



**A webarchiválás elméletének és gyakorlatának alapelemei
A szervezett keretek között zajló webarchiválás kezdetei
Magyarországon**

Egyetemi doktori (PhD) értekezés

a szerző neve: Németh Márton
témavezető neve: Eszenyiné dr. Borbély Mária

DEBRECENI EGYETEM
Természettudományi és Informatikai Doktori Tanács
Informatikai Tudományok Doktori Iskola
Debrecen, 2021.

Ezen értekezést a Debreceni Egyetem Természettudományi és Informatikai Doktori Tanács Informatikai Tudományok Doktori Iskola Az információ technológia és a sztochasztikus rendszerek elméleti alapjai és alkalmazásai programja keretében készítettem a Debreceni Egyetem műszaki doktori (PhD) fokozatának elnyerése céljából.

Nyilatkozom arról, hogy a tézisekben leírt eredmények nem képezik más PhD disszertáció részét.

Debrecen, 2021. június 1.

.....
Németh Márton

Tanúsítom, hogy Németh Márton doktorjelölt 2015.- 2021 között a fent megnevezett Doktori Iskola Az információ technológia és a sztochasztikus rendszerek elméleti alapjai és alkalmazásai programjának keretében irányításommal végezte munkáját. Az értekezésben foglalt eredményekhez a jelölt önálló alkotó tevékenységével meghatározóan hozzájárult. Nyilatkozom továbbá arról, hogy a tézisekben leírt eredmények nem képezik más PhD disszertáció részét. Az értekezés elfogadását javasolom.

Debrecen, 2021. június 1.

.....
Eszenyiné dr. Borbély Mária

A webarchiválás elméletének és gyakorlatának alapelemei

A szervezett keretek között zajló webarchiválás kezdetei

Magyarországon

Értekezés a doktori (PhD) fokozat megszerzése érdekében
az Informatikai technológiák tudományágban

Írta: Németh Márton okleveles informatikus könyvtáros szakos bölcsész –
történelem szakos bölcsész és tanár

Készült a Debreceni Egyetem Informatikai Tudományok doktori iskolája
(Az információ technológia és a sztochasztikus rendszerek elméleti alapjai és
alkalmazásai programja) keretében

Témavezető: Eszenyiné dr. Borbély Mária

A doktori szigorlati bizottság:

elnök:	Dr. Hajdu András	habil. egyetemi tanár
tagok:	Dr. Varga Katalin	habil. egyetemi docens
	Dr. Adamkó Attila	habil. egyetemi docens

Az értekezés bírálói:

Dr.	
Dr.	

A bírálóbizottság:

elnök:	Dr.	
tagok:	Dr.	
	Dr.	
	Dr.	
	Dr.	

Az értekezés védésének időpontja:

Tartalom

Bevezetés	5
A témaválasztás indokai	7
A disszertáció tartalmi szerkezetének főbb elemei.....	10
A számítógépes világháló természete, a webarchiválás eredendő töredék alapú szubjektivitása	13
A web technikai fejlődésének kihívása.....	14
A világháló természetéből fakadó közgyűjteményi archiválási dilemmák.....	19
1. A webarchiválás közgyűjteményi keretei	24
1.1 A webarchiválás a digitális könyvtári szolgáltatások rendszerében....	28
1.2 A webarchiválás mint digitális levéltári feladat.....	32
1.3 A webarchiválás digitális muzeológiai vonatkozásai	35
1.4 Az audiovizuális archívumok és a webarchiválás	37
2. A webarchiválás nemzetközi háttere	39
2.1 A webarchiválás főbb nemzetközi intézményi szereplői és az együttműködés keretei	40
2.2 A webarchiválás különféle nemzeti gyakorlatai – esettanulmányok	44
PANDORA: Az ausztrál webarchiválás modellje	46
Webarchiválás Nagy-Britanniában	49
Netarkivet.dk – dán webarchívum gyűjtőkör és szervezeti keretek	51
Webarchiválás Csehországban	55
Webarchiválás Szlovákiában	57
Webarchiválás Észtországban	59
Webarchiválás Hollandiában.....	61
A nemzetközi áttekintés tanulságai.....	62
2.3. Magyar részvétel a nemzetközi együttműködésekben.....	63
3. A magyarországi webarchiválás kísérleti projektje – a szervezett webarchiválás kezdetei	69

3.1	Előzmények	70
3.2	A kísérleti projekt eredeti koncepciója	74
3.3	A webarchiválás munkafolyamatainak áttekintése.....	81
3.3.1	Az általános webarchiválás keretei	81
3.3.3	Webhelyek válogatásának alapelvei a szelektív webarchiváláshoz ...	86
3.3.4	Eseményalapú webarchiválás	91
3.4	A gyűjtőkör, a gyűjteményfejlesztés alapvető kereteinek tervezete	94
3.4.1	Alapelvek	95
3.4.2	Az elektronikus hungarikumok gyűjtőköre	98
3.4.3	Az online kiadványtípusok gyűjtőköre	98
3.4.4	A könyvtári gyűjtőkör lehatárolása, együttműködés külső intézményi partnerekkel	101
3.5	A kísérleti projekt hardver és szoftver háttere.....	101
3.5.1	A hardver infrastruktúra elemei.....	102
3.5.2	A webarchiválási kísérleti projekt szoftveres háttere	107
3.6	A metaadatok rögzítése	135
3.6.1	A saját metaadat-struktúra kialakításának kezdő lépései	136
3.6.2	Az RDA kompatibilis metaadatmodell megteremtésének első lépései	141
3.7	Minőségellenőrzés, a webarchiválás határai	148
3.7.1	A webarchiválás határai	148
3.7.2	Minőség-ellenőrzés	151
	Összefoglaló gondolatok a minőség-ellenőrzésről.....	157
3.8	Raktározás, a hosszútávú megőrzés kihívásai	158
3.9	Jogi szabályozási szükségletek és kísérletek.....	160
3.9.1	A webarchiválás szerzői jogi kihívásai, jogkezelési eljárások	160
3.9.2	A webarchiválási tevékenység szabályozásának törvényi, jogszabályi háttere	165
3.10	Az OSZK webarchiválási tevékenységének kommunikációs keretei	168
3.10.1	A honlap felépítése és alapvető szolgáltatásai.....	169

3.10.2 A webarchívum mint gyűjtemény	173
3.10.3 A webarchiválási projekt szakmai és szervezeti keretei.....	175
3.10.4 Általános felhasználói információk	176
3.10.5 Kínálatunk a tartalomgazdáknak	176
3.10.6 A közgyűjteményi szakmának ajánlott szolgáltatások	177
3.10.7 A külső kommunikáció keretei	179
3.11 Együttműködések partnerintézményekkel	179
3.12 A jövőre vonatkozó tervek	181
4 A webarchiválás oktatása	182
4.1 Az oktatási tevékenységek nemzetközi háttere	183
4.1.1 A NETLAB kutatócsoport e-learning programja	184
4.1.2 Az IIPC Training Working Group tevékenysége	187
4.2 A Könyvtári Intézettel közösen kidolgozott blended learning koncepció	194
4.3 Jövőbeni oktatási tervek.....	200
5 A webarchívumok tudományos hasznosításának keretei	201
5.1 A WARCnet projekt.....	203
5.2 Webtörténetírás	205
5.2.1 A webarchívumok történeti kutatási célú elemzésének támogatása	209
5.3 A webarchívum mint a nagy mennyiségű adatok forrása, az adattudományi kutatások tárgya	212
6.4 Hiteles webarchívum	216
6 Mikroadatok használata a webarchiválás és a visszakeresés hatékonyságának növelésére	217
6.1 A mikroadatok a dokumentumalapú web és a szemantikus web metszéspontjában	217
6.2 A mikroadatok potenciális felhasználási lehetőségei a webarchiválás során.....	221
6.2.1 A webarchiválás hatékonysága növelésének általános kihívásai	222

6.2.2 A robotokkal végzett webaratás szabályozása mikroadatok segítségével	223
6.2.3 Az archivált állomány megjelenítésének segítése mikroadatok révén	225
6.2.4 A hosszú távú digitális megőrzés támogatása mikroadatok segítségével	226
6.2.5 A kutatástámogatási szolgáltatások segítése mikroadatok segítségével	227
6.3 Mikroadatok előállítása és alkalmazása	229
7 Epilógus - A kutatás értékelése	230
Summary	235
Bibliográfia	237

Bevezetés

A kutatás tárgya, az annak alapjául szolgáló hipotézisek

A Phd dolgozat alapjául szolgáló kutatás a webarchiválás elméleti alapelemeinek bemutatása mellett annak első lépéseit is fel kívánja villantani, hogy az elméleti kereteket miként lehet magyar körülmények között a gyakorlat nyelvére lefordítani, az Országos Széchényi Könyvtárban 2017-ben induló webarchiválási projekt keretei között. A kezdeti kutatási hipotéziseket az alábbiakban öt fő szempont szerint lehet megfogalmazni:

1. A kutatás fő hipotézise, hogy magyar nyelven az újdonság erejét is felvillantva áttekintő elemzés nyújtható a webarchiválás létjogosultságáról, elméleti háttéréről, értelmezési tartományáról, elméleti és gyakorlati korlátairól, közgyűjteményi kereteiről, kutatási célú felhasználásának módozatairól, s emellett feltárhatók a témakör nemzetközi gyakorlatának, illetve szakirodalmi reprezentációjának legfontosabb vonásai is.
2. Fontos kérdésként merült fel a kezdeti hipotézisként, hogy az elméleti és a nemzetközi gyakorlati keretek felvillantása mellett, megteremthető az Országos Széchényi Könyvtár keretei között a szervezett webarchiválás archiválási és szolgáltatási gyakorlata, a megfelelő munkafolyamatok kialakításával, a megfelelő pénzügyi, információtechnológiai, illetve jogi háttér megteremtésével. Amennyiben nem, vagy csak részlegesen kerül erre sor, melyek a fő kihívást jelentő tényezők?

3. Nélkülözhetetlennek tűnik annak tisztázása is, hogy milyen lényeges vonásai vannak a mostanában kialakuló webkönyvtárosi, webkurátori munkaköröknek, s meddig terjed ezen tevékenységek határa a könyvtári keretek között, s mi tartozik a többi közgyűjteményi ág, illetve az informatika keretei közé.
4. Az esetleg felbukkanó szakmai és pénzügyi, információtechnológiai infrastrukturális nehézségek, a jogi, szabályozási háttér megteremtésének nehézkes folyamata nem lehetetlenít-e el, hogy egy kutatás-fejlesztési kísérleti projektből állandó nemzeti könyvtári szolgáltatás válhasson.
5. Meg lehet-e teremteni a webarchiválás témakörének bevonását az akkreditált felsőfokú könyvtáros továbbképzés, illetve a felsőoktatási, egyetemi képzés keretei közé?

A hipotézisekre adott válaszok egyes szakmai elemei a dolgozatba szervesen beágyazva, illetve a kutatás eredményeit összegző tézisek formájában kerülnek kifejtésre, az összkép felvázolására pedig a dolgozat zárófejezetében kerül sor.

A dolgozat további egyedi vonását adja meg, hogy az egyéni kutatási program megvalósítása a hipotézisekben megfogalmazottak megválaszolása kapcsán néhány ponton keveredik azokkal a kollektív erőfeszítésekkel, melyeket az OSZK E-Könyvtári Szolgáltatási, majd 2020. tavaszától önálló Webarchiválási Osztálya keretében munkatársaimmal együtt tettünk az üzemszerű webarchiválási szolgáltatás feltételeinek megteremtéséért. A szövegben az egyes szám első személy használata mindig olyan tartalmi elemre utal, mely az egyéni kutatáshoz, elemzéshez,

kutatási eredmények feltárásához kapcsolódik. A témához tartozó kollektív, gyakorlati erőfeszítések lenyomatait pedig a többes szám első személy használata tükrözi majd vissza. Az egyéni kutatás gyakorlati elemeinek, illetve a kollektív gyakorlati tevékenységek eredményeinek felvillantására több fejezetben is kitérek, az összegzést pedig ezen a téren is a záró fejezetben végzem el.

A témaválasztás indokai

Joggal vetődhet fel kérdésként, hogy milyen főbb tényezők indokolják a kutatási hipotézisekben foglaltak felvetését, miért kell, hogy a webarchiválás kutatási témaként megjelenjen. A webarchiválás, mint tudományos téma előtérbe kerülése számos tényezővel indokolható. Megváltoznak a kulturális örökségünk áthagyományozásának keretei. A közgyűjtemények által őrzött hagyományos írásbeli és tárgyi kultúrkincs jól meghatározható dokumentumhordozókhoz kötött, megfelelő infrastrukturális körülmények biztosításával a hosszútávú megőrzése jó eséllyel garantálható. Kulturális örökségünknek egyre nagyobb szelete azonban manapság már a számítógépes világhálón tűnik fel. Ennek a szegmensnek begyűjtése, feldolgozása, hosszútávú megőrzése kézenfekvőnek tűnik, másfelől viszont igen összetett kihívást jelent.

A web története mintegy harminc évre nyúlik vissza. A svájci CERN intézetben eredetileg a hipertext és az internet összeházasításának fő célja a kutatási anyagokhoz való hozzáférés segítése volt. Harminc év alatt nagyon hosszú utat tettünk meg. A számítógépes világháló az emberi tudás tárolásának és áthagyományozásának alapvető médiumává lett. Amennyiben a kortárs kultúra lenyomatait a közgyűjteményekben nem

tudjuk megőrizni és bemutatni, ezen intézményrendszer gyorsuló ütemben vesztheti el a kultúra átörökítésének képességét, a kortárs társadalmi problémákra történő reflexió képességét. Végso soron tehát a legitimitása kérdőjeleződik meg. A webarchiválással tehát igen súlyos társadalmi jelentősége miatt tudományos szinten is foglalkozni kell. Magyarországon, a szervezett webarchiválási gyakorlat hiánya miatt, a területet érintő tudományos kutatások háttere sem alakult ki, néhány elszórt kísérletről beszélhattünk csupán e területen. A szervezett keretek között zajló webarchiválás kialakulásának ábrázolása, a webarchiváláshoz kapcsolódó tudományos kihívások felvázolásával együtt, magyarországi keretek között, megítélésem szerint igen időszerű feladatot jelent. Fontos megjegyezni, hogy ebben az esetben nem csupán egy kutatás-fejlesztési projekt sorsának figyelemmel kíséréséről van szó (azt a feladatot éves kutatási jelentések sorozatával is megfelelően lehet teljesíteni). A webarchiválás tágabb külső összefüggésrendszerét, az archivált anyagokra épülő kutatás útjait is beépítve a webarchiválás itthoni gyakorlata kezdeteiről való számadás mellé már összetett tudományos igényű elemzés végezhető el. Számos a webarchiválást érintő szakmai kihívás megtárgyalására, a kutatási lehetőségek új útjainak felvázolására is tudományos jelleggel lehet sort keríteni.

A webarchiválás szükségszerűsége információtudományi szempontból sokáig megkérdőjeleződött a számítógépes világhálóról való gondolkodás fényében. Felvetődött, hogy az egész világhálót nem lehet-e valamiféle hatalmas digitális könyvtárként, tudásszervező rendszerként felfogni, amit nem is kell archiválni, mert a rendszeren belül intakt módon

megőrződnek a tartalmak. Mint ahogy arra Niels Ole Finnemann információkutató rámutat, ez a megközelítés több okból is problémás (Finnemann, 2019). Először is, a web egyszerűen túl nagy ahhoz, hogy bármilyen tartalomfeltáró, tartalomkezelő eszközzel átfogható, kezelhető legyen. A második fő kihívás a web állandó jelenidejűsége, efemer jellege. A tartalmak nagyon gyors üzemben követhetetlen sebességgel jelennek meg és tűnnek el a világhálóról. Már a szervezett webarchiválás kezdeteinek idején, 1996-ban is csak 42 napra becsülték egy honlap átlagos élettartalmát, s ez azóta nemhogy növekedett volna, hanem radikálisan le is csökkent. Emellett a webes tartalmak jó részének nincs megfogható szerkesztője, s a sok esetben szinte állandó elvégzett tartalmi változtatások is követhetetlenek. A digitális világ ebben az értelemben radikálisan véget vetett az információhordozó objektum stabilitásának. A harmadik fő indok, amely szerzői oldalról vet véget a stabilitásnak. A világhálón nincs semmiféle kapuőri szerep, bárki bármit publikálhat. Nagyon magas fokú heterogenitás nem képezhető le semmiféle hagyományos közgyűjteményekben használatos taxonómiai eszköztárral. Az áttekinthetetlen összetettséghez ez a jellemző is alaposan hozzájárul (Finnemann, 2019).

A webarchiválás tudományos és társadalmi jelentőségét tehát az eddigiekben is már számos érveléssel támasztottuk alá. Ezzel szemben a tudományos és társadalmi kommunikációban, a kulturális örökséghez kötődő pénzügyi erőforrások elosztásakor globális szinten is csekély figyelmet kap. A nemzetközi tudományos párbeszéd e szakmai területen csupán az utóbbi öt-tíz évben kapott igazán lendületet.

A webarchiváláshoz kötődő tevékenységek és elméleti kutatási irányok, keretek közgyűjteményi szempontú feldolgozására, a legújabb nemzetközi kutatási eredmények bemutatására, ötvözve a saját itthoni kutatói tapasztalatokkal, önálló tudományos munka formájában magyar nyelven még nem került sor. Bármely, a területet érintő résztémakör mélyebb jellegű (magyar tudományos keretek között zajló) tudományos megtagyargalásához előfeltételként szükségesnek érzem egy általános bemutató jellegű elemzés elvégzését. A téma jelentősége a kulturális örökség áthagyományozásának szempontjából véleményem szerint feltétlenül indokolja tehát egy saját kutatási eredményekre is támaszkodó, áttekintő jellegű tudományos munka elkészítését e témakörben.

A disszertáció tartalmi szerkezetének főbb elemei

Ebben az alfejezetben azt mutatom be, hogy miként jelenik meg a fentebb vázolt hipotézisekben foglaltak megválaszolásának igénye a dolgozat tartalmi szerkezetében.

A dolgozat a webarchiválás világának interdiszciplináris jellegű áttekintésére vállalkozik. Felvázolja a legfontosabb tudományos kereteket, továbbá a kapcsolódási pontokat a hagyományos közgyűjteményi begyűjtési, feldolgozási és megőrzési tevékenységek felé, miközben rámutat az egyedi sajátosságokra is. Önálló fejezetben határozza meg a webarchiválás pozícióját a digitális közgyűjteményi szolgáltatások palettáján.

A dolgozat egyik legfontosabb egyedi hozzáadott értékét, az elméleti keretek bemutatása mellett, az első magyar, szervezett keretek között, a szerző tevékeny részvételével zajló kutatás-fejlesztési projekt koncepciójának, céljainak, eredményeinek, kihívásainak elemző bemutatása adja meg. Magyarországon, bár fontos később felvázolandó előzményekről is beszélhetünk, a szervezett intézményi szintű webarchiválási tevékenység csupán 2017-ben indult el. A később jövők előnye, hogy bőven meríthettek a nemzetközi jó gyakorlatokból, s számíthatnak a nemzetközi webarchiválással foglalkozó szakértői közösség segítségére is. Éppen ezért a nemzetközi környezet bemutatása, a fentebb említett jó gyakorlatok rövid bemutatása a tanulságok hazai hasznosításának szemszögéből is a dolgozat fontos eleme.

A webarchiváláshoz kötődő gyakorlati tevékenységek ellátása mellett alapvető fontosságú a webarchiválás jelentőségének tudatosítása, a területet érintő oktatásmódszertani keretek felvázolása, a webarchiváláshoz kötődő kompetenciák széleskörű terjesztése a közgyűjteményi szakmai közeg és a tágabb társadalmi célközönség felé is. Az oktatási dimenzió, s annak itthoni vonatkozásai tehát önálló témakörben jelennek meg. Az oktatási terület fontosságának további indokaként felhozható még, hogy alkalmam nyílt aktívan bekapcsolódni a webarchiválás itthoni oktatási kereteinek formálása mellett a területet érintő nemzetközi kezdeményezésekbe is.

A webarchívumok egyre nagyobb fontosságra tesznek szert úgy is, mint a különféle tudományos kutatások forrásai. Ezzel az izgalmas területtel azért is érdemes foglalkozni, mert hihetetlenül széleskörű az az

összefüggésrendszer, mellyel a webarchiválási tevékenységet és a webarchívumokat mint a kutatások forrását vizsgálni lehet. Egyaránt új utakat kínál történészek, digitális filológusok, informatikusok, adattudósok, illetve információtudományi szakemberek számára is. A webarchiválás gyakorlatának s a területhez kötődő oktatási háttérnek a felvázolása mellett tehát magas hozzáadott értékkel bíró eleme a disszertációnak, hogy a terület tágabb interdiszciplináris környezetéhez kötődő jelenkori és jövőbeni kutatási lehetőségeket, témákat mutasson fel. Egy konkrét javaslat is megfogalmazásra kerül a zárófejezetben, hogy miképpen lehetne a szemantikus technológiákban rejlő lehetőségeket a közgyűjteményi szféra, illetve a tartalomipar összefogásával a webarchiválás terén hasznosítani. Így válik teljessé a webarchiválás környezetét, kapcsolódási pontjait felvázoló összkép. A dolgozat egyediségét annak közgyűjteményi szempontú elemző látásmódja, valamint az itthoni intézményes webarchiválás kezdeteihez tartozó kutatás-fejlesztési munka eredményeinek bemutatása adhatja tehát meg a vonatkozó oktatási és kutatási-hasznosítási háttér alapjainak felvázolásával együtt.

Egy fontos terminológiai megjegyzés is feltétlenül a bevezető fejezetbe kívánczik még. Sokszor elmosódik napjainkban mára már a határ az internet archiválás és a webarchiválás fogalmai között. Ez főként abból fakad, hogy az internet domináns kommunikációs médiumává mára már a számítógépes világháló vált. Az e-mailjeinket is webes felületen nézzük meg, s a médiafogyasztás is webes csatornákon zajlik. Az e-mailek archiválása hagyományos értelemben például nem tartozik a

webarchiválás tárgykörébe, mert egy teljesen önálló platformról van szó. De ugyanezek a dilemmák felvethetők számos, főleg intézményi kommunikációs környezetben használt szolgáltatás esetében is (például fórumok vagy csevegőprogramok használata).

A webarchiválás részleteiben való elmerülés előtt, a bevezetés második részeként, egy tágabb, a világháló fogalmához, természetéhez kötődő összefüggésrendszerben helyezem el a vizsgált területet. Ez, a hipotézisekben is jelzett, általános értelmezési tartomány feltétlenül szükséges a későbbi elemzések megfelelő kontextusban történő értelmezéséhez.

A számítógépes világháló természete, a webarchiválás eredendően töredék alapú szubjektivitása

Az értelmezési keretek kijelölésének első alapvető állítása számomra a kutatásom során, hogy mint rendkívül összetett és strukturált hálózat, a számítógépes világháló a maga egyediségében nem hasonlítható össze igazán egyetlen más olyan hagyományos vagy digitális információforrással sem, amellyel az emberiség történetében eddig találkozhattunk. A második fő állításom, hogy a legnagyobb kihívás talán a közgyűjtemények, illetve bármely más archiválást végző intézmény, magánszemély számára, hogy a világháló annak eredeti természetes szerkezetében, összetettségében, folyamatosan változó dinamikus tartalmi kínálatával a maga teljességében nem archiválható. A web élő szövetéből archiválási céllal csupán önkényes töredékeket (egyres webhelyeket, honlapokat, részben vagy egészben, illetve azok meghatározott gyűjteményét) tudunk kiragadni és megőrizni, ami persze a maga töredékességében is jelentős

fontossággal bír. A válogatott tartalom elemeit körülvevő teljes eredeti kontextus teljes megőrzésének igénye azonban illúzió csupán. A webes archiválás ezen alapvető sajátosságát mindig észben kell tartanunk, ha bármilyen vonatkozásban szóba kerül ez a tudományos terület. A harmadik állítás az utóbbiakból következik: a webarchiválás a maga egyre erősebben automatizált formájában is alapvetően szubjektív emberi tevékenység. Az egységes egészből a töredékek kiszakításának formáját, szempontjait, tartalmi sajátosságait ugyanis az archiválási stratégiát alkotó szakemberek határozzák meg. Ez alapján alkotják meg, illetve módosítják a begyűjtésben segítő szoftverek algoritmusait is.

[A web technikai fejlődésének kihívása](#)

A webarchiválásnak mint tudományos témának a felvetését eleve a technikai fejlődés teremtette meg, s ez a folyamatosan zajló műszaki-társadalmi evolúciós folyamat alapvetően befolyásolja a kutatás tárgyának sorsát is. Drótos László találóan összegezte a jelent tekintve ennek a technikai fejlődésnek az elemeit (Drótos, 2018b). A HTML szabvány már az ötödik verziójánál tart („HTML 5.2”, 2017), a HTTP protokollt lassan mindenhol felváltja a biztonságosabb HTTPS szabvány („HTTPS”, 2020). A webszervereken pedig napjainkban már mindenféle scriptek és egyéb szoftverek futnak. A böngészőprogramok bonyolultsága szinte már elérte az operációs rendszerek szintjét. Elterjedtek az olyan megoldások is, mint a dynamic web page („Dynamic web page”, 2020), a dynamic HTML (DHTML) („Dynamic HTML”, 2020) és a Rich Internet Application vagy Rich Web Application (RIA) („Rich web application”, 2020). Közös jellemzőjük, hogy a korai statikus, egyszerű szerkezetű

weboldalak helyett szerver és/vagy kliens oldalon futó parancsokat és programokat futtatnak, használnak. Valójában már önálló alkalmazásnak számítanak (pl. közösségi oldalak, webáruházak, internetes bankok, online játékok fejlesztésének terén).

Ráadásul, amíg a számítógépes világháló eredetileg alapvetően pc-s böngészőn keresztüli tartalomszolgáltatási és hozzáférési módként jelent meg, mára már külön speciális (mobil) platformok (operációs rendszerek) jelennek meg („Mobile operating system”, 2020). A különféle platformokon olyan plusz kényelmi funkciói jelennek meg például közösségi média alkalmazásoknak, melyek lényegesen más felhasználói élményt nyújtanak, mint a komputeres változatok. Az emberek többsége egyre inkább ezeken a mobilplatformokon keresztül fér hozzá a webhez. A különféle platformokra specializált felbontású tartalmak készülnek, a tartalomkezelő rendszerek legtöbbje már automatikusan fel van készítve a különféle platformok rezponzív jellegű kiszolgálására. Komolyabb kihívást jelentenek azok a speciális alkalmazások pl. mobil applikációk, amelyek szintén értékes tartalmat kínálnak, mentésük módszerei azonban teljesen tisztázatlanok. S akkor még nem is beszéltünk arról, hogy a mobilplatformok teljes hardver és szoftverkörnyezetét is el kellene menteni, emulálhatóvá tenni. Ez a nyílt Android platformon még valamennyire megvalósítható, ám a teljesen zárt IOS architektúrához ilyen céllal hozzáférni szerzői jogi okok miatt, illetve az Apple szolgáltatási politikája miatt sem lehetséges. A töredezettség, töredékesség tehát nem csak az archiválás szempontjából jellemzi napjaink világhálós környezetét,

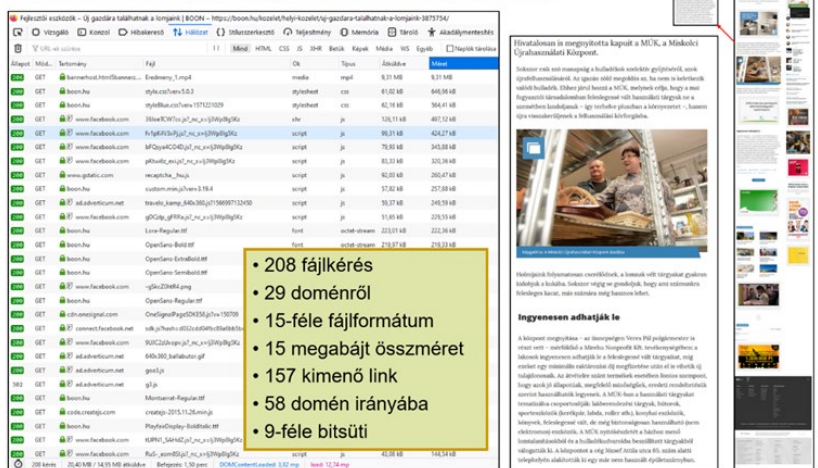
hanem a szolgáltató platformok szerinti széttöredezettség is egyre jellegzetesebbé válik.

De, ha csak a PC-n keresztüli megjelenítő platformnál maradunk, kihívások e tekintetben is bőven jelentkeznek. Drótos László cikke találóan mutatja be, hogy ha megnyitunk egy mai hírportált, a hagyományos újsághoz képest szembetűnő a különbség az összetettség szintjén (Drótos, 2018b). Abban persze (sajnos) hasonlítanak, hogy a tényleges tartalom már szinte elvész a hirdetések között. A világhálós portálon azonban ezek a reklámok gyakran animációk vagy videók, és a szemünk elé ugró ablakokban nyílnak meg, vagy görgetéskor együtt mozognak az oldallal. Sőt személyre szabott ajánlatokat is tartalmazhatnak, a nyomkövető és egyéb websütik által begyűjtött információk révén, a földrajzi helyünk, a korábbi kereséseink és más internetes tevékenységeink, valamint a közösségi oldalakról rólunk gyűjtött információk alapján. Persze a reklámokon túl még számos más elemet is észrevehetünk: websüti- és GDPR-tájékoztatók, időjárési és egyéb aktuális információk, olvasói hozzászólások, további ajánlott cikkekre és webhelyekre vezető linkek, navigációs és értesítő funkciók, beágyazott Facebook, Twitter vagy Google Maps modulok, keresőmező, megosztás gomb, RSS csatorna, forgalommérő kódok... (Drótos, 2018b)

A billentyűzetünk F12 gombjának megnyomásával megnézhetjük, hogy mit is tölt le a böngészőnk, amikor megnyitunk például egy híroldalt, bizony igencsak meglepő szerkezettel találkozunk.

A web egy hipermédia dokumentum?

<https://boon.hu/kozelet/helyi-kozelet/uj-gazdara-talalhatnak-a-lomjaink-3875754/>



The screenshot shows a web browser window with a file list on the left and a sidebar on the right. The file list contains various files and folders, including images, videos, and documents. The sidebar contains a list of links, including a link to a news article about a local council meeting.

- 208 fájlkérés
- 29 doménről
- 15-féle fájlformátum
- 15 megabájt összméret
- 157 kimenő link
- 58 domén irányába
- 9-féle bitsűti

1. ábra Egy weboldal adatai egy mai hírportálon (forrás: Drótos László: Az OSZK webarchívumának újdonságai.

http://mekosztaly.oszk.hu/mia/doc/workshop/Drotos_Laszlo_Az_OSZK_webarchivumanak_ujdon_sagai.pptx)

Tíz-húsz megabájtnyi méretben akár több száz fájlra is szükség lehet egy néhány sornyi hír köré felépített weblap megjelenítéséhez (Drótos, 2018b). A Wappalyzer nevű, a böngészőnkbe beépülő szolgáltatással („Wappalyzer—Identify technology of websites”, 2020) azt is meg tudjuk vizsgálni, hogy milyen technológiák működnek az adott honlap mögött. A legfontosabb állítása tehát a web műszaki fejlődését végigkövető vizsgálódásaimnak, hogy a webes technológiák nagy száma és sokfélesége is azt mutatja, hogy a mai web már inkább olyan, mint egy szoftverbank, nem csupán egyetlen nagyra nőtt hipermédia dokumentum, mint ahogy korábban hivatkoztunk rá.

A web egy hipermédia dokumentum?

CMS: WordPress 5.0.3

Widget: Facebook, Twitter

Analytics: Google Analytics, Gemius, Chartbeat

Blog: WordPress 5.0.3

Captcha: reCAPTCHA

Font Script: Google Font API

Web Framework: Bootstrap

Miscellaneous: Twitter Emoji (Twemoji), HTTP/2

Cache Tool: Varnish 5.0

Programming Language: PHP

CDN: CloudFlare

Database: MySQL

Advertising Network: Advericum

SEO: Yoast SEO 5.4

JavaScript Libraries: jQuery 1.12.4, jQuery Migrate 1.4.1, jQuery UI 1.12.1

2. ábra A hírportál szolgáltatás mögött álló informatikai technológia (forrás: Drótos László: Az OSZK webarchívumának újdonságai.

http://mekosztaly.oszk.hu/mia/doc/workshop/Drotos_Laszlo_Az_OSZK_webarchivumanak_ujdon_sagai.pptx)

Tovább súlyosbítja a helyzetet, hogy a weben használt formátumok gyorsan elavulnak pl. az Adobe Flash („Adobe Flash”, 2020), a Microsoft Silverlight („Microsoft Silverlight”, 2020) illetve a Real Audio („RealAudio”, 2020) multimédiás technológiák eltűnése, a VRML („VRML Virtual Reality Modeling Language”, 1995) háromdimenziós megjelenítő környezet, ami szintén az enyészeté lett mára már. De gondolhatunk azokra a java appletekre („Java applet”, 2020) is, melyeket a manapság használt böngészők már nem hajlandók futtatni. Az egyébként népszerű Java nyelv alkalmazásait biztonsági okokból a weben már nem nagyon lehet használni. Weboldalakból sok-sok milliárd van a felszíni weben, vagyis a Google által is használt programrobotokkal bejárható nyilvános szervereken, és ezek többsége ráadásul változik (akár naponta többször is). A következő fő állításom tehát, amire a kutatás ezen aspektusa során

jutottam, hogy a web jelenleg a legnehezebben megőrizhető médium. Másfelől viszont arra is utaltam már, hogy mindemellett egyben napjaink legfontosabb információhordozó eszköze is egyben (Drótos, 2018b). Tehát az összefoglaló állítást úgy tudom megfogalmazni, hogy ha töredékesen és pillanatfelvételeket készítve is, de mentéseket kell a világhálóról készíteni akár jelenlegi, akár jövőbeli felhasználási lehetőségek, kutatások számára.

A világháló természetéből fakadó közgyűjteményi archiválási dilemmák

Közgyűjteményi szempontból arra jutottam kutatásaim során, hogy a világháló természete, s archiválásának sajátos volta fontos következményekhez vezet. Az első lényeges állítás, melyet Drótos Lászlóval együtt fogalmaztunk meg a magyar kísérleti projekt megalapozása kapcsán a webarchiválást tekintve, hogy alapvető gyűjtőköri alapelveket kell feladnunk (Drótos, 2017a). Nem tudunk például mindent begyűjteni, mindent katalogizálni, feldolgozni és hozzáférhetővé tenni nemzeti könyvtári szinten sem. Ehhez szolgálnak pontos adalékokkal a projekt során szerzett saját tapasztalataim mellett néhány partner nemzeti könyvtár webarchívumainak munkatársai is (Bingham, 2014; Johnson, 2016; Massis, 2016). Több tanulmány arra is utal, hogy ha szükségszerűen válogatni kell, sokszor nagyon nehéz algoritmizálni a rendelkezésre álló szoftveres keretek között a válogatás alapjául szolgáló szempontokat (Ben-David, 2020; Fansler, Gilbertson, & Petersen, 2014). Hiába alkotnak megfelelő jogi szabályozást, a kötelezpéldány törvények vagy rendeletek hatálya alá eső weboldalak

nem fogják bejelenteni soha olyan teljességben a nemzeti könyvtárak számára, mint a könyveket, sajtótermékeket. Erre nemzetközi elemzések is utalnak (Schostag & Fonss-Jorgensen, 2012; Stirling, Illien, Sanz and, & Sepetjan, 2012). Arra pedig végképp hiába várnánk, hogy ezeket a tartalmakat a tulajdonosok saját maguk archiválják, s küldjék be az archiválásra szánt példányt a könyvtárba, mint ahogyan azt a kiadók a digitális könyvek, folyóiratok kapcsán megteszik. A különféle archiválási politikákból számos ország gyakorlatából kaphatunk ízelítőt, nem csupán könyvtári, hanem levéltári kontextusban is (Bingham, 2019; Brügger, Laursen, & Schafer, 2019; Cartledge & Nelson, 2015; Schostag, 2019b; Signori, 2017; Szabo, 2018). Lényeges állításom tehát, amit a gyakorlati webarchiválási munka, s a nemzetközi gyakorlat tanulmányozásából leszűrtem, hogy a webarchiválás nem képzelhető el tehát proaktív könyvtári magatartás nélkül. Az intézménynek magának kell megtalálnia és begyűjtenie a megőrizni kívánt webes tartalmak meghatározó hányadát. Azzal is számot kell vetni, hogy a hagyományos határok szöveges és audiovizuális tartalmak között nagyon könnyen keveredhetnek, elmosódhatnak. Az audiovizuális kulturális kínálat egyre nagyobb hányadát házilag gyártják, a tartalommegosztó portálokon forgalmazott, teljesen kívül kerülve a lineáris média hagyományos keretein. Az intézményes megőrzés a legtöbb országban, így Magyarországon is mégis csupán a lineáris audiovizuális tartalmakra fókuszál. Jól dokumentáltan nyomon követhető, hogy a közösségi, illetve egyéb online médiatartalmak archiválásának feltételei általában nem csupán a magyar gyakorlatban, hanem nemzetközi szinten is teljesen

hiányoznak, szakmai, szabályozási, finanszírozási szinten is (Halbert, Skinner, Wilson, & Zarndt, 2016; Lee, 2017; Thomson & Kilbride, 2015).

A nyilvános internet felületének is különféle nyilvánossági szintjei vannak. A könyvtári megőrzés szintjén sajnos nem lehet mérlegelni azt, hogy a nyilvánosságra hozatal milyen célközönség számára, milyen céllal készült. Különösen igaz ez a közösségi médián keresztül közzétett bejegyzésekre, ahol eddig soha nem látott módon mosódik össze a magánjellegű publikációs felület a nyilvános publikációs formákkal. Ryan Wichtowski (Wichtowski, 2017) szolgál ehhez fontos adalékokkal. A szerzői jogi státusza is igen kérdéses ezeknek a publikálási tevékenységeknek, különös tekintettel az EU GDPR rendeletére. Alejandra Michel ad átfogó képet uniós jogi szempontból e területről (Michel, 2019). Sokféle helyről linkelünk, ágyazunk be, használunk fel különféle tartalmi elemeket, melyek szerzői jogi állapota egyedi szinten szinte nyomon követhetetlen. Az archiváláskor viszont hatványozottan előkerül a szerzői jogi állapot kérdése, a jogtisztázás pedig fokozott nehézségekbe ütközik. A digitális szerzői jogi környezet kapcsán számot kell vetnünk az amerikai és európai jogrendszer közti különbséggel is, ennek kapcsán Joseph P. Liu (Liu, 2016) elemzése mindkét környezetben releváns megállapításokat tartalmaz.

A kutatás, illetve a gyakorlati munka során arra a lényeges következtetésre jutottam, hogy a különféle típusú tartalmak begyűjtése nehezen képzelhető el egy széleskörű együttműködésen alapuló hálózat megszervezése nélkül, amelynek tagjai a közgyűjtemények mellett a jelentős internet-, illetve tartalomszolgáltató szereplők is. Nélkülük nagyon nehéz azokat a releváns webes tartalmi elemeket begyűjteni,

melyeket a közgyűjtemények a gyűjtőkörük részeként meghatároznak. Ráadásul, mint később is szó lesz róla még, nagy szükség lenne a W3C konzorcium által rögzített akadálymentes weboldal szabályrendszerhez (W3C Consortium, 2019) hasonlóan az archívumbarát, illetve robotbarát honlap kritériumokat („Archive-friendly website – MIA”, 2020; „Crawler-friendly website – MIA”, 2020) is megalkotni és széleskörűen meghonosítani. Maguk a fogalmak széleskörűen használtak, mint látható a magyar webarchiválási projekt wiki oldalain is meghatározásra kerültek, de az említett szabványos keretek fájóan hiányoznak. Az elméletben megfogalmazandó előírások gyakorlatba történő átültetése pedig tartalomszolgáltatók partnersége nélkül, meglátásom szerint, illúzió marad csupán.

De ezek a kihívások voltaképpen már a webarchiválás közgyűjteményi keretei tárgyköréhez tartoznak.

Tézisek

[T1.] Kulturális örökségünknek egyre nagyobb szelete a számítógépes világhálón tűnik fel. Ennek begyűjtése, feldolgozása, hosszútávú megőrzése összetett kihívást jelent. E kihívás kezelése számos fontos összetevővel bír az információtudományi, informatikai, társadalomtudományi és bölcsészeti tudományos területeken.

[T2] Amennyiben a mindennapi élet különféle kereteinek webes lenyomatait a közgyűjteményekben nem tudjuk megőrizni és bemutatni, ezen intézményrendszer fokozatosan képtelenné válik a kultúra átörökítésére, a kortárs társadalmi problémákra történő reflexió

képességét is elveszíti. Végző soron tehát a legitimitása kérdőjeleződik meg. A webarchiválással tehát igen súlyos társadalmi jelentősége miatt is tudományos szinten kell foglalkozni.

[T3] Magyarországon jelentős időbeli késéssel indult el a szervezett keretek között zajló webarchiválás a nyugat-európai centrumországokhoz képest. A webarchiváláshoz kapcsolódó tudományos kihívások felvázolása, a nemzetközi kutatás-fejlesztési eredmények implementálása a magyar webarchiválás gyakorlati tevékenységi keretei közé időszerű feladatot jelent. A korszerű archiválási környezet megalapozásával jelentősen csökkenthető a lemaradás következményeinek negatív hatása.

[T4] A webarchiválás tágabb összefüggésrendszerének ábrázolása, az archivált anyagokra épülő kutatások különféle útjainak bemutatása, a webarchiválás itthoni korszerű gyakorlati keretei megteremtésének felvázolásával együtt, összetett tudományos igényű elemzés elvégzését teszi lehetővé. Önálló tudományos tevékenység keretében fontos felvázolni a jövőbe mutató kutatási irányokat is.

[T5] A webarchiváláshoz kötődő tevékenységek és elméleti kutatási irányok, keretek közgyűjteményi szempontú feldolgozására, a legújabb nemzetközi kutatási eredmények bemutatására, ötvözve a saját itthoni kutatói tapasztalatokkal, önálló tudományos munka formájában magyar nyelven korábban nem került sor. Új interdiszciplináris kutatási irány alapozható meg ezen hiányosságok pótlásával a magyar tudományos életben.

[T6] A webarchiválás mint tudományos témakör a magyar tudományos életben történő meghonosításának kritikus előfeltétele az egész kutatási területet átfogóan bemutató hiánypótló tudományos munka elkészítése, az itthoni és nemzetközi tudományos és gyakorlati jellegű erőfeszítések összefoglaló jellegű, elemző bemutatásával együtt. A közgyűjteményi szempontú elemző látásmód, az itthoni intézményes webarchiválás kezdeteihez tartozó kutatás-fejlesztési munka eredményeinek bemutatása, a vonatkozó oktatási és kutatási-hasznosítási háttér alapjainak felvázolásával együtt adhat lendületet az egyes webarchiválást érintő részterületek további kutatásához.

[T7] Az Országos Széchényi Könyvtár webkönyvtárosaként a szerző meg kívánja alapozni a webarchiválás általános elméleti kutatási hátterét, valamint tevékeny részt vállal az összes tárgyalt webarchiváláshoz kötődő munkafolyamat gyakorlati meghonosításában is.

[T8] A webarchiválás általános áttekintése mellett saját kutatási témák is megfogalmazásra kerülnek (például a szemantikus web és a webarchiválás összekapcsolódása révén). Folyamatosan fejlődő dinamikus tudományterületről van szó, ahol csupán pillanatnyi összkép felvázolására nyílik mód, az egyes résztémák alaposabb tanulmányozása folyamatos tudományos feladatot jelent majd a jövőben is.

