Drótos László

Webes tartalmak digitális megőrzése[1]

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 nonprofit és az üzleti világ szereplői fogják elvégezni ezt a munkát[2]. 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 munkafolyamatai, 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 státusza[3] jelent. De az interneten még ezek a lehatároltnak és (legalább ideiglenesen) lezártnak 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 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 MEK[4] a digitális könyvek megőrzését és szolgáltatását vállalta fel, a 2004-től létező EPA[5] az elektronikus periodikumokkal foglalkozik, a 2007-ben alapított DKA[6] 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 MIA[7], 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 Rendszer[8] 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 fájlból[9] terjedt ki egy határtalan, folyamatosan születő és pusztuló világhálóvá, 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 címeken 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, illetve 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 az élő weben már nem elérhetők. Mivel az első hazai webszerver 25 éve, 1993-ban indult el a BME-n[10] 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 lemezeken) 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 KIFÜ[11] 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 X[12], Web ScrapBook[13], WARCreate[14], WAIL[15], Webrecorder[16]) inkább a magáncélú és kis volumenű archiválásra szolgálnak, de például a nagyon felhasználóbarát és még magyar felülettel is rendelkező HTTrack[17] programot mind a mai napig használják az 1996-ban indult ausztrál nemzeti webarchívumot, a PANDORA-t[18] építő könyvtárakban is. Ezeknek a szoftvereknek egy része képes az Internet Archive-nál kidolgozott és 2009-ben ISO 28500 néven szabványosított WARC[19] 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 Heritrix[20] nevű aratógépet és a Wayback[21] 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 Tool[22] és a dán NetarchiveSuite[23]. Szintén dán könyvtári fejlesztés a WARC-ban tárolt weboldalak megjelenítése mellett teljes szövegű keresőt és statisztikai, illetve vizualizációs funkciókat is tartalmazó SolrWayback[24], aminek a tesztelésébe mi is bekapcsolódtunk. Továbbá egy saját kereső prototípusát is elkészítettük SolrMIA[25] 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 „főcímeket" 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 is: a Brozzler[26]. 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 böngésző-modullal (mégpedig a Chrome motorjával), így jobb minőségben lehet vele menteni a modern, dinamikusan generált weboldalakat, mint az eredetileg még az 1.0-ás webhez készült Heritrix-szel.

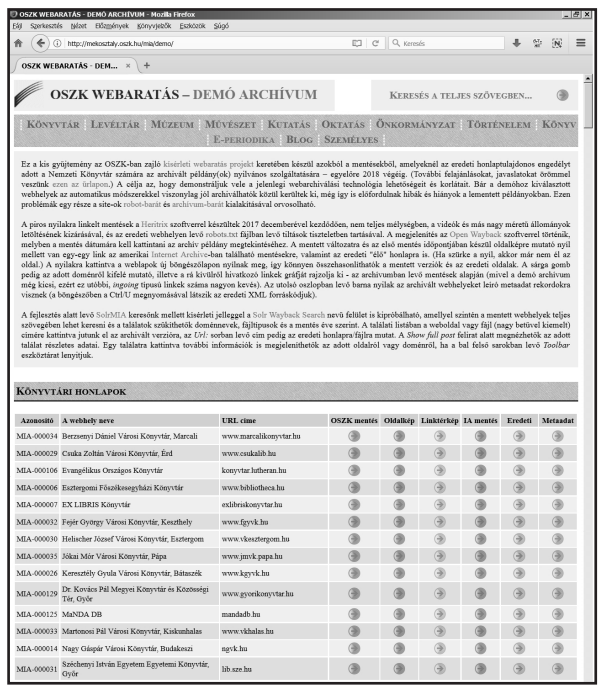
A webhelyek archiválása számítástechnikailag egy meglehetősen bonyolult feladat. Részben a weben használt sokféle formátum, műszaki és design megoldás, program- és parancsnyelv, szerverbeállítás stb. miatt, részben pedig azért, mert a weboldalakat emberek számára fejlesztik, ezért gyakran olyan interaktív 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, illetve 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 robotbarát[27] és archívumbarát[28] megoldásokat is beépítenének a webfejlesztők és webmesterek a szolgáltatásaikba.

