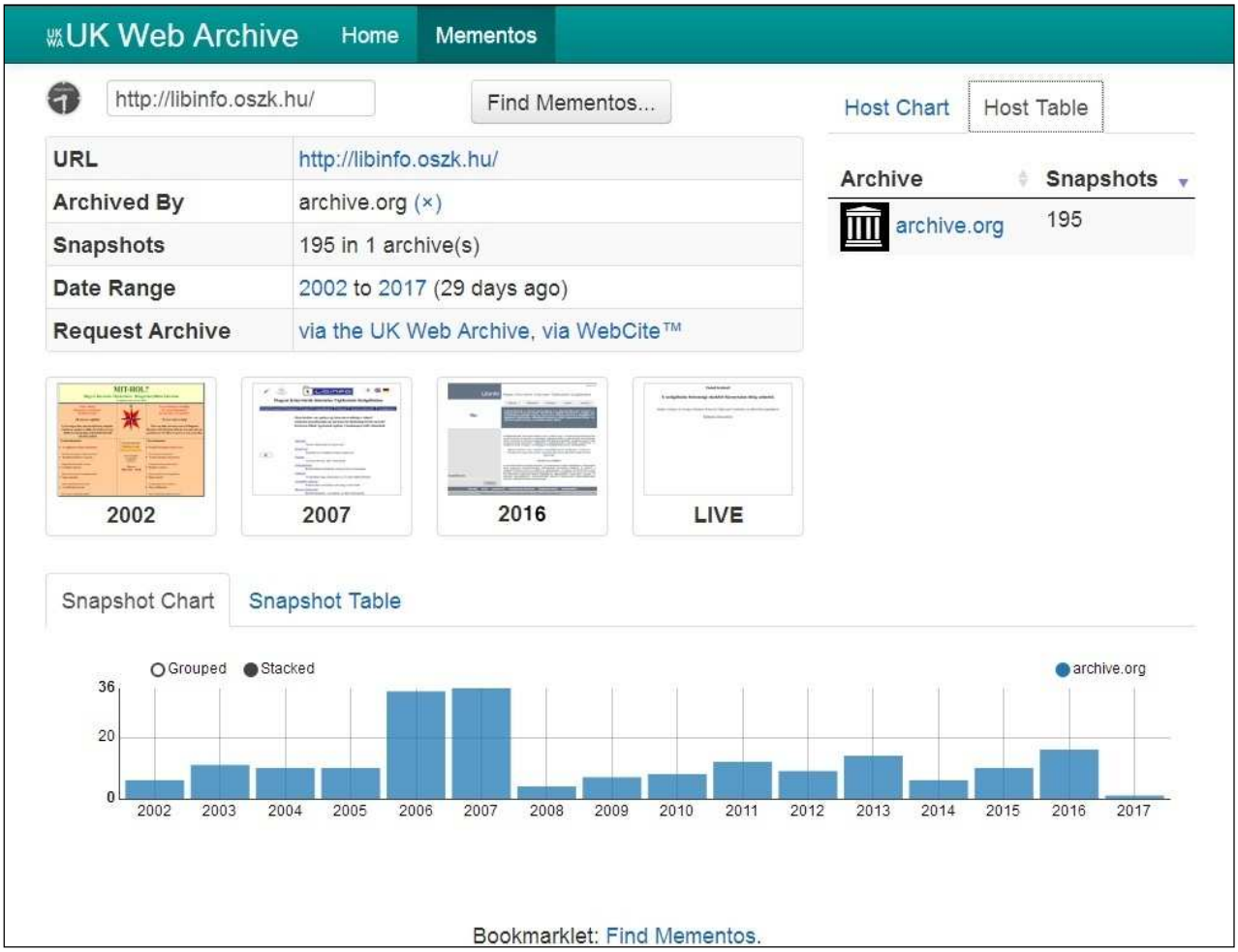
1. ábra **Egy tipikus (eredetileg animált) 404-es hibaoldal**

megnézni. A jó hír, hogy vannak már ilyen megoldások, a kevésbé jó hír pedig az, hogy még sokat kell tenni azért – többek között a könyvtárosoknak is –, hogy ezek megbízhatóan működjenek, és hogy az emberek használják is őket.

Memento

Az egyik legfontosabb fejlesztés a *Memento Project* által javasolt *datetime negotiation* funkció a webszerverek és a kliensek közötti kommunikációt szabályozó HTTP protokollban, melynek köszönhetően a böngészőprogram megadhat egy dátumot is a lekért weboldal címe mellett, és a szerver az ahhoz a dátumhoz legközelebbi mentést, *memento*-t küldi vissza. Ennek a szabványos megoldásnak köszönhetően egyrészt egy webszerver akkor is tud szolgáltatni egy oldalt, ha az már eltűnt az élő honlapról vagy más tartalom került a helyére, de még megvan az eredeti valahol a szerveren, másrészt összekapcsolhatóvá, közösen lekérdezhetővé válhatnak a világ különböző pontjain levő webarchívumok. A projektet a *Los Alamos National Laboratory* és az *Old Dominion University* vezeti, és többek között a *Library of Congress* is támogatja. Az új funkció részletes működését az RFC 7089 jelű dokumentum[1] írja le, a projekt honlapján[2] pedig elérhetők a szükséges kliens- és szerveroldali kiegészítő szoftverek.

Működése legegyszerűbben a *Time Travel* oldalon[3] próbálható ki egy URL cím és egy dátum megadásával, majd a *Find* vagy a *Reconstruct* gomb megnyomásával. Előbbi csak egy találati listát ad, utóbbi pedig magát a weboldalt az adott időpont közelében. A brit webarchívum honlapján[4] is van egy hasonló kereső, de itt a kék hátterű rovatban egy *Find Mementos* nevű Javascript linket is találunk, amit ha lenyomott egérgombbal a könyvjelző eszköztárra húzunk, akkor már be is építettük ezt a funkciót a böngészőnkbe és erre kattintva bármikor meg tudjuk nézni az aktuális weboldal mementóit. (Ilyenkor előbb csak egy összefoglaló táblát és grafikont kapunk. *(2. ábra)* Az egyes mentések a *Snapshot Table* feliratú fülön listázhatók ki és nézhetők meg.) De ennél az egyszerű *bookmarklet*-nél többet tudó modult is adhatunk a böngészőnkhöz, mint például a Memento Time Travel[5] nevű Chrome kiegészítőt, vagy a Firefoxba beépülő Synchronicity-t[6].