1. A webarchiválás közgyűjteményi keretei

A közgyűjtemények szempontjából fontos állításom, hogy a webarchiválás semmiféleképpen nem jelenik meg radikálisan új feladatként. A digitális

könyvtári, levéltári, múzeumi munkafolyamatokat kiterjesztő, azokhoz szervesen illeszkedő tevékenységről van szó. Meglátásom szerint, mivel a kulturális örökség különféle területei új felületen jelennek meg, voltaképpen a hagyományos gyűjtőköri kereteket kell kiterjeszteni a webarchiválás irányában. Ez a fejezet ezekről a kiterjesztési lehetőségekről, a webarchiválás könyvtári, levéltári múzeumi felfogásáról, feladatairól ad áttekintést. Alapvető fontosságú a hipotézisekben megfogalmazott kérdésekre adott válaszok szempontjából is a webarchiválás teljes közgyűjteményi környezetének bemutatása, s hogy milyen értelmezésekben jelenik meg a webarchiválás feladatként az egyes közgyűjteményi ágak szintjén.

Mielőtt azonban erre sort kerítenénk, ki kell térni még arra is, hogy a webes kulturális örökség archiválásának védelmére, fontosságára, teendőire, a közgyűjteményi tevékenységek fontosságára több átfogó nemzetközi dokumentum, tanulmány is rámutat. Ezekből mutatok fel két markáns példát.

Az UNESCO a digitális kulturális örökség megőrzéséről szóló 2013-as chartájában (UNESCO, 2013) már leszögezi, hogy „A világ digitális örökségét a végleges eltűnés veszélye fenyegeti. Ehhez hozzájárul az azt tartalmazó hardware és software gyors elavulása, bizonytalanság a forrásokkal, felelősségekkel, és a fenntartás és megőrzés módszereivel kapcsolatban, valamint a támogató jogi szabályozás hiánya.” A közgyűjtemények felelősségéről pedig így szól: „A nemzeti megőrzési politikák kulcselemeként az archívumokkal kapcsolatos törvényhozásnak, és a könyvtárak, levéltárak, múzeumok, és más nyilvános

gyűjteményeknek fel kell ölelniük a digitális örökséget.” A következőkben tehát azt vizsgáljuk meg, hogy a webarchiválás miként illeszkedik ebbe a kívánatosnak tartott közgyűjteményi digitális megőrzési rendszerbe.

Miért veszíthetjük el a csatát a webarchiválás frontján? Erre a kérdésre keresi a választ a Stanford Egyetem könyvtárának vezető kutatója, David S. Rosenthal a digitális megőrzés napja alkalmából a digitális anyagok megőrzésével foglalkozó egyik legfontosabb civil szerveződés, a Digital Preservation Coalition blogján (Rosenthal, 2017). A szerző felhívja a figyelmet, hogy habár az intézményi erőfeszítések megsokszorozódtak az utolsó évtizedben a webarchiválás terén, a legsúlyosabb akadály az előrehaladásnak az érintett intézmények súlyos alulfinanszírozottsága. Axiómaként állapítja meg, hogy megfelelő pénzügyi erőforrások biztosításával nem lenne probléma az archiválással. Világszinten 30 millió dollárt költenek azon tartalmak archiválására, melyek előállítási költsége százmilliárd dollárokból mérhető!

Az aggasztó valóság másik szelete, hogy például a webarchiválási kutatás-fejlesztés egyik zászlóshajójának számító British Library költségvetése reálértékben az utóbbi tíz évben 45%-kal csökkent!!! A szerző (Rosenthal, 2017) arra is felhívja a figyelmet, hogy a tartalomelőállítási költségek, s így a hirdetési bevételek igénye is folyamatosan növekedő tendenciát mutatnak, s így ezzel párhuzamosan az archiválásra fordított összegeknek is növekedniük kellene. A probléma jelentőségére ennek kapcsán a cnn.com weboldal példájával világít rá. Erről az Internet Archive az utolsó mentését 2017. november elsején tudta elkészíteni. Azóta archiválhatatlan a weboldal, mivel a hatékonyabb működés érdekében

átalakították a grafikus megjelenítést szabályozó JavaScript kódokat. A JavaScript hatásáról az archiválhatóságra még részletesebb adalékokat kaphatunk Brunelle és munkatársainak tanulmányában (Brunelle, Kelly, Weigle, & Nelson, 2016). Lényeges állítás számomra tehát a szakirodalom tanulmányozása alapján, hogy a hatékonyabb szolgáltatás oldalán az archiválás igényei jelenleg tartalomszolgáltatói szinten bármikor bárki számára feláldozhatók.

A továbbiakban Rosenthal a gazdasági tényezőkre s azok technológiai hatására tér rá. A tartalomszolgáltatások költségét alapvetően előfizetésekből, illetve hirdetésekből fedezik. Mindkét bevételi oldal gyenge lábakon áll jelenleg a hirdetésblokkoló programok elterjedt használata, illetve az előfizetési hajlandóság csekély volta miatt. Erre válaszul a tartalomszolgáltató lobbi hatalmas nyomás alá helyezte a W3C szabványalkotó testületet, hogy a digitális másolásvédelem integrálódjon bele a html szabványba, illetve más oldaltól szabványszinten akadályozzák a hirdetésblokkoló programok használatát. Ezek a törekvések, ha megvalósulnának, akkor a számítógépes világháló jó részét egyszerűen használhatatlanná tennék (Rosenthal, 2017).

A végén pedig felmerülhet a kérdés, hogy miben segíthetünk egyszerű állampolgári szinten a webarchiválás ügyén? Erre Rosenthal szerint több mód is adódik. Az Internet Archive Save Page Now szolgáltatásával saját magunk kezdeményezhetjük az általunk hasznosnak tartott oldalak mentését. Az Internet Archive működését pénzügyi adománnyal, illetve tartalmak archiválásra történő felajánlásával is lehet segíteni. Győződjünk meg arról, hogy az adott ország nemzeti könyvtára fontosnak tartja a

világháló archiválását, aktív tevékenységet folytat e téren, s támogassuk erőfeszítéseiket (nálunk az OSZK kísérleti projektje révén vált önálló szolgáltatássá mostanság ez a tevékenység). Civil összefogással széles körű tiltakozás megszervezésével akadályozzuk meg, hogy bármiféle W3C szabványba bekerüljön a digitális másolásvédelem kiterjesztése (Rosenthal, 2017). A tanulmány megállapításai kiegészítődnek még egyéb szerzői jogi kihívásokkal is, mint például annak a veszélyével, hogy az európai szellemi tulajdonjogi szabályozás reformja elérhetetlenné teszi az amerikai Internet Archiveban a webről archivált állomány egy lényeges hányadát, vagy akár a teljességét, mivel a szerzői jogi státusza a közzétett anyagoknak számos esetben nem tisztázott (Rosenthal, 2017). A kutatásnak ebből az eleméből tehát az a fontos állítás adódik, hogy a jogi szabályozás is alapvető jelentőségű a webarchiválás közgyűjteményi megalapozásának szempontjából. A jogi szabályozás kérdéseire egy későbbi fejezetben térek ki kissé részletesebben majd.

1.1 A webarchiválás a digitális könyvtári szolgáltatások rendszerében

Felmerült a tézisekben, hogy egyáltalán meg lehet-e és milyen módon a webarchiválást könyvtári feladatként határozni. Ennek kapcsán arra a megállapításra jutottam, hogy a könyvtárakban végzett webarchiválás gyűjtőköre alapvetően a világhálón nyilvánosan publikált webes tartalomra terjed ki. A webarchívum mint gyűjtemény mozaikszerűen illeszkedik be a digitális könyvtári szolgáltatások környezetébe.



3. ábra A webarchiválás helye a digitális könyvtári szolgáltatások rendszerében

A köteleespéldány szabályok kiterjesztésével egyre inkább lehetővé válik a webnek az egyes országokhoz, nemzeti kultúrákhoz kötődő szeletei archiválásának előírása is. Erre több nemzeti webarchívum (brit, dán, svájci) munkatársai is utalnak, Michaela Mayr és Andreas Predikaka pedig osztrák részről kifejezetten hangsúlyozza a nemzeti szempontok érvényesülését (Bingham, 2014; Mayr & Predikaka, 2016; Schostag & Fonss-Jorgensen, 2012; Signori, 2017). Mint már korábban utaltam rá, meglátásom szerint a hagyományos passzív befogadói szerephez képest a könyvtáraknak sokkal proaktívabb szerepet kell játszaniuk a gyűjtőkörükbe eső webes (sőt tágabb értelemben a digitális) tartalmak begyűjtésében. Miután bárki előállíthat online tartalmakat, aki nincs is tudatában annak, hogy rá is vonatkozhat köteleespéldány rendelkezés, ezért egy jól szervezett könyvtári hálózat, a nemzeti könyvtárral az élén sokat tehet a gyűjtőkörileg leginkább releváns témakörök kapcsán a tartalmak feltérképezéséért és begyűjtéséért. Már hosszú évek óta nemzetközi szinten evidenciaként fogalmazódik meg, hogy azt az

elképzelést, ami a hagyományos kötelespéldány rendelkezések mögött húzódik meg, hogy a teljesség igényével szolgáltatassák be a könyv és sajtótermékeket, a web esetében fel kell adni (Schostag & Fonss-Jorgensen, 2012; Stirling és mtsai., 2012). Mint az idézett dán és francia nemzeti könyvtári példák és a bevezetőben megfogalmazottak is utalnak rá, fogalmunk sincs róla, miből áll voltaképpen a teljesség a világhálón, valamint eszközünk sincs rá, hogy megragadjuk azt. Fontos tanulság, amivel a webarchiválás itthoni gyakorlatának kialakítása kapcsán is szembesültem, hogy egy könyvtár csak arról a webes információforrásról tud értesülni, ami a látókörébe kerül, vagy amiről tájékoztatják.

A saját nemzeti könyvtári webarchívum tervezésekor, a nemzetközi gyakorlatból is adódó tanulságokat leszűrve, hamar tudatosítanom kellett magamban, hogy alapvetően kétféle válogatási elvről beszélhetünk. A tematikus gyűjtemények összeállításáról (ez lehet egy adott téma, illetve esemény, történet, rendezvény is), illetve pillanatfelvételek készítéséről a webnek a .hu domain alatti (illetve egyéb legfelső szintű domaineken található magyar nyelvű) anyagairól, ami tematikus aratásnál jóval felületesebb, áttekinthető jelleggel valósul meg.

A tematikus gyűjtemények kialakításánál meglátásom szerint alapvető fontosságú a rendelkezésre álló keretrendszer kapacitása, amely országonként, intézményenként eltér egymástól. A hardveres és szoftveres korlátok, a rendelkezésre álló munkaerő nagysága, valamint a rendelkezésünkre álló tárhely adta keretek között minél részletesebben archiválni a tartalmat, s visszakereshetővé szeretnénk tenni. A válogatott gyűjtemények anyaga a digitális gyűjtemények egyéb szeleteihez

hasonlóan könyvtári feldolgozásra kerül. A megfelelő azonosítóval és metaadatokkal a válogatott webes állomány megjelenhet az online katalógusban, sőt a nemzeti bibliográfia részévé is válhat, melynek számos megvalósítását is esettanulmányszerűen publikálták már (Lupovici, 2006; Klein, Balakireva, & de Sompel, 2018; Schostag & Fonss-Jorgensen, 2012; Stirling és mtsai., 2012). Nagyon fontos hozzáadott érték az archivált webes anyagok stabil hivatkozhatóságának biztosítása is. Ennek jövőbeni lehetőségeit tekinti át Robin Camille Davis (Davis, 2016) tanulmánya. Ez nem is olyan egyszerű a gyakorlatban, különösen egyes közösségi média tartalelemek esetében, ami a könyvtári jellegű archiválás mellett a személyes tartalmaink archiválására még inkább hatással lehet, ahogy erre Paul-Choudhury utal (Paul-Choudhury, 2011).

Az általános aratások esetében viszont egy reprezentatív összkép felmutatása a cél, hogy egy adott időszakban hogyan nézett ki a webnek egy adott részterülete. Jelen eszköztárunkkal és technikai lehetőségeinkkel ezeknek a mentéseknek nem lehetséges a hagyományos könyvtári szempontú feldolgozása, visszakereshetőségének biztosítása. A honlapok URL címe szerinti keresése érhető csupán el.

A könyvtárak által archivált webes anyagokhoz történő hozzáférésnek szigorú szerzői jogi szabályai vannak. Az archiváláshoz az általános nemzetközi gyakorlat szerint nem kell hozzájárulás az adott tartalom szerzői jogtulajdonosától, bár erre is van ellenpélda, például Hollandia esetében, melyről Barbara Sierman és Kees Teszelszky számolnak be (Sierman & Teszelszky, 2017). Nyilvánosan viszont csak olyan tartalmat lehet szolgáltatni, melyre az intézmény engedélyt kap. Abban már

országoként eltérő az általános gyakorlat, hogy az adott könyvtár épületén belül érhető-e el a jogvédett anyag, vagy pedig zárt dedikált könyvtári hálózatban is. A szervezett magyar webarchiválás kezdeteit, az annak alapjául szolgáló nemzetközi könyvtári webarchiválás általános trendjeit és különféle nemzeti esettanulmányait külön fejezetekben tárgyalom majd.

1.2 A webarchiválás mint digitális levéltári feladat

A levéltári webarchiválás tekinthető véleményem szerint a szó szoros értelmében vett webarchiválási tevékenységnek, bár kiterjedtsége, nagyságrendje jellemzően elmarad a könyvtárakban végzett webarchiválási tevékenység mögött (Drótos & Németh, 2018b). A könyvtári és a webarchiválási tevékenységek közötti legélesebb választóvonalat a gyűjtőkörben és az archiválási módszerekben lehetjük fel, mint ahogy hagyományosan is ezek a jellemzők különítik el a két intézménytípust egymástól. A könyvtári területhez hasonlóan itt is a hagyományos tevékenységek új szempontok szerint történő kiterjesztéséről van tehát szó véleményem szerint, a web, mint információhordozó médium megjelenésével. A levéltárak hagyományos szerepkörükből adódóan elsősorban a nem a széles nyilvánosságnak szánt publikálási szándékkal létrehozott webes tartalmakat mentik, például egy adott intézmény vonatkozásában, melyre brit és holland példákat idézünk Andy Jackson, illetve Suzi Szabo révén (Jackson, 2015; Szabo, 2018). Az intézményi kommunikáció fontos színtereivé váltak például a fórumok, személyes blogok, e-mailek, fotók, videók, személyes webkettes oldalak és csatornák, csevegőprogramok forgalma, zárt csoportok és webhelyek.

Ám a hagyományos levéltári tevékenységhez talán a leginkább illeszkedő gyűjtőkori elem az intézményes és céges belső honlapok és iratok kezelése. Ezt a gyűjtőmunkát igazából el sem lehet választani az adott szervezetre vonatkozó általános levéltári gyűjtési tevékenységektől, azoknak egy szeleteként jelenik meg. A begyűjtött anyagot a levéltári feldolgozási rend szerint digitális fondokba szervezik, az adott szervezetre, személyre, témakörre vonatkozó többi digitális tartalommal együtt.

Némi párhuzamosság is megfigyelhető a könyvtári és levéltári webarchiválási tevékenységek között a kormányzati és államigazgatási honlapok mentésével kapcsolatban. Az idézett forrásokból kiviláglik, hogy nincs általános nemzetközi gyakorlat arra, hogy ezt az anyagot könyvtári (pl. parlamenti könyvtári) vagy levéltári (pl. nemzeti levéltári) keretek között, illetve könyvtári vagy levéltári módszerek szerint dolgozzák- e fel, s hogy bekerül-e az anyag a kötelespéldány szolgáltatás körébe (Espley, Carpentier, Pop, & Medjkoune, 2014; „Government Web& Data Archive”, 2017; Grimshaw, 2016; Holub & Rudomino, 2015; Niu, 2012). Számos országban a digitális állományok archiválására és megőrzésére a teljes államigazgatás körét átfogó kormányzati stratégia készül (Grimshaw, 2016; Paterson, 2015; Storrar & Seaman, 2016). Ez magában foglalja a gyűjtőkör pontos szabályozását, az egyes intézmények általános teendőit, illetve az archiválásért felelős intézmények tevékenységének alapszabályait. A későbbi történeti kutatások számára alapvető fontosságú, hogy az állam, a közigazgatás, az államigazgatás szervei és intézményei miképpen jelentek meg a virtuális térben, milyen online iratokat keletkeztettek. Ahol a levéltárak webarchiválást végeznek, ez a

fajta megőrzési munka áll a digitális webarchiválási tevékenység fő középpontjában, a gyűjtőkör elemei csupán kiegészítő jellegűek. Ebben sincs nagy újdonság, a hagyományos levéltárosi tevékenységek prioritásai is e szerint alakulnak. Például kínálkozik a megőrzés fontosságára az az önálló kormányzati webarchívum szolgáltatás, amelyet a Brit Nemzeti Levéltár üzemeltet 1996 óta (National Archives, 1996). A szolgáltatás így szinte egyidős a legáltalánosabb jellegű webarchiválás tevékenységet végző amerikai Internet Archive-val. Ennek a webarchívumnak hatóköre nem csak a szokásos fent említett államigazgatási webes dokumentumkörre terjed ki, hanem mentik a különféle intézmények Twitteren, Flickr-en, egyéb közösségi média és videomegosztó csatornákon közzé tett információit is. Ám, mint említettem, egyáltalán nem törvényszerű, hogy az államigazgatás online anyagait levéltári keretek között archiválják, megjelenhet az a könyvtári gyűjtőkör részeként is. Arra is van példa, hogy széleskörű együttműködéssel kampányszerűen egy-egy elnöki terminus lezárulta körül gyűjtik egy-egy időszakra vonatkozóan a kormányzati weboldalakat, ez megfigyelhető például az USA-ban a Government Web & Data Archive project kapcsán („Government Web& Data Archive”, 2017; Phillips & Phillips, 2018).

Sajnos Magyarországon levéltári webarchiválás jelenleg nem zajlik, és az államigazgatási, közigazgatási webarchiválásnak is hiányoznak a stratégiai keretei. A 2021-ben hatályba lépett jogi szabályozás az OSZK feladatává teszi a tevékenységet azzal együtt, hogy az állami és önkormányzati szervek, valamint általuk fenntartott intézmények archivált webhelyeit felhasználói engedélykérés nélkül lehet szolgáltatni. Az Országos

Széchenyi Könyvtárban a kísérleti projekt kereteinek meghatározásakor munkatársaimmal együtt arra a meggyőződésre jutottunk, hogy a központi államigazgatás speciális jellegéből adódóan mindenképp szerencsés lenne, ha ehhez a szférához szorosan kötődő, az egyéb típusú államigazgatási anyagok archiválásában szakértelemmel bíró intézmény vállalná velünk együtt ezt a feladatot. Partnerünk lehetne így a Magyar Nemzeti Levéltár, illetve az Országgyűlési Könyvtár is. Újabb példáját láthatjuk tehát az együttműködés fontosságának, hogy milyen fontos a teljes közgyűjteményi összképet a kutatás során feltárni. A nyilvános szolgáltatás viszont segíthet esetleg majd annak áthidalásában, hogy hozzá lehessen illeszteni egy adott közigazgatási intézmény archivált tartalmainak nagyját a más intézményben tárolt webes tartalmakhoz.

1.3 A webarchiválás digitális muzeológiai vonatkozásai

A webarchiválás múzeumi szegmense gyűjtőkörét tekintve erőteljesen eltér a könyvtári, illetve levéltári szegmenstől. A középpontban itt a művészeti, illetve iparművészeti tevékenység webes lenyomatainak megőrzése, illetve megjeleníthető állapotban tartása áll (Drótos & Németh, 2018b). A web, mint utaltam arra korábban már, önmagában rendkívül sérülékeny, s fokozottan igaz ez a webes művészeti alkotásokra. Ezek általában szorosan kötődnek egy adott korszak internetes technológiai elemeihez, a szolgáltatási környezet átalakulásával, elavulásával könnyen elérhetetlenné válhatnak. Épp emiatt az elavuló környezet miatt a megtekinthető állapotban tartás, a megfelelő hardveres, szoftveres emulációs lehetőségek megtalálása is komplex feladatokat állít elénk (Drótos & Németh, 2018b).

A múzeumi webarchiválás két fő szegmensét erre tekintettel véleményem szerint az alábbi módokon határozhatjuk meg. Az első az egyes weben alkotó művészek életművének minél szélesebb körű felkutatása és archiválása. A második pedig az adott múzeum gyűjtőkörének megfelelő szelektív webarchiválási tevékenység. Ennek határai rendkívül szélesek, s igazodnak a múzeumi intézményi terület színes palettájához (Drótos & Németh, 2018b).

Például helytörténeti gyűjteményekben a helytörténeti események webes lenyomata alkothatja a gyűjtőkört, szövegben, hangban, képben egyaránt. A hadtörténeti múzeumok, gyűjtemények érdeklődésének középpontjában a gyűjtőkörükbe eső múltbeli események webes megjelenése, illetve a kifejezetten a kibertérben zajló háborús események lenyomatainak gyűjtése állhat. A sportmúzeumok webes gyűjteménye kiterjedhet a hagyományos sport webes megjelenése mellett az e-sport történetének, eseményeinek dokumentálására is. A fotómúzeumok teret adhatnak a digitális fotóművészet alkotásainak, illetve e művészeti ág történetének webes dokumentálásának. A kereskedelmi múzeumok az egyre bővülő online kereskedelem megjelenési formáit követhetik a kapcsolódó online dokumentumok archiválásával együtt. A műszaki múzeumok pedig az internetes technológiai fejlődésének webes dokumentumait gyűjthetik össze (Drótos & Németh, 2018b).

Magyarországon az online művészet lenyomatai őrzésének van mintegy húsz évre visszanyúló jelentős múltja, mint arról Peternák Miklós, a C3 Kulturális Központ vezetője a kísérleti webarchívum első workshopjának előadásán számot ad (Peternák, 2017). Ez az előadás azzal is szembesít

minket, hogy a webes művészetek múltjára vonatkozóan megfelelő forrásanyag birtokában megalapozott művészettörténeti kutatások is folytathatók. De maguknak a forrásoknak az eltűnése, az egykorvolt múlt nyomainak kutatása is különleges kihívásokat rejt magában. Ez a fajta tevékenység persze ezer szálon összefügg a mostanság formálódó webtörténettel (Kokas & Drótos, 2018), ami külön történeti segédtudományként is meghatározható, s melyről a webarchívumok tudományos hasznosításáról szóló fejezetben lesz részletesebben szó majd.

A Ludwig Múzeum külön kiállítást szentelt az eltűnő internet témakörének, annak, hogy milyen lenyomata maradhat mégis a webes világban a hagyományos tárgyi kultúrát tekintve („THE DEAD WEB – THE END | online kiállítás | Ludwig Múzeum”, 2020).

Sajnos a magyar digitális múzeumi terület különféle tevékenységeinek, a fent említett látványos kivételtől eltekintve, a webarchiválás nem képezi részét. A múzeumok jelenleg nem tartják feladatuknak, hogy a webes kulturális örökség gyűjtőkörükbe eső részét feltárják, dokumentálják, archiválják, ezzel pedig a jelenünk egy fontos szelete vesz el a jövő számára. Reméljük, hogy a jövőben egyre több múzeumi partnert is be lehet majd vonni a muzeológiai alapú webarchiválási tevékenységbe. Ebben segítséget nyújthat talán ez az áttekintő dolgozat is.

1.4 Az audiovizuális archívumok és a webarchiválás

A világ számos országában, így Magyarországon is, a film, illetve a lineáris média archiválásának keretei már régóta adóttak. A médiakörkép

azonban az utóbbi évtizedekben jelentősen kibővült. Ma már a hagyományos lineáris tartalomszolgáltatással egyenrangúvá lett a különféle közösségi videomegosztó platformok által szolgáltatott tartalom. Ez a folyamat az utóbbi években publikációs szinten is szépen nyomon követhető (Acker & Kreisberg, 2019; Elson, 2020; Espley és mtsai., 2014; Thomson & Kilbride, 2015). Ejtettünk már szót arról, hogy fontos jellemzője az online tartalomszolgáltatásnak a belépési küszöb hiánya. A fentebb említett közösségi platformok (pl. YouTube, Facebook) megjelenésével már bárki könnyen megjelenhet az audiovizuális virtuális térben is, s e tartalmak szerves elemévé váltak a populáris kultúrának. Mint a brit kormányzati webarchívum esetében (National Archives, 1996) láttuk, immár az államigazgatás (de általában a társadalmi, kulturális intézmények) társadalmi kommunikációjának is egyre megkerülhetetlenebb elemét alkotják. De maguknak a lineáris médiaszolgáltatóknak is markáns online jelenléte van a közösségi platformokon is. Külön kategóriát jelentenek az audio, illetve video podcastok, melyek sajátos átmenetet képeznek a lineáris, illetve az online műsorszórás világa között, hiszen a virtuális térbe sugárzott, letölthető, offline is fogyasztható médiatartalmakról van szó. Világossá kell, hogy váljon mindenki számára, hogy a világháló ezen szelete is fontos külön szeletét alkotja kulturális örökségünknek. Véleményem szerint az online audiovizuális tartalmak archiválása természetes kiegészítő tevékenységeként kell, hogy megjelenjen a hagyományos médiaarchiválási tevékenységeknek. Másfelől viszont a kutatás során világossá vált számomra, hogy a feladat, illetve a dokumentumtípus speciális jellege miatt ezt a feladatot a könyvtárak, levéltárak vagy

múzeumok általános webarchiválási gyűjtőköri feladatként nem is tudják felvállalni. Van néhány olyan ország, például Dánia (Schostag & Fonss-Jorgensen, 2012) ahol a gyűjtőköri és intézményi keretek megfelelően rögzítésre kerültek, sőt Sabine Schostag felhívta a figyelmet arra, hogy külön útmutató készül például YouTube anyagok archiválására (Schostag, 2019a). Ott szerencsés felállás, hogy a médiaarchívum is a nemzeti könyvtár része, így egységes keretek között tudják kezelni a különféle digitális megőrzési tevékenységeket. Nálunk a 2021. január elsejétől hatályos jogi szabályozás az audiovizuális területet épp annak különleges volta miatt kiemelte az általános webarchiválás szabályozási keretéből, s külön jövőbeni szabályozással a Nemzeti Audiovizuális Archívumhoz (NAVA) rendelné ezt a feladatot. A megfelelő szakmai, infrastrukturális és finanszírozási feltételek megteremtése azonban elengedhetetlen ahhoz, hogy a NAVA tevékenysége erre az intézmény számára természetesnek tűnő irányba is kibővüljön.

E fejezetben már többször szóba kerültek nemzetközi szolgáltatási példák, jó gyakorlatok. A webarchiválás közgyűjteményi alapjainak felvázolását követően tehát e megőrzési tevékenység nemzetközi dimenziója kerül a következőkben napirendre.

2. A webarchiválás nemzetközi háttere

A Magyar Internet Archívum pilot projektjének formálódása közben, egyben a saját webarchiválással kapcsolatos kutatások megalapozásaként is elsődleges feladat volt feltérképezni az internet archiválás főbb nemzetközi szereplőit, tanulmányoztam a különböző nemzeti

webarchiválási modelleket a hasznosítható jó gyakorlatok szempontjából. A projekt előrehaladása közben az OSZK csatlakozni tudott a webarchiválásban érdekelt szereplőket összefogó International Internet Preservation Consortiumhoz (IIPC, 2019), aktívan bekapcsolódtam több munkacsoport tevékenységébe is. Emellett felvettem a kapcsolatot közép-európai partnerintézményekkel későbbi intenzívebb együttműködési formák megalapozása céljából. Ez a fejezet tehát a nemzetközi érintettségű kutatómunka keresztmetszetét tekinti át. A válogatás szempontja Európa-központúságot tükröz, elsősorban abból a szempontból válogattam, hogy honnan lehet a leginkább jó gyakorlatokat átvenni a magyar projekt számára.

2.1 A webarchiválás főbb nemzetközi intézményi szereplői és az együttműködés keretei

Az online világ kezdeti tartalomszolgáltatásaiból (pl. BBS, WAIS, Gopher) nem sok minden maradt meg, de a World Wide Web első publikus webszervere (restaurálás után) még mindig megnézhető az info.cern.ch címen (Drótos és mtsai., 2018). Négy évvel ennek megszületése után, 1996-ban megalakult az Internet Archive (Internet Archive, 2019) nevű nonprofit szervezet azzal a céllal, hogy létrehozza a digitális világ alexandriai könyvtárát, egyebek mellett mára már 380 milliárdnál is több lementett weboldallal, melyek a Wayback Machine szolgáltatásán keresztül úgy böngészhetők, mintha az élő webet használnánk. Sokszínű együttműködést folytatnak az egyes nemzeti keretrendszerek képviselőivel. Az egyes nemzeti modellek (könyvtárakban, levéltárakban, egyetemeken, kutatóintézetekben) tulajdonképpen arra válaszul jöttek

létre, hogy a web fejlődése egyetlen szervezet számára már követhetlenné vált archiválási szemszögből. Ezek egy része eleve csak korlátozott időre szólt, másokat pedig egy idő után újragondoltak és -terveztek, így már több második-generációs webarchívummal is lehet találkozni. A nemzeti szereplők, illetve az Internet Archive egymást kiegészítő tevékenységei adják meg a webarchiválási szolgáltatások teljességét. Jelenleg mintegy 45 nemzeti webarchívum létezik valamivel több mint harminc országban, ugyanis egyes helyeken az államalkotó nemzeti csoportoknak külön archívumuk van (Németh, 2017). A nemzeti webarchiválási tevékenységeket pedig az IIPC nemzetközi konzorcium (IIPC, 2019) fogja közös keretbe, melynek az Országos Széchényi Könyvtár is 2018-tól a tagja lett.



4. ábra Az IIPC tagjainak földrajzi eloszlása (forrás: <https://netpreserve.org/about-us/members/>)

Ennek révén hatékony tudásmegosztási, kompetenciafejlesztési és szakmai együttműködési csatornák válnak elérhetővé számunkra is. A konzorcium 2003-ban alakult meg a Francia Nemzeti Könyvtár kezdeményezésére 12 alapító taggal. Jelenleg a 45 országból összeálló tagság könyvtárakat, levéltárakat, múzeumokat, illetve a webarchiválás területén érdekelt civil és piaci szereplőket is magában foglal. A konzorcium alapvető küldetése az internetes tartalmak archiválásához és hosszú távú biztonságos megőrzéséhez szükséges tevékenységek koordinálása és támogatása. Az IIPC tevékenységét a British Library-ből koordinálják. Külön figyelmet érdemel a 2019-2020-ban létrehozott repozitórium, melynek keretében fokozatosan teszik közzé az IIPC szakmai rendezvényein elhangzott előadások prezentációit, kutatási jelentéseket, cikkeket. Ennek használata szinte azonnal megkerülhetetlenné vált a webarchiválás tanulmányozásához. Az aktuális kutatási témák jóval rövidebb átfutási idővel bukkannak itt fel, mielőtt tudományos publikációkban jelennének meg. Képet kaphatunk arról, hogy mely témák állnak a gyűjteményfejlesztéssel, technikai fejlesztésekkel, illetve a webarchívumok kutatási célú felhasználásával foglalkozó szakemberek érdeklődésének középpontjában („International Internet Preservation Consortium (IIPC) repository”, 2020).

A nemzetközi háttér kutatása során az amerikai és az európai webarchiválás modelljei között jelentős különbségeket tapasztaltam. Európában alapvetően nemzeti közgyűjteményekre jutottak a webarchiválás feladatai, néhány egyetem, kutatóintézet csatlakozásával. Az Egyesült Államokban a legtöbb intézmény, legyen az közgyűjtemény,

egyetem vagy kutatóintézet csupán webkurátori tevékenységet végez, a webarchiválási feladatok technikai lebonyolítását átadják az Internet Archive kereskedelmi üzletágaként működő Archive-IT nevű vállalkozásnak (Leach-Murray, 2018). Ez a cég részletesen paramétereztető szoftveres eszközöket bocsát a szolgáltatásaira előfizető intézmények rendelkezésére. Alaptevékenysége csupán az archiválás lebonyolítására korlátozódik, de számos kisebb előfizető intézmény megbízza a tárolási kapacitás rendelkezésre bocsátásával is („Archive-IT—Web Archiving Services for Libraries and Archives”, 2020). Nagyon fontos megjegyezni azonban, hogy az Archive-IT által learatott archív anyaggal teljes mértékben az előfizető rendelkezik („Archive-IT—Web Archiving Services for Libraries and Archives”, 2020). A nagyobb egyetemek, közgyűjtemények, kutatóintézetek pedig maguk építik ki digitális repozitóriumi rendszereiket, s ebbe integrálják bele a webes állomány megőrzését is. Az IIPC rendezvényeken sokszor hallottam olyan véleményeket amerikai kutatók, webkurátorok részéről, hogy nagyon hálásak, hogy a webarchiválás technikai feladatai lekerülnek a vállukról. Másrészt viszont azért fájjalják, hogy nincsenek olyan szakemberek a soraikban, akik az európai és egyéb kontinensekről érkező kollégák partnereiként technikai kérdésekben, a fejlesztési irányvonalak formálásában kompetens módon meg tudnának nyilvánulni. Legutóbb 2020 végén a belga kollégáktól értesültem arról, hogy az ottani nemzeti könyvtár sem akar saját informatikai infrastruktúrát kialakítani webarchiválási célokra, hanem az Internet Archive-val kívánnak szerződést kötni. Mindenesetre lényeges különbség ez a könyvtári webarchiválási gyakorlatot tekintve, hogy a technikai háttérrel kiszervezik,

s csupán a gyűjteményfejlesztés, szervezés marad egy intézmény hatásköre, vagy a teljes infrastruktúrát is egy adott intézmény bázisán üzemeltetik.

2.2 A webarchiválás különféle nemzeti gyakorlatai – esettanulmányok

Ez a fejezet alapvetően azon a kutatómunka eredményein alapszik, melynek révén a magyarországi pilot projekt első fázisában mértem fel nemzeti webarchiválási jó gyakorlatokat. A munka eredményeként összefoglaló jellegű publikáció (Németh, 2017) is készült. A publikáció elkészítéséhez korábbi tanulmányok is segítségül szolgáltak Hegyközi Ilona és Kornhoffer Mónika révén (Hegyközi, 2014; Kornhoffer, 2011). A kísérleti projekt részeként lezajlott kutatómunka eredményeit néhány aktuális megjegyzéssel egészítettem ki. A fejezet elkészítése során fokozottan tiszteletben kellett tartanom a tartalmi aránytalanságok elkerülését. Jóval több ország nemzeti gyakorlatáról is áttekintést lehetne adni, melyre a kutatás során szintén képbe kerültek, a válogatásnál elsősorban a magyar webarchiválási projekthez történő kapcsolódási pontokat tartottam szem előtt.

Elöljáróban az egyes nemzeti modellek rövid ismertetése előtt, érdemes leszögezнем, hogy a nemzeti könyvtáraknak szinte minden vizsgált esetben központi szerepköre van a webarchiválási szolgáltatások szervezésében. E tevékenységek szervezeti keretei azonban rendkívül változatosak. A siker biztos garanciájának a tartalomszolgáltatók és közgyűjtemények közötti széleskörű összefogás tűnik. A webarchiválás közös ügy, nem egyetlen intézmény felelősségi körébe tartozó

szolgáltatási feladat. Az adott ország közigazgatási szerkezete, a közművelődési intézmény struktúra jellege alapvetően meghatározza a webarchiválási tevékenység szervezeti kereteit (Németh, 2017). Másrészt érdemes visszautalni néhány korábbi megállapításra, hogy a webtechnológia robbanásszerű folyamatos fejlődésével egyre nagyobb kihívást jelent a különféle webes tartalomszolgáltatások archiválhatóságának biztosítása, mind a tartalomszolgáltatók, mind az archiváló intézmények részére. Ígéretes fejlődési irányt jelent, hogy a tartalmakat egyre nehezebben begyűjteni képes aratórobotok helyett a böngészőprogramok működését imitáló, a böngészőmotoron alapuló új szoftverek jelennek meg, ahol már a felhasználó képernyőjén megjelenő tartalom rögzítése válik céljá, ilyen pl. a Brozzler nevű szoftver, melyet a Francia Nemzeti Könyvtár munkatársa mutat be (Sylvie-Rollason-Cass, 2020). Ez részleges paradigmaváltást is hozhat akár a jövőben, a webarchiválás területén alapvetően átalakítva a továbbiakban ismertetett nemzeti modellek gyakorlati kereteit.

A másik jelentős általános nemzetközi kihívás a webarchiválás jogi, szabályozási háttérében mutatkozik meg. Az Internet Archive 1996 óta rengeteg egyedi projekten keresztül gyarapította állományát a tervszerű gyűjtés mellett, ezeknek a jogi feltételeit egyenként szabályozták. Az általános gyűjtéssel bekerült anyag szolgáltatási háttere jogilag pedig teljesen tisztázatlan (Németh, 2017). Ez akkor jelent problémát például, ha valaki (pl. az OSZK részéről, ahogy mi megtettük a munkatársaimmal) kéréssel fordul hozzájuk, hogy egy meghatározott állományrészt szeretne illeszteni a saját szolgáltatásába, el szeretne vinni a saját gyűjteményébe.

A másik jelentős kihívás, hogy míg az amerikai vagy ausztrál jogrendszerben nem követelik meg az előzetes engedélykérést az archivált webes anyagok szolgáltatásához, Európában erre szigorú szabályok vonatkoznak. Mint már említettem, ez kétségesse teszi, hogy jelen formájában meddig lehet szabadon hozzáférni kontinensünk felől az Internet Archive állományához. Amerikában is kiéleződött azonban a kiadókkal való konfliktus, miután a koronavírus járvány miatt az Internet Archive nagymennyiségű jogvédett digitális művet tett közzé, külön gyűjtemény keretei között, a felmerülő jogi vitákról Andrew Albanese adott rövid áttekintést (Albanese, 2020). Az egyes esettanulmányokban ábrázolom, hogy alapvetően a webarchiválás jogi háttere az elektronikus kötelezpéldány rendszerébe simul bele. Olyan példát is látunk majd azonban (Hollandia esetében), ahol történetileg úgy alakult, hogy nem is alakult ki kötelezpéldány rendszer, vagy ugyan létezik, de a webarchiválást annak keretein kívül szabályozzák (Németh, 2017). A különböző nemzeti jogi szabályozásokról átfogó összképet nyújt az IIPC linkgyűjteménye, ahol az adott országok hivatalos nyelvein olvashatók a webarchiválást szabályozó jogszabályi rendelkezések (IIPC, 2020a). Néhányra a nemzeti esettanulmányok tárgyalásakor külön hivatkozom a következőkben.

PANDORA: Az ausztrál webarchiválás modellje

Általános keretek, gyűjtőkör

Az Ausztrál Nemzeti Könyvtár 1996-tól végez webarchiválási tevékenységet a PANDORA: Ausztrál Webarchívum keretei között, a gyűjtőköri alapelvekről részletes dokumentum tájékoztat („Pandora Archive—NLA Selection Guidelines”, 2009).

A webarchiválás keretei Ausztráliában némiképp sajátosan alakultak ki. A tartalmak válogatott archiválása nem egyetlen szervezet, hanem egy intézményi konzorcium keretei között zajlik, melynek tagjai az egyes tagállamok állami könyvtárai, illetve számos tudományos könyvtár, illetve archívum (ideértve az Ausztrál Háború Archívumot, illetve a filmarchívumot is). Egy nemzeti elosztott rendszerbe tölti be mindegyik intézmény a saját maga szelektív módon leartatott tartalmait. Az aratások gyűjtőköre a helyi igények szerint intézményi szinten kerül szabályozásra. Az egységes webarchiválási szabványkörnyezet kialakítása, a rendszer infrastruktúrájának működtetése, az együttműködési keretek kialakítása nemzeti könyvtári hatáskör. A Pandora saját felülete 2019-től beolvadt az Ausztrál Nemzeti Könyvtár Trove nevű központi digitális katalógus szolgáltatásába (National Library of Australia, 2019). A Pandora felülete archív jelleggel működik tovább. („Pandora Archive—Preserving and Accessing Networked Documentary Resources of Australia”, 1999)

A szelektív archiválási tevékenységek mellett 2005-től kezdve a teljes ausztrál domain leartására is sor kerül. Ezt a tevékenységet az Ausztrál Nemzeti Könyvtár megbízásából az Internet Archive végzi el.

A szelektív begyűjtés módszerei

A konzorcium a szelektív mentésekre összpontosít. Ennek keretei között inkább a tartalom begyűjtésének minőségére, a megfelelő tartalmi mélységű aratás biztosítására fókuszálnak a mennyiségi paraméterek helyett. A hivatalos kormányzati dokumentumok tagállamonként kerülnek leartásra, illetve beszolgáltatásra („Pandora Archive—NLA Selection Guidelines”, 2009). Ezt az anyagot egészíti ki az ausztrál élet

teljességéről számot adó szelektív aratások köre (konferenciák, oktatási anyagok, zenei élet webhelyei, vallási oldalak, a napilapok webhelyeinek egyedi webes tartalmi stb.) A kiadókkal történő egyeztetés során törekedni kell arra, hogy a már kereskedelmileg nem értékes szolgáltatásokhoz kapcsolódó weboldalak teljes köre szabadon elérhető legyen a nagyközönség számára. Ugyancsak egyeztetések szükségesek a keresőmotorokkal nem leírási tartalmak begyűjtéséhez is. Az egyes webhelyeket leíró adatokat igyekeznek feldolgozni és beilleszteni az Ausztrál Nemzeti Bibliográfiába („Pandora Archive—NLA Selection Guidelines”, 2009). Sőt az ausztrál webarchívum anyagát a teljes digitálisan született, illetve digitalizált közgyűjteményi dokumentumállományt magában foglaló Trove adatbázisban is elérhetővé tették (National Library of Australia, 2019). Ezen felül a felhasználók adatgazdagítási, pontosítási javaslatokat is tehetnek minden egyes tételhez a rendszerben. Így a webarchiválás tevékenységének társadalmi kiterjesztése felé is jelentős lépést tettek. Ez az egységes keretek között visszakeresés nyújtó tartalomszolgáltató felület a maga szolgáltatási kínálatával mintaszerű, meglehetősen egyedülálló szolgáltatás, ami jó gyakorlatként feltétlenül tanulmányozásra érdemes (Németh, 2017).

Összefoglalásként elmondható, hogy a webarchiválás keretét adó konzorciumi szervezeti forma, bár némiképp a később tárgyalandó britre emlékeztet, mégis egyedi jelenség. A British Library jóval erősebb koordináló szakmai szerepet tölt be, mint az Ausztrál Nemzeti Könyvtár, illetve a konkrét webaratási tevékenységekből is jobban kiveszi a részét,

különösen, hogy a teljes domain aratást az Internet Archive látja el az ausztrálok részére. A gyűjtőkori alapelvek világosan lehatároltak, egységes egészként kezelik a teljes gyűjtőkori tevékenységet, melynek szerves részeként jelenik meg a webarchiválás.

Webarchiválás Nagy-Britanniában

Általános keretek, gyűjtőkör

Nagy-Britanniában a szelektív webarchiválás (British Library, 2019) 2004-ben, a jelenlegi nyilvános szolgáltatófelület pedig 2019-ben indult el a British Library gondozásában tematikus gyűjtemények létrehozásával. A teljes .uk domaint 2013-tól kezdték el aratni, összhangban az akkor megjelent új kötelespéldány szabályozással. A webaratási tevékenység az alábbi intézmények egyenrangú együttműködésével zajlik: The National Library of Scotland, The National Library of Wales, Cambridge University Library, The Bodleian Library in Oxford, The Library of Trinity College Dublin (British Library, 2019; Németh, 2017). A webarchívumhoz csak az említett intézmények fizikai tereiben lehet zárt dedikált hálózaton keresztül hozzáférni a szerzői jogi rendelkezésekkel összhangban. Az intézmények egy olyan szűkített mintát is összeállítanak, melyet igyekeznek nyilvánosan elérhetővé tenni. Az egyes intézmények a törvényi keretek figyelembevételével saját maguk alakítják a webarchiválás gyűjtőkörét. Az általános szempontok az alábbiak: tudományos jelentőség, innovatív webtechnológiai megoldások bemutatása, a brit élet minél teljesebb körű bemutatása, szociális, kulturális, politikai, vallási, tudományos és gazdasági tevékenységek szemszögéből (British Library, 2019; Németh, 2017). Különleges

méltánylást igénylő esetekben kérvényezni lehet egy adott webhely anyagainak teljes törlését is a webarchívumból.