2017 nyarától 2018 októberéig többféle aratást is végeztünk a Heritrix programmal.[29] 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 eseményalapú 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 gyűjteményt[30], 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*[31], 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 kialakításánál az amerikai könyvtári szervezet, az OCLC egyik munkacsoportjának[32] 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 tudjuk. 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

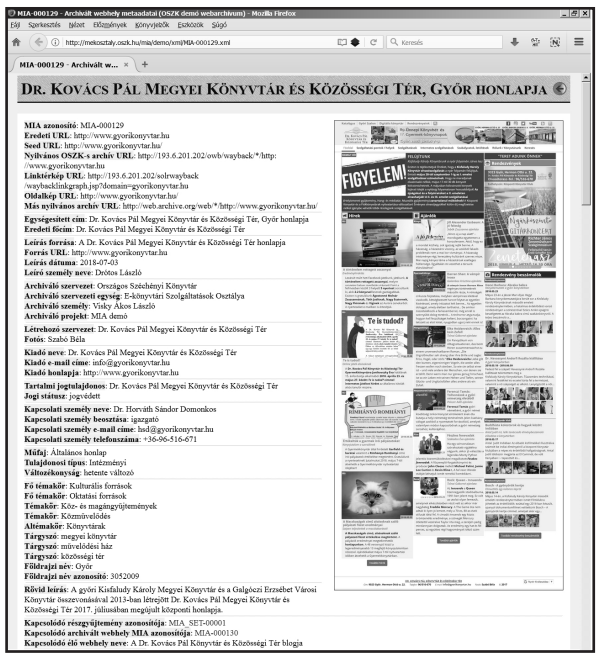
feladat, amit nem tud megoldani egyetlen intézmény és benne néhány ezzel foglalkozó munkatárs. Fontos lenne, hogy minél többen ismerjék meg ennek a szakterületnek az alapjait és kapcsolódjanak be a munkába, akár úgy, hogy megőrzésre érdemes, de kevéssé ismert magyar webhelyeket ajánlanak az erre szolgáló űrlapon[33], 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 honlapja[34], a jelenleg már 30 fős MIA-L levelezőcsoport[35], a közel 600 szócikket tartalmazó MIA wiki[36], a több mint 450 tételes és többféle formátumban is elérhető szakbibliográfia[37], 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.



1. ábra: A nyilvános demó webarchívum részlete

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 részgyűjteményszinten 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 infor-

matikai 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 Consortium[38] munkájában, főként az oktatási munkacsoport keretében.[39] És tovább szorgalmazzuk a hazai együttműködést is a közgyűjtemények között a digitálisan születő, a papíralapú világnál sokkal veszélyeztetettebb és tünékenyebb kultúránk megőrzése érdekében.



2. ábra: Egy archivált honlap „katalóguscédulájának" részlete

Ajánlott irodalom:

Dancs Szabolcs: Webarchiválási politikák. *Könyv, könyvtár, könyvtáros*, 2011. (20. évf.), 10. sz. pp. 14–20.

Drótos László: Az internet archiválása mint könyvtári feladat. *Tudományos és Műszaki Tájékoztatás*, 2017. (64. évf.), 7–8. sz. pp. 361–371.

Drótos László – Kokas Károly: Webarchiválás és a történeti kutatások. *Digitális Bölcsészet*, 2018. (1. évf.), 1. sz. pp. 35–53.

Drótos László – Németh Márton: Az OSZK-ban folyó kísérleti webarchiválási projekt első évének tapasztalatai. *Tudományos és Műszaki Tájékoztatás*, 2018. (65. évf.), 7–8. sz. pp. 389–400.

Németh Márton: A webarchiválásról történeti megközelítésben. *Könyv, könyvtár, könyvtáros*, 2018. (27. évf.), 2. sz. pp. 48–52.