2. ábra **A Libinfo honlap mementói az Internet Archive-ban**

A világ legnagyobb webarchívuma, az *Internet Archive* (IA) pedig a webmesterek számára vezetett be 2013-ban egy új szolgáltatást, *404 Handler*[7] néven. A szerveren levő 404-es hibaoldalba – vicces képek helyett – elég csak ezt beírni: <script src="https://archive.org/web/wb404.js"> </script> és ettől kezdve a szerver az IA Wayback Machine nevű szolgáltatásába irányítja a felhasználókat, ahol jó esetben megtalálhatók az eltűnt weblapok és egyéb fájlok.

Mindezek az okos megoldások viszont csak akkor működnek, ha a nyilvános internetes forrásokról vannak valahol – lehetőleg szintén nyilvános – mentések. A kérdés ezek után az, hogy ki, mit, hogyan, mivel és miért mentsen az internetről? És itt jön a képbe a könyvtárak és a könyvtárosok feladata és felelőssége.

Ki?

Magánemberek

Internetet archiválni mindenki megtanulhat, akár a saját gépén futtatott szoftverekkel, akár valamilyen online szolgáltatással. De hogy egyáltalán lehet-e ilyet csinálni és hogy hogyan kell, hogyan érdemes, abban kézenfekvő segítséget tudnának nyújtani a könyvtárosok, személyes tanácsadással, csoportos oktatással, útmutatókkal és tananyagokkal – amennyiben persze ők maguk már rendelkeznek naprakészen tartott ismeretekkel ezen a területen. Személyes archívumokat az illető érdeklődési vagy kutatási területéhez kapcsolódó online tartalmakból szokás kialakítani, de motivációs tényező lehet például a saját, illetve a családtagjai virtuális életének, internetes jelenlétének megőrzése is. Ezek a magángyűjtemények természetesen leginkább egy saját számítógép vagy mobil

eszköz háttértárán, esetleg valamilyen offline tárolón vagy felhőbeli privát tárhelyen vannak, és így ugyanúgy ki lehetnek téve az eltűnés veszélyének, mint az élő web, továbbá mások számára elérhetetlenek. Viszont léteznek már olyan – nagyrészt ingyenes – szolgáltatások, melyekkel a magáncélra mentett oldalak megoszthatók másokkal is, vagy beküldhetők egy nyilvános webarchívumba. (Ezekről a későbbiekben még szó lesz.)

Vállalatok

A nagyobb vállalatok esetében külföldön már bevett gyakorlat a saját online felületeik és kommunikációs csatornáik megőrzése cégtörténeti, illetve vitás esetekben bizonyítási célokból, valamint a versenytársak és az adott piaci szegmens digitális információinak gyűjtése és kielemzése (adatbányászat, trendkutatás). Az archiválandó tartalmak kiválasztásában, a gyűjtemény szervezésében és metaadatolásában, valamint a ráépülő kereső és adatelemző szolgáltatások kialakításában a vállalati könyvtárosnak vagy egy erre szakosodott infobrókernek fontos szerepe lehet.

Intézmények

Intézmények vagy szervezetek esetében is természetes igény ma már, hogy a történetüket ne csak hagyományos irat- és levéltárral, hanem egy digitális archívummal is dokumentálják, amelynek része kell(ene) hogy legyen a saját honlapjaik és egyéb internetes felületeik időnkénti vagy rendszeres mentése is. Erre már jogszabály is kötelezi őket egyes országokban, sőt például kormányzati szervek esetében a webarchívumot is nyilvánosan kell szolgáltatni az élő honlaphoz hasonlóan, hogy a korábbi – esetleg már érvényüket vesztett dokumentumok – továbbra is elérhetők maradjanak az állampolgárok számára. A saját tartalmak mentése mellett az intézmény szakterületével, feladatával kapcsolatos információforrások megőrzése és kutathatóvá tétele érdekében is sok helyen épülnek már adott témára specializált webarchívumok, általában a helyi könyvtár és/vagy levéltár részvételével.

Közgyűjtemények

A nemzeti, köz- és szakkönyvtárak, levéltárak, múzeumok, audiovizuális archívumok törvényszabta kötelessége a gyűjtőkörükbe tartozó dokumentumok megőrzése és szolgáltatása. Hogy ez a törvény kitér-e a digitálisan születő kultúrára is, és ha igen, akkor annak mekkora körére, illetve milyen előírásokat, jogosítványokat határoz meg a memóriaintézmények számára, az országonként eléggé változó. Mivel a jogalkotás lassan követi a technikai fejlődést, ezért a legtöbb helyen előbb elkezdődött az internetes források gyűjtése ezekben az intézményekben, és egy már létező gyakorlatot szabályoztak utólag, például a kötelespéldány- vagy a levéltári törvény módosításával.

Mit?

Legkönnyebben a web őrizhető meg, annak is a hagyományos formája, amely viszonylag jól begyűjthető és elraktározható automatikus vagy félautomatikus módszerekkel. De ez a „felszíni web" csak kis töredéke a teljes webnek. A „mély web" és a „sötét web" elérhetetlen a keresőgépek és az archiváló rendszerek robotjai számára is. Míg az utóbbit alkotó site-ok esetében érthető a titkolózás, az előbbinél sokszor csak arról van szó, hogy úgy lett kialakítva a webhely, hogy nem járható be véges számú linket követve, illetve az eredeti szoftverkörnyezet nélkül a másolat használhatatlan. Az archiválással foglalkozó intézmények egyik feladata, hogy felhívják az ilyen webhelyek gazdáinak a figyelmét arra, hogy – az akadálymentesítéshez hasonlóan – alakítsák át a szolgáltatásukat a hosszú távú megőrizhetőség érdekében, vagy generáljanak egy *crawler friendly* és *archive friendly* verziót is belőle, és irányítsák oda az archiváló szoftvert a robots.txt[8] fájlban megadott előírásokkal.