Irányelvek

Az archívum weboldalán javaslatokat is lehet megfogalmazni a szolgáltatásba bevonni kívánt webhelyekre vonatkozólag (British Library, 2020a). A dinamikus tartalmak, a csak belső keresőmotor által elérhető tartalmak (linkek nélkül) illetve a Javascript menüszerkezetek némelyike kimarad az aratásból. Így is egy hónap átlagában 28 TB adatmennyiség kerül learatásra. Audio és videoanyagok (pl. YouTube, Flash video) nem esnek bele a web archiválási projekt hatókörébe. Az egyszerűen learatható, a honlapok szöveges-képes tartalomszerkezetébe beágyazódó audio fájlok viszont rögzítésre kerülnek. A könyvtár technikai ajánlásokat is megfogalmaz a weblap üzemeltetők számára az aratás megkönnyítése végett. A .uk domain hatókörébe eső weblapokat általános szabály szerint félévente aratják le, de a fontos weblapok aratására egyéni mérlegelés szerint ennél sűrűbben is sor kerülhet. Nem gyűjtik be azokat a tartalmakat, amelyek jelszóval vannak védve, illetve azokat a tartalomelemeket sem, amelyeket a webhely robots.txt fájlja kizár a megtekintésből. (British Library, 2020b)

Nagy-Britanniában tehát igen jól szervezett webarchiválási tevékenység zajlik, számos közgyűjteményi résztvevővel, illetve a tartalomszolgáltatók bevonásával. Az egységes alapelvek szerint hatékonyan működő regionális szervezetrendszer hatékonyan egészíti ki a British Library országos koordinációs tevékenysége. Vezető szerepet játszanak a szakmai területet érintő szoftveres fejlesztések előmozdításában is.

Általános keretek

Dániában a koppenhágai székhelyű Dán Királyi (Nemzeti) Könyvtár, illetve az Aarhusban található nemzeti könyvtári rangú Állami Könyvtár partnerségén alapult a webarchiválás modellje. Mára már a két intézmény egymással egyesülve, egységes szervezeti keretek között működik. Az internet archiválást érintő gyűjtőköri és szervezeti keretek a két könyvtár együttműködésével külön szabályzatban kerültek rögzítésre (Netarkivet, 2015a).

Irányelvek

A szabályozás irányelvei háromévente áttekintésre (felülvizsgálatra) kerülnek a megadott később részletezett törvényi kereteken belül. Maga a konkrét webarchiválási tevékenység főként az aarhusi intézményben folyik, bár a projekt menedzselését Koppenhágából látják el (Németh, 2017).

A webarchiválással foglalkozó munkacsoportban mindkét intézmény képviselteti magát gyakorlati szakemberekkel (IT és könyvtári oldal egyaránt). Mindkét intézményben külön digitális kurátorok foglalkoznak a begyűjtött, illetve begyűjtésre kijelölt anyaggal, állománnyal (Németh, 2017).

A kulturális minisztérium a nagy médiavállalatok, illetve kutatási szakemberek bevonásával működtet egy tanácsadói munkacsoportot az audiovizuális (nem szövegalapú) tartalmak archiválására. Ez a tevékenység tehát a Netarkiv keretein belül, de eltérő szabályrendszerben, ajánlások alapján zajlik a privát szereplők aktív

részvételével. A videó, illetve audio anyagok learatásával a szokásostól eltérő szoftverkörnyezet is igénybevehető (pl. ftp protokoll használata médiatartalmak aratására a műsorszolgáltató webhelyéről)(Netarkivet, 2015a, 2015b).

Törvényi rendelkezések a begyűjtés vonatkozásában

A gyűjtőköri szabályozás törvényi úton került rögzítésre a 3. fejezet 8-12. paragrafusaiban (Danmarks Kulturministeriet, 2015a). Ebbe épülnek be a webarchiválásra vonatkozó kitételek is. A törvény hatálya kiterjed az összes elektronikus kommunikációs hálózaton nyilvánosan elérhető dán tartalomra. Nem tartoznak tehát a törvény hatálya alá az intraneten, zárt hálózatokon elérhető tartalmak. Az adott tartalomnak a nyilvánosság felé szántnak kell lennie. Egy olyan szolgáltatás például, ahová bárki regisztrálhat nyilvánosan, a törvény hatálya alá esik, a meghívással működő, illetve teljesen zárt regisztrációhoz kötődő tartalmak azonban nem.

A „dán” fogalmának meghatározása az alábbiak szerint történik: 1. Dániában bejegyzett domainekekhez kötődő tartalmak (tehát nemcsak a .dk, hanem az összes dán regisztrátor által bejegyzett domainekekhez kötődő tartalom). 2. A nem dán bejegyzésű domainekek közül azokra terjed ki, melynek dániai célközönsége van. 3. A Kulturális Minisztérium egyedileg is kijelölhet a webarchiválás keretébe tartozó tartalmakat (8. paragrafus).

A Dániában regisztrált domain név tulajdonosának törvény által megszabott kötelezettsége az archiválhatóság követelményeinek megteremtése. Amennyiben nem Dániában regisztrált domainről van szó, akkor a tartalmat publikáló személynek, szervezetnek kötelessége a

tartalmat archiválhatóvá tenni, archiválható formátumot előállítani (például a másolásvédetten szolgáltatott anyagok rendelkezésre bocsátásával). Az eredeti szöveg az „archiválható kiadás” kifejezést használja (9. paragrafus).

Amennyiben szükséges, rendelkezésre kell bocsátani a tartalom archiválásához szükséges hozzáférési jogosultságokat, felhasználóneveket, jelszavakat az archiválást végző intézmény részére. (Ilyenek lehetnek pl. a személyre szabott tartalom eléréséhez szükséges felhasználónevek, jelszavak.) Az adott intézmény köteles ezeket az adatokat bizalmasan kezelni, nem adhat hozzáférést azokhoz külső személyek számára (10. paragrafus).

A dániai illetőségű domain név regisztrátoroknak az általuk regisztrált domain nevek adatait, illetve a domain nevek tulajdonosaira vonatkozó információkat el kell küldeni a webarchiválásért felelős intézmény számára elektronikus formában (11. paragrafus) (Danmarks Kulturministeriet, 2015b).

A beküldés gyakoriságának mértéke nincs törvényileg rögzítve. A nemzeti könyvtári gyakorlatot külön ajánlás rögzíti (Netarkivet, 2015b). Háromféle aratási stratégiát alakítottak ki, melyet az adott honlapok jellegéhez igazítanak. A legfontosabb hírértékű, tudományos értékkel bíró tartalmakat akár naponta aratják. 80 honlap van kiválasztva ebben a körben. Az általános gyakorlat az évi négy aratás elvégzése minden, a törvény hatálya alá tartozó honlapról (cross-cutting harvest). Adott aktuális témakörökhöz kapcsolódó szelektív begyűjtést is végeznek.

Kiválasztanak évente három eseményt, s begyűjtik a releváns weboldalak tartalmait, külön az eseményekhez rendezve (Netarkivet, 2015a).

Fontos kitétel az aratás folyamata kapcsán, hogy minden webhelyről teljes aratás történik. A robotok nem respektálják a tartalomszolgáltató által robot.txt fájlban rögzített korlátozásokat, hanem mindent learatnak (Danmarks Kulturministeriet, 2015b; Netarkivet, 2015a). A törvény utolsó paragrafusa pedig azt rögzíti, hogy a webarchiválási kötelezettség teljesítésével (esetleges járulékos tartalom előállításával) kapcsolatban felmerülő költségeket a tevékenységet végző intézménynek kell fedeznie. (12. paragrafus). A fent említett tevékenységek általában a statikus aratható tartalom előállításának plusz kötelezettségét foglalják magukban a beszolgáltató által a webaratást végző intézmény számára (Danmarks Kulturministeriet, 2015b; Netarkivet, 2015b).

Hozzáférés

A szerzői jogi és a személyes adatokat védő rendelkezések figyelembevételével adhat hozzáférést a két dán nemzeti könyvtár a learatott tartalom egyes szeleteihez a kérelmezők számára. Általában kutatási, oktatási céllal kérvényezett igényeket elégítenek ki. A jogi szabályozás új szelete 2019-ben került bevezetésre, hogy meghatározott időre, adott kutatási célra a szükséges állományrészek a kutatók és doktorandusz hallgatók rendelkezésére bocsáthatók, akik erre a megadott időtartamra el is vihetik a könyvtárból a kutatás tárgyául szolgáló archivált anyagot (Netarkivet, 2020). A teljesen zárt, könyvtáron belüli dedikált munkaállomásokra korlátozódó hozzáférést így sikerült egy kicsit oldani.

Összegzés

Dániában széles társadalmi egyeztetés előzte meg a kötelezpéldány törvény vonatkozó rendelkezéseinek megalkotását, mely igen kemény kitételeket tartalmaz. Ennek révén viszont a webarchiválási tevékenység hatékonyan biztosítható, a technikai feltételek biztosítását a domén név tulajdonosára, illetve a tartalom előállítójára hárítva át! A könyvtárak a webarchiválás tevékenységi kereteit a tartalomszolgáltatókkal aktív partnerségben dolgozták ki, s a rendszeres felülvizsgálat is az érdekeltek párbeszédével zajlik. Így a határozott törvényi rendelkezések betartására az érdekeltek hajlandósága is biztosítottnak tűnik (Németh, 2017). A dán webarchívum munkatársaival az archivált anyag teljesszövegű visszakeresésének megteremtése kapcsán a magyar webarchívum bekapcsolódásával közös fejlesztési projekteknél veszünk részt (Németh, 2019b). Ennek összefoglalója a magyar vonatkozású nemzetközi tevékenységeknél kerül majd megtárgyalásra.

Webarchiválás Csehországban

Törvényi háttér

A kötelezpéldány rendelkezések hazánkhoz hasonlóan Csehországban sem terjednek ki a webarchiválással kapcsolatos tevékenységekre. A webarchiválás feladatát nemzetközi példák alapján a nemzeti könyvtár felé koncentrálni modellezik kísérleti projekt keretében, ami 2001-ben indult el (Czech Webarchive, 2011; Kvasnica & Kreibich, 2013). A cseh webarchiválási projekt alakulását az OSZK különféle szintjein már nagyon régóta figyelemmel kísérték korábban is az itthoni webarchiválási tevékenységek megalapozása szempontjából, elsősorban a Könyvtári

Intézet nemzetközi szemlézési szolgáltatása révén (Ludmilá & Cerbová, 2009). A legújabb helyzetkép pedig Marie Haskovcova jóvoltából 2019 őszén került bemutatásra az ELTE digitális bölcsészeti konferenciáján (Haskovcova, 2019).

Szolgáltatási keretek

A cseh domainre és a cseh vonatkozású tartalmakra vonatkozó teljes domain aratásokat a .cz domainre koncentrálva a CZ. domain nemzeti regisztrátorával való együttműködés révén jelenleg, mint egy 1200000 domain tartalmát aratják le (Czech National Library, 2015a), illetve különféle eseményekhez, tematikákhoz kapcsolódó tematikus, szelektív aratásokat is végeznek (Czech National Library, 2015b; Kvasnica, 2016). A webarchiválási projekt még további kiteljesedéséhez szükség lenne a finanszírozási feltételek rendezésére, illetve a kötelespéldány rendelkezések felülvizsgálatára is. Jogi feltételek akadályozzák a kiválasztott tartalmaknak a Cseh Nemzeti Bibliográfiához történő hozzáadását is. Nagy kihívást jelent a számukra, hogy a robot.txt fájlok tartalmának figyelmen kívül hagyásával rengeteg feleslegesnek bizonyuló adatmennyiség is learatásra kerül. A Cseh Nemzeti Könyvtár 5143 különféle tartalomszolgáltató partnerrel írt már alá megállapodást a webarchiválás biztosításáról. (Czech Webarchive, 2011) A széleskörű együttműködés biztosítása tehát itt is alapfeltétele a webarchiválási tevékenységek kiteljesedésének. Élenjárnak a szelektíven archivált anyagok katalógizálásának terén, a legújabb RDA szabványok szerinti katalógizálási szabályzatot dolgoztak ki, s a teljesszövegű keresés, a nagy mennyiségű adatok kezelése kapcsán további fejlesztési projekteket

terveznek, ahogyan minderről Zdenko Vozár a pozsonyi V4 nemzetközi könyvtáros konferencián beszámolt (Vozár, 2019)

Hozzájárás

A jogtulajdonosokkal történt megállapodások alapján egy szűk adatkészlet elérhető nyilvánosan online (Czech Webarchive, 2011) az egyéb szeletek csupán a nemzeti könyvtár épületében zárt hálózaton. A szerzői jogi törvény az európai szerzői jogi rendelkezésekkel összhangban jelenleg kifejezetten tiltja a szerzői jog által védett leartatott anyagok nyilvános szolgáltatását.

Csehországban igen szilárd alapokon álló, jelentős hagyományokkal rendelkező webarchiválási tevékenység zajlik. Jó gyakorlatokat elsősorban a felhasználói engedélyek beszerzése és az új katalogizálási formák meghonosítása kapcsán érdemes magyar szemmel hasznosítani. Az utóbbi években jól megalapozott együttműködés a következő időszakban remélhetőleg közös fejlesztési projektek kivitelezésében is testet ölt majd.

Webarchiválás Szlovákiában

Alapvető információk

Szlovákiában a pozsonyi székhelyű nemzeti könyvtári ranggal is bíró Egyetemi Könyvtár koordinálásával zajlik a webarchiválási tevékenység (Bratislava University Library, 2016). Az első tesztek még 2006-ra nyúlnak vissza, üzemszerűen pedig 2015-től archiválják a web szlovák szeletét. (Androvič és mtsai., 2017; Androvič & Prókai, 2007) A szolgáltatás megfelelő jogi kereteit is sikerül megteremteni, különös tekintettel a kötelezpéldány szabályozásra. A szlovák webarchiválás is a kötelezpéldány

rendelet kiterjesztésén alapul, amire 2018-ban került sor, ahogy erről Jana Matúšková és Peter Hausleitner a pozsonyi V4 könyvtáros konferencián számot adott (Matúšková & Hausleitner, 2019). Szorosan együttműködnek a cseh partnerekkel, havi szintű személyes szakmai konzultáció zajlik a szakemberek között. Széleskörű együttműködési hálózat kialakítására törekszenek a tudományos és közkönyvtárakkal, illetve a tartalomszolgáltatókkal is (Matúšková & Hausleitner, 2019).

Gyűjtőkör, szolgáltatási keretek

Tematikus, eseményalapú és általános .sk domainre kiterjedő aratásokat egyaránt végeznek. A 352 ezer bejegyzett szlovák domainből 279 ezret sikerült a 2017 februárjában lezajlott második általános aratás során begyűjteni (Bratislava University Library, 2016). A szelektív archiválás keretében 550 intézményt kerestek meg együttműködést kérve, ezek közül odáig 111 intézménnyel sikerült szerződést kötni (Bratislava University Library, 2016; Németh, 2017). A brit példához hasonlóan, a projekt honlapján itt is bárki javasolhat weblapokat archiválásra a megfelelő űrlap segítségével. A webarchívum egyes szeletei a pozsonyi intézményben dedikált munkaállomásokon érhetők hozzá. A nyilvános hozzáférés a tartalom jogtulajdonosának hozzájárulásával biztosítható. Különös jelentőségű a szlovák projekt magyar szempontból azért is, mert a szlovákiai magyar webes tartalmak is a gyűjtőkörbe tartoznak. A HÍD-MOST párt honlapja például elérhető nyilvános szolgáltatás keretei között is (Németh, 2017).

A fiatal, ám dinamikusan fejlődő szlovák projekt feltétlen figyelemre méltó. Bemutakozási lehetőséget biztosítottunk számukra nemzeti

könyvtárunkban Budapesten, s mi is az OSZK részéről előadóként többen rendszeres vendégeik vagyunk szakmai eseményeiknek. Igyekszünk átvenni egymás szakmai tapasztalatait, illetve megteremteni az együttműködési lehetőségeket is.

Webarchiválás Észtországban

Alapvető információk, gyűjtőkör

Az Észt Nemzeti Könyvtár webarchiválási projektjének (Estonian National Library, 2008) alapvető célkitűzése az észt kulturális örökség szempontjából fontos webes tartalmak archiválása (Ruusalep, 2017). A nemzeti könyvtárról szóló törvény nevesíti kötelezettségként a nemzeti könyvtár számára a webarchiválási feladatokat (gyűjtés, feldolgozás, szolgáltatás) (Estonian National Library, 2008). A gyűjtőkör alapelemei a következők: Észt nyelven megjelenő weblapok, Észtországban publikált weblapok, Észtországra vonatkozó információkat tartalmazó weblapok (Estonian National Library, 2008). A webarchiválás szabályozását a köteleispéldány törvény rendelkezései közé is beillesztették. A weben megjelenő publikációk is köteleispéldánynak minősülnek, melyeket le kell aratni és nyilvánosan közzé kell tenni. A tartalom tulajdonosának azonban lehetősége van intézkedni arról, hogy a tartalom csak zárt hálózatban, a Nemzeti Könyvtár épületében legyen elérhető a szerzői jogokra tekintettel. 2017 január 1-től ez már alapértelmezetten így van, az archívum a kijelölt intézményekből érhető csupán el (Riigikogu, 2016).

Szolgáltatási keretek

A pdf alapú kiadványok begyűjtése 2006-ban, a webaratás 2008-ban kezdődött el. Az archívum 2013 óta érhető el a felhasználók számára.

2010-től 2015-ig 56 millió URL-ről arattak le adatot mintegy 4,2 TB mennyiségben. A webarchiválást szelektív aratásokkal kezdték el, ám mivel a releváns tartalomnak csak kis mennyiségét tudták így begyűjteni, 2015-16-ban lebonyolították az észti .ee domain első webes aratását. Az észti webarchívum fejlődése publikációs szinten is nagyon jól nyomon követhető és sok tanulsággal szolgált a magyar szolgáltatási modell kialakítása kapcsán is (Köüts, 2012; Ruusalep, 2017; Tiü, 2018). Egy adott webhelyről maximum 300 MB adatot arattak le, hogy kordában tudják tartani az aratási folyamatot. Összesen 4 TB tömörítetlen adatot arattak le ily módon (Ruusalep, 2017). Az új köteleespéldány törvény szerint, amennyiben a könyvtár nem tudja learatni a gyűjtőkörbe eső tartalmat, akkor a tartalom tulajdonosa kötelezett arra, hogy eljuttassa azt a Nemzeti Könyvtár számára. A beszolgáltatás költségeit a tartalom tulajdonosának kell állnia (Riigikogu, 2016). Öt könyvtár van országosan kijelölve bármiféle köteleespéldány elérésére épületen belül (beleértve a webarchívumot is zárt dedikált hálózaton).

Szervezeti keretek

Az Észt Nemzeti Könyvtárban három teljes állású munkatárs foglalkozik webarchiválási feladatokkal, ketten az archiválásért felelős szakemberek, illetve egy web archiváló alkalmazások kezeléséért felelős adminisztrátor. Az ő munkájukat egy munkacsoport segíti, mely 10 kutatási és kulturális örökség kezelésével foglalkozó intézmény 24 munkatársát tömöríti (Ruusalep, 2017). A munkacsoport tanácsokat ad a learatandó anyag kiválasztására, hozzáférhetővé tételére az aratás számára, emellett képviselik a kutatói szféra webarchiváláshoz kötődő érdekeit. 2011-ben

stratégiai dokumentumban fogalmazták meg a munkacsoport a webarchiválásra vonatkozó gyűjtőköri alapelveket. 2012-től az Észti Nemzeti Könyvtár tagjává vált az IIPC-nek is (Ruusalep, 2017).

Az észti kötelezpéldány szabályozás megítélésünk szerint európai szinten is példaértékűen sikerült. Az Észti Nemzeti Könyvtár megfelelő szervezeti kereteket és erőforrásokat is képes garantálni a webarchiválási tevékenység biztosításához, miközben kialakítottak egy olyan közgyűjteményeket is magukban foglaló szélesebb körű együttműködési hálózatot is, ami a tartalmak begyűjtésének hatékonyságát növeli meg.

Webarchiválás Hollandiában

A holland nemzeti könyvtári webarchiválási modell számos különleges tulajdonsággal bír. Ezekről első kézből értesülhettem az OSZK-ba látogató magyar származású Kees Teszelszkytől (Teszelszky, 2017), aki vezeti a webarchiválással a Holland Nemzeti Könyvtárban foglalkozó osztályt. Kifejtette, hogy Hollandiában történeti okokból nem alakult ki a hagyományos papíralapú dokumentumok vonatkozásában sem kötelezpéldány rendszer, a kiadók önként juttatják el kiadványaikat a Holland Nemzeti Könyvtár részére. Ebből kifolyólag nincs webes kötelezpéldány szabályozás sem. Ennek hiányában viszont kifejezetten tiltott törvényileg az általános célú aratás a holland domainről (Sierman & Teszelszky, 2017). A Holland Nemzeti Könyvtár csak egy a sok közgyűjtemény közül, amelyek szelektív webaratást végeznek, bár egyfajta koordinációs funkciót tölt be a gyűjtőkörök egyeztetésénél, illetve szoftverfejlesztési együttműködésekben is részt vesz nemzetközi partnerekkel (Sierman & Teszelszky, 2017). A holland domain volt az egyik

első, amely kialakult Európában (Teszelszky, 2019b), s ma is nagyon jelentős tartalmi gazdagsággal bír. Igyekeznek különféle eseményekhez, témakörökhöz kötődő aratással minél több szeletét begyűjteni ennek a webes térnek. Sok év helyben járás után e terület most kiemelt figyelmet élvez a digitális könyvtári fejlesztések között, ami pozitív fejleményeket sejtet a jövőre nézve (Sierman & Teszelszky, 2017). Érdekes kísérleti projektjük az önálló felsőszintű domainnel rendelkező Frízöld régió webes anyagainak begyűjtésére irányuló kísérleti projekt, melynek eredményeit a jóval nagyobb méretű holland nemzeti domain keretei között is hasznosítani kívánják (Teszelszky, 2019a).

A nemzetközi áttekintés tanulságai

Az intenzív kutatási munkán alapuló nemzetközi körképet áttekintve megállapíthatom, hogy a jogi, szabályozási kereteket, illetve a webarchiválási tevékenységek szervezeti kereteit tekintve rendkívül vegyes az összkép. Az adott ország állami berendezkedése, a szakmai attitűdök, az együttműködés kultúrájának erőssége erősen befolyásolja a webarchiválási feladatok szervezését. Kritikus elemet jelent a megfelelő, jogi szabályozási keretek megléte. Határozottan képviseltem munkatársaimmal együtt azt az álláspontot a magyar webarchívum kialakításának munkálatai során, hogy kötelezpéldány szabályozásba itthon is bele kellene foglalni a webarchiválási tevékenységet, ami jelentős kihívást jelent szabályozási oldalról. Az észet mintából sokat lehetett meríteni e téren. Végül Magyarországon más jogi keretekbe ágyazva, a kötelezpéldány rendelettől függetlenül valósult meg a webarchiválás jogi státuszának rendezése. Pozitívumot jelentett a magyar webarchívum

kialakítása során, hogy a kísérleti projekt szakmai kereteinek kialakításakor számos példából tudtam meríteni a a közeli országokból is, jó személyes kapcsolatok alakultak ki a ezen országok szakembereivel. A 2017-es IFLA konferencián külön szekcióülés foglalkozott a webarchiválás aktuális kihívásaival, globális előadói körrel (IFLA Library, 2017). Ez az IIPC konferenciák mellett újabb tapasztalatcsere lehetőségét kínálta fel. Reményeim szerint ez a rövid összefoglaló ráirányítja a figyelmet a webarchiválás nemzetközi dimenziójára, elsősorban abból a szempontból, hogy a külhoni tapasztalatok jelentős szeletét az itthoni körülmények közé is sikerrel implementálni lehessen.

2.3. Magyar részvétel a nemzetközi együttműködésekben

Webarchiváló kísérleti projektünk külföldi kapcsolatainak gerincét adja az internet megőrzésével foglalkozó nemzetközi szervezetben, az International Internet Preservation Consortium-ban (IIPC, 2019) kifejtett tevékenységünk. Az OSZK 2018. január elsejétől vált az IIPC teljes jogú tagjává. Ezután nem sokkal a terveinket ismertető bemutatkozó anyag is felkerült a konzorcium honlapjára („IIPC -Országos Széchényi Könyvtár”, 2018). Aktívan közreműködünk az IIPC különféle esemény alapú gyűjteményeinek gyarapításában, s egyben a magyar anyagot a saját archívumunkban is megőrizzuk.

A konzorcium 2013-ban alakult meg a Francia Nemzeti Könyvtár kezdeményezésére 12 alapító taggal. Jelenleg már kb. 45 országból vannak benne könyvtárak, levéltárak, múzeumok, illetve a webarchiválás területén érdekelt civil és piaci szereplők is („About IIPC”, 2019). A konzorcium alapvető küldetése az internetes tartalmak archiválásához és

hosszú távú biztonságos megőrzéséhez szükséges tevékenységek koordinálása és támogatása. Munkáját az egyes tagintézmények képviselőiből megválasztott végrehajtó bizottság fogja össze, a titkárság segítségével.(„Leadership”, 2019) A közös munka nagy része tematikus munkacsoportokban zajlik, ezek jelenleg a tartalomfejlesztéssel, a hosszú távú megőrzéssel, illetve az oktatási kérdésekkel foglalkoznak.(„IIPC Working groups”, 2019) Az OSZK képviseletében ez utóbbi (2017 végén alakult) munkacsoportban (IIPC TWG, 2019), illetve a 2019-ben alapult kutatási munkacsoportban (IIPC, 2020b) veszünk részt.

A tagok számára nagy segítséget jelent az aktuális információk gyors megosztását lehetővé tevő elektronikus levelezőlista, a Twitter hírcsatorna, valamint a Slack nevű csoportmunka alkalmazás segítségével zajló kommunikáció. Így első kézből nyerhetünk áttekintést a technikai újdonságokról, a nemzetközi trendekről és az egyes intézményi gyakorlatokról, projektekről, eredményekről.

Alkalmam nyílt részt venni az IFLA 2017. évi kongresszusán a lengyelországi Wrocławban, ahol önálló szekció foglalkozott az IIPC szervezésében a webarchiválással (IFLA Library, 2017). Itt történt meg 2017. augusztusában az első személyes kapcsolatfelvétel az IIPC tagjaival, vezetőségével. Tanácsokat kaptam tőlük a webarchiválási kísérleti projektünk megalapozásához, még mielőtt hivatalosan taggá váltunk volna.

Az IIPC tevékenységét a British Library-ből koordinálják. Olga Holownia-val, az adminisztráció vezetőjével gyümölcsöző együttműködést sikerült kialakítani. 2018 szeptemberében online találkozhattam az ő segítségével,

az OSZK képviseletében az Internet Archive egyik vezetőjével és az IIPC végrehajtó bizottságának tagjával, Jefferson Bailey-vel. Ezt követően kaptunk tőle szolgáltatási, illetve árajánlatot az Internet Archive magyar anyagának elérésére. A feltételek talán 2021-ben lesznek adottak a megfelelő pénzügyi feltételek biztosítása esetén az előrelépésre.

A már említett konzorciumi titkár mellett a szintén korábban szóba került, a holland nemzeti könyvtár webarchiválásért felelős, részben magyar származású vezetője, Kees Teszelszky nyújtott nélkülözhetetlen segítséget tagságunk kezdetekor a szükséges információk összegyűjtéséhez. Ő irányította rá a figyelmet az oktatási munkacsoport tevékenységére, melynek munkájába már 2018 januárjától bekapcsolódtam. A későbbiekben részletezett saját oktatási tervek formálásához igen jól jött, hogy a munkacsoport keretében fel tudtam mérni az összes tagintézményben a már rendelkezésre álló képzési anyagokat és a jelenlegi gyakorlatot. Ezen felül a konzorciumi tagok számára készült egy kérdőív is (IIPC TWG, 2017), amiben igényeiket, elvárásaikat fogalmazhatták meg a webarchiválás oktatása terén. A kérdőív kiértékelésének előzetes eredményeit is be tudtuk építeni a munkánkba, illetve nyilvános előadásainkba is, melyeket 2018 áprilisában az egri Networkshop konferencián (Drótos & Németh, 2018a), Drótos Lászlóval együtt, illetve októberben a Pozsonyi Egyetemi Könyvtár CDA 2018 (Németh, 2018c) című digitális könyvtári konferenciáján tartottam.

Az IIPC tagságunk másik lényeges eredménye az a kutatás-fejlesztési együttműködés, melyet szerződéses munkatársunk, Vitéz Gábor alakított ki a Dán Nemzeti Könyvtár munkatársaival, elsősorban Thomas Egense-

vel, aki a SolrWayback nevű teljes szövegű kereső és megjelenítő eszközt fejleszti (Németh, 2019a). A szoftver tesztelésébe, egyes hibák javításába és nyilvános tesztelésébe mi is be tudtunk segíteni. A dán fejlesztők az IIPC 2018-as (Egense & Klindt Myrvoll, 2018), én az OSZK képviselőjeként pedig az IIPC 2019-es közgyűlésén (Németh, 2019a) számoltunk be e közös munka eredményeiről, mely az IIPC vezetőségének körében is pozitív visszhangot keltett.

A szintén új IIPC-tag belga webarchiválási projekt munkatársaival a dán Aarhusi Egyetem Netlab kutatócsoportja által 2017 őszén szervezett, a webarchiválással foglalkozó angol nyelvű e-learning kurzus keretében (Netlab, 2018) ismerkedtem meg. Az ott folyó szakmai diskurzus is segítséget jelentett a hazai pilot projektünk megalapozásában. Ez majd szóba kerül még az oktatási témájú fejezetben a későbbiekben.

2018. december 14-én Niels Brügger professzor, a NetLab kutatócsoport vezetője tartott igen érdekes és hasznos webinárium előadást az IIPC webarchívum kutatási lehetőségeket vizsgáló webinárium sorozata keretében, melynek továbbfejlesztett változata a zágrábi IIPC konferencián hangzott el (Brügger, 2018; Brügger & Laursen, 2019) a dán webtér gyűjtése kapcsán szerzett tapasztalataikról. Újra szembesülhettem azzal a kihívással, hogy a webarchiválás csapatjáték, a gyűjtőkori, stratégiai tervezéstől, a leartott tartalomhoz kötődő felhasználási jogosultságok megszerzésén át, a szükséges hardver és szoftver infrastruktúra kialakításán keresztül a kutatási célú hasznosítás tervezéséig. Igen sokszínű, multidiszciplináris az a kompetencia-skála,

amire egy nemzeti webarchiválási projekt kialakításakor szükség mutatkozik.

Személyes találkozásokon alapuló kapcsolatokat sikerült kiépíteni a cseh, a szlovák és az osztrák webarchívumok munkatársaival is. A pozsonyi Egyetemi Könyvtárban több előadást is tartottam, illetve információcserét folytattam a házigazda intézmény és a Cseh Nemzeti Könyvtár webarchiválásért felelős munkatársaival. Az Osztrák Nemzeti Könyvtárban pedig egynapos szakmai tanulmányúton vettünk részt munkatársaimmal együtt, hogy megismerjük az ott zajló webarchiválási tevékenységet. Sikeres szakmai rendezvényt is szervezték fennállásuk 10. évfordulóján, ezen is alkalmam nyílt részt venni, ahol elsősorban a német nyelvű országok webarchiválási projektjeiről lehetett széleskörű áttekintést szerezni, a tapasztalatokat rövid publikáció formájában is rögzítettem (Németh, 2019e).

Csatlakozásunk évében még nem volt módunk magyar részről a IIPC konzorcium éves közgyűlésén, illetve az ahhoz kapcsolódó konferencián való részvételre, mivel ezekre Új-Zélandon került sor. Lehetőségünk nyílt az OSZK részéről viszont egy öt perces videó összefoglalót készíteni a rendezvényen részt vevők számára, bemutatva magunkat és eddigi eredményeinket. 2019-ben viszont részt vállaltam a Zágrában rendezett közgyűlés szervezésében, a programbizottság munkájában is. A kísérleti projektről tartott rövid összefoglaló mellett a dán partnereinkkel közös fejlesztési munkákról adtam elő, a SolrWayback és a saját fejlesztésű SolrMIA keresőkről, és azok találati listáinak relevánsabbá tételéről a metaadatok segítségével (Németh, 2019a). Sikeres együttműködést

alapoztam meg a Web Curator Tool webarchiválási munkafolyamatokat segítő keretrendszer új-zélandi és holland fejlesztőivel, a rendezvényen tartott workshop révén. 2019 őszén ennek tapasztalatai alapján munkatársaimmal egyeztetve eljuttattam kérdéseinket a programcsomag fejlesztése kapcsán a partnereinkhez, így közvetlen befolyással bírunk, informatikus kollégáinkkal együtt, a fejlesztési irányok meghatározására. Megteremtettem az alapjait a közép-európai webarchívumok közötti együttműködésnek is, melynek konkrét szervezeti kereteit közös fejlesztési programokkal, rendezvényekkel a 2021. év során tervezzük partnereinkkel együtt meghatározni. Aktívan részt vállaltam az oktatási munkacsoport munkájában, mely szintén a zágrábi konferencián kapott határozott lendületet. A Digital Preservation Coalition szervezettel együttműködve oktatási anyagok készültek a webarchiválás alapjairól, melyeket be tudunk ágyazni a saját oktatási tevékenységünkbe is (IIPC Training Working Group, 2020). Zágrábban megalakult a kutatási munkacsoport is, mely a webarchívumokban gyűjtött állomány tudományos célú hasznosításával foglalkozik (IIPC, 2020b). Ennek munkájába is bekapcsolódtam saját elképzeléseink megosztása révén, melyek egy későbbi fejezetben kerülnek tárgyalásra.

Jövőre vonatkozó terveink között szerepel a közép-európai webarchívumok együttműködésének szélesítése. Az egymást átfedő gyűjtőkörökbe illeszkedő archivált gyűjteményrészeket közös kereső szolgáltatás segítségével kellene elérhetővé tenni. Emellett főként cseh nemzeti könyvtári partnerünkkel tervezek szorosabb együttműködést az OSZK részéről az általuk javasolt, s saját keretek között különféle

projektekben érintett információ visszakeresése, és vizualizációja területén.

3. A magyarországi webarchiválás kísérleti projektje – a szervezett webarchiválás kezdetei

A nemzetközi példák után ebben a fejezetben széleskörű áttekintést nyújtok a magyar webarchiválási kísérleti projekt megalapozásához szükséges kutatás-fejlesztési tevékenységekről, a webarchiválást érintő állandó jellegű nemzeti könyvtári szolgáltatás megalapozásának folyamatáról. A nemzetközi kapcsolatok megalapozásának munkálatai mellett ezen a téren is számos lehetőség nyílt számomra a projekt keretei között az egyéni kutató és szervezői munkára.

A szervezett formában elinduló webarchiválás itthoni előzményeinek áttekintése után áttekintést adok arról az eredeti koncepcióról, melyet a kísérleti projekt munkálatainak kezdeteikor munkatársaimmal közösen felállítottunk. Ezt követően rátérek arra, hogy a gyakorlatban ehhez képest milyen munkafolyamat rendszert alakítottunk ki a webarchiválás megalapozása kapcsán. Sarkalatos pontja tevékenységünknek a gyűjtőköri szabályozás, illetve a gyűjteményfejlesztési irányelvek kijelölése is. Ezt követően kerül bemutatásra az a hardveres, illetve szoftveres háttér, melyre építve alkottuk meg a kísérleti projekt informatikai kereteit. Nagyon fontos összetevője tevékenységünknek a szelektíven aratott és válogatott állomány megfelelő metaadatsémákon keresztüli leírása, feltárhatóvá tétele, ennek a munkának a kezdeteiről is áttekintéssel szolgálok. Lényeges, önálló kutatások tárgyaként szolgáló tényezőként

merül fel továbbá a (főként válogatott módon leartott) tartalmak minőségellenőrzésének kérdésköre, a webarchiválást érintő taxonómiai keretek taxonómia elméleti és gyakorlati háttere. A kutatási hipotézisekben rögzített témaköröket alapul véve számos megfontolásra érdemes kérdés merül fel nem csupán gyűjtőköri, hanem technikai és közgyűjteményi szakmai, szolgáltatásmenedzselési értelemben is arról, hogy mely vonatkozásokban lehet megállapítani a webarchiválási tevékenységek hatókörének határait. Elengedhetetlen kialakítani a jelen műszaki és szakmai jellegű kompetenciáinkhoz kötött hosszútávú megőrzési kereteket is. Sok nyitott kérdés merült fel a magyarországi webarchiválási tevékenység jogi hátterének rendezése kapcsán, mely szintén áttekintést kíván. Végül, de nem utolsósorban pedig arról számolok be, hogy a projekt szakmai megalapozásához, a kutatás-fejlesztési eredmények ismertetéséhez milyen kommunikációs tevékenységeket rendeltünk munkatársaimmal együtt hozzá.

3.1 Előzmények

A szervezett webarchiválás 2017-től történt elindulását megelőzően a témakör egyes összefüggései már korábban is napirendre kerültek. Drótos László és Kokas Károly foglalta össze ennek legfontosabb vonatkozásait (Kokas & Drótos, 2018). A kilencvenes évek második felében indított magyar webes keresők (Heuréka, Góliát, Altavizsla/Vizsla), majd a 2010 körül megjelent újabb generációs társaik (Bluu, Szöracs, Miner, PolyMeta/Johu, RichPOI) robotjai által gyűjtött magyar tartalomra ezzel szemben nem épült webarchívum. Sajnos, miután nem voltak versenyképesek később a Google által kínált versennyel, ezek a tartalmak mára már nyomtalanul el is tűntek a világhálóról...

Kutatás-fejlesztési szinten jelentős tevékenységet a webarchiválás terén a SZTAKI fejtett ki. 2008 és 2013 között két EU finanszírozású kutatás-fejlesztési projektben vettek részt. A LIWA: Enrichment Tools for Web Archives című projekt (SZTAKI, 2011) 2008 és 2011 között zajlott le. A fő cél a projekt honlap szerint az volt, hogy létrehozzanak a Web tartalom begyűjtését, megőrzését, vizsgálatát és strukturális gazdagítását biztosító új generációs megoldást. Céljuk az archívum tisztaságának biztosítása, a magas minőségű tartalom összegyűjtése és a hosszú távú interoperabilitás megőrzése volt. Az MTA SZTAKI szerepe a web spam szűréssel kapcsolatos Work Package vezetése, illetve szövegbányászati feladatokban való részvétel volt. A LAWA: Longitudinal Analytics of Web Archive Data projekt 2010-2013 között az új web alkalmazások számára akarta feltárni az Internet méretével, eloszlásával, szerkezetével és evolúciójával kapcsolatos információkat. A LAWA projekt (SZTAKI, 2013) honlapja szerint az Európai Internet Archívum központi Web gyűjteményére FIRE (Future Internet Research) szolgáltatást épített, amely virtuális Web obszervatóriumként üzemelve web méretű adatelemzéseket kívánt végezni. Céljuk a nyílt forráskódú, párhuzamos Hadoop környezet kiterjesztése volt újszerű, elosztott adathozzáférési, tárolási és indexelési módszerekkel. Kiemelt feladatot alkotott az időbeli változások vizsgálata, az idő-dimenzió megjelenítése, illetve a tartalom klasszifikációja. Sajnos e projekteknek nem lett folytatása, a területért felelős szakemberek távoztával ezeknek a kutatási-fejlesztési projekteknek az eredményeit nem sikerült becsatornázni a néhány évvel később indult OSZK web archiválás keretei közé.

A 2010-es évek elején még egy kutatás-fejlesztési projektre került sor Kampis György, illetve Gulyás László vezetésével, az ELTE Tudományfilozófia Tanszék tudománymetria munkacsoport szervezésében. Egyetemek és akadémiai kutatóintézetek webhelyeinek mentése állt a projekt középpontjában (Kampis & Gulyás, 2013), s a kísérlet arra is kiterjedt, hogy milyen gyakorisággal frissülnek azok. Fontos megállapításuk volt, hogy sokkal kisebb adatmennyiséget határoztak meg a mentett webes tartalmakból, mint ami hipotézisként bennük felmerült. Továbbá a tartalmak frissülési sebessége is alacsonyabb a becsült mértéknél. Ez abból a szempontból is érdekes volt számukra, hogy az online tudományos diskurzus intenzitásának jóval nagyobb jelentőséget tulajdonítottak, mint ami a vizsgálati eredményekből következett. (Gulyás, Jurányi, Soós, & Kampis, 2014). Kampis Györgynek az első, a webarchiválási projekt gondozásában rendezett, 404 workshopon tartott előadásából (Kampis, 2017) kiderült, hogy ezt a korlátozott szelektív aratási kísérletet ők már egyfajta előjátékának szánták egy későbbi nemzeti webtér szintű webaratási projekt kezdeményezésének, amelynek feladatköre azonban természetesen jóval tágabbra nyúlt a kísérleti kutatóműhelyük kereteinél.

A webarchiválás médiaarchívumi vonatkozásában a Nemzeti Audiovizuális Archívum kísérleti projektje került a látókörünkbe, amit Répászky Lipót 2017. áprilisi személyes közlése alapján ismertetek. Egy olyan megoldást fejlesztettek ki, ami nem relációs adatbázisban történő adattárolás és keresést valósít meg, nyelvi analitikákat felhasználva, valamint egy kidolgozott standard OAI-PMH protokollon megvalósított adatcserét

Dublin Core szabvány segítségével. Ők onnan közelítették meg a webaratást, hogy számukra a webes multimédiaforrások metaadatainak rögzítése volt a fő mentési szempont, a kinézet, az elrendezés másodlagos, nem is törekedtek a teljes szolgáltatási felület mentésére. Akkor tekintették megoldottnak a feladatukat, ha a keresési oldalon a metaadatot megtalálták. Az adatkészlethez külön kísérleti keresőfelületet is fejlesztettek, ami azonban napjainkban már nem érhető el.

A webarchiválás a hazai könyvtári életben 2006-ban merült fel először, amikor Drótos László a Tudományos és Műszaki Tájékoztatás című folyóirat hasábjain javaslatot tett egy Magyar Internet Archívum létrehozására (Drótos, 2006). Számos, az OSZK-ból kezdeményezett szakmai tanulmányút felmerült a webarchiválás hazai meghonosításának kívánalma. Az EU finanszírozású K2 projekt keretében a Könyvtári Intézet és az Informatikai Igazgatóság együttműködésével sor került korlátozott mértékű szoftvertesztekre, magam is részt vettem magyar nyelvű rövid belső felhasználású tájékoztató anyagok összeállításában. Az OSZK munkatervébe már 2007-től bekerült a webarchiválás mint szakmai feladat (OSZK, 2007), de forrásokat tíz évig nem sikerült találni hozzá. Közben a digitalizált, illetve digitálisan született könyvek, folyóiratok, képanyagok gyűjtése és szolgáltatása állandó munkafolyamatként fokozatosan beépült az OSZK tevékenységi körébe. 2017-től jelenhetett meg a webarchiválás az Elektronikus Könyvtári Szolgáltatások Osztálya tevékenységei között az előbb említett területek mellett az OKR projekt keretében (Drótos, 2017a).

A webarchiválási projekt előzményeként, a tanulási és kutatás-fejlesztési folyamat fontos előfeltételeként meg kell még említeni az itthoni könyvtári szaksajtóban futó publikációs tevékenységet, melynek révén a kétezres évek második felétől jelenik meg a webarchiválás az itthoni szakmai témák között. Ábrázolásra kerültek például a norvég (Nuys, Albertsen, Pedersen, & Szalóki, 2005), brit (Bailey, Thomson, & Szalóki, 2006), a cseh (Ludmilá & Cerbová, 2009), illetve a szlovák (Androvič & Prókai, 2007) jó gyakorlatok is. A 2010-es évek elején már a különféle webarchiválási politikákról is született rövid összehasonlító tanulmány Dancs Szabolcs révén (Dancs, 2011). Néhány évvel később pedig már Hegyközi Ilona nagyobb ívű szakirodalmi szemlében mérte fel 2014-ben a webarchiválás helyzetét számos európai ország és Ausztrália szolgáltatásain keresztül (Hegyközi, 2014). Ez utóbbi már közvetlen előzményének tekinthető a nemzetközi környezetet feltáró kutatótevékenységnek. Fel kell hívnom még a figyelmet arra is, hogy tudományos munkák terén a 2010-es évek elejétől merült fel témaként először a webarchiválás Kornhoffer Mónika révén, OTDK dolgozat, szakdolgozat, illetve publikáció szintjén is (Kornhoffer, 2010b, 2010a, 2011).