Németh Márton: Nemzetközi körkép a webarchiválás gyakorlatáról. *Könyvtári Figyelő*, 2017. (63. évf.), 4. sz. pp. 575–582.

Jegyzetek

1. A *Born Digital – Digitális tartalom, digitális szolgáltatás* című K2 műhelynapon, 2018. október 10-én, az OSZK-ban elhangzott előadás szerkesztett változata.
2. Lásd pl. az amerikai nonprofit 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 képgyűjteményekben vegyesen vannak digitalizált és digitálisan született művek.)
3. Például: mi tekinthető kiadványnak? Mi esik a kötelespéldány szabályozás alá? Mennyiben más, mint a nyomtatott kiadása? Ki az illetékes jogtulajdonos? Milyen feltételekkel szolgáltatható?
4. Magyar Elektronikus Könyvtár: http://mek.oszk.hu
5. Elektronikus Periodika Archívum és Adatbázis http://epa.oszk.hu
6. Digitális Képarchívum: http://dka.oszk.hu
7. Drótos László: Mi a MIA? – Javaslat egy Magyar Internet Archívum létrehozására http://mek.oszk.hu/html/irattar/eloadas/2006/mia.htm
8. OKR-projekt: http://www.oszk.hu/okr-projekt
9. CERN – Home of the first website: http://info.cern.ch
10. BME Irányítástechnika és Informatika Tanszék: http://www.fsz.bme.hu/www/other_h.html
11. Kormányzati Informatikai Fejlesztési Ügynökség: http://kifu.gov.hu
12. http://mekosztaly.oszk.hu/mediawiki/index.php/ScrapBook
13. http://mekosztaly.oszk.hu/mediawiki/index.php/Web_ScrapBook
14. http://mekosztaly.oszk.hu/mediawiki/index.php/WARCreate
15. http://mekosztaly.oszk.hu/mediawiki/index.php/WAIL
16. http://mekosztaly.oszk.hu/mediawiki/index.php/Webrecorder
17. http://mekosztaly.oszk.hu/mediawiki/index.php/HTTrack
18. http://mekosztaly.oszk.hu/mediawiki/index.php/PANDORA_(ausztr%C3%A1l)
19. http://mekosztaly.oszk.hu/mediawiki/index.php/WARC
20. http://mekosztaly.oszk.hu/mediawiki/index.php/Heritrix
21. http://mekosztaly.oszk.hu/mediawiki/index.php/Wayback
22. http://mekosztaly.oszk.hu/mediawiki/index.php/WCT
23. http://mekosztaly.oszk.hu/mediawiki/index.php/NetarchiveSuite
24. http://mekosztaly.oszk.hu/mediawiki/index.php/SolrWayback
25. http://webadmin.oszk.hu/solrmia/
26. http://mekosztaly.oszk.hu/mediawiki/index.php/Brozzler
27. http://mekosztaly.oszk.hu/mediawiki/index.php/Crawler-friendly_website
28. http://mekosztaly.oszk.hu/mediawiki/index.php/Archive-friendly_website
29. Általában csak a kezdőoldaltól számított két-három szint mélységig ment le a robot és videofájlokat többnyire nem töltöttünk le.
30. http://mekosztaly.oszk.hu/mia/demo/
31. http://mekosztaly.oszk.hu/mediawiki/index.php/Memento
32. http://mekosztaly.oszk.hu/mediawiki/index.php/OCLC_WAM
33. https://goo.gl/forms/Y1qIIxcM7APPiq443
34. http://mekosztaly.oszk.hu/mia/
35. http://mekosztaly.oszk.hu/cgi-bin/mailman/listinfo/mia-l
36. http://mekosztaly.oszk.hu/miawiki
37. http://mekosztaly.oszk.hu/mia/doc/webarchivalas-irodalom.html
38. http://mekosztaly.oszk.hu/mediawiki/index.php/IIPC
39. A 2003-ban alapított IIPC-nek kb. 45 országból vannak tagjai és 2018-ban csatlakozott hozzá magyar részről az OSZK is.