A dinamikusan változó oldalakból álló webkettes felületek (pl. Facebook, Twitter, Instagram, Flickr, Tumblr) nemigen beszélhetők rá ilyen változtatásokra, de ezekhez rendszerint van valamilyen API, amin keresztül – a megfelelő jogosultságok megléte esetén – legalább maga a tartalom letölthető. Ha az oldalak elrendezését és külalakját, valamint a bennük levő linkek működőképességét is szeretnénk megőrizni, az csak emberi közreműködéssel vagy emberi viselkedést szimuláló szoftverekkel oldható meg. Utóbbi esetben egy ún. *headless browser*-t használnak, amely mindent tud, amit a ma használatos böngészők, ugyanúgy tudja értelmezni a weboldalak HTML kódját és végrehajtani a bennük levő JavaScript/AJAX parancsokat, de nincsen grafikus felülete, hanem parancssorból vagy scriptekkel vezérelhető.

Szintén nagy kihívás a hang- és videotartalmak, főként a sugárzott média mentése és kereshetővé, szolgáltathatóvá tétele – már csak a hatalmas tárhelyigény miatt is (pl. YouTube, Vimeo, Twitch, Ustream, Facebook Live, SoundCloud, internetes

rádiók és televíziók). Ezekkel a műfajokkal a webarchívumok gyakran nem is foglalkoznak, hanem ezt a feladatot az adott ország audiovizuális archívuma látja el a hagyományos rádió- és tévéműsorok, illetve filmek gyűjtése mellett.

A rengeteg mobil és asztali alkalmazás (pl. WhatsApp, Skype), amelyek gyakran saját „szabvány" szerint kommunikálnak a központi szerverrel, valamint a szerver nélküli (peer-to-peer) rendszerek tartalmának megőrzése pedig megint egy másfajta technikai problémát jelent, de ezek már kívül esnek a közgyűjtemények érdeklődési körén.

Hogyan?

Alapvetően kétféle megközelítés jöhet szóba. Az első esetben valamilyen szempontrendszer, gyűjtőkör alapján emberi közreműködéssel vagy automatikus/félautomatikus módon határolják le az internetes forrásoknak azt a halmazát, amelyet egyszeri alkalommal vagy időről-időre begyűjtenek, learatnak. A válogatási szempont lehet intézménytípus (pl. kutatóintézetek, kormányhivatalok), műfaj (pl. blogok, e-folyóiratok), téma (pl. emberi jogok, helyismeret és helytörténet), esemény (pl. választások, olimpia), híres ember (pl. a halála vagy valamilyen évfordulója esetén). Az archivált források lehetnek teljes webhelyek, vagy azok részei, esetleg csak egyes weboldalak, vagy azokról letölthető egyedi dokumentumok. A másik esetben nincsenek ilyen speciális szelekciós szempontok, a gyűjtés körét csak adott aldoménra (pl. .gov.uk), vagy doménra (pl. .at), vagy a nemzeti webtérre (pl. a finn nyelvű vagy finn közönségnek szánt tartalom), vagy a globális webtérre (pl. a robotok számára is hozzáférhető nyilvános webtartalom) korlátozzák és azon belül igyekeznek legalább egy reprezentatívnak tekinthető méretű és kiterjedésű aratást végezni évente néhány alkalommal.

A tárolási mód szerint négy típusba sorolhatók az archívumok és az archiváló szoftverek:

- Fájlrendszerbe mentés: a webhelyet alkotó fájlok egyenkénti tárolása, az eredeti fájlnevek és alkönyvtárak megőrzésével vagy átnevezésével, és többnyire a linkek relatívvá, lokálissá tételével, hogy az archív példány is navigálható maradjon.
- Archív állományba mentés: a weboldalakat alkotó objektumoknak és azok technikai metaadatainak szabványos szerkezetű csomagokba mentése. Ezek a „konténerek" lehetnek például az Internet Archive által is használt ARC vagy WARC állományok, vagy az egyes böngészők által is támogatott MAFF (Mozilla Archive Format), illetve MHTML (MIME HTML) formátumú fájlok.
- Egységes formátumba mentés: a weboldalak tartalmának és/vagy kinézetének megőrzése azok eredeti szerkezetének megtartása nélkül, például egységesen XML formátumra konvertálva, vagy PDF/A fájlba „nyomtatva" őket, vagy PNG képeket készítve róluk.
- Adatbázisba mentés: elsősorban nem webhelyek, hanem például elektronikus levelek, tweetek, blogbejegyzések, Facebook posztok, hírportálokról letöltött cikkek és képek stb. adatbázisrekordokként való tárolása.

Gyakoriság és időmenzió szerint háromféle módszerről beszélhetünk:

- Ismétlődő mentések hosszú távú megőrzéssel: egy-egy website rendszeres mentése, lehetőleg a webhely változékonyságához optimalizálva, adott időpontbeli állapot rekonstruálásának lehetőségével, hosszú távra tervezve.
- Ismétlődő mentések az utolsó állapotot megőrizve: egy-egy website rendszeres mentése, de a korábbi változatok megőrzése nélkül, például a keresőrendszerek számára szükséges indexeléshez, vagy egy piaci szegmens aktuális állapotát kutató adatbányászathoz.
- Egyedi vagy alkalmi mentések: egy-egy webhely, vagy weblap, vagy webkettes tartalom, vagy dokumentum egyszeri vagy alkalomszerű mentése (pl. hogy egy publikációban stabil URI-val lehessen rá hivatkozni, vagy hogy bizonyítékként felhasználható legyen egy jogi eljárásban).

Az aktív gyűjtés, az aratás vagy letöltés mellett meg kell említeni, hogy ún. *push* technikával is szoktak internetes archívumokat építeni például cégek vagy nemzeti könyvtárak. Ilyenkor a tartalmat vagy az eredeti szolgáltató szerver küldi be az archívumba, valamilyen szabványos adatcsere-protokollon keresztül, vagy egy proxy szerver küld be egy másolatot az archívumba minden rajta áthaladó, a felhasználók kliensei által lekért digitális objektumról. Könyvtárak és levéltárak esetében az is előfordulhat, hogy önkéntes depozitként (pl. digitális hagyatékként) kapnak webhelyeket vagy egyéb internetes tartalmakat egy letölthető/feltöltött csomagban, vagy valamilyen offline hordozón.

A „Hogyan?" kérdésre még egyéb szempontokból is választ lehet és kell adni. Egyrészt *szabályozottan* lenne jó az internetes források archiválását

végezni, különösen a közgyűjteményekben. Ez azt jelenti, hogy a kötelespéldány- és a szerzői jogi törvényben, továbbá a könyvtárakat, levéltárakat és más archiváló intézményeket érintő egyes további jogszabályokban, valamint ezen intézmények belső szabályzataiban foglalkozni kell ezzel a területtel, kitérve a személyi és üzleti adatok védelmére, s a copyright és a szabad felhasználás kérdéseire is a begyűjtés, a megőrzés és a hozzáférés esetében egyaránt.