3.2 A kísérleti projekt eredeti koncepciója

A kísérleti projekt megvalósításának első lépése 2017 áprilisában Drótos László révén, egy olyan, a célkitűzéseket és alapvető szakmai feladatokat, elvárásokat, időbeli ütemezést és munkaerőszükségletet rögzítő koncepció elkészítése volt, mely kiindulópontul szolgált a projekt végrehajtásához, a webarchiválás mint állandó nemzeti könyvtári

munkafolyamat előfeltételeinek megteremtéséhez (Drótos, 2017c). A magam részéről 2017 késő tavaszán, nyár elején kezdtem el vele együtt dolgozni a célkitűzések végső formába öntésében és gyakorlati megvalósításában. A koncepció alapvetően befolyásolta a gyakorlati munkával kapcsolatos hipotézisek megalkotását is. Lényegében azt lehet megvizsgálni, hogy a következőkben vázoltakból milyen elemek valósultak meg. Az alábbiakban az idézett koncepció főbb elemei kerülnek kifejtésre:

A kísérleti kutatás-fejlesztés céljaként egy leendő magyar internet archívum (webarchívum) koncepciójának, szervezeti kereteinek és informatikai infrastruktúrájának megteremtését, valamint egy tesztarchívum felállítását foglalmaztuk meg. Ennek segítségével tapasztalatok gyűjtését kívántuk előtérbe helyezni egy majdani nagy kapacitású, üzemszerűen működő rendszerhez. Végső célként egy olyan rendszer létrehozatalának kívánalma fogalmazódott meg, amely az interneten nyilvánosan megjelenő magyar és magyar vonatkozású kulturális örökség hosszú távú megőrzésének feladata mellett képes kiszolgálni az oktatás, a tudományos kutatás, az állami szervek, az üzleti szféra és az egyes internet használók igényeit is (Drótos, 2017c). A rendszer megteremtésének fő indokának azt jelöltük meg, hogy az addig csupán jelen időben létező magyar internetnek „múltja” is legyen. Olyan lehetőségek megnyitását helyeztük előtérbe a jelen és a jövőbeni időszakok felhasználói számára, amelyek egy webarchívum hiányában nem, vagy csak nehézkesen valósíthatók meg (pl. megszűnt weboldalak megtalálása, régi vagy jelenkori, de feltört ill. véletlenül törölt webhelyek helyreállítása, weboldalak időbeli változásának elemzése és vizualizálása,

stabil hivatkozhatóság, idődimenziót is tartalmazó szöveg- és adatbányászati alkalmazások futtatása, internettörténeti kutatások, hiteles másolatok szolgáltatása, igény szerinti webarchiválás megrendelésre történő szolgáltatása) (Drótos, 2017c). A rendszert a külföldi legjobb gyakorlatok tapasztalatainak felhasználásával terveztük kialakítani, együttműködésben a hazai közgyűjteményi és informatikai szféra egyes szereplőivel. Fő célkitűzésként szabtuk meg, hogy alkalmas legyen nagy tömegű webaratásra, egyes webhelyek vagy weboldalak szelektív lementésére, valamint online tartalmak önkéntes beadására is (Drótos, 2017c). Az ehhez szükséges infrastruktúra megteremtése mellett rögzítésre került a webarchiválás és az archivált anyag metaadatolásának módszertana kialakításának kívánalma; az archívum és az eredeti tartalomgazdák, valamint az archívum és annak felhasználói közötti jogviszonyt szabályozó szerződésminták előállítása is célként fogalmazódott meg; továbbá ajánlások, tanulmányok, előadások és tanfolyamok segítségével kívántuk megismertetni az érintett szereplőkkel a webarchiválással és a webarchívum-használattal kapcsolatos ismereteket (Drótos, 2017f).

A koncepció általános céljai mellé a kutatás fejlesztési-munka konkrét formába öntésére egy 13 tételből álló feladatlista is hozzá lett rendelve, melyeknek végrehajtása 2018 végéig lett előirányozva (Drótos, 2017f). Ezáltal alkalom nyílik a kutatás során annak felmérésére is, hogy az előirányzott célok mennyiben valósultak meg. Az egyes tételek sorrendje egyben időbeli sorrendet is jelöl, az egyes feladatok teljesítésének

sorrendjét illetően. Két fő feladatot és tizenegy azokhoz kötődő háttérfeladatot határoztunk meg.

1. Tesztarchívum létrehozása. Az elméleti ismeretek megszerzése mellett gyakorlati tapasztalatok megszerzésének céljából felállítani terveztünk egy kezdetben néhány száz (főként kulturális és tudományos) webhelyből, valamint az Elektronikus Periodika Archívumban csak távoli forrásként nyilvántartott kb. 1700 elektronikus periodika, néhány alkalommal történő mentéséből álló tesztarchívumot. Ez sikeresen meg is valósult. Továbbá meg kívántuk kísérelni a .hu domain alatti webszerverek kellően nagy halmazának learatását egyszeri vagy kétszeri alkalommal (mivel egy ekkora aratásnak a rendelkezésre álló infrastruktúrától függően akár több hónapig történő futása is előre jelezhető volt az azt megelőző külföldi tapasztalatok szerint). Ezek a teszt jellegű, terheléses próbák is sikerrel zárultak, segítettek felmérni a domain szintű aratás infrastruktúra szükségleteit. Komolyabb zárt vagy nyilvános hatókörű szolgáltatást nem terveztünk kialakítani erre a gyűjteményre, egyrészt a szerzői jogi háttér hiánya miatt, másrészt pedig, miután a szoftveres infrastruktúra folyamatos beállítását, finomhangolását is el kívántuk végezni, ezért az első néhány aratásból nem is vártunk szolgáltatásban felmutatható eredményeket. Egy kis demo gyűjteményt azonban a szerzői jogi engedélyek megszerzését követően a jó minőségű mentésekből közzé kívántunk tenni pl. learatott könyvtári webhelyekből, bemutatva a webarchívum lehetséges szolgáltatási alapfunkcióit. -

Erre a kísérleti projekt honlapjába beágyazva sor is került, a projekt végeztéig pedig a gyűjtemény folyamatosan bővült újabb kategóriákkal és tételekkel.

2. A tesztarchívum létrehozása mellett el kívántuk kezdeni a szelektíven (is) archiválandó webhelyek vagy webhelyrészek kellően reprezentatív méretű halmazát is feltérképezni. Egy mintegy tízezres nagyságrendű URL-lista kialakítását határoztuk meg, kiegészítve néhány további metaadattal (pl. a webhely neve, fenntartója, műfaja, témaköre, mélysége, változékonysága, veszélyeztetettsége, arathatósága, jogi előírásai, a tartalomgazda elérhetősége). Ezt az adatbázist jelöltük meg a majdani üzemszerűen működő archívum gyarapítási forrásaként, ez alapján terveztük indítani - optimális gyakorisággal és konfigurációs beállításokkal - az egyes webhelyek mentéseit. A címlista bővítésébe és karbantartásába a projekt iránt érdeklődő partnerintézmények munkatársait is be kívántuk kapcsolni, sőt a magyar internetes társadalom szélesebb rétegeinek figyelmére is számítani kívántunk (felhasználókéra és szolgáltatókéra egyaránt), akiknek lehetővé kívántuk tenni, hogy szintén javasolhassanak majd megőrzésre érdemes tartalmakat. - Maga az alapgyűjtemény létrejött, s most már az állandó szolgáltatás részeként bővül dinamikusan újabb és újabb témakörökkel, valamint a tematikus alapú ütemezett aratások anyagával.
3. A külföldi jó gyakorlatok megismerése, a téma aktuális szakirodalmának feldolgozásával együtt. Ennek eredményeit már egy korábbi fejezetben összefoglaltam.

4. Csatlakozás a webarchívumokat tömörítő nemzetközi konzorciumhoz (IIPC). A csatlakozásra 2018 elejétől sor került, az előnyökről, eredményekről is szó esett már.
5. A magyar webarchívum gyűjtőkörének meghatározása. - Egy később részletezett javaslat készült el ennek kapcsán, melynek véglegesítése és hivatalos formába öntése még nem történt meg.
6. A magyar webtér lehatárolási szempontjainak felmérése. - A felmérés sikeresen lezajlott, alapjául szolgál a domain szintű aratások tervezésének.
7. A metaadatstruktúra kialakítása a szelektív ill. a tömeges mentések esetében is, és a szelektív mentések folyamatos metaadatolási munkafolyamatának kialakítása. - Az XML alapú metaadatséma kialakításra került, a későbbiekben szó esik róla még. Az RDA alapú metaadat séma megteremtése még folyamatban lévő feladat.
8. A tartalomszolgáltatókkal kötendő szerződés tartalmának és az engedélyt kérő e-mail szövegének megfogalmazása. - A munkafolyamat kialakításra került. A szükséges emberi erőforrás hiánya az adminisztráció terén fokozottan jelentkezik, hátráltatva a szabadon szolgáltatható jogvédett állomány dinamikusabb gyarapítását is.
9. Az archiválási tevékenység és az archívumhoz való hozzáférés jogi keretének megteremtése. A jogi keretek 2020. január elseji hatállyal megteremtődtek.
10. A kísérleti időszak tapasztalatai alapján az üzemszerű működésre alkalmas, jól skálázható informatikai infrastruktúra megtervezése. - Több hullámban került sikeresen megtervezésre.
11. A webarchiválással, a webarchívumok működtetésével, használatával kapcsolatos ismeretek oktatása az OSZK, illetve egyéb könyvtárak, közgyűjtemények érdeklődő munkatársai számára. - Belső tanfolyamokat, tájékoztatókat tartottunk, illetve 2018-ban elindult az államilag akkreditált könyvtáros továbbképzési program is a Könyvtári Intézettel együttműködve. Ennek részletei az oktatással kapcsolatos fejezetben kerülnek kifejtésre.

12. A magyar online tartalomszolgáltatók és a sajtó folyamatos tájékoztatása a tervekről és a projekt előrehaladásáról. - A kommunikációs keretek kialakítása megtörtént, amelyet a későbbiekben részletezem.
13. A kísérleti időszakot követő állandó szolgáltatás fenntarthatóságának megtervezése. -Az állandó szolgáltatás keretei jogilag is alátámasztva kialakításra kerültek, 2020. január elsején történt indulással.

A kísérleti projekt kereteinek kialakításakor már a kezdetektől építeni kellett a társ digitális gyűjteményekben (Magyar Elektronikus Könyvtár, Elektronikus Periodika Archívum) felhalmozott tapasztalatokra, a gyarapítási munka során történő együttműködésre.

Az eredeti koncepció mellé számba vettem a személyi feltételeket (humán erőforrás) is. Az osztályvezető mellett megneveztük a kísérleti projekt összehangolását, irányítását gondozó témafelelős pozícióját, ezt a feladatkört a kezdetektől Drótos László látja el, legalább egy webkurátort, aki a gyűjtemény gondozásáért, az ezzel összefüggő jogi-adminisztratív teendők ellátásáért felel, Visky Ákos László dolgozik most ebben a pozícióban. További munkatársakat is szeretnénk bevonni a kurátori, jogkezelési tevékenységekbe. A webkönyvtáros feltárja a nemzetközi jó gyakorlatokat, a téma szakirodalmának összegyűjtését gondozza, részt vesz a nemzetközi kapcsolatok alakításában, a projekt szakmai tervezésében, az oktatási tevékenységek megalapozásában. Ezt a pozíciót töltöm be a kezdetektől fogva, jelen pillanatban is. Ez a webkönyvtárosi tevékenység adott gyakorlati kereteket számomra a jelen disszertációt megalapozó kutatás-fejlesztési tevékenységnek. Informatikai oldalról a hardver és szoftverstruktúra kiépítéséért és üzemeltetéséért felelős

rendszergazda, illetve egy alkalmazásfejlesztési és optimalizálási feladatokat ellátó programozó felvételét vettük tervbe. Ebből annyi valósult meg, hogy egy belső és egy külső foglalkoztatásban álló kolléga (Kovács Péter, illetve Vitéz Gábor) segíti negyed- illetve félállásban a munkánkat, napi szinten leginkább utóbbi segítségére tudunk támaszkodni.

3.3 A webarchiválás munkafolyamatainak áttekintése

Az általános koncepcionális keretek felvázolását követően az alábbiakban áttekintem azt a munkafolyamat rendszert, amit a webarchiválási kísérleti projekt kapcsán leszűrt tapasztalataink, valamint a nemzetközi jó gyakorlatok kapcsán alakítottunk ki. Ezzel adom meg a gyakorlati webarchiválási tevékenység alapvető kereteit. A kutatási hipotézisekben említett szolgáltatási környezeti kívánalomnak ez adja meg voltaképpen a lelkét.

3.3.1 Az általános webarchiválás keretei

Az általános, nemzeti domaint érintő webarchiválás munkafolyamata meglehetősen kevés lépésből áll. A nagymennyiségű begyűjtött adatot a jelenlegi technikai, illetve információfeldolgozási módszerekkel indexelni, visszakereshetővé tenni, minőségellenőrzési munkafolyamatokat végezni nem vagyunk képesek. Nem csupán mi, mint nemzeti könyvtár, de a nagyságrendekkel nagyobb infrastruktúrára támaszkodó domináns nemzetközi szereplők sem. A katalogizálás kereteinek terveit azonban megalkottuk, a kellő személyi és tárgyi feltételek megvalósulása esetén el is tudjuk kezdeni a megvalósítását (Németh, 2018b). Az aratás

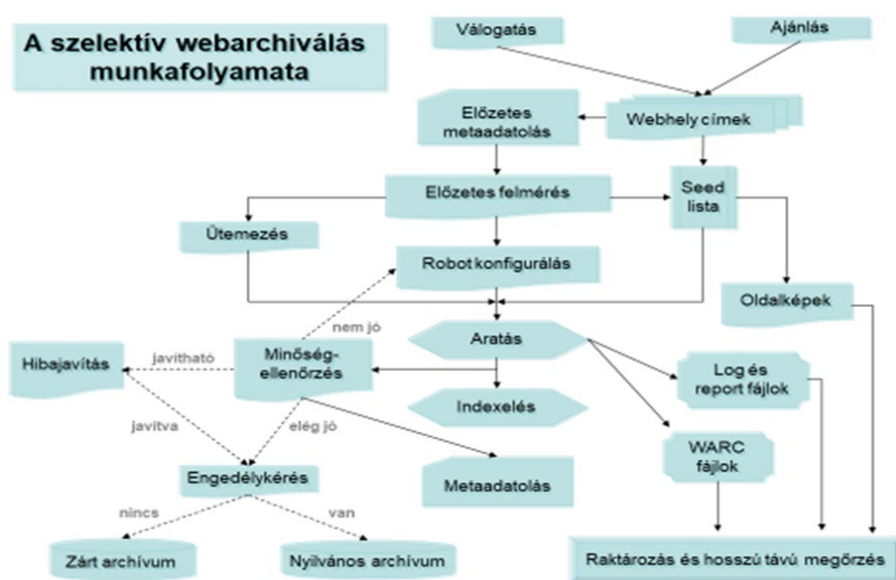
bemeneteként szolgáló seedlista segítségével, a munkafolyamat lezárultát követően az URL címek közötti keresést lehet biztosítani. A teljes szövegű indexelés lehetőségének kialakításával kísérletezünk (Németh, 2019a), ám a teljes feladatot eddig a megfelelő mennyiségű tárhely hiánya, illetve a ráfordítandó munkaidő szűkössége miatt még nem végeztük el. Tervezzük a hosszútávú hardver infrastruktúra kialakításával összhangban a teljesszövegű keresési szolgáltatás megvalósítását is (Németh, 2019a). Mivel a 2000-es évek elején a kifejezetten magyar tartalmakat feltáró keresők megszűntek, anyaguk eltűnt. Egy ilyen teljesszövegű kereső szolgáltatás hiánypótló szerepet tölthetne be a 2017-től rendelkezésre álló anyag vonatkozásában.

A leírt állomány pedig tömörített konténerállományokba, úgynevezett WARC fájlokba rendezve kerül eltárolásra a zárt archívumban. A WARC egy nyílt ISO szabványformátum (ISO 28500), amelynek legfontosabb jellemzői magyarul is elérhetők a vonatkozó angol nyelvű hivatkozásokkal együtt („WARC – MIA”, 2017). Miután a nyilvános szolgáltatásra a gyűjtemény kapcsán nem szerezzük be a megfelelő szerzői jogi engedélyeket, ezért ez az anyag csak dedikált könyvtári hálózaton, speciális munkaállomásokon keresztül lesz szolgáltatható, elsősorban kutatási célokra. A munkaállomások konfigurációjának garantálnia kell, hogy a megjelenített archivált anyagot ne lehessen lemásolni semmiféle külső adathordozóra.

3.3.2 A szelektív webarchiválás munkafolyamatai

A szelektív webarchiválás menedzselésének munkafolyamata sokkal összetettebb (Németh, 2018b), melyről az 5. ábra nyújt vizuális

áttekintést. A tematikus gyűjtemények alapvető hatóköre az adott témához illeszkedő webhelyek, illetve tematikus webhelyelemek (cikkek, rovatok, címkék). A gyűjtés jellegét tekintve inkább a teljes webhelyek gyűjtése a meghatározóbb. A munkafolyamatok rendszerének megalkotása lényeges részét képezte a webarchiválási projekthez kötődő kutatás-fejlesztési tevékenységemnek, melyet webkurátor és témafelelős munkatársammal együtt öntöttük a jelenlegi, lentebb ábrázolt formájába.



5. ábra A szelektív webarchiválás munkafolyamata

A webarchívumba kétféle módon lehet egy adott webhely bevitelét kezdeményezni. A webkurátorok tematikus listákat állítanak össze, s abban rögzítik az adott tételt, illetve javasolni is lehet webhelyeket megőrzésre az ajánló úrlapon keresztül. Miután ellenőriztük, hogy nem szerepel-e már az archívumban, s tematikus illeszkedést találunk valamely gyűjtemény felé, bekerül az aratandó címek közé (Drótos, 2018a).

A webhelyekről az ajánlásban megadott vagy a webkurátor által rögzített alapvető azonosító adatok (honlap neve, címe, tematikai besorolása) alapján a legfontosabb jellemzőket előzetes metaadatként rögzítjük. Ezzel párhuzamosan bekerül az adott tétel általában az adott témakörhöz kötődő jellemzői alapján a gyűjtemény webhelyeit nyilvántartó tematikus táblázatba. A táblázatokból állítjuk össze a robotszoftver által kezelhető aratási (seed) listákat. Az egyes webhelyekről célszerű egy előzetes gyors felmérést készíteni, hogy tartalmi szerkezetük, design elemeik alapján előreláthatólag igényelnek-e valamiféle speciális konfigurációs lépést az aratórobot szoftvert illetően. Az előzetes felmérést követően, illetve azzal párhuzamosan határozzuk meg az aratások időbeli ütemezését is. Az egyes webhelyek kezdőoldalairól képernyőképeket is készítünk, hogy az eredeti szerkezet, vizuális elrendezés minden esetben megtekinthető legyen. Ezek a képernyőképek, elkészültük után, egyenesen a hosszútávú megőrzést is biztosítani hivatott archív raktárba kerülnek. Az aratásokat az OSZKbot névre hallgató szoftver végzi a megadott seed listák és konfigurációs paraméterek alapján. Az aratás végeredményeként létrejött log és riport fájlok a raktárba kerülnek, a begyűjtés produktumaként létrejött WARC fájlokkal együtt. A begyűjtött anyag annak hatékony visszakeresése érdekében indexelésre is kerül. Az aratást követő minőség-ellenőrzés keretében döntjük el, hogy megfelelőnek minősítjük-e a létrejött mentéseket. Amennyiben az eredmény negatív, fel kell mérni, hogy az egyes hibák javíthatók-e, szükség van-e a robot újra konfigurálása után az aratási művelet megismétlésére. Amennyiben az eredmény megfelelő, akkor egymással párhuzamos munkafolyamatokra kerül sor. Egyrészt ellátjuk metaadatokkal a begyűjtött állományt,

másrészt amennyiben nyilvánosan is szolgáltatni szeretnénk azt, akkor el kell indítani a szolgáltatási engedély beszerzésére irányuló eljárást. Ennek eredményeként kerül az adott tartalom a zárt, illetve a nyilvános archívumba. A zárt archívumhoz a már említett módon dedikált könyvtári hálózaton keresztül, kutatási célból lehet hozzáférni. A nyílt archívumban pedig bárki számára nyilvánosan hozzáférhető a mentett tartalom. Fontos megjegyezni még, hogy a nyilvános archívum elérését megtiltjuk Google és egyéb keresőszolgáltatások robotjai számára. Ez azért fontos, mert a tartalomtulajdonos elsődleges érdeke, hogy a forgalom az eredeti, élő webhelyre irányuljon (amennyiben elérhető az még a világhálón) az archivált másolat helyett.

A munkafolyamatok legszűkebb keresztmetszetét egyértelműen a rendelkezésre álló emberi erőforrás jelenti (Drótos, 2018a). Bár számos feladat automatizáltnak elvégezhető, a webkurátor munkája elengedhetetlenül szükséges az egyes munkafolyamatok összehangolása, felügyelete terén. A legmunkaigényesebbek persze azok a munkafolyamatok melyeket nem lehet automatizálni, például az aratási típusokat, a begyűjtendő weboldalak körét meghatározó tartalmazó táblázatok elkészítését, vagy a jogi engedélykérési folyamatot. Az egyes munkafázisokon elegendő munkaerő hiánya esetén gyakran párhuzamosan kell dolgozni, nem jut elég idő, energia a teljes munkafolyamat rendszer összehangolására, az egyes feladatok magas színvonalú elvégzésére. A munkafolyamat szerkezetének és egyes komponenseinek meghatározása mellett tehát azzal egyenrangú szakmai kihívást jelent a napi szintű tevékenységek menedzselése is. A korlátozott

mértékű emberi erőforrással sok esetben szinte patikamérlegesen kell gazdálkodni, hogy minden szükséges feladat legalább a minimálisan megkívánt mértékben elvégezhető legyen.

3.3.3 Webhelyek válogatásának alapelvei a szelektív webarchiváláshoz

Az alábbiakban a szelektív webarchiváláshoz kötődő válogatási szempontokról, a listák összeállításának módszeréről, az összegyűjtött címek kezeléséről adok összefoglalást.

3.3.3.1 A válogatás szempontjai

A legfontosabb kiválasztási szempont, hogy az elsődleges gyűjtőkörünkbe (kultúra, tudomány, oktatás, közélet) tartozik-e az adott webhely, illetve kiemelt műfajba sorolható-e be (pl. időszaki kiadványok). Másrészt mérlegelni kell, hogy mennyi, lehetőleg egyedi saját, tartalom található a webhelyen, milyen színvonalon szerkesztett és mennyire széles körnek érdeklődésére tart számot. Fontos szempont, hogy mennyire jól archiválható (pl. ne legyen a robotok által nem lekérdezhető szolgáltatófelülettel bíró adatbázis) az adott webhely, illetve, hogy frissül-e még a tartalma vagy már lezárt tartalmi egységről van szó. Ezek nem egymást kizáró szempontok, csak éppen az általános gyakorlattól eltérő kezelést igényelnek: például egy adatbázist is nyilván lehet tartani, mint fontos elemét az adott témakörnek, de nem kell aratni, hanem majd esetleg bekérni a tulajdonosától; illetve a már lezárt, de értékes tartalmú webhelyekről elég évente csak egy biztonsági mentést végezni, nem érdemes az általános negyedéves ütemezéssel aratni.

3.3.3.2 Címgyűjtemény létrehozása

A címgyűjtemény létrehozásakor többféle módszert érdemes párhuzamosan használni a minél alaposabb eredmény elérése érdekében, igénybe véve az adott témában jártas forrásmunkák, illetve szakemberek segítségét is. Ennek során igénybe lehet venni már korábban készített összeállításokat is, de ezek adatait csupán újabb ellenőrzés, szűrés és egyeztetés után lehet felhasználni.

1.) A legegyszerűbb forrás valakinek az ajánlása, archiválási javaslata. Ezzel az eszközzel bárki élhet, de jellemzően azok teszik meg, akik valamilyen okból érintettek a tartalom tematikáját és/vagy létrejöttét tekintve (tulajdonos-közreműködő, felhasználó, érdeklődő).

2.) Hasznos forrás, nagyobb merítésre ad lehetőséget valamilyen (link)gyűjtemény átnézése, illetve bővülésének nyomon követése. Jó példa erre az elektronikus időszaki kiadványok esetében az Elektronikus Periodika Archívum és Adatbázis (EPA) által folyamatosan végzett katalogizálás: minden új tétel továbbításra kerül a webarchívum számára is.

	A	B	C	D	E	F	G	H
1	Szín-kód	A webhely neve (eredeti megnevezés, esetleg zárójeles kiegészítésekkel)	Seed URL (domén vagy aldomén, gyökérkönyvtár vagy alkönyvtár, kezdőlap)	Alternatív URL (szintén aratandó címváltozat, az ütemény részének a MIA azonosító)	Egyéb URL (számokat tartalmazó fájlok, ill. nem archivált)	MIA_ID		EPA URL (a kiadvány oldala az EPA)
2						MIA-904757		
61	GY	Prospecto.hu – Akciós újság, szórólap.	http://prospecto.hu/					-
62	GY	Retroújság - Sok régi ismerős egy új he	http://retrojsag.com/					-
63	GY	Régió Lapkiadó Kft.	http://www.regiolapkiado.hu/					-
64	GY	Sajtóforrás.hu - Magyarországon terjesz	http://sajtoforras.hu/					-
65	GY	scripta			http://www.c3.hu/s			-
66	GY	Szülővilág - Magazinok	http://szulovilag.hu/					-
67	GY	Tudományos közlemények (Liceum Kiad	http://webshop.uni-eszterhazy.hu/hu/terms					-
68	GY	Ujsagomat.hu Akciós újságok online	http://www.ujsagomat.hu/					-
69	GY	Újság lap			http://ujsag.lap.hu/			-
70	GY	Újságosstand	http://www.pc.shop.hu/ujsagm.htm					-
71		MAGAZINOK:						
72	M	10/10 Magazin - Az online talponálló	http://www.tizperiz.hu/			MIA-900001		-
73	M	108 Magazin	http://www.108magazin.hu/			MIA-900002	http://epa.oszk.hu/01900/	
74	M	4x4 online	http://www.4x4.hu/			MIA-900003		
75	M	5perc Angol Magazin	https://www.5percangol.hu/5percangol_ma	https://www.5percangol.hu/		MIA-904662	https://epa.oszk.hu/03800/	
76	M	777 - A hit nem magánügy!	http://777blog.hu/			MIA-900004		-
77	M	7óra7 - Színház és pont	http://7ora7.hu/			MIA-900005	http://epa.oszk.hu/03600/	
78	M	A Földgömb - A Magyar Földrajzi Társa	http://www.heiling-media.hu/			MIA-900006	http://epa.oszk.hu/03000/	
79	M	A Földgömb (A Földgömb világa) - A Ma	http://www.afoldgomb.hu/			MIA-900007	http://epa.oszk.hu/01300/	
80	M	A hetedik sor közepe	http://hetediksor.hu/			MIA-900008	https://epa.oszk.hu/03800/	
81	M	A kontroll	http://www.akontroll.hu/			MIA-900009	http://epa.oszk.hu/01900/	
82	M	A Muzikus Világa - A Muzikus Club H	http://www.muzikus.hu/			MIA-900010	http://epa.oszk.hu/03600/	
83	M	active beauty - A szépség, divat és élet	http://www.activebeauty.hu/			MIA-900011	http://epa.oszk.hu/03600/	
84	M	Addiktológia			http://www.mat.org	MIA-900012	http://epa.oszk.hu/01300/	
85	M	Aerobicmix Mix Magazin - minden hír	http://www.aerobicmix.hu/			MIA-900013	https://epa.oszk.hu/03800/	
86	M	Aeromagazin	http://www.aeromagazin.hu/			MIA-900014	http://epa.oszk.hu/00000/	
87	M	Agrár Élet - Az agrár magazin - input-gé	http://www.agrarelet.hu/			MIA-900015	http://epa.oszk.hu/03600/	
88	M	AgrárUnió - az egyik legkedveltebb és le	http://www.agrarunio.hu/			MIA-900016	http://epa.oszk.hu/01300/	
89	M	Agro Napió - A mezőgazdasági hírportál	http://www.agronapio.hu/	https://issuu.com/agronapio		MIA-900017	http://epa.oszk.hu/00000/	
90	M	Ahimsa - életmód magazin útkeresőkne	http://www.ahimsa.hu/			MIA-900018	https://epa.oszk.hu/03800/	
91	M	AIRportal.hu - Légiközlekedési és légiut	http://www.airportal.hu/			MIA-900019	https://epa.oszk.hu/03700/	
92	M	AikiMag - Az Aikido Magazin - Harcműv	http://aikimag.blog.hu/			MIA-900020	https://epa.oszk.hu/03800/	
93	M	Akvárium Magazin	http://www.akvariummagazin.hu/			MIA-900021	http://epa.oszk.hu/03700/	

6. ábra Az elektronikus periodikák nyilvántartása

3.) Célzott címfeltárássra ad lehetőséget - amennyiben találunk ilyet - egy domén listából szűréssel leválogatni a releváns URL-eket. Ebben az esetben sajnos hibaforrás, hogy csak adott kifejezésekre tudunk keresni, és amennyiben a keresett tartalom nevében vagy leírásában az nem szerepel, kimarad a találatok közül.

4.) Végül nem hiányozhat az internetes keresés módszere sem. A találati halmaz nagysága függ a keresőkifejezés pontosságától, szűkíthetőségétől. Emiatt nem általános gyűjtésre érdemes használni (pl. rákeresni az "irodalom" szóra, de a "kortárs irodalom" kifejezés már kezelhetőbb eredményt adhat), hanem kiegészítésképpen, pontosítás céljából (pl. a

linkgyűjteményben tárolt URL már nem él, ekkor a név alapján próbálhatjuk megkeresni az új elérhetőséget). Hasznos lehet nem ismert linkgyűjtemény keresése a találati listában.

3.3.3.3 Címek kezelése

Az összegyűjtött címeket rendezni, csoportosítani szükséges a gyorsabb áttekinthetőség céljából. Minél nagyobb egy-egy gyűjtemény, annál részletesebb, több szempontú besorolást lehet alkalmazni, de ügyelni kell arra, hogy ne merev, hanem az adott gyűjteményhez illeszkedő elválasztást alkalmazzunk. Csoportképző lehet a webhely műfaja (honlap, blog, stb.), jellege (pl. egyházi), területisége (pl. országos, helyi), és adott esetben, átfedés esetében mérlegelni kell, melyik a fontosabb jellemző (pl. levéltáraknál egy egyházmegyei levéltár alapjellemezője az egyházi volta, nem a területisége, de annak blogját már érdemes lehet műfaj szerint elkülöníteni).

1.) Az első és legfontosabb feladat a talált címek ellenőrzése: működő címről van-e szó, van-e releváns tartalom az adott webhelyen?

2.) A gyűjteménybe kerülő webhelyek egyedi azonosítót kapnak (MIA-000000 struktúrában). Nevüket pedig egységesített módon kell felvenni (a névhez hozzátesszük a webhely jellegét is, pl. honlapja, blogja, stb.), hogy elkülöníthetők legyenek az azonos néven futó, de eltérő jellegű webhelyek), de külön adatként rögzíteni kell az eredeti formát is. Több névalak esetében a hivatalos változatot, annak hiányában a leginkább kifejezőt kell felvenni - ezek meghatározásához a szükséges forrásokat kell felhasználni (pl. a webhelyen előforduló alakok, egyedi azonosítás, külső források stb.). Intézményi nevek esetében teljes névalak (címstruktúra),

rövidítés és székhely szükséges a pontos azonosításhoz. Végül rögzíteni kell a felvétel dátumát is.

3.) Meg kell határozni a webhely legmagasabb URL címét (ez nem feltétlenül azonos a böngészőben megjelenő formulával!), mely kiindulási címként szolgál majd, ahonnan legteljesebben bejárható az adott tartalom. A legmagasabb cím általában a domén vagy aldomén szintjén ér véget, az utána következő alkönyvtárakat és fájlokat (pl. /hu, /index.html stb.) már nem tartalmazza. Speciális esetekben ettől el lehet térni (pl. olyan idegen nyelvű honlap esetében, ahol csak a magyar tartalom fontos számunkra). Egyúttal fel kell tární és rögzíteni az adott doménen belüli aldoméneket, az alternatív URL-eket és címvariánsokat is (pl. http, https, www).

4.) Érdemes rögzíteni a webhelyen fellelhető egyéb fontos információkat is, ami később hasznos lehet, pl. kapcsolati személy(ek) és elérhetőségek feljegyzése.

5.) Meghatározott időközönként érdemes frissíteni a szelektív archiválásba vont webhelyek körét. Hasznos lehet új forrás felbukkanása, a témával közelebbi kapcsolatban lévő partnerek ismeretanyagának bevonása (pl. helyi honlapok), beérkező ajánlások besorolása.

6.) Az adatok nyilvántartásának legegyszerűbb eszköze egy táblázat. A kísérleti szakasz során Google-táblázatokat (pl. MUZGAL) használtunk erre a célra, mert könnyen hozzáfér több munkatárs is. Hátránya, hogy a szigorú adatkezelés miatt személyes adatot nem célszerű tárolni benne. Ezeket az adatokat vagy külön, biztonságos helyen tárolt változatban lehet

gyűjteni (erre nyújthat lehetőséget például az OSZK-ban is bevezetésre kerülő MICROSOFT OFFICE 365 platform), vagy egy, hosszabb távon eleve használhatóbb megoldást jelentő adatbázisban. Ez utóbbi remélhetően az Országos Könyvtári Platform kialakításának munkálatai során születik majd meg a későbbiekben.

7.) Az archiválás során figyelni kell az egyedi szempontok (pl. sürgősség) érvényesülésére is. Amennyiben pl. veszélyeztetett webhely jut tudomásunkra – akár már ismert, akár ismeretlen –, nem elég felvenni egy majdan aratandó listára, hanem egyedi mentést kell készíteni róla.

3.3.4 Eseményalapú webarchiválás

Az eseményalapú webarchiválásnak, bár az alapvető munkafolyamat keretrendszere lényegében azonos a szelektív webarchiváláséval, s a válogatási szempontok jó része is érvényesül, néhány lényeges ponton mégis eltér attól. Más a válogatási módszer, hiszen itt nem egy általános témához kötődő webhelyek számítanak alapvető egységnek, hanem egy adott eseményhez, történéshez szorosan kötődő címkék, rovatok, cikkek alkotják a gyűjtőkört. Eltér az aratás módszere is, mivel adott tartalmakra koncentrálunk, így csupán maximum két szint mélységig mentjük le a konkrét eseményalapú tartalmakat. Az aratás ütemezésében is találunk eltéréseket, hiszen itt hetente kell az aktuális friss információkat begyűjteni. A gyűjtemény anyagának hatóköre is lényegesen eltér a szelektív aratásokétól. Utóbbi esetben ugyanis nem használjuk fel az általános hírportálok anyagát, hiszen tárhely és módszerek sincsenek egyelőre azok teljes aratására, melyhez egyébként a nemzetközi gyakorlat szerint az adott tartalomszolgáltató aktív közreműködése is szükséges

lenne, hiszen akár naponta kellene menteni a teljes aktuális tartalmat. 2020-ban kezdtünk el egy kísérletet a Brozzler nevű szoftver segítségével a legjelentősebb hírportálok címlapjainak napi szintű mentésére.

Az eseményalapú aratásnál azonban az adott eseményhez kötődő cikkekhez, rovatokhoz, címkékhez tartozó tartalom a hírportálokról is természetesen a gyűjtőkör részét képezi, hiszen korlátozott mennyiségű célzott információgyűjtésről van szó. A tematikus gyűjtemények között találunk olyanokat, amelyek csupán a zárt archívumban érhetők el (például a parlamenti és önkormányzati választásokkal, vagy a mostanában gyűjtött koronavírus járvány online visszhangjához kötődő anyagok).

Részben nyilvános gyűjteményként jelenik meg viszont, attól függően, hogy a tematikusan gyűjtött tartalom mekkora hányadának nyilvános szolgáltatására kapunk engedélyt, a 2019-ben a Közgyűjteményi Digitalizálási Stratégia (KDS) keretei között kialakított Rákóczi gyűjtemény. A gyűjtemény, illetve az ahhoz kötődő mintaalkalmazás részletes ismertetését webkurátor és programfelelős kollégáim végezték el (Visky & Drótos, 2020). Ennek külön tematikus honlapot hoztunk létre, melyen nem csupán a webarchívum Rákóczi- emlékévéhez kötődő eseményalapú gyűjtése jelenik meg, hanem társgyűjteményeink, a Magyar Elektronikus Könyvtár, illetve az Elektronikus Periodika Archívum tematikus anyaga is.



Ismerető



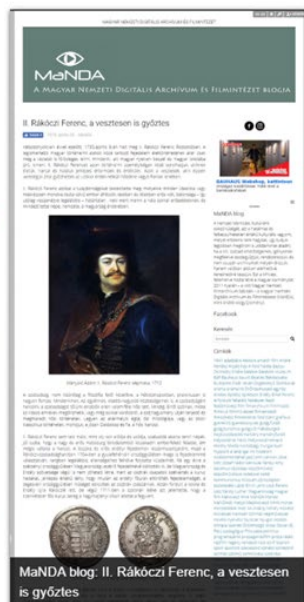
Ez a digitális archívum 2019-2020-ban a **II. Rákóczi Ferenc Emlékév** alkalmából készült az Országos Széchényi Könyvtárban, a Közgűjteményi Digitalizálási Stratégia pályázat keretében. Elsődleges célja a webarchiválás technológiájának és a webarchívum más digitális gyűjteményekkel való integrálásának bemutatása egy mintaalkalmazás keretében. Az Emlékévvvel, a Fejedelemmel, a Szabadságharcra és a Rákóczi-családdal kapcsolatos weboldalak és webhelyek mentései mellett a Magyar Elektronikus Könyvtárban levő digitális vagy digitalizált könyvek, az Elektronikus Periodika Archivumban található cikkek és a Digitális Képtárházból származó fotók, illusztrációk és egyéb képi dokumentumok is böngészhetők ezen a honlapon. Így egyben a 17. és 18. század magyar történelmének, társadalmi- és hadtörténetének gazdag forrásként oktatási és kutatási célokra is használható ez az összeállítás.

A böngészés és a keresés a webarchívum mellett a másik három gyűjtemény metaadatai szerint is lehetséges, a találati listákban a színes sávok jelzik,

hogy az adott tétel melyik állományból származik, a lakat ikonok pedig azt, hogy a dokumentum nyilvánosan megtekinthető-e. Piros vagy sárga lakat azoknál a weboldalakkal fordul elő, amelyek archivált példányainak szolgáltatására a tartalomgazda nem, vagy még nem adott engedélyt, de a metaadatok, a kis méretű oldalkép és az Internet Archive mentései ezekben az esetekben is megnézhetők. A teljes szövegű keresés csak a webarchívumra terjed ki és annak is csak a nyilvánosan szolgáltatható részére.

Mivel a web egy komplex médium, ezért a webes tartalmak lementése, hosszú távú megőrzése és megtekinthető állapotban tartása komoly informatikai kihívás. Az OSZK webarchívuma által készített mentések többféle szoftverrel készültek és nézhetők vissza, így összehasonlíthatók az archiváló és a megjelenítő eszközök képességei. Két másik nyilvános részgyűjtemény is elérhető további tanulságos példakkal: az egyik a nemzeti könyvtár saját online szolgáltatásairól készült, a másik pedig egyéb webhelyekből áll.

Az Országos Széchényi Könyvtárban 2017-ben indult webarchiválási projektről a <http://mekosztaly.oszk.hu/mia> honlapon lehet tájékozódni, további információk pedig a mia@mek.oszk.hu levélcímen kérhetők.



Archiválásra javaslom ...



7. ábra A Rákóczi webarchívum gyűjtemény kezdőlapja

Így egyben arra is jó példát kapunk, hogy miként lehet tág hatókörű tematikus szolgáltatást, többé digitális könyvtári gyűjteményre

támaszkodva megvalósítani (Visky & Drótos, 2020). A Trianon-emlékévhöz kötődő tematikus gyűjtés kapcsán is hasonló tágabb körű tartalomszolgáltatást kívánunk majd a későbbiekben kialakítani, nyilvánosan szolgáltatva a tartalomtulajdonos által engedélyezett archivált webhelyeket.

Egyes eseményalapú gyűjtemények létrehozatalának legfontosabb indoka az internet archiválásért felelős nemzetközi konzorcium (IIPC) által kezdeményezett gyűjtési kezdeményezésekhez történő csatlakozás volt. Erre szolgált példaként például az egyes olimpiai játékok, vagy az Új-Zélandon bekövetkezett christchurchi merényletsorozat online lenyomatának feltárása és archiválása. A szakmai koordinációt az IIPC tartalomfejlesztési munkacsoportja látja el, a technikai hátteret pedig az Archive-IT biztosítja („IIPC Content Development Working Group”, 2019). Ezekben az esetekben elsődlegesen saját táblázatokban bonyolítottuk le a gyűjtést, majd annak címlistáit megosztottuk nemzetközileg megosztott Google táblázatok segítségével partnereinkkel is. Ezekben az esetekben természetesen néhány ponton módosítani kellett a rögzített adatok tartalmán, hiszen az itthoni rögzítési konvenciók, illetve az egyes webhelyekről nemzetközileg kért adatok nem minden esetben estek egybe egymással.

3.4 A gyűjtőkör, a gyűjteményfejlesztés alapvető kereteinek tervezete

A munkafolyamatok kialakítása mellett egy webarchívum (s bármilyen könyvtári jellegű gyűjtemény) megalapozásának alapvető kiinduló lépése a gyűjtőkör meghatározása. Mint már látható volt, a projekt kezdetén

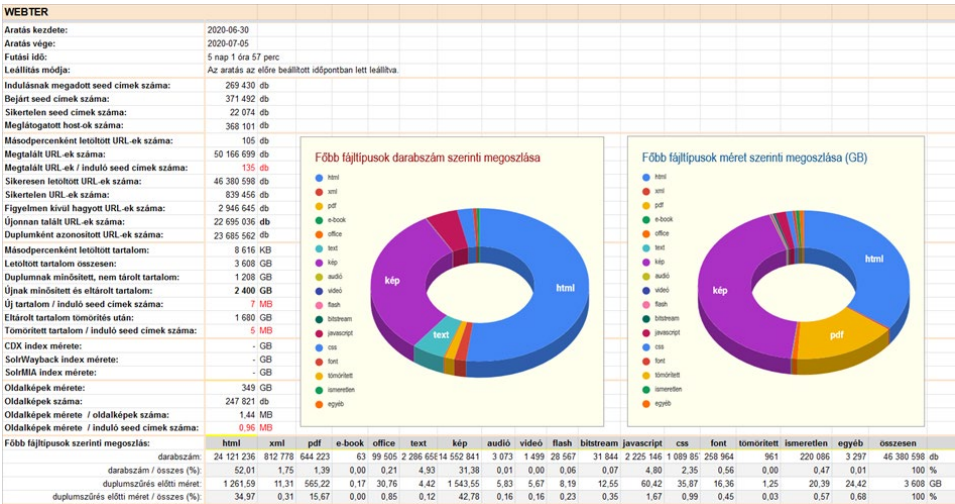
megfogalmazott követelménylistában is kiemelt helyen szerepelt. Az alábbiakban arról adok áttekintést, hogy az Országos Széchényi Könyvtár általános gyűjtőköri szabályozásának figyelembevételével milyen általános alapelv tervezet alakítottunk ki munkatársaimmal együtt 2017 júniusában a webarchívum vonatkozásában (OSZK Webarchiválási munkacsoport, 2017). Az itt felsorolt gyűjtőköri irányelvek a kísérleti projekt időszakára érvényesek. A webarchiválást szabályozó jogi háttér megszületését követően 2021-től válnak majd ezek az irányelvek remélhetőleg az OSZK-ra vonatkozó külső, illetve az állandó szolgáltatások gyűjtőkörét szabályozó belső szabályozások alkotó elemeivé. A tervezéshez szükséges nemzetközi összehasonlítási alapot az IIPC összehasonlítása adta, amelyben összegyűjtik néhány nemzeti könyvtár gyűjtőköri alapelveit („IIPC - Collection development policies”, 2018). Ezek feldolgozása, illetve a magyar irányelvek kialakítása is fontos feladat volt a kísérleti projekt munkálatai során.