A másik fontos szempont, hogy *szervezetten* kellene ezt a tevékenységet folytatni, mert a feladat – mind a megőrzendő tartalom mennyiségét, mind pedig változatosságát tekintve – olyan hatalmas, hogy ezzel egyetlen könyvtár, de még egy egész könyvtári hálózat sem lesz képes megbirkózni. Munkamegosztásra van szükség, és nemcsak az egyes közgyűjtemények között, de együtt kell működni a tartalom- és internetszolgáltatókkal, egyes informatikai cégekkel is, sőt az internethasználók széles köre is bevonható mondjuk az archiválandó webhelyek összeválogatásába (*crowdsourcing*). Az olyan szabványos megoldások használata pedig, mint a már említett WARC archív formátum vagy a Memento protokoll, lehetővé teszik az internetarchívumok összekapcsolását országon belül és országok között is, így nagyobb az esély arra, hogy valamelyikben megtalálható a keresett digitális objektum.

És végül: egy archívumnak akkor van értelme, ha hosszú ideig létezik, ezért csak *fenntarthatóan* érdemes csinálni. Maga a technikai fenntarthatóság, a gyorsan növekvő tárhelyigény, a rendkívül sokféle és részben szintén gyorsan avuló fájlformátum megjeleníthetőségének megoldása migrálással vagy a régi szoftverek emulációjával, önmagában is hatalmas kihívás. De még nagyobb probléma a finanszírozhatóság, mivel erre még nincsenek kialakult mechanizmusok a legtöbb országban. Mindenképpen többféle forrásból, például állami és EU-s költségvetésből, tudományos kutatási alapokból és alapítványi támogatásokból, pályázatokkal és szponzorálással, illetve az archivált tartalomra ráépített fizetős szolgáltatások bevételeivel lehet biztosítani azt, hogy mind a hardveres és szoftveres infrastruktúra, mind pedig a hozzáértő szakembergárda hosszú évtizedekig rendelkezésre álljon.

Mivel?

Bár az internetezéshez használt szoftverek (pl. böngészők, levelezők, csevegőprogramok) is rendelkeznek saját mentési, exportálási vagy naplózási funkcióval, amelyekkel lementhetők illetve archiválhatók egyes fájlok, weboldalak, levélmappák vagy beszélgetések, de egy sor, ezeknél sokkal többet tudó kiegészítő modul, önálló szoftver, komplett rendszer, illetve online szolgáltatás, felhőalapú megoldás közül választhatunk, ha személyes, intézményi, vagy nemzeti archívumot szeretnénk létrehozni – ráadásul sok közülük ingyenes. Ebben a fejezetben néhány webarchiválásra alkalmas eszközre szeretném felhívni a figyelmet.

ScrapBook[9]

Japán programozók által 2004 óta fejlesztett ingyenes Firefox plug-in modul weboldalak, webhelyek letöltésére és a mentések menedzselésére, melyek teljes szöveggel kereshetők, sőt még szerkeszteni is lehet a mentett oldalakat (pl. törölni vagy átrendezni egyes oldalelemeket, színekkel kiemelni szövegrészeket és jegyzeteket vagy linkeket fűzni hozzájuk). Az egyes mentések össze is fűzhetők, egy saját „webhelyet" alakítva így ki belőlük. A projektet 2016-ban egy tajvani programozó vette át és átnevezte ScrapBook X-re. Ez már tud – további kiegészítők telepítése után – MAFF fájlba is menteni, és konvertálni is oda-vissza MAFF, EPUB, ZIP és egyéb tárolási formátumok között. Sok nyelvre lefordították, magyarítás is van hozzá.

Webrecorder[10]

Az amerikai Rhizome nonprofit szervezet által 2016 óta fejlesztett ingyenes szolgáltatás webböngészések videomagnószerű rögzítésére. De nem videofájlokba ment, hanem WARC csomagokba, vagyis a weboldalakat alkotó fájlokat tárolja el, így a mentett példány ugyanúgy navigálható marad, mint az eredeti. Mivel csak a felhasználó által – a *Recording* gomb bekapcsolása és kikapcsolása közt – megnézett oldalakat menti, ezért nem egy teljes webhely archiválására, hanem annak valamilyen szempontból releváns részeinek mentésére alkalmas. Ideális megoldás olyan, regisztrációhoz kötött, interaktív, dinamikus, médiagazdag oldalakhoz (pl. Facebook), amelyekkel a hagyományos *crawler*-ek és egyéb letöltők nem boldogulnak. A „felvétel" először a *webrecorder.io* szerver 5 gigabájtos ingyen tárhelyére kerül, ahonnan meg tudjuk osztani másokkal is, de WARC formátumban le is tölthetjük a saját gépünkre, ahol megnézhetjük a Webrecorder Player[11] programmal, ami egy lokális webszervert indít el és azon keresztül szolgáltatja nekünk az archív példányt.

HTTrack[12]

Ingyenes, nagy teljesítményű webhelyletöltő szoftver Linux, Mac OS X és Windows rendszerekre, utóbbihoz magyar felület is van. A letöltés előtt – és részben közben is – rengeteg paraméter beállítható. Egy .txt fájlban több kiinduló URL-t is megadhatunk és a letöltések elindítását scriptekkel időzíthetjük is. Fájlrendszerbe ment, az eredeti alkönyvtárstruktúra megőrzésével. Hogy mennyire professzionális eszköz, azt jól jelzi, hogy a *National Library of Australia* által vezetett konzorcium keretében 1996 óta épülő webarchívum, a PANDORA[13] is ezt használja letöltő szoftverként. 2016 októberéig már több mint 48 ezer teljes vagy részleges webhelyet, illetve egyedi dokumentumot mentettek le vele legalább egyszer, 25.7 terabájt összméretben.

A továbbiakban bemutatott megoldások elsősorban a *link rot*[14] elleni küzdelemhez használhatók, ami az internetes információforrásokra, dokumentumokra mutató URL hivatkozások, linkek és könyvjelzők tönkremenésének jelensége, mivel idővel törlődnek, máshová kerülnek, vagy megváltoznak a mögöttük levő tartalmak. Ez a folyamat nemcsak a tudományos publikációk és az oktatási anyagok esetében jelent komoly problémát, hanem például a keresőrendszerek találati listáinál és a személyes könyvjelző-gyűjteményeknél is frusztrációt okoz. A jelenség sebességére különböző mérési adatok vannak, attól függően, hogy mikor és milyen jellegű linkeket vizsgáltak: az éves linkromlásra 5-20% közötti értékeket kaptak, a felezési időt pedig 5-10 év között becsülik. Sokféle módszerrel lehet csökkenteni a problémát, például stabil azonosítókkal (URN, DOI, Handle stb.), a webszerveren beállított átirányításokkal, az eltűnt lapokat megkereső szoftverekkel, de az igazi megoldást az igény szerint archiváló szolgáltatások jelentik.

archive.is[15]

Ingyenes weboldal-archiváló szolgáltatás, amely egy *bookmarklet* segítségével böngészőbe is beépíthető. A felhasználó kezdeményezésére lementett weblapok stabil URL-eken hivatkozhatók és kereső is van hozzájuk. A mentett oldalakról 1024×768-as méretű képernyőfotó is készül. Az archív példány címe megosztható, sőt akár egy wikibe is bemásolható. A maximális méréthatár 50 megabájt oldalanként (képekkel együtt), megőrzési határidőkorlát nincs. A szöveges tartalomról három másolatot tárol, de a képanyag is duplikálva van különböző európai adatközpontokban. A Memento Project tagja.

Perma.cc[16]

Sok – főként amerikai – könyvtár által támogatott *link rot* elleni szolgáltatás, melyet a *Harvard Law School Library* egyik munkacsoportja fejlesztett ki. A rendszer a felhasználó által megadott URL címen levő weboldalt vagy egyéb dokumentumot lementi (és egy PNG képernyőfotót is készít róla), majd egy stabil azonosítót ad neki, amellyel hosszú távon is hivatkozható marad. Ha a mentés nem sikerülne valamiért, maga a felhasználó is feltölthet egy képet vagy egy PDF fájlt az adott dokumentumról. A rendszer elosztottan működik a könyvtárak szerverein, így a fennmaradására nagyobb az esély, mint a hasonló, de egyetlen céghez kötődő szolgáltatásokéra. 2017. április 25-én 450 ezer mentett dokumentumhoz tartozott ilyen *perma link* és 887 intézmény (ebből 213 könyvtár), illetve 14 587 felhasználó vette igénybe a szolgáltatást. A használat regisztrációhoz kötött és havi 10 mentésig ingyenes mindenkinek, de könyvtárhasználók, folyóiratok szerkesztői, egyetemi oktatók, bíróságok és más szervezetek tagjai korlátlan hozzáférést kaphatnak. Fejlesztők számára API-t is biztosítanak a rendszerhez.

WebCite[17]

Elsősorban szerzőknek, szerkesztőknek stb. szánt ingyenes *on-demand* archiváló szolgáltatás (de intézményi partnerprogramjuk is van, pl. könyvtáraknak), amely archiválja és stabil URI-val látja el a felhasználó által javasolt publikációkat és egyéb online forrásokat, így biztosítva, hogy az ezekre való hivatkozások hosszú távon is működőképesek maradjanak. Böngészőbe beépíthető könyvjelző-alkalmazás is van hozzá.

Komolyabb céges, intézményi vagy közgyűjteményi webarchívumhoz komplett rendszert vagy felhőalapú SaaS (Software-as-a-Service) szolgáltatást is kínál ma már néhány külföldi vállalkozás. Ezek közül itt most csak egyet emelek ki:

Archive-It[18]

Az Internet Archive 2006-ban indított előfizetéses archiváló szolgáltatása könyvtáraknak és más intézményeknek. Az Egyesült Államokon kívül további 16 országból több mint 400 megrendelője van. Az archiválandó webhelyek körét a megrendelő határozza meg és kap egy adminisztrátori, valamint egy szolgáltatási felületet az IA szerverein tárolt lementett anyaghoz.

A webarchívumok között külön kategóriát jelentenek a nemzeti szintűek, melyeknél nagy méretük és hosszú távú céljaik miatt különösen fontos a költséghatékony és szabványos megoldások használata, valamint az, hogy az archivált tartalom ne egy külföldi szerveren legyen. Egyre több nemzeti könyvtár használja az *International Internet Preservation Consortium*[19] által is támogatott open source szoftvereket, mint amilyen a Heritrix[20] aratószoftver (crawler), az OpenWayback[21] megjelenítő, a NutchWAX[22] kereső és a Web Curator Tool[23] nevű adminisztrációs, ütemező és metaadatoló keretrendszer.

Miért?

- Hogy legyen múltja is az internetnek, ne csak jelene;
- hogy kutathassuk a virtuális világ történetét, valamint a valódi világ elmúlt eseményeinek internetes lenyomatait;
- hogy elemezni és ábrázolni lehessen nagy mennyiségű digitális tartalmakat;
- hogy megbízhatóan tudjunk hivatkozni tudományos publikációkban és tananyagokban online forrásokra;
- hogy helyreállíthatók legyenek elveszett webhelyek;
- hogy vitás esetekben bizonyítható legyen, hogy mi jelent meg egy weboldalon;
- hogy a 404-es hibákra más megoldás is legyen, ne csak a vicces képek.

A fenti érvek közül a másodikra és a harmadikra szeretném külön is felhívni a figyelmet. A „webhistoriográfia", vagyis a webarchívumok történettudományi célú felhasználása lassan önálló segédtudománnyá növi ki magát (lásd pl.: Web Archives for Historians[24], The Web as History[25]). De emellett a nyelvészettől és a politológiától kezdve, a művészettörténeten és a gasztronómián át, a média- és családfakutatásig mindenféle szakterület előtt egészen új lehetőségek nyílnak az internet-archívumokban halmozódó sok milliárdnyi fájlban található információk *big data* módszerekkel való elemzése és vizualizálása révén. Utóbbira néhány érdekes példa:

What Did It Look Like?[26]

A Memento Project keretrendszerére épülő szolgáltatás, amely véletlenszerűen választott, illetve a felhasználók által javasolt weboldalak kinézetének változását mutatja meg képernyőfotókból álló *slideshow*-k formájában. A képeket az archívumokból összeszedett mementókról a PhantomJS nevű *headless browser* készíti, majd az ImageMagick szoftver gyártja le az animált GIF-eket. A korábbi válogatások is visszanézhetők.