3.4.1 Alapelvek

Az Országos Széchényi Könyvtár gyűjtőköri szabályzatában rögzített feladatai közé tartozik a magyar webes tartalmak gyűjtése, reprezentatív képet nyújtva ily módon a múltban egy adott időpillanatban elérhető, a magyar közönségnek szánt, a kulturális örökség részét képező online tartalomkínálatról, a hungarikumok körébe tartozó és nyilvános szolgáltatási felületekről letölthető elektronikus dokumentumokról.

A dokumentumok gyűjtése háromféle módon történik: válogatva a legfontosabb magyar webhelyekről, kiemelt eseményekhez kötődve a főbb hírforrásokból, illetve általános jelleggel a magyar webtérrel.

Szelektíven kerül gyűjtésre a tudományos, kulturális, oktatási, közéleti jellegű tartalmak meghatározott köre. Az általános gyűjtés a .hu domén alatt regisztrált, vagy egyéb doménhez tartozó, de magyar közönséget megcélzó nyilvános webhelyekre terjed ki, az illegálisnak, illetve pornográfnek minősülő tartalmak kiszűrésével, valamint a nagyméretű multimédia és egyéb állományok kihagyásával.



8. ábra A magyar webtér aratásának nyilvántartása

A webaratás csupán azon webhelyeket érinti, ahonnan technikailag biztosítható a tartalom automatikus begyűjtése. Az aratási művelet során a könyvtár figyelembe veszi a begyűjtő robot számára az adott webhely tulajdonosa által beállított korlátozásokat (robots.txt).

Az archivált webhelyek esetében a könyvtár elsősorban azok tartalmának hosszú távú megőrzésére törekszik. A begyűjtött tartalom szolgáltatása során teljes mértékben garantálni kell, hogy a jogtulajdonosok szellemi tulajdonjoga ne sérüljön. Nyilvános szolgáltatás csak azokról a

webhelyekről végezhető, amelyek esetében a tartalomgazda erre engedélyt adott, a többi szelektíven archivált honlap csak a könyvtár zárt szolgáltatási felületén érhető el, az általános webtér aratás anyaga pedig a törvényi szabályozás megszületéséig, illetve a technikai feltételek megteremtéséig nem hozzáférhető.

Az OSZK gyűjtőköri szabályzatának 10.12. pontja előírja az „interneten közzétett, nyilvános szolgáltatási felületről letöltött, a könyvtár megfelelő szolgáltatási felületén szabadon szolgáltatatható, hungarikumkörbe tartozó elektronikus dokumentumok; egyéb, a könyvtár nemzeti könyvtári gyűjtőkörébe tartozó, távoli hozzáférésű (online) dokumentumok; online és digitalizált időszak kiadványok - archiválás és szolgáltatás céljára” (Az Országos Széchényi Könyvtár gyűjtőköri szabályzata, 2015) történő gyűjtését.

Az online dokumentumok elektronikus gyűjtésére az alábbi részletes előírások vonatkoznak: Az online dokumentumok mennyisége, tárhelyigénye, műfaj- illetve színvonalbeli változatossága miatt gyűjtésük válogatott mind a hazai dokumentumokból, mind a külföldi hungarikumokból - kivéve a kötelespéldányként beszolgáltatott online dokumentumokat. Online dokumentumok esetén a könyvtár elsősorban a dokumentumok tartalmának hosszú távú megőrzésére törekszik. Abban az esetben, ha az aktuális információhordozó formátumhasználata lehetetlenné válik, a benne lévő tartalom szinte kinyerhetetlen, a könyvtár más, arra alkalmas formában, formátumban igyekszik azt megőrizni. A különböző - változó - formátumban meglévő dokumentumok használata

érdekében a könyvtár törekszik egy megfelelő szoftverarchívum kialakítására is.

3.4.2 Az elektronikus hungarikumok gyűjtőköre

A Könyvtár a felsorolt online elektronikus kiadványok közül a hungarikumok körébe tartozóan a következőket gyűjti (Az Országos Széchényi Könyvtár gyűjtőköri szabályzata, 2015):

- Magyarországon bejegyzett doménnevek alatt megjelenő, elektronikus formában létrehozott és az interneten nyilvánosságra hozott kiadványok, akár ingyenesen, akár térítés fejében tették hozzáférhetővé;
- hazai költségvetési forrásból létrehozott vagy támogatott, határon túli, hungarikumnak minősülő, online elektronikus kiadványok;
- Magyarországon vagy külföldön korábban előállított és belföldön kiadott, többszörözött formában (nyomtatásban vagy valamely elektronikus adathordozón) megjelent és terjesztésre szánt kiadványból készült online elektronikus kiadványok.

3.4.3 Az online kiadványtípusok gyűjtőköre

Statikus online elektronikus kiadványok (Az Országos Széchényi Könyvtár gyűjtőköri szabályzata, 2015)

- audiovizuális kiadvány
- folytatódó kiadvány (e-periodikum)
- hírlevél

- kartográfiai kiadvány
- konferenciaanyag
- kotta
- könyv (e-könyv)
- kutatási jelentés
- multimédia kiadvány
- műszaki jelentés
- statikus tartalmú adattár (online szótár, lexikon stb.)
- szabályzat

Dinamikus online elektronikus kiadványok (Az Országos Széchényi Könyvtár gyűjtőköri szabályzata, 2015)

- kulturális célú, esetleg közpénz igénybevételével készült, a nemzeti kulturális örökség részének szánt, dinamikus tartalmú adattár (online szótár, lexikon, szövegkritikai kiadás stb.);
- online tudományos és kulturális tartalmú időszaki kiadványok;
- nem hosszú távú fennmaradásra szánt, dinamikus tartalmú adattár - egyedi mérlegelés alapján ezeket csak akkor gyűjti a könyvtár, ha az adatbázisként strukturált dokumentum a struktúra megőrzésével archiválható.

- Az ide sorolt dokumentumtípusok egy online nemzeti könyvtár középponti gyűjteményét képezik, amelyek hosszú távú megőrzése az OSZK feladata. E feladatnak a Könyvtár csak akkor tud eleget tenni, ha az archiválásnak, ill. az archivált változat használatba bocsátásának technikai feltételei is adóttak.”

Online dokumentumok automatikus gyarapítása (webaratás)

A következő dinamikus online kiadványtípusok webaratás útján kerülhetnek a könyvtár gyűjteményébe (Az Országos Széchényi Könyvtár gyűjtőköri szabályzata, 2015):

- híroldal, hírcsoport
- portál
- honlap
- weblog, blog, napló
- levelezőlista, e-postai üzenetek
- faliújság
- társalgási szoba,
- nem platformfüggetlen, illetve nem tartalomszolgáltatási céllal létrejövő futtatható állományok (program, szoftver, játék stb.)

3.4.4 A könyvtári gyűjtőkör lehatárolása, együttműködés külső intézményi partnerekkel

A könyvtár nem gyűjti a következő online elektronikus dokumentumtípusokat, mivel ezek gyűjtése és archiválása más intézmények feladata (Az Országos Széchényi Könyvtár gyűjtőköri szabályzata, 2015):

- kormányzati dokumentumok, jogszabályok (megőrzésükről az Országgyűlési Könyvtár gondoskodik),
- szabványok (megőrzésükről a Magyar Szabványügyi Testület gondoskodik),
- szabadalmi dokumentumok (megőrzésükről az Szellemi Tulajdon Nemzeti Hivatala gondoskodik),
- rádió vagy televízió által sugárzott műsorok (megőrzésükről a Nemzeti Audiovizuális Archívum gondoskodik),
- mozgófilmek (megőrzésükről a MaNDA gondoskodik).

Az OSZK webarchívumra vonatkozó gyűjtőkörének alapvető szabályozása, a 2021. január elsejétől hatályos jogi szabályozás keretében jelenik meg (1358/2020. (VII. 1.) Korm. Határozat, 2020). A fentebb felsorolt alapfeladatok archivált webes anyagok kezelésére történő kiterjesztése, a gyűjtőköri szabályzati elemek részletes kidolgozása további intézményi szintű szabályozási intézkedéseket követel majd meg.

3.5 A kísérleti projekt hardver és szoftver háttere

A munkafolyamat-modell, illetve a gyűjtőköri tervezet felvázolását követően az alábbiakban arról a hardver és szoftver háttérrel adok

áttekintést, melyet a webarchiválási kísérleti projekt lebonyolításához, illetve a leendő állandó szolgáltatás alapjainak megteremtéséhez kellett kialakítanunk.

3.5.1 A hardver infrastruktúra elemei

A teljes magyar webtér learatása hardverigényéhez kötődő tervezési – tesztelési munkálatok 2017 elejétől indultak el. A konkrét hardver paraméterek az OSZK új technikai infrastruktúrájának kiépítésének függvényében folyamatosan változnak, a lentebb leírt állapot a 2019. év végi helyzetet tükrözi.

Az OSZK Informatikai Igazgatóságának munkatársa, Kovács Péter már a korábbi években is tesztelte a különféle szoftvertermékek képességeit, hardverkövetelményeit, amikor még nem volt világos, hogy milyen keretek között lehet a webarchiválási tevékenységet az intézményben elindítani, de már egyre gyűltek az információk a környező országok jó gyakorlatairól. Végül a pilot projekt szerver háttérét a KIFÜ, illetve az OSZK bázisán alakítottuk ki. A KIFÜ szerveren kapott helyet a zárt archívum, az OSZK szerveren pedig egy kisebb teljesítményű szerveren a nyilvános demo archívum, a metaadatok, illetve a teljes szövegű keresés szolgáltatás került beüzemelésre. 2018 végén egy másik szervert is kialakítottunk melyen a Rákóczi tematikus gyűjtemény, illetve a leendő állandó szolgáltatás háttérül szánt Wordpress tartalomkezelő rendszerben épített honlap kapott helyet. (OSZK Webarchiválási munkacsoport, 2018).

Tájékoztató a honlapról

Az Országos Széchényi Könyvtár a 2000-es évek elejétől kezdett foglalkozni a digitálisan születő dokumentumok megőrzésével, előbb a könyvekkel foglalkozó Magyar Elektronikus Könyvtár, majd az időszaki kiadványok számai archíváló Elektronikus Periodika Archivum és Adatbázis és a képi dokumentumokat gyűjtő Digitális Képtárház keretében. Már 2006-ban felmerült, hogy más országokhoz hasonlóan egy webarchívumot is létrehozzon a nemzeti könyvtár, ami 2015-ben a **gyűjtőkori szabályzatában** is rögzítésre került. A szükséges informatikai és személyi feltételek végül 2017-ben, az Országos Könyvtári Rendszer Projektnek köszönhetően valósultak meg és egy két és fél éves tanulási és teszt időszak után elkezdődött a magyar webterületen közzétett tartalmak egy részének időszakos lementése és gyűjteménybe szervezése. Az OSZK webarchívuma azzal a céllal jött létre, hogy reprezentatív képet nyújtson az egy adott időszakban nyilvánosan elérhető, a magyar közönségnek szánt és a kulturális örökség részét képező online tartalomkínálatról, a hungarikumok körébe tartozó elektronikus dokumentumokról.

A dokumentumok gyűjtése háromféle módon történik: **válogatva** a legfontosabb magyar webhelyekről, kiemelt **eseményekhez kötődve** a főbb hírforrásokból, illetve **általános jelleggel** a magyar webterületről. Szelektíven kerül gyűjtésre a tudományos, kulturális, oktatási, közéleti jellegű tartalmak meghatározott köre. Az általános gyűjtés a .hu domén alatt regisztrált vagy egyéb doménhez tartozó, de magyar közönségnek megcélzó nyilvános webhelyekre terjed ki. A webarchívum csupán azon szervezeteket érinti, ahonnan technikailag biztosítható a tartalom automatikus lementése. Az archíválás során a könyvtár figyelembe veszi a begyűjtő szoftver számára az adott webhely tulajdonosa által beállított **korlátozásokat**.

Az archívált webtartalom esetében a nemzeti könyvtár elsősorban annak hosszú távú megőrzésére törekszik. A szerzői és személyiségi jogok tiszteletben tartása érdekében a gyűjtemények csak egy kis része tekinthető meg **nyilvánosan**, olyan webhelyek, amelyeknél a tartalom tulajdonosa erre engedélyt adott. Az archívum többi része csak a könyvtár zárt szolgáltatási felületén lesz elérhető – elsősorban kutatói célokra –, a jogi szabályozás megszűlése után.

Mivel az internetes tartalmak archiválásával kapcsolatban hazánkban még kevés a tapasztalat, így ez a honlap egyben tudásbázisként is kíván szolgálni ezen a területen. Ezért a webarchívum és a projekt bemutatása mellett a főmenüből külön csoportosításban érhető el a felhasználóknak, a tartalomgazdáknak, a téma iránt érdeklődő szakembereknek és újságíróknak szóló tájékoztatók, dokumentumok és egyéb információforrások.

Figyelem! A honlap jelenleg még fejlesztés alatt van, egyes kereső és a böngésző funkciók még nem működnek.

Gyorslinkek:

Keresés

Demó archivum

OSZK-s archivum

Rákóczi archivum

Archiválásra javasolm...



Hírek, események

2020-02-29

Új honlap

Elkészült a webarchívum új honlapja a webarchivum.oszk.hu címen (alternatív URL: <https://webarchive.hu>). Ezentúl már csak itt jelennek meg a projekttel kapcsolatos hírek és információk, bár egy ideig még a régi honlap is elérhető marad. Az új...

2020-02-10

Üdvözlő videó a spanyol webarchívum születésnapjára

A spanyol webarchívum fennállásának 10. évfordulóján, február 20-án tartandó rendezvényhez az ottani nemzeti könyvtárban dolgozó kollégák egy-egy rövid videót kértek az IIPC tagoktól. A magyar webarchívumot bemutató és a nemzetközi kapcsolatok fontosságát hangsúlyozó összeállításunk...

● ● ● ● ●

2020-02-04

Rákóczi Emlékév Archivum

A II. Rákóczi Ferenc Emlékév alkalmából a Kögyűjteményi Digitalizálási Stratégia pályázat keretében elkészült egy mintaalakítás, amely a webarchívum integrálásának lehetőségét mutatja be más könyvtári digitális gyűjteményekkel: <https://rakoczi2019.webarchivum.oszk.hu/>

Hirarchívum

Hónap kiválasztása ▾

1014 Budapest, Szent György tér 4-5-6. mia@mek.oszk.hu +36-1-487-8688

OSZK honlap Jogi közlemény Honlapértékp és -kereső Régi honlap Kapcsolat

Az első szerver paraméterei az alábbiak voltak (webadmin. oszk.hu): Webadmin (webadmin.oszk.hu): 4 CPU mag + 16 GB memória + 6.5 TB iscsi tárhely (5 + 1.5 TB külön kiosztva) ebből 3.5 TB szabad. Erről sincsen egyelőre biztonsági mentésünk. 2018 őszén egy Samba-alapú megosztott tárhely lett kialakítva ezen a szerveren, ami így virtuális meghajtóként felcsatolható a Windows gépekre, megkönnyítve ezzel a fájlok cseréjét.

A második szerver címe webarchivum.oszk.hu, ennek paraméterei a következők voltak: 8 CPU mag + 16 GB memória + 300 GB iSCSI tárhely / 270 GB szabad

A hosszabb távú szolgáltatási keretek között is meg kívánjuk tartani ezt a kettősséget, a nyilvános archívum, illetve a honlap és tartalomszolgáltatások szerverének háttere terén. A hosszabb távú működést biztosító szerverek a Kormányzati Informatikai Ügynökség (KIFÜ) bázisán kerültek beüzemelésre 2020 végén.

A kísérleti projekt idejére a KIFÜ biztosított egy Linux-alapú szerveret (IP cím: 193.224.22.162) a nagy méretű aratások memória- és tárhelyigényének tesztelésére. Ennek a gépnek az eredeti paraméterei: 16 magos CPU, 64 GB RAM, 100 GB rendszerlemez, 20 TB tárhely (RAID6, iSCSI), 10 Gbps sebességű hálózati csatlakozás. A szerveren terhelési tesztek futottak több tíz- és több százezer tételes URL listákkal 2017 májusában, augusztus végén, szeptemberben és októberben. Ezeknek az aratási próbáknak az anyaga természetesen nem lett megőrizve. A már megőrzendőnek szánt mentések során keletkezett tömörített WARC fájlok

és a hozzájuk tartozó indexállományok összesített mérete 2018 végére meghaladta a 10 TB-ot. Erről az anyagról jelenleg még nincsen biztonsági másolat az OSZK-ban, részben azért, mert a nagy tárolókapacitású digitális raktári rendszer kialakítása még nem történt meg, részben pedig azért, mert a 2018 októberében tett 5 TB-os másolási kísérlet annyira leterhelte a FreeNAS storage-ot, hogy az már az ott futó egyéb szolgáltatásokat is túlságosan lelassította. A KIFÜ-től ígéretet kaptunk arra, hogy ezt a tesztszervert tovább használhatjuk addig, amíg a nagy kapacitású infrastruktúra beszerzése és beüzemelése megtörténik, és 2018 novemberében további 20 TB tárhelyet igényeltünk ehhez a szerverhez, ami elegendő volt kb. 1 évig (OSZK Webarchiválási munkacsoport, 2018). Az OSZK hosszú távú szolgáltatási keretekre szolgáló szerverparkjának beüzemelésével újra napirendre került a KIFÜ szerver anyagának migrálása, megfelelő nagykapacitású OSZK szerverkörnyezetbe, hogy így egy biztonságosabb megduplázott szolgáltatási környezet valósuljon meg. Végül 2020 második felében mégis azt a döntést hozták vezetői szinten, hogy a webarchívum technikai infrastruktúrájának alapjait a KIFÜ biztosítja. A migráció lezajlott az OSZK környezetébe, majd vissza a KIFÜ platformjára, ahol a hosszútávú üzemelést biztosító hardveres keretek kialakítására 2020 végétől került sor.

A 2017-es terhelési tesztek alapján a KIFÜ informatikai szakembere 2017. október 29-i levelében ezt az infrastruktúrát javasolta az üzemszerű működéshez a következő 3-4 évre: 32-44 processzormag (2-3 GHz közötti teljesítményű magok), 256 GB memória, 500 TB adat-optimalizált lemez, 2 PB szalag. Egy ilyen vagy ehhez hasonló hardver beszerzése folyamatban

van, üzembe állítása 2020 végén történt meg (OSZK Webarchiválási munkacsoport, 2018). A legnagyobb tárhelye a tömeges aratásokra használt és a zárt archívumot tartalmazó webharvest szervernek van. Az ehhez hozzárendelt 150 terabájtól eddig kb. 60 foglalt, az éves növekedés 20 terabájt körüli. Az őszi költözködés jó alkalom volt arra, hogy kialakítsunk ezen a tárhelyen egy raktározási rendszert, és az év végére sikerült megoldani, hogy az archív anyagot tartalmazó WARC fájlok, az oldalképek és az egyes aratások konfigurációs-, napló- és report-állományai részgyűjtemények és időrend szerint külön almappákban jelennek meg.

A webarchiválással kapcsolatos könyvtárosi feladatokra felvett két munkatárs 2018 közepén új asztali gépeket kapott, ezek paraméterei: HP Prodesk PC, 3,4 GHz-es Intel i5 processzor, 8 GB RAM, 250 GB-os Western Digital merevlemez, 24"-os Benq LCD monitor. A PC-ken 64 bites Windows 10 operációs rendszer fut és ezeken zajlik a webarchiválással kapcsolatos levelezés és adminisztráció mellett a megőrzésre javasolt honlapok archiválhatóságának tesztelése, illetve a speciális beállítást vagy letöltő programot igénylő webhelyek mentése, valamint a Linux-os szervereken automatikus módszerrel készült, de hibás oldalképek újra generálása windowsos szoftverek segítségével (OSZK Webarchiválási munkacsoport, 2018). 2020 márciusától a munkatársak a COVID járvány miatt feladataikat távmunkában látják el. A könyvtári munkaállomásokhoz, illetve szerverekhez biztosított a biztonságos távoli elérés, illetve minden egyes munkatársunknak a megfelelő otthoni szélessávú internet és számítógépes infrastruktúra is rendelkezésre áll a munkájuk elvégzéséhez.

3.5.2 A webarchiválási kísérleti projekt szoftveres háttere

A nemzeti könyvtárakban működő webarchívumok esetében általános módon az OSZK részéről is alapvetően olyan nyílt forráskódú szoftverekből tervezzük felépíteni a rendszert, amiket az Internet Archive és az IIPC nevű nemzetközi konzorcium is támogat. Ezeket fogjuk kiegészíteni saját fejlesztésű modulokkal vagy önálló alkalmazásokkal, melyek speciális igényeket elégítenek ki, illetve integrálják a webarchívumot az Országos Könyvtári Rendszer keretében bevezetésre kerülő új könyvtári rendszerrel, a metaadat-nyilvántartással, a digitális raktárral, a névterekkel, a statisztika-készítő, a felhasználói azonosítást végző, a jogosultságokat kezelő, a dokumentum-hitelesítő, a partneryilvántartó és még további szóba kerülhető rendszerekkel.

3.5.2.1 Arató szoftverek

Az aratószoftverek alapvetően két csoportba sorolhatók be. Az első csoportba tartozó szoftverek egyszerűen letöltik az oldalakat. Ezek régebbi szoftverek, jobban teszteltek, kisebb erőforrás igényük van, gyorsabban töltik le a tartalmat. Hátrányuk, hogy a sok dinamikus programozási elemet (pl. JavaScript) csak korlátozott eredményességgel képesek menteni, mint ahogyan Brunelle és munkatársai számot adnak erről (Brunelle, Weigle, & Nelson, 2017).

A második csoportba tartozó programok böngésző és a hozzá tartozó proxy szerver kombinált megoldáson alapulnak. Grafikus felület nélküli Chrome böngésző vezérlése történik a hibakereső (debug) interfészen

keresztül, a forgalom szűrésére pedig beépül egy proxy szerver, az összes azon áthaladó forgalom mentésre kerül.

3.5.2.1.1 Heritrix

Az első csoportba tartozik. Webhelyek tömeges aratásához a legfontosabb eszköz az Internet Archive által 2003 óta fejlesztett és használt crawler vagy robot („Heritrix—IA Webteam Confluence”, 2020; Osborne, 2020). Moduláris felépítéssel bír. Egy adott funkciót ellátó komponensnek több fajta megvalósítása létezhet. A különböző komponenseket az XML job konfigurációs fájl segítségével lehet összedrótozni (Jurányi, 2013). Ez szinte magától működik, ám szerencsére nem szükséges nagyon a mélyére látni. Java alapon fejlesztett program, forráskódja egy nagy monstrum. Az XML alapú kézzel szerkeszthető konfigurációs fájljának megismeréséhez és beállításához intenzív, időigényes tanulási folyamat szükséges. Az erőbefektetés viszont megtérül, megbízhatóan, konzisztensen működő szoftvert kapunk végeredményként. A jól beállított Heritrix képességeit összegzi (Jackson, 2018). A job konfigurációs fájl XML szöveges fájl, kézzel szerkesztendő. Ez tartalmazza a seed listát (hivatkozást rá külön fájlban) és egyéb opciókat. (Vitéz, 2020)

Az alapfunkciók grafikus interface segítségével is vezérelhetők, valamint nyomon követhetők az aktuális jobok is, a pontos leírás különféle gyakorlati példákkal együtt több forrásban is rendelkezésre áll (Maeda & Oyama, 2017; Osborne, 2020).

Fontosabb opciói a maxHop (hány link távolságra kalandozzon el a robot a kezdőponttól), különféle szempontok szerinti kizárások beállítása az aratásból, illetve a Robots.txt szabályainak figyelembevétele. Beállítható

a párhuzamos letöltések száma, újrapróbálkozások közti idő és egyéb opciók is. A Heritrix nagy előnye, hogy egy külön telepíthető modullal deduplikációra is képes, vagyis nem tárolja el újra azokat a fájlokat, amelyek egy korábbi mentés óta nem változtak. Tud URL- és tartalom-alapú deduplikálást kezelni. Az első esetben az adott URL-en elérhető tartalmat eleve csak akkor kéri le, ha az megváltozott. A második esetben letölti a teljes tartalmat, de csak azt menti warc fájlba, ami a tartalomra számolt hash alapján még nem volt leartatva. Mindkettő csökkenti az ismételt archiválások helyigényét (OSZK Webarchiválási munkacsoport, 2018).

HERITRIX
EngineJob DirConfigurationCopy JobScripting ConsoleBrowse Beans

Job 1538135550559

(1 launches, last 52s165ms ago)

buildlaunchpauseunpausecheckpointterminate teardown

Job Log [more](#)

```

2018-09-28T11:52:55.515Z INFO RUNNING 20180928115252
2018-09-28T11:52:54.844Z INFO PREPARING 20180928115252
2018-09-28T11:52:50.658Z INFO Job launched
2018-09-28T11:52:50.337Z INFO Job instantiated

```

Job is Active: RUNNING

Totals

17 downloaded + 35 queued = 53 total
150 KiB crawled (150 KiB novel, 0 B dupByHash, 0 B notModified)

Alerts

none

Rates

0.12 URIs/sec (0.35 avg); 0 KB/sec (3 avg)

Load

1 active of 25 threads; 1 congestion ratio; 35 deepest queue; 35 average depth

Elapsed

47s932ms

Threads

25 threads: 24 ABOUT_TO_GET_URI, 1 ABOUT_TO_BEGIN_PROCESSOR; 24 noActiveProcessor, 1 fetchHttp

Frontier

RUN - 2 URI queues: 1 active (1 in-process; 0 ready; 0 snoozed); 0 inactive; 0 ineligible; 0 retired; 1 exhausted

Memory

57038 KiB used; 88576 KiB current heap; 232960 KiB max heap

10. ábra Heritrix job

A Heritrix szabályrendszere meghatározza, hogy egy adott URL letöltendő-e. Egy szabálynak két állapota lehet, három lehetséges állapot közül REJECT/PASS illetve ACCEPT/PASS. Az utolsó opció, ami nem PASS, kerül végrehajtásra. Például: REJECT mindenképp (első, feltétel nélküli szabály); ACCEPT, mert aratandó doménre mutat a link (acceptSurts); PASS, mert a tartalom pl. nem kizárandó videó tartalom (nem REJECT/PASS), végül az eredmény az ACCEPT. A tipikusan alkalmazható szabályok az alábbiak: Szűrés domainre (SURT); Regexp alapú URL szűrés; MIME típusra szűrés; maxHop – seed listától hány linket követhet maximum; Már letöltött HTML-hez tartozó CSS, képek stb. (Osborne, 2020) Első ránézésre bonyolult, de végül átlátható, nagyon finom beállítást tesz lehetővé. Például kínálkozik az oszk.hu aratása teljes mélységben, az oldalról való „elkalandozás” nélkül. A használatába érdemes beletanulni, ha nem csak egyszer-egyszer kell pár oldalt lementeni. Megbízhatósága és jó kompatibilitása miatt ritkák a meglepetések. Tömeges aratások esetén viszont memória-igényes lehet (Vitéz, 2020). Legújabb, 3.2-es verzióját telepítettük a KIFÜ-s szerverre, viszont mivel a később ismertetendő WCT keretrendszerhez eredetileg az 1.14-es verzió volt integrálva, ezért az OSZK-s szerveren kezdetben ezt használtuk, majd 2017 őszén ott is áttértünk az újra. Bár a Heritrix eredetileg alapvetően még a “hagyományos”, 1.0-ás webre lett kitalálva, de később sokat fejlesztettek rajta olyan irányban, hogy a 2.0-ás, JavaScriptekben gazdag, dinamikus megoldásokat használó weboldalakat is ki tudja elemezni és le tudja menteni (Brunelle és mtsai., 2016, 2017). Ennek ellenére a tapasztalataink szerint a 3.2-es Heritrix is sokat hibázik, elég gyakoriak a hiányosan vagy hibásan lementett oldalak. A problémák egy része speciális konfigurációs

beállításokkal orvosolható (persze tömeges aratásoknál erre nincs emberi kapacitás), más része viszont az eredeti webhelynél használt technológia módosítását igényelné, vagy egy más típusú eszközt (pl. Brozzler vagy Webrecorder) kell használni ilyen esetekben (OSZK Webarchiválási munkacsoport, 2018).

Hozzáadott funkcióként építhető be a DecideRules nevű szabályok használata, amivel sokkal finomabban hangolható a robot működése. (Utóbbiakat még nem próbáltuk ki, de tervbe vettük már a tesztelését).

3.5.2.1.2 Brozzler

A Heritrix korlátai miatt az utóbbi években több intézményben elkezdődtek a böngészőn keresztül való archiválással való kísérletezések. A béta fázison már túljutott szoftverek közül mi a Brozzler (browser + crawler) nevű eszközzel végeztünk néhány próbát, („Brozzler – MIA WIKI”, 2020), melyet 2016-ban mutatott be az Internet Archive, és azóta már be is építették az előfizetéses Archive-IT szolgáltatásukba („How and when to use Brozzler”, 2020; OSZK Webarchiválási munkacsoport, 2018; Sylvie-Rollason-Cass, 2020). Előnye, hogy Javascriptet intenzíven használó oldalakat is le tud menteni. Ehhez egy Chrome böngésző példányt indít, amit a Chrome debug interfészén keresztül irányít. Működésmódja a következőképpen foglalható össze: A Chrome megnyitja az archiválandó oldalakat. A Brozzler része egy, a Chrome elé beépülő proxy, amit warcprox-nak hívnak. Ez WARC fájllokba menti a http forgalmat, ami áthalad rajta. Így végül az oldalhoz szükséges minden css/js/kép/videó/stb. fájl letöltésre kerül. Sajnos ez a szoftvertermék még kevésbé kiforrott, a főbb hátrányai az alábbiakban foglalhatók össze: Job

indítás csupán parancssoros felületen érhető el. Grafikus felhasználói felület csupán a futó jobok megnézésre használható, a szoftver működésébe így beavatkozni így nem lehet. Rethinkdb adatbázist használ, ami azért jelent gondot, mert a Rethinkdb fejlesztése mögött álló cég 2016-2017 körül csődbe ment, így nehéz a szoftvert nehézkes működésre bírni. A Warcpox pedig alapértelmezésben a párhuzamos aratások anyagát egyetlen WARC állományba menti, holott nem szeretnénk, ha minden művelet eredménye ebben az egyetlen WARC állományban jelenne meg. Megfelelő konfigurálással azonban a műveletenként külön WARC fájlba mentés funkció viszont aktiválható. (Vitéz, 2020)

A MANDA blogján teszteltük először a Brozzlert, és bár nem kaptunk jobb eredményt, mint a Heritix esetében, amely csak a kezdőoldalt tudta lementeni, de miután az egyes éves blogarchívumok URL-jeit is megadtuk kiinduló (seed) címnek, a 434 blogposztból 296-ot sikerült lementeni, viszont a Javascriptes navigáció továbbra sem működött mindenhol. Bár a szoftver parancssoros módban indítható, de az adminisztrátori felületen, illetve az általa indított Chrome böngészőben követhető a robot tevékenysége (mely oldalakat jár be). Kisebb szépséghibája, hogy egy weboldaltól több bejegyzés is megjelenik az OWB-ben, amit a program sajátos működésére lehet visszavezetni (az oldal aratása után mindig visszaugrik az elejére, s ennek biztosítására kiegészíti saját bejegyzéssel a forráskódot is). Létrehoztam egy Linux alapú virtuális gépet 2020 nyarán, mely windowsos Virtualbox környezetben fut 40 GB tárterülettel és 4 GB memóriaigénnyel. Ennek segítségével számomra, Drótos László témafelelőssel együtt a távmunkában használt saját otthoni gépeinken

könnyen tesztelhetők a különféle konfigurációs beállítások, jobban értelmezhetővé vált a program működése. Az ily módon kikísérletezett konfigurációs beállításokat alkalmazzuk a hírportálok címdoldalainak aratásánál. A 2020. évi 404 workshopon Drótos László témafelelős kollégámmal együtt bemutatót tartottunk az egyéni webarchiváláshoz használható és a virtuális gépre telepített programok működéséről. Ennek eredményeképpen több partner könyvtár számára is tesztelési célból használatra bocsátottuk ezt az eszközt. A bemutató anyagából publikáció is készült, amely a Könyv Könyvtár Könyvtáros folyóirat hasábjain jelent meg (Németh & Drótos, 2020). Ennek révén is megfelelő áttekintést lehet szerezni a mindennapokban bárki által használható archiváló szoftverek használatának alapjairól.

3.5.2.1.3 WAIL

Eredetileg alapvetően a Heritrixre épült a Web Archiving Integration Layer (WAIL) (Kelly, 2013/2020; „WAIL - MIA WIKI”, 2017), aminek mi a windowsos változatát próbáltuk ki, mert bár Linux alá is telepíthető, elsősorban kisebb méretű, személyes webarchiválásra szolgáló szoftver. A két említett platform mellett Macintosh operációs rendszer alá is telepíthető. Ez voltaképpen egy olyan szoftvercsomag, mely telepít egy saját szervert, Heritrix-szel, pywb archivált dokumentumokat visszanező szoftverrel, valamint a grafikus felület nélküli Chrome böngészőre épülő Electron (Berlin, 2017) nevű proxy arató szoftver megoldással. A grafikus felületen választhatjuk ki, hogy az Electront vagy a Heritrixet próbáljuk futtatni az egyes feladatokhoz. A beépített Heritrix sajnos nem szabható olyan mértékig testre, mintha önállóan telepítenénk. Ez a szoftvercsomag

azoknak hasznos, akik egy komplett, könnyen beállítható szoftverkörnyezetet szeretnének a saját gépükön látni, ahol a saját gépen induló webszerver mellé mind a két fő aratószoftver típus be van integrálva, s vissza is nézhetők a lementett anyagok. Az egyes webhelyek azonban csupán 3 szint mélységig menthetők vele és a robot viselkedése sem paraméterezhető, a címen és egy rövid ismertetőn kívül egyéb metaadatok nem adhatók meg, viszont a mentések részgyűjteményekbe szervezhetők. A WARC fájlok megjelenítéséhez egy Pythonban írt Wayback verziót (pywb) használ úgy, hogy előbb egy lokális webszervert indít a gépen. A WAIL érdekessége, hogy Twitter csatornák mentésére is fel van készítve, továbbá 2017-ben a Chrome böngészőmotorját használó WARCcreate egy módosított változatát (Electron néven) is beleépítette a fejlesztője, így a dinamikusan generált, bonyolult felépítésű weboldalak jobb minőségben menthetők vele, mint a Heritrix-szel (Berlin, 2017). Ezt a mi tesztjeink is megerősítették például a hírportálok és az OSZK virtuális kiállításai esetében, viszont mivel a linkeket ilyenkor nem követi a WAIL, ezért csak egyes oldalak mentésére alkalmas ez az üzemmód. További tapasztalat volt, hogy a WAIL-be integrált Heritrix-szel végzett aratások némileg más eredményt adtak ahhoz képest, mint amikor a KIFÜ-s vagy az OSZK-s szerveren mentettük le ugyanazt a webhelyet. Ezt a konfigurációs fájlokban levő eltérések okozhatják (OSZK Webarchiválási munkacsoport, 2018).

3.2.1.4 Webrecorder

A Webrecorder („Webrecorder—MIA WIKI”, 2018) eredetileg csupán egy online böngészőn keresztül elérhető alkalmazásként volt elérhető

(„Conifer”, 2020). Mostanra már önállóan is telepíthető, Linux, Macintosh, illetve Windows operációs rendszerekre érhető el. A kétfajta termék fejlesztése szét is vált, különféle fejlesztői csapatokhoz került. Az online megjelenítő programot átnevezték Coniferre. A tesztelés alapján a telepíthető programnak a linuxos verziója fut a leggyorsabban, így a közösségi média mentési kísérleteket a már korábban említett Linux alapú, ám Windows alatt futó virtuális gépünkön célszerű elvégezni (Németh & Drótos, 2020). A Webrecorder Desktop program működésének kihívása, hogy JavaScript alapú parancsfájlokkal vezérli az egyes közösségi média szolgáltatásokról automatizált algoritmusok segítségével (autopilot funkció) a tartalmak begyűjtését, a felület folyamatos fejlesztése miatt azonban ezek a fájlok gyorsan elavulnak. Valamiféle közösségi alapú kezdeményezésre lesz szükség az IIPC keretei között ezen parancsfájloknak a folyamatos gondozására, mert csak így lehet garantálni a közösségi médiatartalmak archiválását. A program működését leginkább az egykor létezett videomagnók számítógépes világhálóra alkalmazott változataként lehet megragadni. Elindít egy egyszerű kezelőfelülettel rendelkező böngészőt (Németh & Drótos, 2020). Az online felületen többféle emulált konfiguráció közül választhatunk, a telepített verzióban egyfajta alapértelmezett változat fut. Megadjuk a lementeni kívánt weboldal címét, s minden anyagot, ami a böngészőfelületen megjelenik, a háttérben futó program WARC fájlba ment el. Hátránya, hogy nem automatizálhatók a mentések, csupán a böngészőablakban lefelé történő automatikus görgetést lehet beállítani. Nagy előnye viszont, hogy más módszerekkel nem aratható, bonyolult szerkezetű weboldalak is menthetők a segítségével. Főleg közösségi

média felületek, illetve bonyolult programozási elemekkel teli weboldalak mentésére hasznos a számunkra. Olyankor is jól használható, amikor eseményalapú mentésnél egyetlen weboldalt, kívánunk menteni (vagy annak egy részletét), s nem akarjuk, hogy a robot bármilyen, az oldalról linkelt egyéb tartalom felé tovább haladjon (Németh & Drótos, 2020). A lementett WARC fájlokban tárolt webhelyek visszanézésére önálló program áll rendelkezésünkre Webrecorder Player néven, mely Windows, Linux, illetve Macintosh környezet alá telepítve is elérhető. 2020-ban az online és telepítve használható programok szétválását követően, megjelent egy gyorsabb visszakereséseket biztosító Windows-alapú és online is használható WARC nézegető szoftver is a Webrecorder fejlesztőinek kiadásában ReplayWeb.page néven (Németh & Drótos, 2020). A WARCIT (*Webrecorder/warcit*, 2020) nevű program segítségével pedig számos fájltypusból készíthetünk WARC fájlcsomagokat, melyeket aztán feltölthetünk az online futó Conifer szolgáltatásba, vagy visszanézhetjük az említett WARC fájlnevező programokkal.

3.2.1.5 HTTrack

Windows, Linux és Macintosh környezetekben futtatható, magyar nyelvű kezelőfelülettel is az aratási paraméterek részletes beállítását lehetővé tévő program („HTTrack Website Copier—Free Software Offline Browser (GNU GPL)”, 2020; „HTTrack—MIA WIKI”, 2017). Be lehet állítani rajta akár egy adott webhely hozzáféréshez kötődő felhasználónevet és jelszót is, ezáltal nem nyilvános oldalak is menthetők segítségével, például személyes webarchiválás céljára. Nincs saját böngészőfelülete, a fájlrendszerbe mentett anyag bármely böngészővel visszanézhető. A

kísérleti projekt keretében a windowsos felületét teszteltük. Széleskörű használata sajnos nem kerülhet szóba. Legnagyobb hátránya ugyanis, hogy WARC konténer fájlformátum helyett fájlrendszerbe ment, és így sem az OWB vagy a pywb visszanező programok, sem a SolrWayback teljesszövegű indexelő és böngésző alkalmazás alá nem tehető be a lementett tartalom. A Warcit nevű segédprogrammal (*Webrecorder/warcit*, 2020) sikerült néhány esetben használható WARC fájlokat konvertálni a HTTrack mentésekből. Egyéni webarchiválásra viszont kisebb webhelyek esetében, ahol nem akadály a fájlrendszerbe történő mentés, jól használható. Ezért oktatjuk használatát a könyvtárosok továbbképzésében, illetve a tanfolyamot záró gyakorlati vizsgafeladatok terén is szerepet szánunk neki.

3.2.2 Az archivált tartalom megjelenítését nyújtó szoftverek

A következőkben olyan szoftvereket mutatunk be, melyek segítségével az archivált anyag tárolási egységeként szolgáló WARC konténerformátumban tárolt fájlokból megjeleníthetők az archivált weboldalak. Ezek tartalmazzák a letöltött anyagot lementett HTTP válaszok formájában (HTTP fejléc és adat). A warc fájlokhoz egy segédprogrammal egy index fájl készül, különben a .warc fájlokat mind végig kellene keresni az éppen lekért URL feltárására. Minden visszanező szoftvernek megvan a saját index-típusa és saját segédprogramjai. A WARC fájl egymás után tartalmazza a HTTP válaszokat. A WARC fájl tömörített állapotban warc.gz formátumban is tárolható. Ez nem a hétköznapi gzip tömörítés, hanem warc rekordonként készül, amiket aztán összefűz az archiváló szoftver. Másképp nem lehetne gyorsan egy

adott pozícióra ugrani a fájlokban, s biztosítani a hatékony visszakeresést (Vitéz, 2020).

Minden egyes szoftver saját index típussal bír:

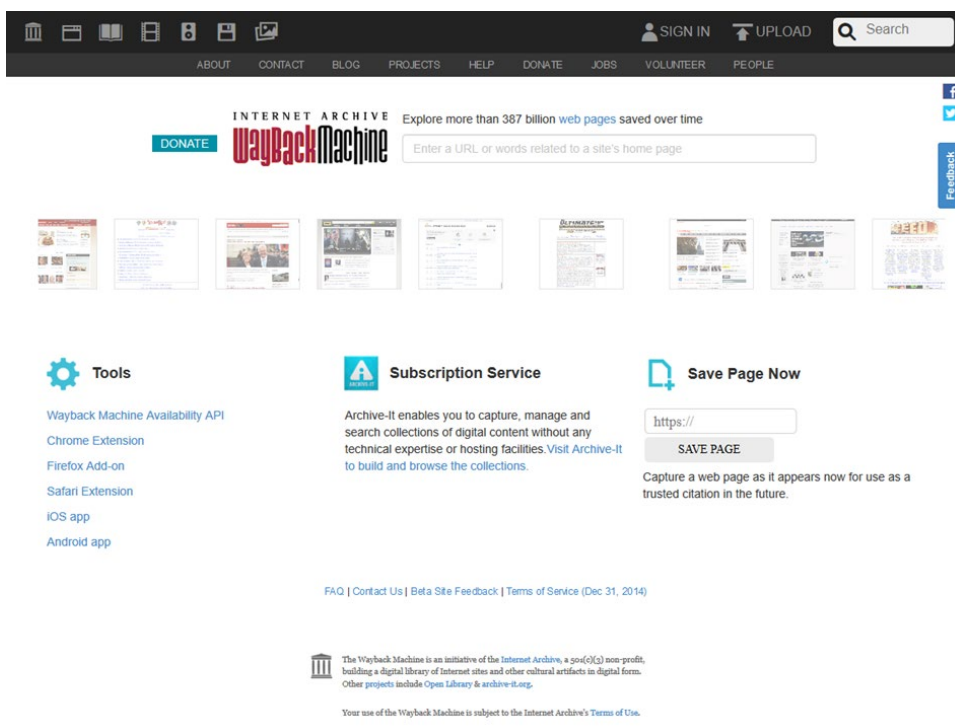
PyWb: cdxj, szöveges

OpenWayback: CDX (szöveges) / BDB (bináris)

SolrWayback: a Solr index tartalmazza az URL-hez tartozó warc fájlneveket és pozíciókat

[3.2.2.1 OpenWayback](#)

Az imént bemutatott aratóprogramok (a HTTrack kivételével) ún. WARC formátumú konténerekben tárolja a robot által talált linkekről letöltött fájlokat. (Drótos, 2017b; IIPC, 2017) Ezek megjelenítéséhez a CDX indexekre épülő, az Internet Archive által fejlesztett OpenWayback (OWB) szoftvert (IIPC, 2017) használjuk mindkét tároló szerveren. Ez az Internet Archive Wayback Machine megjelenítőjének nyílt forráskódú változata.