A tajvani nemzeti webarchívum idővonala[27]

A *National Taiwan University Library* webarchiváló projektje 2006-ban indult és 2008 áprilisától érhető el a könyvtár honlapján. A NTUWAS nevű rendszerben HTTrack-kel mentenek szelektíven webhelyeket illetve weblapokat. 2017 tavaszán már közel 9600 site volt visszakereshető a nyilvános felületen, ami saját fejlesztés, látványos megoldásokkal (pl. időskálára és térképre vetítések).

Trendelemzés a brit webarchívumban[28]

A SHINE egy, a brit UKWA webarchívum által a Big UK Data Arts and Humanities projekt számára fejlesztett teljes szövegű kereső (facettás találati listával), de egyben egy prototípusként létrehozott szolgáltatás is. Utóbbihoz az Internet Archive-tól kapott, az .uk domén aratásával 1996 és 2013 között gyűjtött WARC fájlokat indexelték le, melyek mintegy 3.5 milliárd objektumot tartalmaznak. Az egyszerű és összetett keresőűrlap mellett van egy *Trends* nevű aloldal is, ahol a keresett szó vagy szavak előfordulásának időbeli változását nézhetjük meg egy grafikonon. A trendvonal valamely pontjára kattintva max. 100 véletlenszerű weblapot is kilistáz, ahol az adott időpontban előfordult a keresett szó, s ezek archivált verzióit is megtekinthetjük.

Kanadai pártok webhelyeinek mérete[29]

2005−2015 közötti webaratások eredményeiből készített grafikon, mely a kanadai politikai pártok és érdekcsoportok webhelyeinek méretét (a weboldalak számát) mutatja egy időskála mentén. A legalsó oszlop mindig a legnagyobb site-ot jelzi, felette a második legnagyobb következik és így tovább. (Az első 20 szervezet után minden további az „egyebek" kategóriába lett összevonva.)

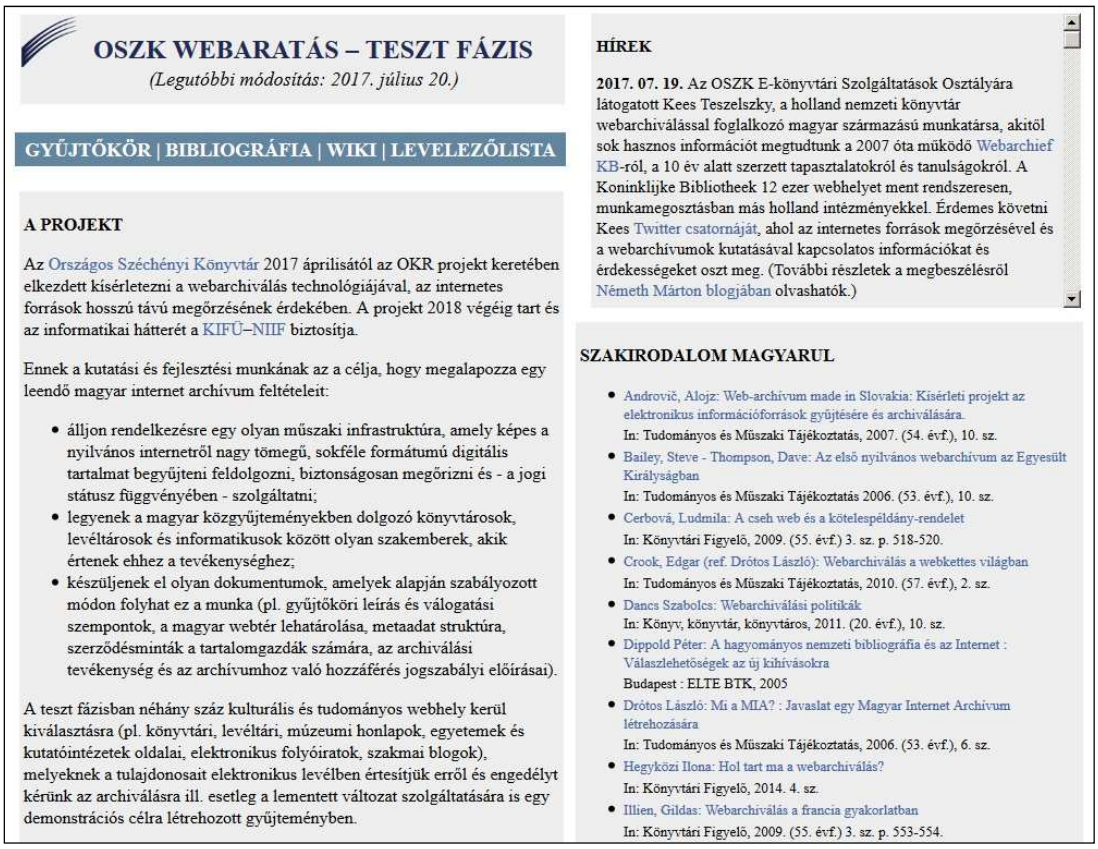
Hol tartunk?

Az Internet Archive, ez az 1996-ban San Franciscoban alapított nonprofit szervezet, a szöveg-, kép-, hang-, videó- és szoftvergyűjteménye mellett a globális webet is archiválja. 2017 júniusában már 284 milliárd weboldalt lehetett visszanézni a Wayback Machine[30] segítségével.

A kilencvenes évek második felétől kezdve kb. 40 nemzeti szintűnek tekinthető webarchívum indult el harmincegynéhány országban. Az Egyesült Államok és a nagy nyugat-európai országok mellett van már például portugál, baszk, katalán, holland, osztrák, cseh, horvát, szlovén, ukrán, észt, lett, izlandi, finn, svéd, dán, kínai, japán, tajvani, szingapúri projekt is a nemzeti web megőrzésére. Több ilyen rendszer már a második generációnál tart: néhány éves működés után újragondolták és az időközben kialakult szabványos megoldásokra építették át őket.

És létezik vagy létezett sok kisebb-nagyobb internetarchiválási kezdeményezés külföldi könyvtárakban, levéltárakban, állami hivataloknál, tudományos intézetekben, egyetemeken, vállalatoknál, ahol szelektíven mentenek/mentettek le számukra fontos webhelyeket és egyéb online forrásokat hosszú távú megőrzési vagy rövidebb távú kutatási célból.