11. ábra A Wayback Machine kezdőlapja

Mivel a KIFÜ-s szerveren található mentések többségének nyilvános szolgáltatásához nincs engedélyünk, ezért ott a hozzáférést tűzfal korlátozza az OSZK IP címtartományára. Az OWB az egyik legrégebben használt Java servlet alapú, könnyen konfigurálható visszanezítő program (Drótos, 2017b). Szöveges (CDX), vagy bináris (BDB) indexet tud használni, az előbbi bír széleskörűbb támogatással. A bináris indexelés esetében maga a Java servlet végzi el a feladatot, figyelve azt, hogy mikor jelenik meg új WARC fájl az adott mappában. A gyakorlatban azonban ez az észlelés általában csak az OWB újraindítása után aktiválódik. A CDX indexelést külön szoftver nyújtja az OWB számára. A WARC fájlokból kézzel vagy megfelelő script segítségével kell legyártani a CDX index fájlt. Az OWB képes több CDX indexből is dolgozni. Ez így összességében igen

rugalmas megoldás, több ugyan vele a munka, mint a BDB esetében, de bizonyos idő elteltével ez kiegyenlítődik, a jól automatizálható munkafolyamatok révén. Fontos megjegyezni, hogy mindkét esetben jól működik az inkrementális indexelés. Ez azt jelenti, hogy csak az új warc fájlokat kell feldolgozni. Nagy archívumok esetén amúgy is csak ez a megoldás a járható út (Vitéz, 2020).

Az OWB időgépszerűen teszi lehetővé a lementett webtartalom böngészését egy URL cím megadása után. Először naptári nézetben mutatja meg az adott URL-ről készült mentések dátumait, majd egy időpontot kiválasztva megjeleníti a fájlt, és ha vannak benne linkek, akkor azokra kattintva ugyanúgy böngészhető az archívum, mint az élő web – mindaddig, amíg egy olyan linkre nem kattint a felhasználó, ami már nincs benne az archívumban. A linkeket követve a szoftver mindig a kiinduló időponthoz legközelebbi mentést jeleníti meg, de így is előfordulhat néhány kattintás után, hogy akár több hónappal korábbi vagy későbbi fájlok jelennek meg, ha nincsen közelebbi dátumú példány. További probléma lehet, hogy a linkek lokalizálásával a felhasználót megpróbálja az archívumon belül tartani, de ez nem mindig sikerül és egyes linkek kivezethetnek az élő webre. Az OpenWayback egy saját fejléctet tesz a mentett weboldalak fölé, ami egyrészt jelzi, hogy még az archívumban vagyunk, másrészt mutatja azt is, hogy mikor lett az éppen nézett weboldal lementve és milyen más időpontbeli mentések vannak még róla. Ez a fejléc a felhasználó által kikapcsolható, és erre néha szükség is van, ha például eltakarja az oldal főmenüjét. 2018 májusában belefutottunk egy olyan problémába, hogy az ékezetes betűket tartalmazó fájlnevek

esetében a dupla átkódolás miatt az OWB nem találta meg a fájlt és hibaüzenetet küldött, de ezt az Nginx webszerver konfigurációjának módosításával sikerült megoldani (OSZK Webarchiválási munkacsoport, 2018).

3.2.2.2 Pywb

A neve is elárulja, hogy Pythonban íródott programról van szó. („Pywb”, 2019) Kevésbé erőforrásigényes a használata, mint az OpenWaybacké. Az IIPC konzorcium egyre inkább ennek használatát ajánlja, miután az OWB fejlesztése lényegében leállt. Másrészt viszont kevésbé tesztelt, kiforrott termékről van szó. Erre utal az is, hogy csupán fejlesztői módban sikerült az OSZK-s szerveren elindítani (Országos Széchényi Könyvtár, 2018). Az inkrementális indexelés a gyűjteménybe került új warc fájlokat csak akkor veszi észre, ha a Pywb közben fut. A teljes indexelés megbízhatóan működik (Vitéz, 2020). A megjelenítéskor a grafikus felülete egyszerűbb, mint az OWB-é. Nincs időgép és fejléc sem. Lehet viszont keresni egy URL cím részleteire is. A visszakeresési lehetőségek rendelkezésre állnak MIME típus, STATUS és URL-cím szerint, s vannak reguláris kifejezés alapú szűrési lehetőségek is. Arról nem tudunk tájékozódni, hogy mikor készült a mentés, de rákereshetünk egy adott URL alapján, hogy egy adott időszakban készültek-e mentések. Szűkíthetünk prefix, host és domén szerinti keresésre. Még általunk nem tesztelt funkcionálitása a programnak, hogy az indexből kinyert metaadat elemek is visszakereshetők lehetnek. Ez a része azonban még nagyon kiforratlan a programnak, a későbbiekben e keresőfunkciók fejlesztésében jelentős potenciál húzódhat meg.

3.2.2.3 SolrWayback, SolrMIA

Az OpenWaybacknek és a Pywb-nek is megvannak a maga korlátai (például csupán URL cím szerint lehet keresni vele, az egyes mentett weboldalakról pedig nagyon kevés információt ad), és ezért alternatív megoldásokat kezdtek el fejleszteni. A legígéretesebb ezek közül a dán SolrWayback (Egense & Klindt Myrvoll, 2018; Netarkivet, 2018), ami a British Library webarchívumához készült WARC indexelőre és az általános célú Solr keresőplatformra épül, vagyis ez nem CDX indexfájlokat használ. Egy java servletről van szó, így a működéséhez kell egy Servlet Container is (tomcat/jetty/stb.) A működéséhez szintén szükség van egy SOLR példányra is (Vitéz, 2020).

Jó a kompatibilitása a különböző archivált anyagokkal, a megjelenítés szinte tökéletesnek mondható. A kereső felületet persze folyamatosan fejleszteni kell a változó webes környezettel összhangban. Az indexelő szoftver is Java-ban íródott, a hagyományos indexelés mellett a teljes szöveget is indexelteti a Solr-rel. Apache Tika segítségével elemzi a warc fájlokban található tartalmat. Így például .pdf, .doc és egyéb fájlformátumokban is lehet az indexelés végeztével keresni. Teljes szövegű- és képkereső funkció is van benne, utóbbinál a JPEG képfájlokban levő GPS koordináták alapján térképre vetíthetők a találatok, sőt egy, a felhasználó által feltöltött fájl előfordulását is meg tudja keresni az archívumban.

Móra múzeum

SEARCH

☐ Grouping (slower response)

Clear search

☒ Image search
 ☒ Geo search
 ☐ URL search
 Search with uploaded file
 Search for HTML-tags
 Domain stats

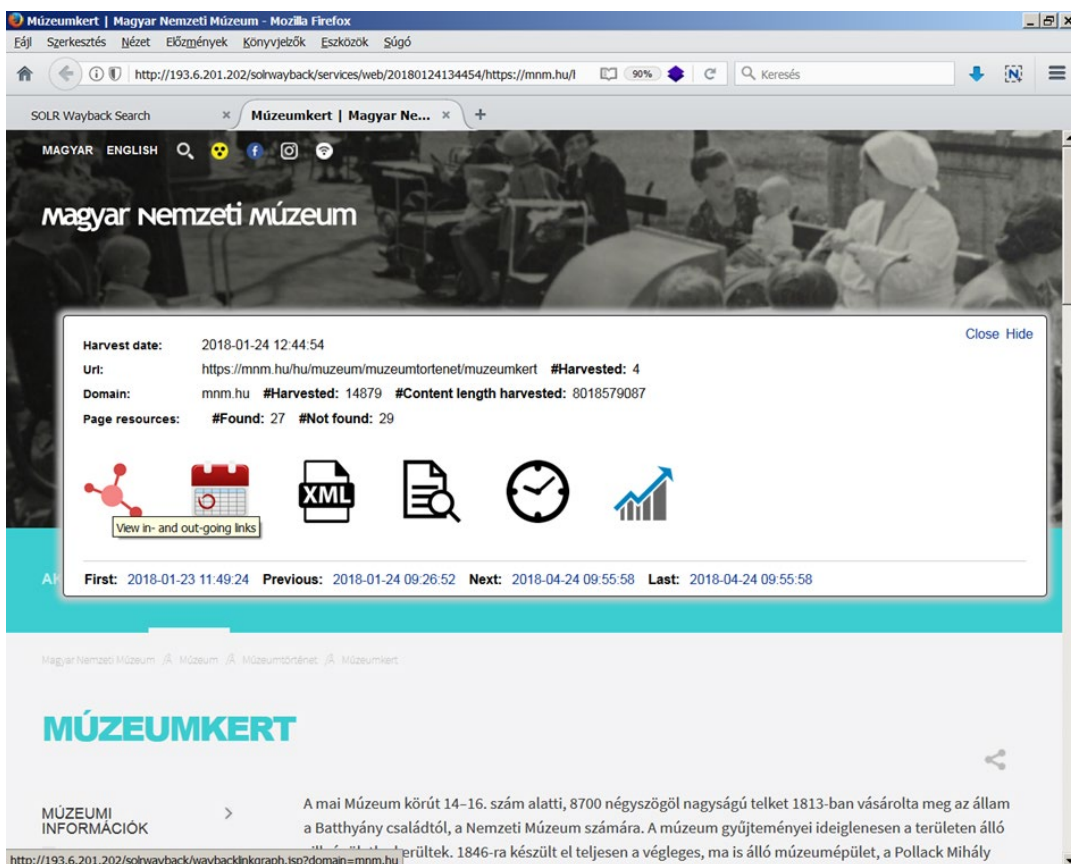
Térkép

Műhold

Google

Adjust radius

Radius in km: 10



14. ábra A SOLR Wayback eszköztára

A találati listában a „Show full post” feliratra kattintva pedig megnézhető az adott dokumentum összes technikai metaadata.

SOLR Wayback Search

The screenshot displays the SOLR Wayback Search interface. At the top, there is a search bar containing the text 'Múzeumkert' and a 'SEARCH' button. Below the search bar, there are links for 'Clear search', 'Image search', 'URL search', 'Search with uploaded file', 'Search for HTML-tags', and 'Domain stats'. The main content area is titled 'Limit results' and shows a list of domains and their respective hit counts. To the right, there is a section titled 'Showing 1-20 of 1 260 hits' with 'Previous' and 'Next' buttons. Below this, the search results for 'Múzeumkert | Magyar Nemzeti Múzeum' are displayed, including content type, domain, URL, score, and highlighted content. The highlighted content includes 'Keresés', 'Akadálymentes', 'Facebook', 'Instagram', 'Free WiFi', 'Bejelentkezés', 'Magyar Nemzeti Múzeum', 'Múzeum Múzeumtörténet', 'Múzeumkert', 'Múzeumkert', and 'Facebook'. Below the search results, there is a grid of six images, each with a 'Search for image' and 'Pages linking to image' link.

Limit results

Domain

- mnm.hu : 1235
- csukalib.hu : 3
- forumhungaricum.hu : 3
- gyoriszaon.hu : 3
- mandadb.hu : 3
- munkacsy.hu : 3
- dornyaymuzeum.hu : 2
- ldm.hu : 2
- sztaki.hu : 2
- deske.hu : 1

Content Type Norm

- html : 1227
- pdf : 19
- other : 14

Type

- Web Page : 1227
- Document : 19
- Other : 14

Crawl Year

- 2018 : 1260

Showing 1-20 of 1 260 hits [Previous](#) [Next](#) [Download result as CSV](#)

Múzeumkert | Magyar Nemzeti Múzeum

Content type: text/html
Domain: mnm.hu
Url: https://mnm.hu/hu/muzeum/muzeumtortenet/muzeumkert
Score: 120.32439
Highlighted content: KeresésAkadálymentesFacebookInstagramFree WiFi Bejelentkezés Magyar Nemzeti Múzeum Múzeum Múzeumtörténet **Múzeumkert** **Múzeumkert** Facebook

[Show full post](#)


[Search for image](#)
[Pages linking to image](#)


[Search for image](#)
[Pages linking to image](#)


[Search for image](#)
[Pages linking to image](#)


[Search for image](#)
[Pages linking to image](#)


[Search for image](#)
[Pages linking to image](#)


[Search for image](#)
[Pages linking to image](#)

15. ábra Találatok megjelenítése a SOLR Waybackben

Hátrányaként említhető meg, hogy bonyolultsága miatt az indexelési folyamat lassú, erőforrásigényes. Az indexelő komponens törekvése miatt a WARC fájlokat egyesével érdemes indexeltetni. Az inkrementális indexelés scriptelve megoldható, a SOLR indexből le lehet kérni a már indexelt WARC fájlok listáját, így könnyen kideríthető, hogy mit kell még indexeltetni vele (OSZK Webarchiválási munkacsoport, 2018; Vitéz, 2020).

A SolrWayback keresőjét 2018 márciusában beépítettük a nyilvános gyűjteményünkbe („Solrwayback kereső”, 2018), és mivel voltak különböző problémáink vele, Vitéz Gábor, a projekt rendszergazdája

felvette a kapcsolatot a dán fejlesztővel, aki készségesen segített, és jó munkamegosztás alakult ki: mi tesztelhattuk elsőként az új funkciókat, Thomas Egense pedig az OSZK-s szerveren futó példánnyal mutatta be több nemzetközi rendezvényen is a SolrWayback képességeit (Egense & Klindt Myrvoll, 2018). 2018 őszén a KIFÜ-s szerveren levő néhány nagy aratási halmazt is leindexeltünk vele, hogy felmérjük a szoftver idő-, memória- és tárhelyigényét.

A SolrWayback által létrehozott Solr index közvetlenül is kereshető. Emellett azonban más keresési felület illeszthető hozzá, pl. a projekt keretében fejlesztett SolrMIA felület („SolrMIA”, 2018). 2018 tavaszán kezdtük el saját keresőfelületként fejleszteni. A Solr indexre épül, de a találatok az OpenWaybackben jelennek meg. További különbségek még, hogy ez használja a magyar nyelvhez készült szótövező algoritmust, szűrőfeltételként felajánlja a webhelyek metaadatai között rögzített főtémakört, témakört, altémakört, műfajt és típust is, az egyes találatok alá kiírja a webhely egységesített nevét (szintén a metaadatokból átvéve), valamint, ha egy weboldalnak vagy egyéb fájlnek nincs önálló címadata, akkor helyette a fájlnevet jeleníti meg, ami jó esetben hordoz valamiféle jelentést (OSZK Webarchiválási munkacsoport, 2018). A SolrMIA fejlesztése 2019-ben leállt, a benne megvalósított plusz funkciók a tervek szerint a SolrWayback jövőbeli verzióiba lesznek beépítve. 2020-ban egy jelentősen átalakított, több új funkciót is tartalmazó verzió készült a SolrWayback-ből, melyet a jó szakmai kapcsolatnak köszönhetően a világon mi tesztelhattunk először. Ennek bevezetése a szolgáltatási keretek közé jelenleg is folyamatban van.

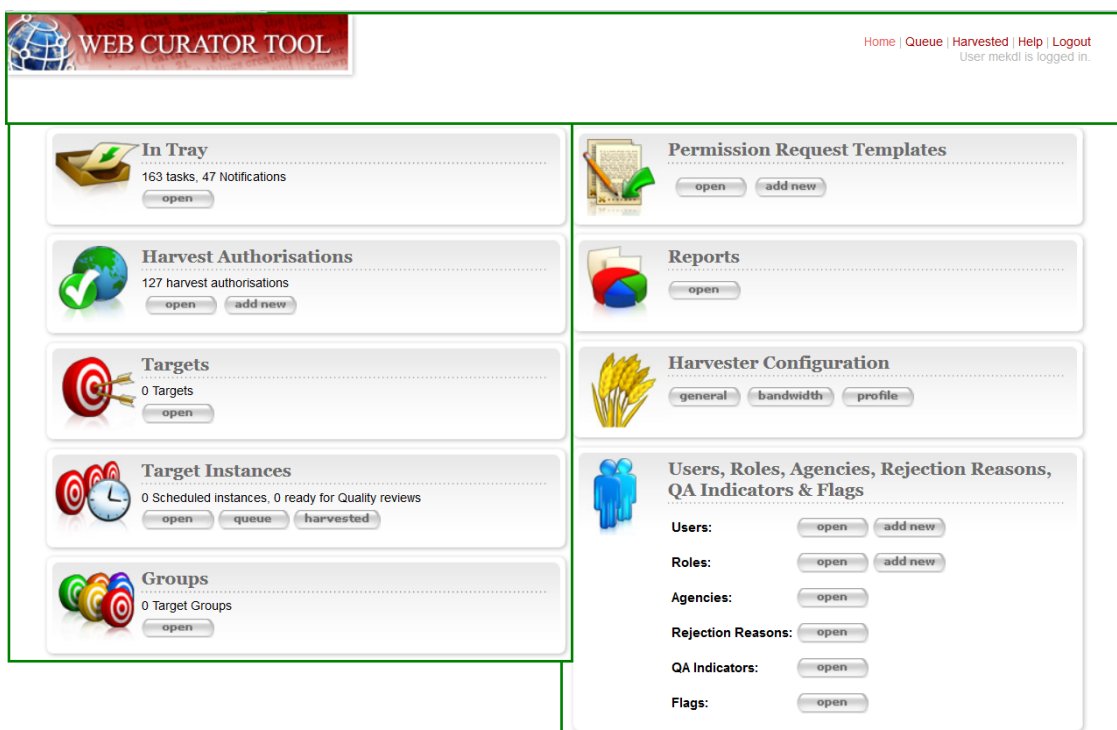
3.2.3 A webaratási munkafolyamatok menedzselését segítő keretrendszerek

Mivel a Heritix és az OWB vagy a Pywb vezérlése komolyabb informatikai ismereteket igényel és időnként a szerverhez való közvetlen hozzáférésre is szükség van, ezért az évek során több olyan keretrendszert is kifejlesztettek, amelyekkel felhasználóbarátabb felületen indíthatók és állíthatók le az aratások, valamint indexelhetők és nézhetők vissza a lementett fájlok. Ezek a szoftverek általában még további funkciókat is tartalmaznak, amelyek az archiválási folyamat néhány egyéb munkafázisát is támogatják. Az alábbiakban két Európában elterjedt, általunk is tesztelt és oktatott keretrendszert mutatunk be. Ezek közül a projektben a Web Curator Tool (WCT) eszközt vettük használatba, de a Netarchive Suite is elterjedtnek mondható az európai webarchívumok körében.

3.2.3.1 Web Curator Tool

A 2006-tól fejlesztett Web Curator Tool (Beresford, 2007; „WCT manuals”, 2019; „Web Curator Tool”, 2018) mely eredetileg az új-zélandi és a brit nemzeti könyvtár közös projektje volt, és több más nemzeti webarchívumnál is elkezdtek használni. A British Library azonban egy idő után kiszállt a fejlesztésből, az új-zélandiak pedig nem fejlesztették tovább. Emiatt az általunk 2017 júniusában a MEK tükörszerverén kipróbált WCT még nem támogatta a Heritrix 3-as verzióit. Az egy hónappal később megjelent és novemberben a webadmin.oszk.hu gépre is feltelepített WCT v1.6.3 ugyan már képes volt az új Heritrixet is elindítani, de az aratási paramétereket nem adta át, így csak kézzel

lehetett konfigurálni a robotot. Mivel időközben a holland Koninklijke Bibliotheek (Nemzeti Könyvtár) is bekapcsolódott a szoftverfejlesztésbe az új-zélandiak mellé, 2018 novemberére elkészült a WCT 2.0-ás változata („WCT manuals”, 2019), amit jelenleg mi is tesztelünk.



16. ábra A Web Curator Tool keretrendszer kezdőképernyője

Ez már teljesen integrálva lett a 3-as Heritrix-szel, és a tervek szerint idővel más crawlert is támogatni fog. A WCT az aratások paraméterezésén és ütemezésén kívül számos egyéb funkciót is tartalmaz. Például metaadatokkal láthatók el és csoportokba sorolhatók a szelektív archiválásra kiválasztott webhelyek. Kiküldhetők vele engedélykérő levelek, majd nyilvántarthatók a kapott engedélyek. Többféle módon segíti a lementett tartalom ellenőrzését, illetve a megismételt aratásoknál

a minőség javítását, a hiányok pótlását, a robot olyan csapdába kerülését, ahonnan nem tud tovább haladni (crawler trap) stb. Kifejezetten csoportmunkára készült, így akár egy intézményen belül, akár más intézményekből többen is használhatják, mert szigorúan elkülöníti a jogosultságokat, különböző szintű felhasználó-csoportokat lehet benne definiálni. Ugyancsak nagyon szigorúan kezeli az engedélyeket: csak olyan webhely menthető le vele, amelyre vagy a webhely tulajdonosa adott engedélyt, vagy a kötelespéldány szabályozás alá esik, és így a nemzeti könyvtár jogosult az archiválásra.

WEB CURATOR TOOL

Home | Queue | Harvested | Help | Logout
User mekdi is logged in.

In Tray | Harvest Authorisations | Targets | **Target Instances** | Groups | Management

Target Summary

Search

ID: From: To: Agency: Owner:

Name: Flag: State: QA Recommendation:

☐ Scheduled ☐ Queued ☐ Running ☐ Paused ☐ Harvested ☐ Archive ☐ Reject ☐ Investigate ☐ Delist ☐ Failed

☐ Non-Display Only: ☐ Aborted ☐ Endorsed ☐ Rejected ☐ Archived ☐ Archiving

Results

Multi-select Action: ☐ delist ☒ archive ☐ endorse ☐ delete ☐ reject

☐	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
☐	Bölcsészettudományi Kutatóközpont (MTA BTK) honlapja	07/02/2018 12:37:17	Harvested	Á. Visky	00:00:13:38	67 MB	3331	0.0	2		View
☐	Bölcsészettudományi Kutatóközpont (MTA BTK) honlapja	08/02/2018 10:29:23	Harvested	Á. Visky	00:00:22:56	742.96 MB	7207	0.0	2		View

Results 1 to 2 of 2
Page 1 of 1
Rows per page: 100

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management

17. ábra Aratások vezérlése a Web Curator Tool keretrendszerben

Ez a része kissé túl is van bonyolítva a rendszernek, elég sok adminisztrálást igényel, mire elindítható egy aratás (OSZK Webarchiválási munkacsoport, 2018).

2019-ben a zágábi IIPC kongresszuson alkalmam nyílt személyesen is megismerkedni a teljes fejlesztői csapattal. Virtuális gép segítségével tesztelhettem a legújabb fejlesztési elképzeléseket. Részben ennek folyamányaként is, különféle nemzeti könyvtárak bevonásával ígéretes fejlesztői közösség alakult ki. Rendszeresen eljuttatjuk az OSZK részéről finomítási igényeinket a fejlesztők felé, ők pedig minden felvetésünkre igyekeznek válaszokat találni. Minden reményünk megvan tehát arra, hogy tovább fog bővülni a keretrendszer funkcionalitása, figyelembe véve a közösség tagjainak egyedi kéréseit is az egyes funkciók fejlesztése kapcsán.

3.2.3.2 Netarchive Suite

A másik általunk tesztelt Heritrix keretrendszer a dán fejlesztésű NetarchiveSuite (NAS) („NetArchiveSuite”, 2017; „NetArchiveSuite documentation”, 2018), amit szintén használnak más (pl. osztrák és francia) nemzeti könyvtárak is. Ebben nincs engedélykérő és -nyilvántartó funkció, kevésbé részletesen lehet metaadatokkal ellátni webhelyeket (valójában csak domén-szintű leírásra van lehetőség, de itt legalább bővíthetők az adatmezők) és az aratási paraméterek közül is csak a fontosabbak definiálhatók úrlapon keresztül („NetArchiveSuite documentation”, 2018). Viszont feltölthetünk a rendszerbe tetszőleges számú, eltérő beállításokat tartalmazó konfigurációs fájlt, melyek közül bármelyiket hozzákapcsolhatjuk egy-egy doménhoz. Az egyedileg

ütemezhető mentések mellett a Snapshot Harvest funkcióval az adatbázisba bevitt valamennyi domén aratása egyszerre is elindítható. A mentett tartalom az OWB proxy üzemmódjával nézhető vissza. Ilyenkor az OWB nem lokalizálja a linkeket, hanem az archívumot üzemeltető szerver megkapja az eredeti webszerver doménnevét. Ehhez proxy módba kell kapcsolni a böngészőt, és egyszerre csak egy webhely nézhető így vissza, viszont az élő webre való áttévedés és az onnan történő tartalom-beszívargás kizárt. A NAS működésével kapcsolatban elég sok problémába ütköztünk: ha nem a localhost-ról, vagyis nem magáról a webadmin.oszk.hu szerverről, hanem PC-ről indítottunk el egyes funkciókat (pl. CDX index generálás, hiányzó linkek kigyűjtése, az aratásról készült report megnyitása), akkor ezek hibaüzenetekre vezettek, valószínűleg amiatt, mert ilyenkor olyan portokat használ a szoftver, amelyek le vannak tiltva a tűzfalon. A NAS különlegessége az ArcRepository modul a WARC fájlok hosszú távú megőrzésére és az ezek sérülésmentességének ellenőrzésére használható Bitpreservation menüpont. Mivel nálunk nincsen ilyen tárolórendszer, ezért ezeket a funkciókat sem tudtuk tesztelni (OSZK Webarchiválási munkacsoport, 2018).

Összességében a bővebb munkafolyamat támogatása miatt döntöttünk a WCT mellett, de osztrák és dán partnereink segítségével igyekszünk figyelemmel követni a NAS fejlesztéseit is.

3.2.3.3 Saját fejlesztések

A saját fejlesztéseink a metaadatok kereshetővé tétele mellett a rendszeres munkafolyamatok automatizálására irányultak. Ilyen a Kaptafa

nevű webes felület, amivel az „egy kaptafára készülő” tömeges aratásokat paraméterezzük a Heritrix robotja számára; valamint az a szkript, amivel a webhelyek kezdőlapjairól készítünk oldalképeket.

3.2.4 Oldalkép készítő programok

Az archívumban való böngészéshez, illetve az archivált verzió és az eredeti weboldal közötti eltérések összehasonlításához, valamint a mentés idején használatos böngészőkben való megjelenés megőrzéséhez fontos, hogy oldalképek is készüljenek legalább a kezdőoldalról és esetleg néhány fontosabb aloldalról. Erre a célra először a Grab Them All (GTA) („Grab Them All”, 2017) nevű Firefox kiegészítőt teszteltük, mellyel többségében jó minőségű, a teljes weboldalt mutató PNG vagy JPG képek készíthetők egy megadott URL címlista alapján. A képek mérete és a várakozási idő jól paraméterezhető, de a fájlnevnél nincs sok lehetőség, így ezeket utólag kell átnevezni az általunk használt formára. Az elvégzett műveletek eredményéről a program Excelbe betölthető CSV fájl generál. Mivel parancssorból is futtatható, alkalmas lenne automatizált működtetésre is, de sajnos évek óta nem fejlesztik, ezért a Firefox 2018-ban megjelent Quantum változataival nem működik, egy régebbi ESR verziót kell telepíteni hozzá. (Mivel ilyenből van “portable” típus is, ezért mindkét Firefox változat használható ugyanazon a PC-n.) További hiányossága, hogy nem görgeti végig a weboldalt, így ha azon olyan elemek (pl. képek) vannak, amelyek csak akkor töltődnek le a szerverről, amikor a felhasználó lejjebb görget, akkor ezek az elemek hiányozni fognak az oldalképről (OSZK Webarchiválási munkacsoport, 2018).

Ezt a hiányosságot küszöböli ki a Nimbus Screen Capture („Nimbus screen capture”, 2018) kiegészítő, aminek mi elsősorban a Chrome-ba beépülő változatát teszteltük. Ha a teljes oldal fotózását kérjük, akkor a Nimbus alulról felfelé végiglapozza a megnyitott weboldalt, így általában egészen pontos képet készít az oldal egészéről (bár ennél is belefutottunk néha hibákba, pl. a Flash-alapú weboldalaknál). Az elkészült kép utólag még szerkeszthető is, majd elmenthető a saját gépre vagy egy felhőtárhelyre. Hátránya viszont, hogy nem adható meg egy URL lista, így nem alkalmas tömeges képkészítésre, ezért inkább csak olyankor használjuk, amikor az automatikus megoldások nem jó eredményt adtak (OSZK Webarchiválási munkacsoport, 2018).

A windowsos szoftverek korlátai miatt és azért, hogy az oldalképek készítése minél jobban integrálható legyen az aratási folyamattal, 2018 szeptemberében elkezdtünk Linux-alapú megoldást is keresni. Négy-öt különböző eszközt teszteltünk ugyanazokkal a weboldallal és ezután választottuk ki azt, amelyik a legkevesebb hibás képet gyártotta. A választott megoldás a Puppeteer nevű programmal („Puppeteer”, 2020) vezérelt Chromium (a Chrome böngésző motorja) lett, amihez Vitéz Gábor a projekt rendszergazdája írt egy scriptet, hogy egy szövegfájlban általunk megadott URL címekről töltődjének be a weboldalak, majd pedig az elkészült képek vagy a szintén előre megadott MIA azonosítókat tartalmazó, vagy pedig az URL címekből kódolt és az aznapi dátummal kiegészített nevű fájlokba kerüljenek. Az első nagyobb képhalmazokat részletesen ellenőriztük, és a típushibákat a program paramétereinek finomhangolásával néhány hét alatt részben sikerült kiküszöbölni. A

problémák egy része automatikusan nem javítható, mert például van, ahol előbb rá kell kattintani egy figyelmeztető ablakra és csak azután jelenik meg rendesen a weboldal. Ilyen esetekben a Nimbusszal kell kézzel elkészíteni a megfelelő oldalképet (OSZK Webarchiválási munkacsoport, 2018).

3.2.5 Informatika tanulságok

A webarchiválás informatikai hátterének megteremtése az egyik legjelentősebb kihívásnak bizonyul napi munkánk során. A hagyományos web átrendeződését a platform alapú, bonyolult programozási hátterű, a közösségi média által nagyban befolyásolt környezet irányába az archiváló szoftverek fejlődésének is követnie kell. Egy alapvető generációváltás időszakában élünk. A régebbi eszközök, bár még megbízhatóak és részlegesen jól használhatók, már nem elégítik ki minden archiválási igényünket. Az újabb szoftverek pedig még nem elég kiforrottak. Néhány éven belül talán tisztul majd a kép, amikor az új webes környezethez kapcsolódó szoftvergeneráció is megjelenik kiforrott funkcionalitással. A fejlődés azonban sohasem áll meg. Amikor az éppen aktuális hátteret foglaljuk össze, nagyon tudatában kell lenni annak, hogy csupán pillanatképeket készítünk, s akár egy-két éven belül is olyan változások következhetnek be, hogy az itt leírtak jó részének csupán történeti jelentősége marad majd.

3.6 A metaadatok rögzítése

Egy webarchívum metaadat-struktúrájának kialakításakor meg kell ismerni először is magának a forrásnak a természetét. Ezt követően kapcsolódási pontokat kellene találnunk a bibliográfiai világban a

metaadatszerkezet megtervezéséhez, ahogyan ezt a modellt a Belga Nemzeti Könyvtár munkatársai kézzelfogható módon bemutatják (Di Pretoro & Geeraert, 2019). Első lépésben egy saját XML alapú metaadat-struktúrát alkottunk meg a kísérleti projekt keretei között (Németh & Drótos, 2019b). Ezt követően indult el egy, a legújabb RDA bibliográfiai leírási szabványkörnyezethez kötődő metaadatmodell megalkotása, az OSZK Szabványosítási Irodájában tevékenykedő Ilácsa Szabina segítségével, aki számot is adott már publikált formában a közös munka részeredményeiről (Ilácsa, 2020). Ennek kapcsán a munkálatok még nem zárultak le, de az alapvonalakról már áttekintést lehet adni.

3.6.1 A saját metaadat-struktúra kialakításának kezdő lépései

Egy könyvtári webarchívumnál jogos elvárás, hogy a teljes szövegű keresés mellett metaadatok alapján is kereshető, illetve böngészhető legyen. Az archiválási munkafolyamat és a hosszú távú megőrzési kötelezettség pedig azt is megkívánja, hogy a bibliográfiai leírások mellett adminisztrációs és technikai metaadatokat is rögzítsünk. A metaadatokkal történő ellátás történhet gyűjtemény, részgyűjtemény, webhely, webhelyrész és fájl szintjén (Németh & Drótos, 2019b) (Utóbbi természetesen csak a fájlokból automatikusan kinyerhető vagy generálható adatokból, hiszen több száz millió fájl esetében más megoldás nem jöhet szóba.). A projekt keretében egyelőre a részgyűjtemény és a webhely szintű metaadat rögzítés lehetőségeivel foglalkoztunk (Drótos & Visky, 2018a). Az IIPC próbált először lépéseket tenni, egy, a tagok által használandó alap metaadat struktúra kialakítása felé, de a munka az előkészítő szakasz után elakadt (Clifton, 2005; „IIPC

Web Archiving Metadata Set—UNT Digital Library”, 2005). Az amerikai OCLC (Online Computer Library Center) Web Archiving Metadata Working Group nevű munkacsoportjának 2018 februárjában közzétett (Jackie & Bowers, 2018), Dublin Core-alapú ajánlását (a leíró alapú metaadatok rögzítendő típusainak meghatározásában), valamint a Magyar Elektronikus Könyvtár, illetve Elektronikus Periodika Archívum már létező gyakorlatát (elsősorban az XML alapú adatstruktúra szerkezete mintájaként) figyelembe véve dolgoztuk ki 2018 áprilisában a saját XML alapú adatstruktúránkat (Drótos & Visky, 2018a; Németh & Drótos, 2019b), amelynek egyes mezőit, amennyiben lehetséges volt, megfeleltettük a MARC21 adatszerkezeti elemeinek is. Az OCLC ajánlás (Jackie & Bowers, 2018) és a saját MIA sémánk (Drótos & Visky, 2018a; Németh & Drótos, 2019b) is 14 leíró elemet tartalmazott, de ezek közül csak 11 feleltethető meg egymásnak egy az egyben (collector, contributor, creator, date, description, language, relation, rights, source of description, subject, title), az extent-et a MIA nem vette át, a publisher az OCLC ajánlásban nem található meg. Két elem a MIA-ban bővebb értelmű, mint az OCLC ajánlásban (genre/form < dc_type; URL < dc_identifier). Fontos korlátja az OCLC ajánlásnak, hogy a 14 elem csak a leíró adatokra vonatkozik, viszont (szinte) mindegyik tovább tagolódik alelemekre (Ilácsa, 2020).

A honlap, illetve részgyűjtemény szintű metaadatsémáról, a metaadatok elkészítésének szabályairól részletes útmutatót állítottunk össze, munkatársaimmal együtt több előadásban, illetve publikációban is feltártuk a metaadatkészítés részleteit (Drótos & Visky, 2018a, 2018b;

Németh & Drótos, 2019b). Az egyedi webhelyekhez készült metaadatolási útmutató 91 leíró alelemet különböztet meg, az összes metaadattípus alelemeinek száma pedig 133, melyeket három csoportba soroltunk: leíró, adminisztrációs és technikai. A mezők többsége ismételhető vagy csoportosan ismételhető (vagyis almezőkből áll), és ahol lehetett, ott választható, kötött listákat definiáltunk, hogy egyértelmű és egységes legyen ezeknek a mezőknek a kitöltése. Egy webarchívumnál a nagy számosság miatt nem várható el, hogy minden webhelyről azonos részletességű leírások készüljenek, mert ez a tevékenység nagyon élőmunka-igényes, és a legszűkebb keresztmetszet mindig a rendelkezésre álló munkaerő. Ezért a webhelyek esetében csak 3 kötelező mezőt határoztunk meg: azonosító, név, URL, minden további adat kitöltése opcionális, illetve reményeink szerint részben majd automatizálható lesz (Drótos & Visky, 2018a; Németh & Drótos, 2019b).

Bár a WCT és a NAS is képes bizonyos metaadatok nyilvántartására, de ezen a téren egyik sem elég rugalmas és fejlett (a WCT-ben jelenleg még nem is bővíthetők szabadon az adatmezők). Így egyéb megoldásokat is számításba vettünk. Itt jött jól az az XML alapú adatszerkezet minta, mely egyéb digitális gyűjteményeinknél már kidolgozásra került. Annál is inkább igaz ez, mert ugyan készítettünk néhány tesztrekordot az OSZK Tudástárak alprojektjéhez készült adatrögzítő programmal, valamint a KOHA nevű ingyenes könyvtári rendszert is megnéztük ilyen szempontból, ám sajnos egyikhez sem kaptunk olyan informatikai segítséget, amivel az elképzelt adatstruktúrát ki tudtuk volna alakítani.

2018 júniusában egy XSD fájlban definiáltuk a webhelyeket leíró metaadatsémát, amit a következő hónapokban a tapasztalataink alapján folyamatosan finomítottunk, és később a részgyűjtemény szintű leíráshoz is létrehoztunk egy hasonló állományt (Drótos & Visky, 2018a). A fejlesztést témafelelős és webkurátor munkatársaimmal együtt végeztük el. Az XSD fájlok alapján a Windows alatt futó ingyenes XML Notepad szerkesztővel tudjuk elkészíteni a metaadatokat tartalmazó XML fájlokat. Az egész munkafolyamathoz részletes útmutatókat írtunk, hogy minél egységesebbek és pontosabbak legyenek a leírások, melyeket minden esetben még egy utólagos ellenőrzésnek is alávetettünk a hibák, hiányok és következtelenségek kiküszöbölése érdekében (Drótos & Visky, 2018a, 2018b). Azért is választottuk a fent említett szerkesztőprogramot, mert ebbe mezőkre lebontva bele tudtuk integrálni a vonatkozó súgó, útmutató szövegeket. Ahhoz, hogy ezek a metaadat rekordok a weben is megnézhetők legyenek, XSL stílusfájlokat is készítettünk, melyek alapján a böngészők szépen formázott HTML oldalakká tudják alakítani az XML fájlokat.

A metaadatoláshoz kapcsolódó valamennyi segédfájl és dokumentum, valamint a folyamatosan készülő XML rekordok elérhetők a projekt webszerveréről (Drótos & Visky, 2018a; Németh & Drótos, 2019b). A KIFÜ-s szerveren levő zárt anyagról pedig 2018 szeptembere óta részgyűjtemény szintű leírásokat készítünk, de ezek nem érhetők nyilvánosan (Drótos & Visky, 2018b).

Az XML alapú metaadatok kereshetősége egy ideiglenes keresőmegoldás segítségével is napirendre került a Solr alapokon. Így az ideiglenes

adattároláshoz egy ugyancsak ideiglenes metaadatkeresőt is társítani tudunk.

ÓBUDAI MÚZEUM BLOGJA

MIA azonosító: MIA-000063
Eredeti URL: <http://obudaimuzeum.blogspot.hu>
Seed URL: <http://obudaimuzeum.blogspot.hu/>
Nyilvános OSZK-s archív URL: http://193.6.201.202/ovb/wayback/*http://obudaimuzeum.blogspot.hu/
Linktérkép URL: <http://193.6.201.202/solrwayback/waybacklinkgraph.jsp?domain=obudaimuzeum.blogspot.hu>
Más nyilvános archív URL: http://web.archive.org/web/*http://obudaimuzeum.blogspot.hu/

Egységesített cím: Óbudai Múzeum blogja
Eredeti főcím: Óbudai Múzeum - helytörté.net
Alcím: Egy múzeum élete

Leírás forrása: Az Óbudai Múzeum blogja és a múzeum honlapja
Forrás URL: <http://obudaimuzeum.blogspot.hu>
Forrás URL: <http://www.obudaimuzeum.hu>
Leírás dátuma: 2018-07-05
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: Óbudai Múzeum

Kiadó neve: Óbudai Múzeum
Kiadó e-mail címe: titkarsag@obudaimuzeum.hu
Kiadó honlapja: <http://www.obudaimuzeum.hu>

Tartalmi jogtulajdonos: Óbudai Múzeum
Szolgáltatási/technikai jogtulajdonos: Google Inc.: Blogger
Jogi státusz: jogvédelem

Kapcsolati személy neve: Népessy Noémi
Kapcsolati személy beosztása: múzeumigazgató
Kapcsolati személy e-mail címe: titkarsag@obudaimuzeum.hu

Műfaj: Blog
Tulajdonosi típus: Intézményi
Változékonyság: nem változó

Fő témakör: Kulturális források
Fő témakör: Tudományos források
Fő témakör: Oktatási források
Témakör: Köz- és magángyűjtemények
Témakör: Társadalomtudomány
Témakör: Ismeretterjesztés
Alttémakör: Múzeumok
Alttémakör: Történelem, helytörténet
Tárgyszó: helytörténet
Tárgyszó: várostörténet

18. ábra XML alapú metaadat rekord megjelenítési nézetben

Reményeink szerint az ezekben az XML fájlokban levő adatok betölthetők lesznek majd az OSZK leendő új digitális gyűjteménykezelő rendszerének adatbázisába, és azzal még a jelenleginél is jobban strukturált adatszerkezetet tudunk majd kialakítani, valamint megoldódik vele a metaadatokban való keresés és a leíró típusú adatok MARC21-re való konverziója is. Bár 2019-2020 folyamán többször is egyeztettünk a Hermész-Soft Kft. munkatársaival arról, hogy hogyan lehetne a jelenleg Google táblázatokban és XML fájlokban tárolt metaadatainkat egy

adatbáziskezelőbe áttenni és a továbbiakban abban menedzselni őket, érdemi előrelépés nem történt ezen a téren. Az általuk javasolt OTR rendszer képességei jelenleg nem elegendők a webarchívum speciális igényeinek (pl. az időben változó adatoknak) a kezelésére, és bár az egyeztetések során ígértek fejlesztéseket, előrelépés nem történt és arról sincs információnk, hogy haladnak ezek megvalósításával.

3.6.2 Az RDA kompatibilis metaadatmodell megteremtésének első lépései

Az OSZK Szabványosítási Irodájának segítségével Ilácsa Szabina bevonásával indultak el azok a munkálatok, melyeknek segítségével a fentebb vázolt XML alapú metaadatmodellre is épülve be tudjuk építeni a webarchívum archiválását a már létező szabványos bibliográfiai környezetekbe (Ilácsa, 2020). Az első és legfontosabb feladat egy metaadat tipizálási modell kiválasztása volt. Ez voltaképpen nézőpont választást jelent, s annak pontos meghatározását, hogy egy adott típus környezetében hogyan határozzuk meg a metaadat kategóriákat (Ilácsa, 2020; Németh, 2018b).

A munkálatok során a NISO kategóriarendszerre esett a választásunk. Legfőképpen azért, mert a webarchívumokban történő adattárolás és az archivált anyagok hozzáféréseinek szempontjából egyaránt kiemelten fontos jogi jellegű metaadatok külön kategóriát képeznek, a fentebb tárgyalt bibliográfiai leíró és adminisztrációs jellegű metaadatok mellett. Utóbbi kategórián belül a technikai metaadatok a fájlok dekódolásához, használatához szükségesek, a megőrzési metaadat a fájlok hosszú távú kezelését, esetleges migrációját segítik. A jogi metaadatok pedig a tartalomhoz kapcsolódó szellemi tulajdonjogokat írják le. A

webarchiváláshoz kötődő metaadatléírás sajátossága, hogy az általános gyakorlattal szemben, a fentebb vázolt kategóriában rögzített metaadatok nagyobb fontossággal bírnak a leíró metaadatoknál, hiszen így tudjuk meghatározni az archivált állományhoz való hozzáférés, illetve használat kereteit. Arra is gondolnunk kell, hogy a felhasználók számára igen lényeges információkat jelent az archiválási folyamat részleteinek taglalása, például az, hogy milyen mélységű a mentés, vagy azon elemek listája, amit nem sikerült lementeni. A kutatási célú webarchívum használatánál a kutató számára kulcskérdés, hogy az archivált anyag mennyire ad hű képet a forrás akkori állapotáról. Arról pedig már korábban is szó esett, hogy a tartalomhoz kapcsolódó szerzői jog tulajdonosának hozzájárulása nélkül nincs lehetőség az archivált webhely nyilvános szolgáltatására. Így a forrás jogi helyzetéhez kapcsolódó adatok megfelelő nyilvántartása a webarchívum mint szolgáltatás alapját képezik (Ilácsa, 2020).

Érdeemes utalni arra is, hogy külön kihívást jelentett a nemzetközi gyakorlat számbavételekor, hogy miután a könyvtárak és a levéltárak is foglalkoznak webarchiválással, s a két intézménytípus meglehetősen eltérő feldolgozási, s ebből adódóan metaadat leírási gyakorlatot is követ, ezért dönteni kellett, hogy melyiket tekintjük fő kiindulópontnak. Ha a levéltári gyakorlatot vesszük mintának, a felhasználót már alapértelmezett kategorizálással segíthetjük, ahol egy forrás többfajta kategóriába is besorolható, számolni kell azonban e besorolás alapvetően szubjektív voltával is. Az egyedi forrás feltárásának hiánya a feldolgozónak, illetve a felhasználónak egyaránt kihívást jelenthet. Új

kategóriák létrehozásakor a válogatást külső források használatával kell elvégezni (Ilácsa, 2020). Amennyiben könyvtári forrásként tekintünk egy weboldalra, akkor az adott forrás több szempontból kereshetővé válik, nem csupán szubjektív kategóriák szerint. A felhasználó a keresés eredményeként kapott találati halmaz elemeit tetszés szerint csoportosíthatja, több információt kaphat egy adott forrásról. A metaadat modell megalkotásakor kezdetektől arra törekszünk, hogy mindkét szemlélet előnyeit ötvözni tudja (Ilácsa, 2020). A metaadatok előállításának részleges automatizálási lehetőségeivel az egyes webhelyekbe beépíthető szemantikus jellegű mikroadatok, illetve az archívumbarát webhely ajánlások alkalmazásának segítségével a zárófejezet foglalkozik majd.

A metaadatok leírása kétféle szinten történik meg: gyűjteményi, illetve egyedi webhelyek. A gyűjteményi szinten a válogatás alapja szerint határozhatunk meg kategóriákat, ezek tartalmát is meghatároztuk már korábbi fejezetekben (tematikus, eseményalapú, intézmény, személy, földrajzi hely szerinti, vegyes, illetve a magyar webtérre koncentráló kis mélységű általános mentések). Az egyedi leírások tárgya értelemszerűen maga az adott webhely. Az egyes kategóriák között nincs formálisan kifejezett kapcsolatrendszer. A webtér szintű gyűjtemény gyűjtőköri megközelítés szerint ugyan minden egyéb tematikus, illetve eseményalapú gyűjtemény felettese lehet, azonban a gyakorlatban ez nem jelent rész-egész kapcsolatot. Az esemény alapú, illetve a tematikus gyűjtemény gyűjtőköri különbségeit pedig korábban vázoltuk már. Egyedi szintű metaadat leírások jelenleg azon webhelyekről készülnek, melyeket

a megfelelő engedélyek birtokában nyilvánosan is szolgáltatunk, de persze ennek gyakorlata később még változhat. A másik fontos szempontja az egyedi leírásnak a gyűjtőkör szerint kiemelten fontos tartalmak elkülönítése. A jövőben formálódó gyűjteményszervezési gyakorlat alakítja ki ezen webhelyek körét. Elképzelhető továbbá, hogy olyan alkategóriákat alkotunk később, melyekhez különféle részletességű leírási gyakorlat lesz rendelve, akkor majd ezek részletes gyűjtőköri lehatárolását és metaadatolási szabályozását is meg kell majd alkotni. A metaadatmodell megalkotását mindenesetre az egyedi webhelyekre mint fő kategóriára támaszkodva tesszük meg, az IFLA könyvtári referencia modelljére szabott RDA szabvány segítségével (Ilácsa, 2020).

Az IFLA Könyvtári Referenciamodell (LRM) a tág értelemben vett bibliográfiai adatokra szabott entitáskapcsolati modell (Riva, Boeuf, Žumer, & Szerkesztőbizottsága, 2017), mely nem tesz különbséget az adatok között aszerint, hogy milyen típusú rekordokban tároljuk azokat. Fel lehet térképezni segítségével az adott forrás belső metaadatszerkezetét, illetve kapcsolati hálóját, külső források, entitások felé, így az adatok egymás közti viszonyrendszere is ábrázolásra kerülhet. A modell magyar fordításban is megtalálható az IFLA honlapján. Az LRM szemléletmódja adja meg az alapját az RDA katalogizálási szabályzatának. Az RDA az LRM koncepciójára épülő adatelemek, irányelvek és instrukciók összessége (Ilácsa, 2020). Ennek segítségével áttekinthetjük a metaadatok felhasználási lehetőségeit, majd az ideális állapotként felvázolt célokat hozzárendelhetjük jelenlegi rögzítési lehetőségeinkhez és kapacitásainkhoz. Ennek keretei persze majd az időben előre haladva

dinamikusan változhatnak, a tervezésnél tehát kissé proaktívnak kell lennünk. Amennyire lehet, nekünk kell formálnunk a külső környezetünket annak érdekében, hogy ne a jelenlegi korlátokból kiinduló, ám gyorsan elavulni képes gyakorlat mellett kössünk ki.

Az LRM egy új szempontot jelenít meg a webarchívumokhoz kötődő metaadatolás kapcsán, mellyel a kísérleti modellünk megalkotásakor korábban nem számoltunk. El kell különíteni koncepcionálisan ugyanis azt az oldalt, amelyről a mentés készül, illetve azt, ami az archivált példányként jelenik meg. Bonyolult kérdéseket vet fel e két oldal egymás közötti viszonya. Ugyanazon megjelenési formának két külön kifejezési formájáról beszélünk, vagy két külön műről? A digitális másolatok esetében általában az előbbi gyakorlatot követjük. Az archivált webhely azonban számos tulajdonságában eltér a digitális másolatok alapjellemzőitől. Itt nem stabil tartalmú fizikai példányról készül a másolat, hanem dinamikus, folyton változó környezetről. Elég csak visszautalniom e munka bevezető fejezetére, ahol a web, mint dokumentumforrás sajátosságait ecseteltük már korábban. A mentett állapot szinte soha nem tükrözi a web önkéntesen kiragadott szegmensének eredeti szellemi és művészi tartalmát, nem is beszélve az azt körülvevő környezetről. Az aktív és az archív webhelyeket szerencsésebb tehát olyan külön entitásoknak tekinteni, melyek között szoros kapcsolat áll fenn. Két külön leírást is kell készíteni róluk, mert attól függetlenül, hogy elméletileg minek tekintjük őket, a megjelenési forma szintjén mindenképpen elkülönülnek. A megjelenési formák és a példány

szintű leírások külön kezeléséhez tehát egyedi leírásokra van szükségünk (Ilácsa, 2020).

Az LRM alapú RDA adatmodell kiindulási pontjaként a MIA kísérleti modellben publikált leírások és adatelemek szolgáltak. A rendelkezésre álló nemzetközi szakirodalomban vázolt módszerek közül az OSZK RDA munkacsoportban dolgozó kollégánk által a Zeng-Quin Metadata című könyvében (Zeng & Quin, 2016) található módszer lett kiválasztva tulajdonképpen sorvezetőnek (Ilácsa, 2020). A szerzők többféle módszer és a saját tapasztalataik alapján dolgozták ki a maguk módszerét. Azért erre esett a választás, mert az általa megfogalmazott lépések könnyebben voltak igazíthatók az elképzeléseinkhez. Más kérdés, hogy miután itt már egy meglévő metaadatstruktúra új környezetbe történő implementálásáról (átstrukturálás, kiegészítés, interoperabilitás megteremtése) van szó, s nem a szakirodalomban felvázolt nulláról történő építkezésről, ezért egyes elemeket értelemszerűen elhagytunk, illetve módosítva használtunk fel (Ilácsa, 2020).

A gyakorlati munkát három nagy szakaszra bontottuk. Ebből az első már teljes egészében kész van, hiszen itt a rendelkezésre álló gyűjtemény elemzéséről és megértéséről volt szó, illetve már korábban felvázoltuk a használni kívánt elemeket is (Ilácsa, 2020). A második szakaszban elemkészletet, illetve alkalmazásprofilt kell kidolgozni a webarchívum gyűjteményhez. Először a funkcionális követelményeket kellett megfogalmazni. Ezt követően kellett felvázolni egy olyan szakterületi modellt, mely rögzíti, hogy milyen entitásokat akarunk leírni s milyen kapcsolat lesz ezek között (Ilácsa, 2020). A kívánt metaadatelemeket be

kellett azonosítani, le kellett írni, magyarázatokkal kellett ellátni, példákat rendelni a használatukhoz, illetve egyedileg feltüntetni, hogy kötelezően használható, választható, illetve ismételhető elemről van-e szó (Ilácsa, 2020). A következő lépés annak eldöntése, hogy az imént leírt elemek értékei ellenőrzöttek-e. Ha igen akkor meg kell határozni annak módját (kifejezések listája, értékszótár/séma, már létező jó gyakorlat átvétele). Át kellett azt is tekinteni, hogy a már létező névterekből milyen elemeket lehet átemelni a saját modellünkbe, mintegy újrahasznosítva azokat (Ilácsa, 2020). A jövőbeni harmadik szakaszban az esetleges névtér átvételek után a végleges állapotát elnyert elemlista tételeit egyedileg dokumentálni kell (név, címke, URI, definíció, implementálás státusza (kötelező vagy opcionális), illetve ismételhető-e az adott elem, megjegyzések benne az esetleges névtér kontroll adatokkal). Az elemkészlet elkészülte után meg kell feleltetni azt például egyéb szabványos adatcsere formátumokkal. Ezt követi az átfogó útmutató elkészítése a teljes elemkészlethez. Minden egyes elemnél megjegyzésben kell rögzíteni a használatáról szóló alapvető irányelveket, tanácsokat. Ezeket a megjegyzéseket aztán egy önálló dokumentumban kell összefoglalni, mely gyakorlati mellékleteként szolgál az átfogó útmutatónak. Az utolsó lépés pedig egy XML séma vagy egy OWL ontológia elkészítése, illetve az elemek kifejezése RDF Schema nyelven (Ilácsa, 2020). Ennek gyakorlati hasznáról, alkalmazásáról a szemantikus webhez történő illeszkedés kapcsán majd a webarchiválás kutatási célú felhasználása kapcsán a későbbiekben esik szó.

Az elemkészlet rögzítésének véglegesítése folyamatban van. Ezt követően kell egyedileg dokumentálni az összes elemet, majd elkészíteni a részletes dokumentációt és a szemantikus megfeleltetéseket a harmadik fázis részeként. Amennyiben a szabványosítási irodával együtt eredményesen le tudjuk bonyolítani ezt a projektet, akkor nemzetközi értelemben is számottevő kutatás-fejlesztési eredményeket érhetünk el a közeljövőben. A metaadatmodell, illetve annak szemantikus kifejezése referenciaként szolgálhat, jó gyakorlatként egy nemzetközi ajánlás rendszer kidolgozásához is.

3.7 Minőségellenőrzés, a webarchiválás határai

3.7.1 A webarchiválás határai

Fontos kutatási témaként merült fel már a hipotézisek megfogalmazása kapcsán is kutatói munkám során a webarchiválás hatóköri lehatárolásának kérdése, illetve a begyűjtött anyag minőségellenőrzésének kihívásai. A robotokkal történő automatikus archiválás a web fejlődésével egyre nagyobb korlátokba ütközik, mint ahogy már a nyitó fejezetben is szó esett erről. Arra is utaltam már korábban, hogy a web élő szövetéből eleve csak töredékeket tudunk kiragadni, amelyek önmagukban csak részlegesen állják meg a helyüket. Egy aratórobottal csak megadott mélységben és megadott domain tartományban tudjuk a webes tartalmakat archiválni. A robotnak tehát megadhatjuk, hogy egy adott webhely határai között mozogjon, de e külső webes célpontokból beágyazott interaktív tartalmakat nem fogja lementeni. Ide kapcsolódik a felhőben történő dokumentumtárolás kihívása is. Számos kiadó a nyílt hozzáférésű elektronikus folyóiratának

számait felteszi a felhőbe, s onnan szolgáltatja. A robotok számára, melyek az adott folyóirat honlapját mentik, nehezen specifikálható az a felhőtartomány, amelyből ezeket a dokumentumokat menteni kellene. A külső interaktív elemek mentését nem csupán nagyon nehéz specifikálni a robotok számára, hanem az is előfordulhat, hogy a mentett tartalomba olyan kód kerül be, ami az élő webről vesz tartalmakat, ami a lementett anyag történeti (illetve adott esetben akár jogi) hitelességét kérdőjelezi meg. A beszűrődési jelenséget hívjuk angol nyelvű szakkifejezéssel live web leakage-nek („Live web leakage – MIA”, 2017), az időbeliség problémáját pedig a temporal drift („Temporal drift – MIA”, 2017) fogalmával jellemezhetjük.

Külön kihívást jelent több szempontból a közösségi média tartalmainak aratása is. Az egyes platformok tartalmainak jó része csak (általában ingyenes) regisztráció ellenében érhető el. Ez már rögtön egy jelentős probléma, hiszen a robot nem jut át a regisztráció falán. Éppen ezért erre a célra különleges célszoftvereket elehet alkalmazni. Ilyen például a Webrecorder, melynek működése az offline világból ismert videomagnóéhoz hasonlít. Böngészővel járjuk az adott platform tartalmait körbe, s közben a program WARC fájlba rögzíti a megtekintett elemeket. Eredményesen alkalmazható bizonyos határig Instagram, Facebook, illetve Twitter tartalmak mentéséhez is. Sajnos azonban ez a mentési folyamat sem hajtható végre maradéktalanul sikeresen. A rögzítő programban ugyanis nem lehet beállítani azt, amikor a folyamatos lefelé görgetés közben késleltetetten szolgáltat adatokat a szerver. Ilyenkor azt érzékeli a rögzítő, hogy véget ér az aratás és zárja a munkafolyamatot.

Külön problémát jelentenek azok az extrém módon jelenorientált tartalmi elemek, mint például a Facebook történetek vagy az Insta történetek, amelyek adott esetben csak néhány percig érhetők el a világhálón, s utána törlődnek is az éterből... Számos esettanulmány járta az utóbbi években körül ezeket a kihívásokat (Acker & Kreisberg, 2019; Anderson, 2020; Elson, 2020; Littman és mtsai., 2018; Macnaught, 2018; Thomson & Kilbride, 2015).

Az adatbázisok mentése a robotok számára szintén megoldhatatlan kihívást jelent. Egy szöveges, illetve multimédia adatbázis maradéktalan lekérdezéséhez az összes létező keresőkérdést le kellene futtatni, amely az adott adatbázisban tárolt elemek visszakeresésére vonatkozik, ami persze nyilvánvaló képtelenség. Archivumbarát megoldást jelenthet például, ha az adott adatbázis gazdája XML-ben kiexportálja egy adott helyre az adatbázis tartalmát, mely így már aratható lenne a robotok által. Arról már a zárófejezetben lesz szó, hogy ennek az elérési útját mikroadatként akár bele lehetne illeszteni az adott webhely forráskódjába is, a robotok által értelmezhető módon.

A nyitófejezetben esett már szó azokról a nem hagyományos dokumentumtípusokról, melyek egyre nagyobb szeletét hasítják ki a világhálónak. A mobilplatformok s az azokra épülő alkalmazások archiválása teljesen új kihívást jelent a közgyűjtemények számára. Jelenleg a British Library foglalkozik önálló kutatás-fejlesztési projekt keretében azzal, hogy ezekhez a zárt platformokhoz, s azok alkalmazásaihoz hogyan lehet archiválási céllal hozzáférni, s egyáltalán milyen módszerrel ragadható meg a tartalmak begyűjtése.

Szintén szó esett már arról, hogy a szöveges tartalom rovására a világhálón egyre nagyobb a szerepe a multimédia tartalmaknak. Az online műsorszórás (streaming), illetve az on-demand alapú médiaszolgáltatások tartalmainak mentéséhez is külön speciális célszoftverekre van szükség. Igen összetett feladatkörrel van szó, az eredményesség hatékonysága pedig nagyban függ az archiválni kívánt tartalmak tulajdonosainak és szolgáltatóinak aktív együttműködésétől is. De az online média archiválásának területe kívül is esik az Országos Széchényi Könyvtár webarchiválási tevékenységének gyűjtőkörén, s akár önálló disszertáció témája is lehetne. Épp ezért a továbbiakban erről nem fogok szót ejteni.

3.7.2 Minőség-ellenőrzés

A minőség-ellenőrzés szakirodalma a webarchiválás kapcsán nem mondható bőségesnek, az intézményi példák közül Nicola Bingham révén British Library gyakorlatát tudtam áttekinteni a saját projektünk tervezésekor (Bingham, 2014, 2016), illetve a 2010-es évek elejének SHARK projektjéig lehetett visszanyúlni (Denev és mtsai., 2009, 2011).

A szelektíven mentett webhelyeknél a minőség-ellenőrzés történhet szűrőpróbaszerűen (nagy seed listák aratásakor), illetve szisztematikusan (például a demó gyűjteménybe válogatott webhelyek esetében). Az ellenőrzés időpontja szerint lehet azonnali (az aratás lefutása után 1-2 órán vagy napon belül), vagy visszamenőleges (akár több hónappal a mentés után). A minőség-ellenőrzés szintje szerint pedig három kategóriát különböztetünk meg: minimális, gyors és részletes. Ez a három szint egymásra épül, ami azt jelenti, hogy a gyors ellenőrzés esetében is mindenképpen el kell végezni a minimális szint feladatait (ha az nem

történt meg korábban), a részletes ellenőrzés pedig a minimális és a gyors szint teendőit is magába foglalja, viszont az egyes munkafolyamatokat alaposabban végezzük el. Minőség-ellenőrzést már a webhely első teszt és/vagy éles mentéséről érdemes végezni. Ezt követően pedig olyan későbbi mentések esetében is sorra kell kerülnie, amelyeket más paraméterekkel vagy más szoftverrel arattunk, illetve olyankor is el kell végezni, ha jelentősen átalakult a webhely (pl. tartalomkezelő rendszert váltott az adott webhely). Ha több, különböző időpontban készített mentésünk van ugyanarról a webhelyről, amely lényegileg nem változott, webhelyről, akkor az ellenőrzés tárgyául a legfrissebb mentés szolgál. Az ellenőrzéseket az erre rendelkezésre álló munkaidő függvényében és a webhely fontosságát figyelembe véve, az észszerűség határáig érdemes csupán folytatni.

A webarchiválást támogató keretrendszerek támogatják a minőségbiztosítási folyamatok ellátását. A WCT fejlett beépített funkciókat tartalmaz a minőség-ellenőrzési feladatokhoz. Ezek működése a felhasználói segédlet (User Manual) "Target Instance Quality Review" című fejezetében részletesen rögzítve van (*NetArchiveSuite documentation*, 2018).

Az ellenőrzés legegyszerűbb, minimális változata csak a Heritrix naplófájljaira terjed ki, hogy egyáltalán sikerült-e valamennyi tartalmat lementeni a megadott seed cím(ek)ről, és ha nem, akkor miért nem (pl. hibás a seed cím, nem működött a webszerver, ki vannak tiltva a robotok). Ezek az információk részben a Heritrix által létrehozott crawl-report.txt és a seeds-report.txt áttekintésével, részben pedig a seed URL cím és az illető

szerveren levő robots.txt fájl böngészőben való megnyitásával deríthetők ki. A crawl-report.txt-ben az "URI successes" és "URIs processed" értékek aránya például megmutatja, hogy a lekért fájlok hány százalékát sikerült ténylegesen letölteni. Érdekes megnézni még a mimetype-report.txt fájlt is, hogy valóban kimaradtak-e a letöltésből azok a fájltypusok (pl. videók), amelyeket nem akartunk lementeni, illetve letöltöttük-e azokat a fájlokat (pl. képek), amelyeket viszont igen. Az első néhány tematikus mentés során valószínűleg rossz sorrendben lettek rögzítve a szűrőfeltételek a Heritrix konfigurációs fájljába, éppen ezért jelentős mennyiségű multimédia anyag is aratásra került, melyet ki szerettünk volna zárni a gyűjtésből. Ha gyanúsan kevés pl. a képfájlok száma és/vagy összmérete a mimetype-report.txt-ben, akkor újra a robots.txt-t kell megnézni, nincs-e benne kitiltva a képeket tartalmazó alkönyvtár (pl. img/, kepek/, uploads/). Ha pedig irreálisan nagy a teljes mentésnek vagy valamelyik fájltypusnak az összmérete, akkor a hosts-report.txt-ben kell ellenőrizni, hogy nem tévedt-e át valamelyik külső doménre a robot, ahonnan nagy mennyiségű tartalmat szedett le. Erre példa, amikor egy olyan domént arattunk le, amin régen magyar tartalom volt, viszont időközben kínai tulajdonba ment át. Kiugróan sok tartalom töltődött le róla, s a minőség-ellenőrzés során tisztázódott a probléma, hogy ez a honlap már kívül esik a gyűjtőkörön.

A következő szint a gyors ellenőrzés, ahol már az OWB-ben (és/vagy a NAS ill. WCT beépített megjelenítőjében) is meg kell nézni a legalapvetőbb dolgokat maximum 10-15 perc alatt. Az archivált webhelyek kezdőlapjainak gyors átnézéséhez nagyon hasznos két Chrome-kiegészítő:

Open Multiple URLs és a Revolver Tabs. Előbbivel akár száz fület is megnyithatunk, utóbbival pedig néhány másodpercenként váltogathatjuk ezeket. Az alábbiakban főleg az Open Wayback megjelenítési kihívásaival foglalkozom. Emellett azonban meg kell jegyeznem, hogy célszerű többféle megjelenítővel (pl. PYWB) is tesztelni az adott oldalt, mert számos hiba magának a megjelenítő programnak a konfigurálásából adódhat, s a másik program képes lehet jóval kevesebb hibával visszaadni ugyanannak a begyűjtött webhelynek a tartalmát.

Az ellenőrzés főbb szempontjai az alábbiak:

Megjelenik-e rendesen az OWB fejléce?

Ha az OWB fejléce nem látszik, vagy csupán részben látszik (mert a weboldal egésze vagy annak egyes elemei eltakarják), akkor ezt jegyezzük fel problémaként, mivel így sem azt nem látja a felhasználó, hogy mikori mentést néz éppen, sem pedig az időgép funkcióval nem tud lépegetni az egyes mentések között. Ha maga az OWB fejléc takarja el vagy torzítja el az oldal valamely fontos részét (pl. a főmenüt), az is hiba, de csak kisebbfajta, mert a fejléc kikapcsolásával rendszerint orvosolható. (Bár olyan is előfordul néha, hogy nem lehet kikapcsolni a fejlécet, illetve, ha minden aloldalon egyenként kell kikapcsolgatni azért, hogy navigálni lehessen a webhelyen, akkor az azért elég zavaró.)

Megmaradt-e a külalak vagy legalább nem torzult-e el nagyon?

Ha nincs vagy szétesett a dizájn, akkor az eredeti weboldal forráskódjában (Ctrl+U) rá kell keresni a “.css” betűcsoportra és megnézni, hogy milyen

alkönyvtárból töltődik be a stílusfájl, majd előbb a robots.txt-ben ellenőrizni, hogy nincs-e ez kitiltva, és ha nincs, akkor pedig az OWB-ben rákérteni a .css fájlra és/vagy annak alkönyvtárára, hogy vajon a Heritrix letöltötte-e egyáltalán?

Működik-e a főmenü?

A felső és/vagy az oldalsó menüben néhány pontra rá kell kattintani, különösen azokra, amelyek almenüket nyitnak meg vagy le, mert ezeknél a Javascript vagy Flash gyakran nem működik.

Lehet-e navigálni a webhelyen?

Meg kell nézni egy-két olyan aloldalt, amelyeknél lapozás van (pl. hírek, galéria), hogy működnek-e a lapozógombok, és ha igen, akkor az utolsóra kattintva is megjelenik-e még tartalom, vagy ilyen mélységig már nem követte a linkeket a Heritrix. Ha van címkefelhő, akkor abban is érdemes egy-két kiemelt szóra rákattintani. Ugyancsak rá kell kattintani (ha van) az idegen nyelvű és az akadálymentes felületre, ill. hibaként feljegyezni, ha esetleg magától téved át ezekre a magyar kezdőlapról az OWB.

Megjelenik/letölthető-e a nem szöveges (nem HTML) tartalom?

Az már rendszerint a kezdőoldalon is látszik, hogy a képek le lettek-e mentve (nincs-e kizárva az alkönyvtárunk a robots.txt-ben), de keressünk legalább egy aloldalt is, ahol feltehetően vannak képek (pl. hírek, rendezvények, galéria), és nézzük meg, hogy a kis képekre kattintva nagyban is megjelennek-e. Ugyancsak ellenőrizzük a PDF fájlok letölthetőségét legalább egy helyen (ilyenek pl. a beszámolók,

publikációk, kiadványok, pályázatok stb. aloldalakon található), mert van, amikor ezeket is elzárják a robotok elől, vagy csak túl mélyen vannak.

A részletes ellenőrzés lényegében a minimális és a gyors munkafázisainak alaposabb elvégzését jelenti, több időt (mintegy 15-30 percet) szánva ezekre. Ilyenkor meg kell próbálni a webhely minden lényeges alegységét és elemét megnézni, a kevésbé súlyos vagy zavaró hibákat/hiányokat is feljegyezni, és lehetőleg az okaikat is kideríteni, hogy később esetleg kiküszöbölhetők legyenek. Mivel egy webhely (még ha méretben nagy is) csak néhány technikai megoldást használ a dizájnhoz, a képmegjelenítéshez, a navigációhoz stb., így, ha ezek működését egy-két helyen leellenőriztük, akkor a webhely legnagyobb részét már fölösleges végignézni. A részletes ellenőrzésnek inkább a specialitásokra kell kiterjednie, azokra az egyedi megoldásokra, amik lehet, hogy csak egy aloldalon fordulnak elő. Ha van honlaptérkép a webhelyen, akkor az sokat tud segíteni az általános szerkezetből kilógó, illetve a legmélyebben levő aloldalak beazonosításában (utóbbiak azért érdekesek, mert oda már lehet, hogy nem ment le a robot). Ilyen speciális oldalelem, weboldal vagy honlaprész lehet például:

- bármilyen felhasználói interaktivitást igénylő megoldás (pl. csak legörgetéskor, vagy csak kereséskor betöltődő tartalom)
- eseménynaptár
- címkefelhő
- animált/képváltós banner/menü
- diavetítő
- játék
- beágyazott médialejátzó
- beágyazott widget (pl. GoogleMaps, Facebook, blogroll)

- bármilyen egyéb Flash/Silverlight/Java/Javascript/AJAX alkalmazás
- reklámcsík/banner/fejléc (ami pl. eltakarja az OWB saját fejlécét)
- híroldal
- képgaléria vagy virtuális kiállítás
- digitális/digitalizált dokumentumgyűjtemény
- időszaki kiadvány (hírlevél, folyóirat)
- adatbázis
- cookie-t/session ID-t/tokent használó webhely/aloldal/funkció
- reszponzív felület (ami az OWB-ben nézve már nem a képernyő aktuális méretéhez és a böngészőhöz igazodik)
- akadálymentes változat (amiből pl. kiugrik vagy amibe átugrik az OWB)
- a honlaphoz tartozó, de önálló aldomén alatt levő és onnan belinkelt vagy beágyazott saját tartalom
- a mentett verzióba az élő webről “beszivárgó” tartalom (pl. aktuális időjárás vagy névnap)
- felhasználói azonosítást (pl. captcha ellenőrzést, jelszót, Google/Facebook fiókkal való belépést) igénylő webhelyrész

Összefoglaló gondolatok a minőség-ellenőrzésről

A fejezetből kiderül, hogy rengeteg kihívással lehet küzdeni már eleve annak kapcsán, hogy be lehessen gyűjteni a lehető legteljesebb módon az adott intézményi gyűjtőkörbe tartozó webes tartalmakat. Emellett pedig külön kihívásokat rejt az archivált tartalom visszanézése a különféle célszoftverek segítségével, illetve az archivált tartalmak minőségének ellenőrzése is számos tanulságot rejt magában.

3.8 Raktározás, a hosszútávú megőrzés kihívásai

Számos elméleti, illetve gyakorlati kihívással lehet szembesülni a webarchívumok raktározási logikájának kialakítása kapcsán. Ezek áttekintése is lényeges részét képezte a webarchiválással, illetve az OSZK webarchívum kialakításával kapcsolatos kutatómunkámnak, a témakör a kutatási hipotézisek megfogalmazásakor is felvetődött már.

A webarchívumban tárolt anyagoknak elmosódott a tárolási struktúrája. Nincsenek egyedileg jól lehatárolható tartalmi egységek. Ezeket csak akkor lehet meghatározni, ha deduplikáció nélküli egyedi webhelyeket mentünk. Az általános gyakorlatban azonban tömeges aratásokat végzünk, s a tárolt anyag mennyisége miatt, a deduplikáció használata is nélkülözhetetlen. Kisebb gyűjteményeknél persze alkalmazni lehet egyedi webhelyeken alapuló tömörített warc fájlokat, melyekhez hozzá lehet rendelni az adott tartalmi egységhez tartozó információkat, metaadatokat is. Ez azonban csak olyan kisebb gyűjteményeknél képzelhető el, ahol szelektív, tematikus vagy eseményalapú webarchiválást folytatnak. A leírt webhelyek adatait pedig egyedi fájlokban tárolva más digitális forrásokhoz (könyvek, folyóiratok) hasonlóan tudják feldolgozni a saját intézményi digitális raktári rendszerben.

Összességében azonban meg kell állapítanom, hogy a begyűjtés módszerei és a tárolás módja a legtöbb gyűjtemény esetében nem teszi lehetővé, hogy a webarchívumot beillesztjük egy hagyományos digitális tároló rendszer keretei közé, mely szilárd, jól körülhatárolható tartalmi egységekre épül. Vannak már kísérletek arra, például Új-Zélandon, hogy a könyvtár által a webarchiválás adminisztrálására használt rendszert

(WCT) összekötik a nemzeti könyvtár Rosetta repozitóriumi rendszerével, kialakítva azt a nem hagyományos raktározási környezetet, melyet a webarchiválás igényel. Dániában, Franciaországban, Spanyolországban pedig a Net Archive Suite keretrendszer köti össze a nemzeti könyvtár integrált könyvtári rendszerével. Finnországban és Szlovéniában is történtek kísérletek a hosszú távú digitális megőrzés kereteinek kialakítására (Decman, 2011; Veikkolainen, Lager, Lehtonen, & Kaunonen, 2016). A hosszú távú megőrzés repozitóriumi kereteinek meghatározása azonban a legtöbb országban, így hazánkban is meglehetősen sok nyitott kérdést vet fel. A közeljövőben az Országos Könyvtári Platform kereteiben megvalósuló integrált rendszer raktári moduljának funkciói között kell specifikálni a webarchívum anyagainak hosszú távú tárolását.

Az integrált könyvtári rendszerhez, illetve intézményi repozitóriumokhoz történő illeszkedés mellett egyéb kihívásokkal is találkozhattam. A leírt WARC fájlokat hiába tárolják el egy raktári rendszerben, ha nincsenek kialakítva munkafolyamatok arra, hogy ezeket hosszú távon kezelni legyenek képesek. Az informatikai hardver-szoftver környezet elavulása miatt konverziós szükségletek léphetnek majd fel, emulációs környezetekhez kell majd illeszteni a tárolt állományt. Teljesen megoldatlan az, hogy miként lehetséges az állomány integritását szavatolni, kiszűrni az esetleges sérüléseket, elkerülni az ebből adódó adatvesztés kockázatát. Általánosságban elmondható, hogy a múlt webes emlékeit a közgyűjtemények globális szinten jelenleg elsősorban a jelennek gyűjtik, s alakítják ki a rövid távú megőrzési kereteket. Ez természetes is, hiszen a szolgáltatások elindításának ez kellett, hogy

legyen az első fázisa. Meglátásom szerint azonban a begyűjtött anyag növekedésével, s az idő múlásával, az informatika fejlődésével összhangban azonban kiemelt figyelmet kell érdemelnie most már a hosszútávú megőrzés keretei kialakításának is.

3.9 Jogi szabályozási szükségletek és kísérletek

A webarchiválás jogi háttere kapcsán két nagy témakört szükséges érinteni. Először a webarchiválási tevékenységet érintő gyakorlati kihívásokat tekintem át szerzői jogi szempontból. Ezt követi a szabályozási keretek megteremtésének különféle formáinak, illetve a fokozatosan formát öltött és végül hatályossá vált magyar szabályozási elképzeléseknek az ismertetése.

3.9.1 A webarchiválás szerzői jogi kihívásai, jogkezelési eljárások

A technikai problémák mellett a jogi korlátozások is számottevően nehezítik a webarchiválási tevékenység ellátását. Mindenképpen szükséges ezeket a tényezőket röviden önállóan is számba venni. A webhelyek szerzői jogi szempontból védett és nem védett alkotásokat is tartalmaznak. Gyűjteményes műnek minősülnek jellegük szerint így felhasználás tekintetében, a tulajdonjog mellett, ez a szabályozási elem az iránymutató. Bár egy szerzői jogilag védett alkotásról a másolatkészítés is engedélyköteles, az általános európai (s így a magyar) szabályozási gyakorlat szerint a közgyűjtemények szabad felhasználás keretében, nem haszonszerzés céljából végezhetnek többszörözést a védett művekről is, ami a webarchiválásra is vonatkozik. Nagyon fontos megjegyezni azonban, hogy a hazai szabályozási keretek a digitális dokumentumokat érintően (ideértve a webarchiválás révén begyűjtött anyagot is) csupán a

megőrzésre és helyben vagy zárt hálózaton történő szolgáltatásra adnak lehetőséget, nyilvános közzétételre nem! Csupán a jogtulajdonos engedélyével tehető közzé nyilvánosan a begyűjtött anyag. Ráadásul a szerzői jogi törvény szerint az engedélyezés csak kétoldalú szerződéssel jöhet létre. A webhelyek szerzői jogi szempontból gyűjteményes műnek minősülnek, ezért nem kell külön-külön szerződni minden szerzői jogtulajdonossal. Tulajdonjogi szempontból viszont lehet több tulajdonos, akikkel külön kell megállapodni.

A kísérleti projekt keretében a demonstrációs célból létrehozott gyűjtemény esetében merült fel elsődlegesen a nyilvános szolgáltatás igénye. Ehhez illeszkedve tehát egy speciális jogkezelési eljárást kellett kialakítani. Később ehhez a gyűjteményhez csatlakoznak tematikus digitális szolgáltatások, például a Rákóczi gyűjtemény. Előbbi esetben a teljes webhelyekre kérünk az OSZK részéről szolgáltatási engedélyt. A tematikus szolgáltatások kapcsán azonban igényelhetünk csupán egy-egy weblapra is, mivel főleg cikkek kerültek be a gyűjteménybe. A kísérleti időszak lejártával, s az állandó szolgáltatási keretek kialakításával összhangban a korábbi kísérleti időszakra szóló szerződések kezdeményezése helyett a határozatlan idejű szerződések megkötésére tértünk át. A szabad felhasználás minősítéssel bíró webhelyek esetében, amennyiben ezt a vonatkozó jogi licenz előírásai lehetővé teszik, nem kérünk felhasználási engedélyt.

További kihívás, hogy tisztázni kell a szerződés megkötésének céljából a jogtulajdonos személyét, s kapcsolatot kell létesíteni vele. Ahogyan a hagyományos kiadványok esetében is problémás a jogtulajdonos

személyének felderítése és a kapcsolat felvétele, nincs ez másként a webhelyek esetében sem. Egy hagyományos online publikálási rendszerben megjelenő webhely esetében még jó eséllyel találni lehet impresszumot, illetve legalább egy kapcsolati e-mail címet, netán a jogvédelemre vagy CC-licencre utaló kitételt. Blogok vagy közösségi média felületek esetében ezek már gyakran hiányoznak. Szerencsés esetben van lehetőség üzenetküldésre valamilyen formában, rosszabb esetben csak egy bejegyzéshez kötődő hozzászólás megtételére van lehetőség, ami nem a legmegfelelőbb módja a hivatalos kapcsolatfelvétel kezdeményezésének... Kerülő utakon lehet információhoz jutni, amennyiben például egy Facebook-fiók nevéből vagy adatlapjáról kiderül a tulajdonos kiléte, és róla alternatív forrásból lehetséges találni kapcsolati adatot.

Sokszor azonban a kapcsolatfelvétel során kiderül, hogy nem az adott személy a tartalom tulajdonosa... Az internet technológiája a különböző tartalmak összekapcsolására épül, aminek alapvetően háromféle formája lehet: hivatkozás, beágyazás és tényleges átmásolás. Bár tartalmilag és archiválás szempontjából integráns tartalmi egységet alkothat az adott webhely, a felhasználó számára sem okoz problémát, hogy ezek a tartalmak nem különülnek el élesen, mivel egy felületen látszanak, jogi szempontból azonban ezek külön tartalmak is lehetnek. Ebben az esetben minden egyes jogtulajdonos kilétét ki kell deríteni és velük is fel kell venni kapcsolatot a nyilvános szolgáltatáshoz. Paradox módon előfordulhat olyan eset is, hogy akár egy archivált tartalom valamely meghatározott részére szól a szolgáltatási engedély, míg bizonyos részére (pl. beágyazott

videóra) nem. Ilyenkor csak az engedélyezett rész látszódhat nyilvános tartalomként, aminek technikai előfeltétele, hogy a résztartalmak külön kerüljenek archiválásra. Ez kihat a webarchiválás munkafolyamatainak sorrendiségére is. Ideális esetben a jogosultság tisztázása meg kellene, hogy előzze az archiválás műveletét. Számos elvi kérdés is felvetődhet ennek kapcsán. Mennyiben tekinthető az eredeti tartalom kompilált másolatának? Mi történik akkor, ha a kiegészítő tartalomra van felhasználási engedély, de arra a tartalmi környezetre nincs, amelybe be volt ágyazva, s tartalmilag, illetve megjelenésében elválaszthatatlan attól?

Engedélyeztetés céljából partnereinkkel elektronikus úton próbáljuk felvenni a kapcsolatot, egy részletes tájékoztató levélben röviden írunk a webarchiválásról, annak technikai hátteréről, a szükséges jogi tudnivalókról, valamint a nyilvános szolgáltatáshoz szükséges szerződést is mellékeljük (Visky & Drótos, 2020). Sajnos leveleink többsége válasz nélkül marad. Mintegy egyharmadára kapunk választ, ez nem kirívó, más webarchívumok munkatársai is ilyen arányról számoltak be. Arról viszont nem rendelkezünk információval, hogy a válasz elutasítás, feledékenység, esetleg céltévesztés miatt marad el. Bizonyos idő után szoktunk emlékeztető levelet küldeni, az ezekre adott reakciók azért sejtetik, hogy hangsúlyos a feledékenység a válaszadás elmaradásában... A megkeresésre adott válaszok túlnyomó többsége pozitív, kategorikus elutasítás ritkán fordul elő elvi okból, inkább jellemző az, hogy jogi tisztázatlanság miatt nem kapunk felhasználási engedélyt. Sok pozitív visszajelzést is kapunk, és többen ajánlják archiválásra is webhelyeiket.

A legtöbb probléma abból adódik, hogy a szerződést csak papír alapon lehet megkötni, eredeti példányokkal, mert a szerzői jogi törvény csak ezt a formát ismeri a felhasználási engedély kapcsán. Ezt sokat kifogásolják, életszerűtlennek tartják a mai világban, és olyanra is volt példa, hogy bár elektronikus úton elküldték nekünk a kitöltött szerződést, papíron már nem, ami így természetesen nem érvényes. A Rákóczi archívum esetében az is problémát okoz, hogy mivel az egész emlékévk alatt bővítjük a gyűjteményt, ezért folyamatosan kell engedélyeket kiküldeni az új tartalmakra, így például egy híroldalnál könnyen előfordul, hogy több szerződést is kell kötnünk. Igyekszünk feleslegesen nem terhelni a partnereinket, ezért nem egyesével kérünk engedélyt, hanem időszakosan összegyűjtve több cikkre egyszerre. Sajnos így is előfordult, hogy míg az első engedélykérésünkre kimondottan pozitív reagálás érkezett, az újabb megkeresésre már elmaradt a válasz.

A jogkezelés legszűkebb keresztmetszete megítélésem szerint egyértelműen a rendelkezésre álló erőforrások szűkösége. A nyilvános szolgáltatás kezdete óta nem sikerült minden szolgáltatni kívánt webhelynek elküldeni a szerződést, sem a kísérleti gyűjtemény, sem a Rákóczi archívum kapcsán. Míg az archiválási folyamatok nagy része automatizálható, a szerződéskötések adminisztrálása, a szükséges levelezés intézése aránytalanul sok időt igényel az egyéb feladatok mellett. Ez az oka annak, hogy a Rákóczi gyűjteményben az archivált weboldalak közel fele (még) nem szabadon hozzáférhető. A böngészhető listákban látszik, hogy milyen státuszban van a webes tartalom engedélyeztetési folyamata: zöld lakat jelezni a szabad hozzáférést, sárga

azt, ha az még nem zárult le, és piros jelezné az elutasítást, eddig ilyen szerencsére még nincs.

Összegzésül elmondhatom, hogy a szerzői jogi szabályozási környezet a webarchívumokban tárolt anyagok túlnyomó többségének nyilvános szolgáltatását nem teszi lehetővé. A könyvtár fizikai falain kívüli hozzáférés hiányát a jövőben esetleg dedikált terminálokra épülő országos hálózat révén lehetne áthidalni, mint ahogyan az a NAVA pontok esetében is történt. A jogkezelés bonyolult, elavult adminisztratív szabályozása a felhasználási engedélyek beszerzésének vonatkozásában, melyhez nem kapcsolódik elegendő humán erőforrás sem, súlyos akadályokat gördít az archivált anyagok nyilvános szolgáltatásának útjába. A megoldás valószínűleg európai uniós keretek között lenne elképzelhető. Legalább a nyilvánosan korlátozások nélküli elérhetőséggel begyűjtött webes anyagokat a közjó (public domain) keretei közé kellene utalni. A legutóbbi uniós szerzői jogi reform szabályozásai munkálatai során ez a témakör sajnálatosan szóba sem került...

A továbbiakban arra térek rá, hogy a webarchiválási tevékenységet milyen törvényi keretek között tervezték Magyarországon szabályozni, s milyen nemzetközi példákból lehet ennek vonatkozásában kiindulni.

3.9.2 A webarchiválási tevékenység szabályozásának törvényi, jogszabályi háttere

A webarchiválás törvényi szabályozása alapvetően nemzeti jellegű s igen különböző jogi szabályozási mintázatokat mutat. Van, ahol semmilyen konkrét jogi előírás sincs, annak ellenére, hogy már évek óta működik nemzeti szintű webarchívum, így csak a tisztességes felhasználás és az

archiválás jogtulajdonos részéről való megtagadhatósága elveinek mentén, vagy az elektronikus publikációkra vonatkozó kötelességtörvény szabályokat a webre is kiterjesztve végzik ezt a tevékenységet. Az egyik legkülönlegesebb példa Hollandiáé, ahol kötelességtörvény sincs, jelenleg is csupán csak azokat a weboldalt tudják archiválni, melyekre szerződést kötnek a tartalom tulajdonosával. Az egyes országokban a kötelességtörvény-, a könyvtári-, a média- vagy egyéb szaktörvény tartalmaz valamilyen részletességű szabályozást kifejezetten a webes vagy általában az internetes tartalmak megőrzésére vonatkozóan. A jelenlegi gyakorlatokról az IIPC honlapján található összeállítás igazít el a legjobban (IIPC, 2020a).