Magyarországon még nincs komolyabb webarchívum. A 2010-es évek első felében az *ELTE Tudománytörténet és Tudományfilozófia Tanszék*én volt egy webaratási kísérlet: kb. 400 tudományos és oktatási intézet honlapját, valamint hírportálok anyagát mentették. A NAVA pedig néhány éve az MTVA számára gyűjt online sajtóhíreket. Az *Országos Széchényi Könyvtár*ban az internetről (is) válogatott egyedi dokumentumok, kiadványok mentése és metaadatolása történik a MEK (könyvek – 1994 óta), az EPA (periodikák – 2004 óta) és a DKA (képek – 2007 óta) keretében. Bár már 2006-ban felmerült a webhelyek archiválásának a terve is, ehhez hosszú ideig nem sikerült forrást találni. 2017 márciusától viszont az *Országos Könyvtári Rendszer* fejlesztése keretében végre elindulhatott egy kísérleti fázisú webaratási projekt 2018 végéig, azzal a céllal, hogy megalapozza egy leendő, üzemszerűen működő magyar internetarchívum feltételeit. Egyelőre a technológia tesztelése, a külföldi jó példák megismerése, a szükséges elméleti és gyakorlati ismeretek megszerzése folyik. A projekt weboldalán[31] lehet tájékozódni a tervekről és az eddig elért eredményekről *(3. ábra)*. Van itt egy wiki[32] is, amely az internetes források megőrzésével kapcsolatos fogalmakat, projekteket, szolgáltatásokat, szoftvereket, formátumokat, rendezvényeket, szervezeteket stb. ismerteti



OSZK WEBARATÁS – TESZT FÁZIS

(Legutóbbi módosítás: 2017. július 20.)

GYŰJTŐKÖR | BIBLIOGRÁFIA | WIKI | LEVELEZŐLISTA

A PROJEKT

Az Országos Széchényi Könyvtár 2017 áprilisától az OKR projekt keretében elkezdett kísérletezni a webarchiválás technológiájával, az internetes források hosszú távú megőrzésének érdekében. A projekt 2018 végéig tart és az informatikai hátterét a KIFÜ–NIIF biztosítja.

Ennek a kutatási és fejlesztési munkának az a célja, hogy megalapozza egy leendő magyar internet archívum feltételeit:

- álljon rendelkezésre egy olyan műszaki infrastruktúra, amely képes a nyilvános internetről nagy tömegű, sokféle formátumú digitális tartalmat begyűjteni feldolgozni, biztonságosan megőrizni és - a jogi státusz függvényében - szolgáltatni;
- legyenek a magyar közgyűjteményekben dolgozó könyvtárosok, levéltárosok és informatikusok között olyan szakemberek, akik értenek ehhez a tevékenységhez;
- készüljenek el olyan dokumentumok, amelyek alapján szabályozott módon folyhat ez a munka (pl. gyűjtőköri leírás és válogatási szempontok, a magyar webtér lehatárolása, metaadat struktúra, szerződésminták a tartalomgazdák számára, az archiválási tevékenység és az archívumhoz való hozzáférés jogszabályi előírásai).

A teszt fázisban néhány száz kulturális és tudományos webhely kerül kiválasztásra (pl. könyvtári, levéltári, múzeumi honlapok, egyetemek és kutatóintézetek oldalai, elektronikus folyóiratok, szakmai blogok), melyeknek a tulajdonosait elektronikus levélben értesítjük erről és engedélyt kérünk az archiválásra ill. esetleg a lementett változat szolgáltatására is egy demonstrációs célra létrehozott gyűjteményben.

HÍREK

2017. 07. 19. Az OSZK E-könyvtári Szolgáltatások Osztályára látogatott Kees Teszelszky, a holland nemzeti könyvtár webarchiválással foglalkozó magyar származású munkatársa, akitől sok hasznos információt megtudtunk a 2007 óta működő Webarchief KB-ről, a 10 év alatt szerzett tapasztalatokról és tanulságokról. A Koninklijke Bibliotheek 12 ezer webhelyet ment rendszeresen, munkamegosztásban más holland intézményekkel. Érdemes követni Kees Twitter csatornáját, ahol az internetes források megőrzésével és a webarchívumok kutatásával kapcsolatos információkat és érdekességeket oszt meg. (További részletek a megbeszélésről Németh Márton blogjában olvashatók.)

SZAKIRODALOM MAGYARUL

- Androvič, Alojz: Web-archívum made in Slovakia: Kísérleti projekt az elektronikus információforrások gyűjtésére és archiválására.
  In: Tudományos és Műszaki Tájékoztatás, 2007. (54. évf.), 10. sz.
- Bailey, Steve - Thompson, Dave: Az első nyilvános webarchívum az Egyesült Királyságban
  In: Tudományos és Műszaki Tájékoztatás 2006. (53. évf.), 10. sz.
- Cerbová, Ludmila: A cseh web és a kötelespéldány-rendelet
  In: Könyvtári Figyelő, 2009. (55. évf.) 3. sz. p. 518-520.
- Crook, Edgar (ref. Drótos László): Webarchiválás a webkettes világban
  In: Tudományos és Műszaki Tájékoztatás, 2010. (57. évf.), 2. sz.
- Dancs Szabolcs: Webarchiválási politikák
  In: Könyv, könyvtár, könyvtáros, 2011. (20. évf.), 10. sz.
- Dippold Péter: A hagyományos nemzeti bibliográfia és az Internet : Válaszlehetőségek az új kihívásokra
  Budapest : ELTE BTK, 2005
- Drótos László: Mi a MIA? : Javaslat egy Magyar Internet Archívum létrehozására
  In: Tudományos és Műszaki Tájékoztatás, 2006. (53. évf.), 6. sz.
- Hegyközi Ilona: Hol tart ma a webarchiválás?
  In: Könyvtári Figyelő, 2014. 4. sz.
- Illien, Gildas: Webarchiválás a francia gyakorlatban
  In: Könyvtári Figyelő, 2009. (55. évf.) 3. sz. p. 553-554.

3. ábra **Az OSZK-s webaratási pilot projekt ideiglenes weboldala**

rövid szócikkek formájában, valamint egy válogatott bibliográfia[33] a téma idegen nyelvű szakirodalmából. Elindult továbbá egy levelezőcsoport MIA-l[34] néven, melyre várjuk a téma iránt érdeklődő kollégák jelentkezését.

Hová kellene eljutni?

Legyen egy közgyűjtemények, intézmények és cégek közötti munkamegosztással működő, nagy teljesítményű, fenntartható nemzeti internetarchívum, amely képes:

- rendszeresen menteni sok ezer fontos magyar webhelyet;
- alkalomszerűen menteni kiemelt eseményekhez kapcsolódó hírforrásokat;
- évente kétszer egy reprezentatívnak tekinthető mentést csinálni a magyar webtérről;
- kötelespéldányként és önkéntesen beadott webes és más internetes tartalmakat befogadni;
- mindezeket hosszú távon megőrizni és megtekinthető állapotban tartani;
- szolgáltatásokat nyújtani az internetezők, a tartalomgazdák, a tudományos, oktatási, kormányzati és üzleti szféra számára.

Mindezek elérésének előfeltétele, hogy:

- legyenek nálunk is a webhelyek és az egyéb online források megőrzéséhez értő könyvtárosok, informatikusok és egyéb szakemberek, akik képesek akár magánszemélyek, akár más intézmények, akár a saját könyvtáruk, levéltáruk vagy múzeumuk számára kisebb-nagyobb archívumokat létrehozni;
- legyen egy olyan jogi környezet, amely a magyar közgyűjtemények számára is lehetővé teszi, hogy a nyilvános internetről archiváljanak tartalmakat, valamint azokat – a szerzői és a személyiségi jogi korlátozások figyelembevételével – nyilvánosan, vagy helyben, vagy egy zárt hálózaton szolgáltassák.

A végső célt pedig így lehetne röviden megfogalmazni: *Inkább a 404-es hibák tűnjenek el, ne a weblapok.*

Magyar nyelvű ajánlott irodalom

ANDROVIČ, Alojz: Web-archívum made in Slovakia: Kísérleti projekt az elektronikus információforrások gyűjtésére és archiválására.
In: Tudományos és Műszaki Tájékoztatás, 2007. (54. évf.), 10. sz.

BAILEY, Steve – THOMPSON, Dave: Az első nyilvános webarchívum az Egyesült Királyságban
In: Tudományos és Műszaki Tájékoztatás 2006. (53. évf.), 10. sz.

CERBOVÁ, Ludmila: A cseh web és a kötelespéldány-rendelet
In: Könyvtári Figyelő, 2009. (55. évf.) 3. sz. p. 518-520.

CROOK, Edgar (ref. Drótos László): Webarchiválás a webkettes világban
In: Tudományos és Műszaki Tájékoztatás, 2010. (57. évf.), 2. sz.

DANCS Szabolcs: Webarchiválási politikák
In: Könyv, könyvtár, könyvtáros, 2011. (20. évf.), 10. sz.

DIPPOLD Péter: A hagyományos nemzeti bibliográfia és az Internet : Válaszlehetőségek az új kihívásokra
Budapest : ELTE BTK, 2005

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

HEGYKÖZI Ilona: Hol tart ma a webarchiválás?
In: Könyvtári Figyelő, 2014. 4. sz.

ILLIEN, Gildas: Webarchíválás a francia gyakorlatban
In: Könyvtári Figyelő, 2009. (55. évf.) 3. sz. p. 553-554.

JODELIS, Remigijus: Elektronikus források begyűjtése és archiválása Litvániában: úton egy virtuális könyvtár felé
In: Tudományos és Műszaki Tájékoztatás 2004. (51. évf.), 6. sz.

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

KORNHOFFER Mónika: A világhálón található információk gyűjtésének és megőrzésének hazai és nemzetközi áttekintése
Pécs : PTE FEEK, 2010

NUYS, Carol Van – ALBERTSEN, Ketil – PEDERSEN, Linda et al.: A Paradigma projekt.
In: Tudományos és Műszaki Tájékoztatás, 2005. (52. évf.), 11-12. sz.

Hivatkozások

[1] https://tools.ietf.org/html/rfc7089

[2] http://mementoweb.org/about/

[3] http://timetravel.mementoweb.org

[4] http://webarchive.org.uk/mementos

[5] https://chrome.google.com/webstore/detail/memento-time-travel/jgbfpjledahoajcppakbgilmojkaghgm

[6] https://addons.mozilla.org/hu/firefox/addon/synchronicity/

[7] https://blog.archive.org/2013/10/24/web-archive-404-handler-for-webmasters/

[8] https://en.wikipedia.org/wiki/Robots_exclusion_standard

[9] https://addons.mozilla.org/en-US/firefox/addon/scrapbook-x/

[10] https://webrecorder.io

[11] https://github.com/webrecorder/webrecorderplayer-electron/releases/latest

[12] https://www.httrack.com

[13] http://pandora.nla.gov.au

[14] https://en.wikipedia.org/wiki/Link_rot

[15] http://archive.is

[16] https://perma.cc

[17] http://www.webcitation.org

[18] http://archive-it.org

[19] Az IIPC-t 2003-ban a francia nemzeti könyvtár és 12 partnerintézmény alapította. Jelenleg már több mint 45 országból vannak tagjai (főként könyvtárak és levéltárak). A célja az internet megőrzésével foglalkozók közötti tapasztalatcsere, az ehhez szükséges technológiák közös fejlesztése, a szabványosítás. Honlap: http://www.netpreserve.org

[20] http://crawler.archive.org

[21] http://netpreserve.org/openwayback

[22] http://archive-access.sourceforge.net/projects/nutchwax/

[23] http://dia-nz.github.io/webcurator/

[24] https://webarchivehistorians.org

[25] http://www.ucl.ac.uk/ucl-press/browse-books/the-web-as-history

[26] http://whatdiditlooklike.mementoweb.org

[27] http://webarchive.lib.ntu.edu.tw/eng/

[28] https://www.webarchive.org.uk/shine/graph

[29] http://lintool.github.io/warcbase/vis/crawl-sites/

[30] http://web.archive.org

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

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

[33] http://mekosztaly.oszk.hu/mia/doc/webarchivalas-irodalom.html

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

Beérkezett: 2017. VI. 5-én.

Drótos László
könyvtáros
OSZK – E-könyvtári Szolgáltatások Osztály.
E-mail: *mekdl@iif.hu*