Mivel Magyarországon évek óta napirenden volt a kötelességtörvény rendelet módosítása, ezért kézenfekvőnek tűnt, hogy ezt kellene kiegészíteni a tartalomszolgáltatók és webarchívumot működtető nemzeti könyvtár jogait és kötelességeit definiáló részekkel. Ennek elősegítése érdekében 2017 májusában összefoglalókat készítettem néhány nemzeti webarchívum (dán, cseh, észt, brit, ausztrál) gyűjtőköreiről, jogi és szervezeti kereteiről (Németh, 2017), majd az észt kötelességtörvény angol nyelvű fordításából kiemeltem és kommentáltam azokat a részeket, amelyeket a magyar jogalkotók számára is megfontolandónak ítéltam. 2017 nyár végén pedig egy, az EMMI-be eljuttatott dokumentumban a projektünk szintjén összefoglaltuk a szerintünk legfontosabb szempontokat, melyeknek szerepelniük kellene a szabályozásban.

A Közgazdasági Digitalizációs Stratégia megvalósításáról 2018 márciusában megjelent kormányhatározatban előírt jogszabály-

módosítások előkészítése során eldőlni látszott, hogy a webarchiválás a kötelespéldány-szolgáltatáson túlmutató tevékenység, ezért önálló törvényben kerül majd szabályozásra, melynek előkészítése a nyár folyamán elég intenzíven meg is indult a minisztériumban. Mi is készítettünk az OSZK részéről ehhez egy koncepciótervet 2018 augusztusában, továbbá véleményeztük az EMMI-től kapott szövegváltozatokat és fogalommagyarázatokkal is segítettük a törvénytervezet előkészítését. A társadalmi vita előtt álló törvénytörvény végül ezekre a főbb területekre terjedt ki: a webarchiválás elhatárolása a kötelespéldány szolgáltatási kötelezettségtől, a webarchiválás módszerei, a webarchiválás ütemezése, a tartalomszolgáltatók bejelentési kötelezettsége, a webarchívumhoz való hozzáférés szintjei, az adatvédelem, a webarchiválás költségei. Belekerült még a szövegbe, hogy a webarchiválási tevékenység egyes részleteit egy miniszteri rendelet fogja szabályozni, és eldőlt az is, hogy a webarchiválás a szerzői jogi törvény módosítását nem teszi szükségessé. A jogalkotó 2019 folyamán arra az elhatározásra jutott, hogy nem önálló törvényben, hanem a kulturális törvény módosításában, illetve vonatkozó kormányrendeletben, ahhoz kötődve pedig miniszteri szinten kiadott végrehajtási utasításban fogja megadni a webarchiválás jogi háttérét. 2020. május 19-én megszavazta az országgyűlés a kulturális törvényt módosító javaslatokat, ezek között szerepel az is, amely a nemzeti könyvtár feladatává teszi a webtartalom megőrzését (2020. évi XXXII. törvény, 2020). Az előírásoknak 2021. január 1-től kell megfelelni, a törvénymódosítás onnantól válik hatályossá. A webarchiválás részletesebb szabályozásáról külön kormányrendelet készült, amely szintén 2021. január elsejétől lépett

életbe. A 1358/2020. (VII. 1.) sz. kormány határozat rendelkezik az OSZK webarchiválási feladataihoz rendelt költségvetéséről a 2021-es évre, valamint az azt követő időszakra (1358/2020. (VII. 1.) Korm. Határozat, 2020). A kiadványok kötelespéldányainak szolgáltatásáról, megőrzéséről és használatáról szóló, az online elektronikus kiadványokra is kiterjedő 195/2019. (VIII. 1.) kormányrendelet is tartalmaz a webarchiválást érintő rendelkezéseket (A Kormány 195/2019. (VIII. 1.) Korm. Rendelete, 2020). Ennek hatálybalépése többször halasztást szenvedett, a 2021. év elején került erre is sor.

Mint már utaltunk rá, 2019-ben elkészült annak a szerződésmintának a szövege is, melynek alapján azokkal a tartalomtulajdonosokkal szerződést kötünk, akik hozzájárulnak archivált webhelyeik nyilvános szolgáltatásához. Az Országos Széchényi Könyvtár, illetve az intézményt fenntartó EMMI is mindent megtett azért, hogy 2021-től valóban új fejezet kezdődjön a magyarországi webarchiválásban. Világos jogi felhatalmazással, állandó nemzeti könyvtári szolgáltatásként, a megfelelő forrásokat is biztosítva folytatódhasson tovább a munka.

3.10 Az OSZK webarchiválási tevékenységének kommunikációs keretei

A webarchiválás kommunikációs kereteinek kialakításakor központi szerep jutott eleinte a kísérleti projekt honlapjának. Már a kezdetekkor támaszkodhattunk a Drótos László által létrehozott wiki szolgáltatásra, melynek frissítését 2020-ban én vettem át, emellé kezdtem fejleszteni a webarchiválás nemzetközi bibliográfiáját, illetve létrehoztuk a munkatársaimmal együtt a kísérleti projekt honlapját, ahol

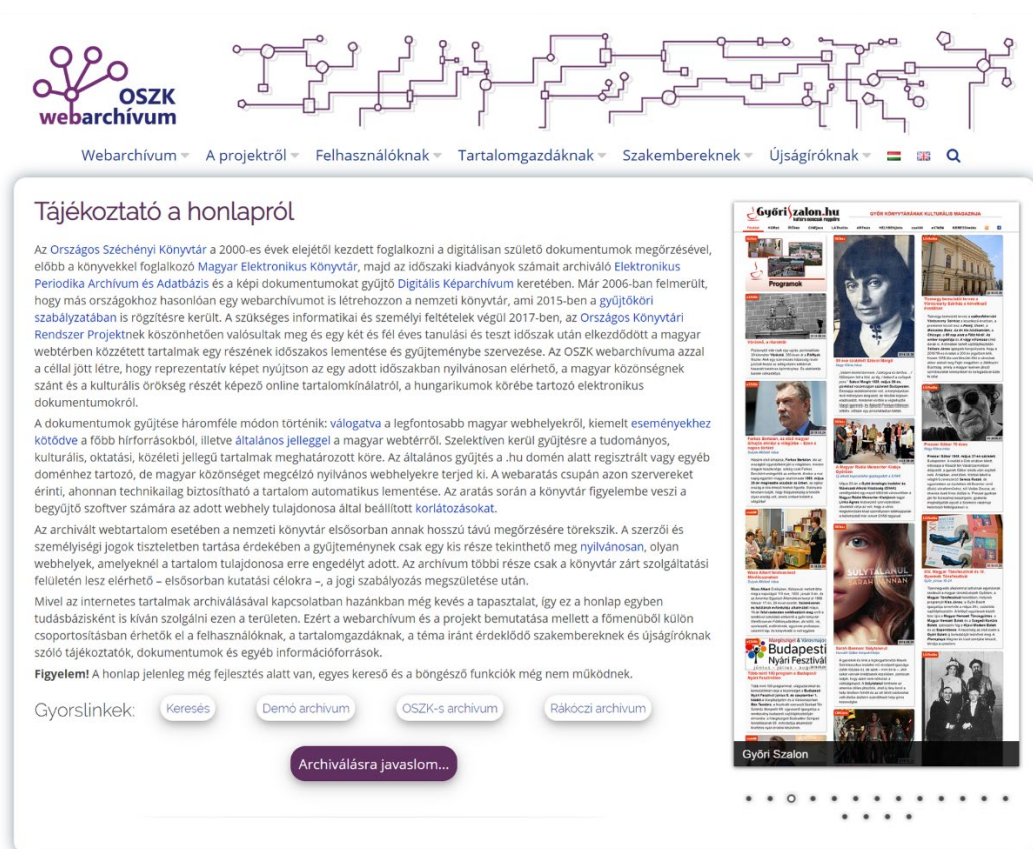
tevékenységeink előrehaladását is nyomon lehetett követni (Németh, 2020a). A projekt előrehaladásával lassanként, az évek folyamán megteremtődtek a webarchiválás mint állandó könyvtári tevékenység körvonalai s annak beillesztésének feltételei az OSZK szakmai kínálatába. A kísérleti projekt keretein túllépve egy olyan új, már a leendő szolgáltatási kínálat egyes elemeit számba vevő honlapra lett szükségünk, mely keretbe foglalja a különféle célcsoportok számára általunk végzett tevékenységeket. Természetesen mindezt oly módon terveztük meg, hogy a kísérleti projekt keretében szolgáltatott információk se vesszenek fel, hanem új szakmai környezetben kiteljesedve találjanak teljesebb otthonra. Ez a rövid áttekintés számot ad az új honlap koncepcionális felépítéséről, s arról, hogy különféle célcsoportoknak milyenfajta szolgáltatásokat tudunk kínálni (Németh, 2020a).

3.10.1 A honlap felépítése és alapvető szolgáltatásai

Fontos szempont volt az új honlap elkészítésekor, hogy az illeszkedni tudjon majd a még formálódás alatt lévő központi nemzeti könyvtári tartalomszolgáltatási keretrendszerhez. Éppen ezért az Információtechnológiai Igazgatóság illetékeseivel konzultálva döntöttünk a Wordpress tartalomkezelő rendszer használata mellett. Az OTR is erre a platformra épül, emiatt remélhetőleg minimális nehézségek árán lesz majd integrálható a szolgáltatás. A honlap kialakítása számunkra is egy tanulási folyamat volt, hiszen a Wordpress-szel azelőtt főként, mint blogszolgáltató platformmal találkoztunk. Ez alkalommal pedig az ezen a szerepkörön túlnyúló általános tartalomkezelő funkciók kezelésének

elsajátítására volt szükségünk. Fontos szempont volt az is, hogy a tartalmak jó része magyar és angol nyelven is elérhető legyen.

Az új honlap koncepciójának megtervezésekor magunk elé képzeltük egyrészt az általunk kínálni kívánt szolgáltatásokat, másrészt pedig azokat a fő célcsoportokat, melyek azoknak elsődleges felhasználói lehetnek. Ilyenek lehetnek a webarchívum felhasználói, a webes tartalmak gazdái, a közgyűjteményi szakemberek, illetve az újságírók, ennek figyelembevételével alkottuk meg az új honlap szerkezetét (Németh, 2020a).



19. ábra Az OSZK webarchívum honlapja

A kezdőlapon az archívumunkról szóló alapvető információk mellett a jobb oldali hasámban válogatott oldalképek diavetítésével ajánlunk néhány nyilvánosan szolgáltatott webhelyet a felhasználók figyelmébe. Az ismertető alatt közvetlen linket helyeztünk el a honlapajánló úrlapra, mely a gyarapítási javaslatok rögzítésére szolgál. Gyorslinkekkel ugyancsak elérhető közvetlenül a keresőfelület, a demó archívum, a Rákóczi gyűjtemény, valamint az OSZK honlapok gyűjteménye is. A kezdőlap alján áttekinthetők legfrissebb híreink, elérhető közvetlenül a hírarchívum is. Az elérhetőségeinek közzététele mellett itt szolgáltatjuk a honlapra vonatkozó jogi információkat, közvetlen hivatkozás mutat anyaintézményünk, az OSZK honlapjára, valamint projektünk régi honlapjára is. A saját honlapunk áttekinthetősége és archiválhatóságának szempontjából fontos funkció a honlaptérkép, a felhasználók tájékozódását pedig a honlap anyagához illesztett szabadszöveges kereső is segíti, mely szintén közvetlen hivatkozással érhető el a kezdőlapról (Németh, 2020a).

A Webarchívum menüpontban felvázoljuk a gyűjteményszervezés kereteit. Emellett böngészési és keresési lehetőséget igyekszünk biztosítani sokféle szempont szerint az archívumban őrzött nyilvánosan, illetve nyilvánosan nem szolgáltatható anyagokhoz is (utóbbi esetben szerzői jogi okok miatt a mentett archivált webhelyek csak az OSZK épületén belül dedikált terminálon lesznek elérhetőek, a NAVA pontok mintájára, ennek szolgáltatási koncepciója is fejlesztés alatt áll még). Emellett statisztikai adatokat is kínálunk az archívum összetételéről, illetve annak felhasználásáról.

Ezt követi a webarchiválási projekttel kapcsolatos információkat összefoglaló menüpont. Itt kaptak helyet aktuális híreink, a szervezeti keretek bemutatása, publikációs és ismeretterjesztő tevékenységeink részletezése, együttműködő partnereink bemutatása, valamint innen lehet elérni a kísérleti projekt már nem frissülő régi honlapját is.

A harmadik menüpont az általános felhasználási információkat összegzi. Tájékoztatót nyújtunk a honlapról, a gyakran ismétlődő kérdések is itt kaptak helyet, valamint innen lehet gyarapítási javaslattal is élni a webarchívum számára.

A következő menüpontban a webes tartalomgazdákat igyekszünk megszólítani. Innen érhető el az a keretszerződés szövege, mely szükséges az archívum nyilvános szolgáltatásához. Itt helyeztük el azokat a pecséteteket, melyeket a nyilvános szolgáltatásba bekerült partnereink tehetnek ki a honlapjukra., valamint technikai információkkal és ajánlásokkal is szolgálunk a webhelytulajdonosok számára, továbbá elérhető a vonatkozó jogszabályok listája is.

A szakemberek menüpont a közgyűjteményi szakmai közeget szólítja meg elsősorban. Itt adunk információt az általunk rendezett korábbi eseményekről, a szakmai együttműködés lehetőségeiről, a különféle szakmai kommunikációs csatornákról és oktatási tevékenységeinkről. Innen érhető el a nemzetközi szakbibliográfiánk, illetve a wiki alapú tudásbázis szolgáltatásunk is. Webhelyeket is ajánlunk, valamint igyekszünk becsatornázni a webarchiválással foglalkozó legfrissebb híreket a Twitterről is.

Az újságíróknak menüpont a külső kommunikáció során szükséges alapvető tényeket, információkat foglalja össze tevékenységeinkről. Szabadon felhasználható képeket, szórólapokat, plakátokat kínál, valamint áttekintést nyújt médiamegjelenéseinkről is.

Az alábbiakban az imént felvázolt szerkezethez illeszkedve nyújtok rövid áttekintést az egyes szolgáltatási elemekről.

3.10.2 A webarchívum mint gyűjtemény

A webes tartalmak megőrzése a már korábban felvázoltak szerint több szinten és többféle munkafolyamattal történik, és ahogy korábban már említettem, az archívum hozzáférés szempontjából is kétfelé oszlik: vannak nyilvánosan is szolgáltatott részei és van az egyelőre “sötét archívum”, amely majd a könyvtári hálózathoz lesz elérhető, elsősorban kutatóknak. Ezt a kettősséget tükrözi az új honlap első, Webarchívum nevű menüpontja, mely alatt a részgyűjtemények ismertetőinél, és a böngészési, illetve keresési funkcióknál is szét lett választva a nyilvános és a zárt gyűjteményrész (Németh, 2020a).

Előbbibe jelenleg három szolgáltatás tartozik: A Rákóczi-emlékév alkalmából készített, a webarchívum és másik három digitális gyűjtemény anyagából válogatott önálló honlap, saját böngésző- és keresőfelülettel. Részletesebb leírásával önálló tanulmány foglalkozik Visky Ákos László és Drótos László révén (Visky & Drótos, 2020). A másik kettővel pedig elsősorban a webarchiválás technológiájának lehetőségeit és korlátait szeretnénk demonstrálni. Ezekben az OSZK saját online szolgáltatásainak, illetve más intézmények vagy magánszemélyek honlapjainak, blogjainak, időszaki kiadványainak különféle szoftverekkel készült és megjelenített

mentéseit lehet böngészni. Mivel az általunk mentett példányok mellett az archiválás idején készült oldalképek, továbbá az amerikai Internet Archive mentései, valamint természetesen az eredeti webhelyek is megnézhetők, ezért össze lehet vetni, hogy milyen különbségek vannak közöttük (Németh, 2020a).

A zárt archívumban levő, témák vagy műfajok szerint összeválogatott, több mint 27 ezer magyar webhelyhez viszont nincsenek ilyen részletes leírások, ezekről csak a részgyűjtemények szintjén rögzítünk adatokat. Ezért a honlapon mindössze az egyes webhelyek nevének vagy URL címének részletére lehet rákeresni, és a találati listában csak az látszik, hogy melyik részgyűjteményben tartjuk őket nyilván és hogy azt mikor arattuk utoljára. Bár a mentett verziókat szolgáltatási engedély hiányában nem lehet nyilvánosan szolgáltatni, de legalább kis méretben megnézhetők a kezdőlapokról készült oldalképek, valamint itt is belinkeltük az Internet Archive nyilvános mentéseit és természetesen az élő weben levő eredeti szolgáltatást is (Németh, 2020a).

A tematikus és műfaji válogatások mellett eseményekhez kötődő archiválást is végzünk, ezekről is lehet tájékozódni az új honlapon. Mivel ezek a részgyűjtemények többnyire nem webhelyekből, hanem egyedi weboldalakból, hírekből, blogbejegyzésekből és Wikipédia szócikkekből állnak, ezért nem volt célszerű közösen kereshetővé tenni őket a fent említett 27 ezres címlistával, úgyhogy csak táblázatos formában böngészhetők (Németh, 2020a).

A magyar webtér egy jelentősebb részére kiterjedő, pillanatfelvétel-szerű aratásból eddig kettőt csináltunk: 2018 őszén 291 ezer, 2019 végén pedig

közel 247 ezer címről indítottuk el a robotot. Ezeknek csak az összesített adatai nézhetők meg a Webarchívum menü Statisztikai adatok pontja alatt. Ugyancsak itt találhatóak a többi részgyűjtemény legutóbbi aratásairól készített táblázatok és diagramok, valamint a honlapunk látogatottságát mutató grafikon (Németh, 2020a).

Végül egy érdekesség a honlapnak ebből a részéből: a Böngészés almenüben található egy Megszűnt webhelyek oldal, ahol elkezdjük gyűjteni azokat a webhelyeket, amelyekről sikerült viszonylag jó mentést készítenünk, mielőtt azok eltűntek volna az internetről. Ilyen például a Lánchíd Rádió honlapja, a CriticalHit videójátékos magazin, a HAMLET nevű erdélyi színházi portál, vagy a Tolna Megyei Levéltár régi weboldala (Németh, 2020a).

3.10.3 A webarchiválási projekt szakmai és szervezeti keretei

Az imént felvázolt, a gyűjteményre épülő szolgáltatások mellett nagy szükségét éreztük annak is, hogy azokat kereteket is bemutassuk a honlapon, melynek révén a gyűjteményfejlesztési munka zajlik. Voltaképpen arról a szervezeti, illetve szakmai bázisról adunk itt számot, melyre alapozva végezzük tevékenységünket. A kezdőoldal mellett itt kapnak helyet a munkánkkal kapcsolatos legfrissebb hírek, illetve a hírarchívum is, melynek segítségével a kísérleti projekt kezdeteiig vissza lehet követni, hogy miként formálódtak különféle teendőink. Nagyon fontosnak tartjuk egységes keretek között közzétenni a kísérleti projekt kezdeteitől nagy hangsúlyt kapó összes szakmai megjelenésünket (konferencia előadások, prezentációk, hazai és nemzetközi publikációk révén). Az egyes eseményekről, publikációkról természetesen folyamatos

tájékoztatást adunk a hírek rovatban, itt azonban egységes rendszerben áttekinthetővé válik tevékenységeink teljes vertikuma. A szakmai tevékenységünkről való számadás fontos fejezete ezen felül a projekttel kapcsolatos nyilvános dokumentációs források közzététele is. Innen lehet az előadásokban, illetve publikációinkban megfogalmazottaknál mélyebben is tájékozódni az egyes szakmai részterületekről, melyek tevékenységeink középpontjában állnak. Természetesen röviden bemutatjuk saját magunkat, illetve együttműködő partnereinket is (Németh, 2020a).

3.10.4 Általános felhasználói információk

Rövidnek tűnik tartalmilag a menüpont kínálata, ám tartalmilag annál fontosabbak, sőt elengedhetetlenek az itt közreadott információk. Minden európai webarchívum biztosítja felhasználói számára a lehetőséget gyarapítás javaslatok megtételére, s ez a funkció már a kísérleti projekt honlapján s rendelkezésre állt. (6. ábra) Két irányból is el lehet jutni ide, a menürendszerből, illetve közvetlenül a kezdőlapról is. Ez a szolgáltatás adja az általános felhasználók felé szóló kínálat magvát, amellet persze, hogy általános tájékoztatást nyújtunk a honlapról. A gyakran ismétlődő kérdések menüpontban pedig olyan kérdéseket igyekszünk megválaszolni, melyek a gyakorlati munka során sokszor előkerültek a saját tapasztalataink szerint (Németh, 2020a).

3.10.5 Kínálatunk a tartalomgazdáknak

A honlap tervezésekor igen fontosnak tartottuk egy olyan menüpont megtervezését, ahol minden, a tartalomgazdát érintő információ egységes keretek között megtalálható lesz. A velük kötendő minél

szélesebb körű partnerség a webarchiválás sikeressége szempontjából kulcsfontosságú tényező. Egyrészt szeretnénk arra bátorítani minél több, a látókörünkbe került honlap tulajdonosát, hogy szerződjenek velünk az archivált webhelyek nyilvános szolgáltatásának céljából. Ennek segítségével kész szerződésmintát nyújtunk, melybe már csak az illető tartalomtulajdonos, illetve webhely adatait kell beilleszteni. Tájékoztatást nyújtunk továbbá az általunk végzett archiválási és szolgáltatási tevékenység pontos jogi háttéréről is. Másik oldalról a tartalomgazdákkal kötendő szövetség másik sarokpontja, hogy jól archiválható tartalmakhoz férhessünk hozzá. Ennek előmozdítására különféle szakmai ajánlásokat is megfogalmaztunk, melyeket a webhelyek fejlesztésekor igen hasznos lenne figyelembe venni. További pozitív motiváló erőként tervezzük azoknak a pecséteknek az alkalmazását, melyek segítségével egy -egy webhely tulajdonosa, aki fontosnak érzi, az felmutathatja, hogy archiváljuk a webhelyét a jövő számára (Németh, 2020a).

3.10.6 A közgyűjteményi szakmának ajánlott szolgáltatások

A gyűjteményalapú, illetve a tartalomgazdáknak nyújtott szolgáltatások mellett voltaképpen ez a szolgáltatási kínálat képezi az új honlap harmadik sarokpontját. Könnyen kezelhető wiki alapú tudásbázist kínálunk, szinte minden összefüggésben a webarchiválással kapcsolatos fogalmakról, munkafolyamatokról. Ezt egészíti ki az immáron harmadik éve általam épített nemzetközi szakbibliográfia. Ezt az EBSCO, a PROQUEST, illetve az ACM informatikai szakadatbázis által feltárt szócikkek, konferencia előadások alapján állítom össze, de helyet kapnak benne az EPA által nyilvántartott belföldi anyagok is. Az anyagot előbb a Mendeley, majd

most már a Zotero bibliográfiakezelő rendszerben gondozom, onnan végzem Drótos László segítségével, a folyamatos (3-4 havonkénti rendszerességgű) gyarapítást szabványos bibliográfiai bemeneten keresztül. Az egyes tételekről részletes információkat tartalmazó nézetet kínálunk. A teljes anyagról a legfontosabb bibliográfiai adatokat nyújtó áttekintési mód is elérhető, valamint exportálhatjuk azt különféle szabványos bibliográfiai formátumokban saját felhasználásra is. A bibliográfia, illetve a tudásbázis mellett elérhetővé tesszük az általunk évente megrendezett workshopok prezentációját, illetve az események felvételeit a Videotorium rendszerén keresztül. Ezeket a felvételeket is mi töltjük oda fel s látjuk el metaadatokkal. Nagyon fontos ezen felül bemutatni azokat az oktatási-ismeretterjesztő tevékenységeket is, melyeket a webarchiválás kapcsán kifejtünk. Információkat kínálunk a Könyvtári Intézetben akkreditált webarchiválás tanfolyamunkról, illetve e-learning alapú tartalomfejlesztéseinkről is. Kifejezett célunk az, hogy a webarchiválás iránt érdeklődő kollégákkal minél intenzívebb szakmai együttműködést alakítsunk ki. Az OSZK-nak, bár jogilag is rögzített feladata a webarchiválás, külső szakmai partnerek nélkül képtelenség feltárni a begyűjtendő tartalomforrások teljes vertikumát. Külön levelezőlistát hoztunk létre az érdeklődő szakmai partnerek tájékoztatására, illetve a honlapon is felvázoljuk az együttműködési lehetőségek fő irányait. A KDS pályázat számos megyei-városi közkönyvtár felé mostanában nagyon ígéretes lehetőséget nyújt ezen együttműködési formák elmélyítésére. Nagyon reméljük, hogy a közrebocsátott információk katalizátorai lesznek egy széles szakmai együttműködési háló kialakításának. A tájékoztató

szolgáltatások mellett ez a legfontosabb célja az itt felvázoltaknak (Németh, 2020a).

3.10.7 A külső kommunikáció keretei

A honlap tervezésekor nagyon fontosnak tartottuk egy olyan menüpont beillesztését is a szolgáltatási kínálatba, mely elsődlegesen a sajtó érdeklődő munkatársait, illetve a webarchiválást széles körben népszerűsíteni kívánó partnereinket célozza meg. Itt tömören összegezzük az újságírók számára könnyen feldolgozható módon a tevékenységeinkről szóló alapvető információkat. A szöveges tartalom mellett pedig szabadon felhasználható képeket, szórólapokat, plakátokat kínálunk népszerűsítési céllal. Az újságírók munkáját jelentősen segítheti az az összeállítás is, melyben a webarchiválási projektről, illetve tágabban a webarchiválásról szóló nyomtatott és elektronikus sajtómegjelenéseinket tartjuk nyilván, illetve megosztjuk a cikke teljes szövegét is, amennyiben az lehetséges (Németh, 2020a).

3.11 Együttműködések partnerintézményekkel

A webarchiválás feladatainak mérete és összetettsége nyilvánvalóvá teszi, hogy (bár a hatályos jogszabály az egész területet a nemzeti könyvtár feladatkörként határozza meg) nem lehet csupán egyetlen intézmény tevékenységi körében hagyatkozni rá. A nemzetközi példák is arra utalnak, hogy széleskörű együttműködési hálózatot kell kialakítani könyvtárakkal, egyéb közgyűjteményi szereplőkkel, egyetemi tanszékekkel, kutatóintézetekkel, az üzleti szférával (internetszolgáltatók, illetve a tartalomipar szereplői). Néhány esetben spontán módon is megalapozódnak együttműködések, például a Könyvtári Intézetben zajló

akkreditált tanfolyamunkon részt vett vajdasági kollégákat be tudtuk később vonni az ottani magyar vonatkozású címek gyűjtésébe. Az első szervezett együttműködési forma kerete a Közgyűjteményi Digitalizálási Stratégia pályázatán nyertes intézményekkel alakult ki.

A 2019-ben kiírt pályázat egyik alfeladatként határozták meg a webarchívummal történő együttműködést. Többféle feladatot határoztunk eredetileg meg alfeladatként, ám a pályázat késedelmes lebonyolítása miatt reálisan csupán a címek gyűjtésében való részvétel valósulhatott meg. A partnereink megyei és városi hatókörű könyvtárak voltak (a Budapesten és Pest megyében működő intézmények kivételével)(Visky, 2020).

Partnereink mintegy 3400 webhelyet gyűjtöttek össze, illetve ajánlottak az OSZK webarchívumának figyelmébe, ezeket a javaslatokat még egyszer átválogattuk, kiszűrtük az archívumban már megtalálható elemeket. Végeredményül elmondhatjuk, hogy a webarchívum addigi gyűjteménye mintegy 7%-os gyarapodást ért el így. Az egyes intézmények által gyűjtött címek mennyisége nagy szóródást mutatott 100 és 500 között, az átlag 310 volt. Az összegyűjtött címek mintegy 70 százalékát tudtuk új tételként hasznosítani (Visky, 2020).

A különféle címek tartalmi jellemzőinek vizsgálata kapcsán beigazolódott az előzetes hipotézisünk, hogy a helyismeret nagy előnyt jelent a helyi kulturális, közéleti jellegű tartalmak, illetve közintézményi szféra felmérésekor. A lokális jellegű közösségi média (Facebook) tartalmak gyűjtése is lendületet kaphat a megfelelő technikai feltételek megteremtése esetén. Új, eddig jelentősebb címgyűjteménnyel még nem

bíró témakörök is előtérbe kerültek lokális fontosságuk alapján, mint a turizmus-vendéglátás, illetve a sport. Nagy előny számunkra, hogy így partnereinkkel együtt új szempontokkal is tudjuk a gyűjtőkört bővíteni. Igen fontos előny továbbá, hogy a helytörténeti gyűjtés révén számos eddig rejtőzködő periodika, illetve kulturális intézmény is a látóköreinkbe kerülhetett (Visky, 2020).

Eddig három intézmény jelezte, hogy szívesen folytatná az együttműködést a webarchiválás kapcsán. Velük szerződést fogunk kötni és további munkafázisokba is bekapcsolódhatnak majd például a minőség-ellenőrzés és a metaadatolás területén.

3.12 A jövőre vonatkozó tervek

A vonatkozó jogszabályi környezet megszületése után fontos feladatunk lesz, hogy a magyar illetőséggel bíró honlaptulajdonosok számára előírandó bejelentési kötelezettségnek egy megfelelő űrlap segítségével eleget tudjanak majd tenni (Németh, 2020a).

A könyvtárakkal való együttműködés elmélyítése során szükségessé válhat egy kollaboratív munkafelület kialakítása is valamilyen platformon, s ezt is illeszteni kellene majd a honlap keretei közé (Németh, 2020a).

Utaltunk már rá, hogy annak elkészültével ki kell építeni a kapcsolatokat az Országos Tartalomszolgáltatási Rendszer (OTR) felé, illetve be kellene mélyebben integrálni a honlapba a remélhetőleg folyamatosan készülő e-learning tartalomfejlesztéseket is. Ez utóbbi elsősorban technikai kihívást jelent, az e-learning tartalomkezelő szabványokon alapuló tartalmak integrálása a Wordpress keretei közé számos kihívást rejt magában.

Elképzeltető, hogy ha bekeröl a honlap az OTR keretei közé, annak révén erre is könnyebben megoldást lehet majd találni (Németh, 2020a).

A megfelelő finanszírozási feltételek biztosításával megnyílna a lehetőség arra, hogy az Internet Archive által begyűjtött magyar webhelyeket is a honlapunkba integrált keresőfelületen tudjuk elérhetővé tenni. Ez nagyon jelentős hozzáadott értéket jelentene a magyar web múltjának feltárása terén (Németh, 2020a).

A zárt archívum elérhetőségének biztosítása, valamint az Internet Archive által gyűjtött anyag szolgáltatása felértékelheti a kutatástámogató funkciók fejlesztését is. Elképzeltető, hogy külön menüpontban kell majd tájékoztatást adni a kutatási lehetőségekről a tudományos szakemberek számára, a vonatkozó nemzetközi gyakorlat bemutatásával (Németh, 2020a).

Középtávú terveink közé tartozik a közép-európai webarchívumok Memento protokollon alapuló közép- európai közös keresőfelületének megteremtése, egy közös portál keretében, ahol együtt meg tudnánk jelenni, természetesen megfelelő kapcsolódási pontokkal a nemzeti webarchívumok honlapjaihoz is. A koronavírus-járvány miatt sajnos 2020-ban már nem tudtuk ennek a szervezését elkezdni, de 2021-ben a meglévő személyes kapcsolatainkra építve mindenképp szeretnénk elkezdni a megvalósítását (Németh, 2020a).

4 A webarchiválás oktatása

A webarchiválás oktatásának külön fejezetet szentelek, mivel megítélésem szerint az oktatási tevékenység hatóköre túlnyúlik azokon a

projekt kereteken, melyeket az előző fejezet tárgyal. A hipotézisek megfogalmazásánál is külön egységként emeltem ki. Ez volt az egyik olyan legfontosabb önálló kutatási szegmens a disszertáció készítése során, mely az Emerald által kiadott Digital Library Perspectives című folyóiratban Drótos Lászlóval közösen publikált cikkünk révén (Németh & Drótos, 2019a) a nemzetközi webarchiválási közösségben is széles figyelmet és elismerést keltett. A fejezetben elsőként azokkal a nemzetközi mintákkal, együttműködési formákkal foglalkozom, amelyekből ötleteket tudtunk meríteni a saját webarchiválási tevékenység megalapozásához. Ezt követően rátérek a Könyvtári Intézettel, illetve a webarchiválási projekttel párhuzamosan futó E-learning projekt munkatársaival együtt kidolgozott tanfolyami keretek felvázolásához. A közgyűjteményi szakemberek továbbképzésének biztosítása mellett fontos jövőbeni célként jelenik meg a közoktatás szereplői felé történő nyitás, illetve a webarchiválás iránti széles társadalmi érdeklődés felkeltése. A jövőre utaló tervek között tehát egy olyan formálódás alatt lévő tananyagról, illetve képzési keretrendszerrel ejtek szót, amely a bárki által végezhető személyes webarchiválás kompetenciáinak elsajátításában, a kapcsolódó szoftveres eszköztár megismerésében és kezelésében segít.

4.1 Az oktatási tevékenységek nemzetközi háttere

A nemzetközi háttér felvázolása kapcsán olyan projektekről ejtek szót, melyek főszerepet játszottak az itthoni webarchiválással kapcsolatos pedagógiai szemlélet megalapozásában, s a konkrét oktatási tevékenységek formálódásában. Röviden bemutatásra kerül az Aarhusi

Netlab kutatócsoport nemzetközi továbbképzés programja, illetve az IIPC tréning munkacsoportjának tevékenysége.

4.1.1 A NETLAB kutatócsoport e-learning programja

A dániai Aarhus egyetemén az országos DIGHUMLAB kutatási infrastruktúra részeként működő, Niels Brügger professzor vezette NetLab kutatócsoport (Netlab, 2018) online egyetemi kurzusokat kínál PhD hallgatók, ill. a témában érintett szakemberek, kutatók számára a webarchiválás témakörében. 2017 őszén első alkalommal angol nyelven, nemzetközi keretek között is meghirdették a kurzust, ráadásul a részvétel ingyenes volt (Netlab, 2018). A szervezési-moderátori feladatokat Asger Harlaug, a kutatócsoport helyettes vezetője látta el. A szakmai tartalom mellett nagyon fontos volt a részvétel abból a szempontból is, hogy áttekintést kaptam arról, hogy miként kell egy webarchiválással foglalkozó kurzust felépíteni, megszervezni, gondozni. Szembesülhettem a távoktatási felületen zajló kollaboratív munka módszertani jellegzetességeivel is. Külön növelte a kurzus értékét, hogy az OSZK kísérleti projektjével nagyjából egyidőben indult belga webarchiválási program munkatársaival is alkalmam nyílt megismerkedni, a dán szakemberekkel zajló közös munka mellett. Mély benyomást tett a résztvevőkre a konstruktivista pedagógiai háttér. Nem hiányoztak a segédanyagok, a kurzus menetrendje, a közös munkafelület. Emellett azonban alapvetően a résztvevők szakmai irányultságán, lelkesedésén múlt, hogy milyen tartalommal lehetett megtölteni a közös munkát. Nagy tanulság volt ez a későbbi tantervfejlesztési munkámra nézve is, hogy végső soron a tanfolyam sikerének talán kisebbik hányadát adja az

előzetes felkészülés, a tananyag kidolgozása. A sikeresség a tartalom megfelelő felvázolásán, a résztvevőkkel zajló osztálytermi, illetve virtuális kommunikáció minőségén áll vagy bukik (Németh & Drótos, 2019a).

Az online szeminárium egyéni felhasználó név/jelszó párossal védett, Moodle-alapú távoktatási felületen zajlott. Itt publikálták a szervezők a kurzus segédanyagait, a feladatok szövegeit, és itt volt lehetőség a válaszok megfogalmazásra és az egymással történő társalgásra is a kurzusfórum keretében (Drótos & Németh, 2018a). Ez a felület csak a tanfolyam ideje alatt volt elérhető, viszont a kurzus végeztével lehetőségünk volt az összegyűlt anyag lementésére.

Öt fő terület került megtárgyalásra. A kurzus mellé adott, Jan Nielsen által írt kézikönyv szolgált a segítségünkre az egyes témák kidolgozása során (Nielsen, 2016). Először annak felmérése zajlott, hogy a résztvevőknek milyen kapcsolatuk van a webarchiválással és milyen elvárásaik fogalmazódnak meg a kurzussal szemben? Már ebből is következtetni lehetett arra, hogy a konstruktivista pedagógiai alapon szerveződő szeminárium minden egyes félévben más és más profillal bír, a tartalmát erőteljesen befolyásolja a résztvevők érdeklődési köre, szakmai irányultsága (Németh & Drótos, 2019a).

A második téma körül zajló társalgás éppen arra irányult, hogy az általános érdeklődési kör megfogalmazása után mindenki elmondja, hogy miért hasznos számára webarchiválással foglalkozni és felvázoljon a kurzus további teljesítéséhez kapcsolódó valamiféle konkrét munkatervet vagy kutatási tervet (Drótos & Németh, 2018a).

A harmadik témakörben keresni kellett három, az egyéni érdeklődési területhez tartozó honlapot, melyek legalább egy éve üzemelnek már, s áttekintést kellett adni ezek archiválásának kihívásairól. Mintaként lehetett használni az Internet Archive vagy bármelyik nemzeti webarchívum nyilvánosan elérhető mentéseit. Itt lehetett igazán először szembesülni a webarchiválásra jellemző technikai-szakmai kihívások sokaságával és szinte áttekinthetetlen összetettségével. A kurzus résztvevői a saját munkájuk során szembesülhettünk azzal, hogy egyáltalán milyen keretek között lehet webarchiválásról beszélni, hogy az archiválás sikerességének kritikus tényezője a megfelelő szoftveres eszköztár kiválasztása, mely sokszor csupán a tartalom szerkezetének alapos feltérképezése után végezhető el. Persze szembesülhettünk a korlátozott szakmai-emberi erőforrások természetével is, hogy a webarchiválási tevékenység szempontjából kritikus jelentőségű a rendelkezésre álló kapacitások hatékony felhasználásának művészete is (Németh & Drótos, 2019a).

A negyedik témakörben a kutatási terület (gyűjtőkör) előzetes felmérése, valamint a megadott idő alatt egy adott személy által elvégezhető munka kereteinek operatív meghatározása után meg kellett fogalmazni a kutatási célunkhoz kötődő további honlapok összeválogatásának stratégiáját. Ezt követően ki kellett választani a megfelelő szoftveres hátteret, kísérleti aratásokat végezni, majd mini riportban összegezni a tapasztalatokat (Németh, 2018c). Számomra ez bizonyult a leghasznosabb pontnak, hiszen a belgák is és mi is elő tudtunk állni konkrét tartalmak kísérleti mentéseinek tapasztalataival, s meg tudtuk beszélni egymással az

aratáshoz használt egyes szoftverek előnyeit, hátrányait, illetve értékeltük a végeredményt. A társalgásba több ponton bekapcsolódtak a dán nemzeti webarchívum munkatársai is, akik saját tapasztalataikkal színesítették a kommunikációt. Az volt a fő cél, hogy mindenki választ kapjon a kérdéseire a konkrét kihívások kapcsán, s ezeket hasznosítva tudja tovább építeni a saját projektjét (Németh & Drótos, 2019a).

Mindezek után pedig az ötödik témakör keretében összesítettük a kurzussal kapcsolatos tapasztalatainkat, s ezzel zárult a szeminárium, melynek elvégzéséről a résztvevők bizonyítványt is kaptak (Drótos & Németh, 2018a).

4.1.2 Az IIPC Training Working Group tevékenysége

Az IIPC konzorciumon belül 2017 végén alakult meg az oktatási és képzési munkacsoport (IIPC TWG, 2019). Első projektjük keretében kérdőíves vizsgálatot végeztek (IIPC TWG, 2017). Arra irányult a kutatás fókusza, hogy az egyes országokban milyen típusú szervezetek, mekkora létszámban foglalkoznak webarchiválással, valamint ezeknek a szakembereknek milyen igényeik vannak az oktatás, a szakmai továbbképzés terén. A kérdőívet 2018 januárjában lehetett kitölteni. Az eredményekről egy gyors összegzés látott az IIPC tagok körében napvilágot, az eredményeket nyilvánosan nem publikálták. Az alábbiakban ennek megállapításai kerülnek terítékre, elsődlegesen az oktatási tevékenységek megalapozásához szükséges nézőpontokból vizsgálva (Drótos & Németh, 2018a).

A kérdőívet 224-en töltötték ki, öt kontinensről. A válaszokból kitűnik, hogy döntően egyetemi, kutatóintézeti, illetve kisebb, de azért jelentős

részben nemzeti könyvtári keretek között végeznek webarchiválási tevékenységet. Jelentősnek nevezhető még a levéltári szféra súlya, de jelen vannak a múzeumok, az audiovizuális archívumok, illetve a piaci szereplők is. A webarchiválással foglalkozók létszámát firtató kérdésre adott válaszokból kiderült, hogy az intézmények mintegy felében kevesebb mint 1 teljes státuszú munkatárs foglalkozik e területtel! A válaszadók mintegy negyede 1 és 3 között adta meg a létszámot. A másik végletet, a legalább 10 főt alkalmazókat az intézmények kilenc százaléka jelenti, a többi eloszlik a minimum 3 ill. 5 munkatársat foglalkoztatók között. A harmadik kérdés a webarchiválás során végzett tevékenységek felsorolására irányult. A válaszokból kiderült, hogy legtöbbször az archiválendő anyagok válogatásával, illetve szabályzat-alkotással foglalkoznak. Majdnem azonos a súlya a metaadatszerkesztési, a minőségbiztosítási, a kommunikációs és a konkrét webábrák elkészítését lebonyolító tevékenységeknek, a legkevesebben pedig a szoftverfejlesztők vannak. A kitöltők között többségben voltak a közgyűjteményi háttérű szakemberek az informatikusok rovására. A legtöbbször arról számoltak be, hogy csak nemrég kezdtek munkahelyükön webarchiválással foglalkozni, illetve, hogy jelentős kihívást jelent beilleszteni ezt a területet az általános szolgáltatási palettába. Jónéhányan utaltak arra is, hogy egyelőre csak tervezik az ezzel a szakterülettel való foglalkozást, de konkrét gyakorlati tevékenységet még nem végeznek. Ebből is kiderül, hogy még nagyon az út elején tartunk, ami különösen felértékeli az oktatási, képzési tevékenységek súlyát. Amikor a különféle oktatási, képzési igények megfogalmazása kerül terítékre, a webarchiválás általános háttérét fentebb felvázoltakhoz illeszkednek a válaszok is. A

legtöbbször jelenleg online forrásokra támaszkodnak, ha szakmai kompetenciáik bővítésére vágnak, s ettől jelentősen elmarad bármilyen szervezett oktatási tevékenység hozzáférhetősége, illetve relatív módon jelentős a súlya a semmiféle képzésben részesült munkatársaknak is. Messze a legkevesebben vesznek részt hivatalos tantervi alapú képzésekben. Itt mindenképpen érzékelhető egy olyan hiányterület, mely alapvető prioritást jelent a tréning munkacsoport tevékenységében. Elsődlegesen olyan segédanyagok előállítására fogalmazódott meg célként, melyek a webarchiválást érintő alaptevékenységekbe nyújtanak bevezetést. Erre annál is inkább szükség van, mert az elsődleges vizsgálatok arra utaltak, hogy az oktatási tevékenység, ha elérhető, akkor workshopok keretében, illetve informális keretek között vagy valamely szakmai szervezet által koordinálva zajlik. Az új segédanyagok elsődleges felhasználási területe tehát ezekben a képzési formákban képzelhető leginkább el (Németh & Drótos, 2019a).

A képzési igényeket felmérő kérdésre a legtöbbször azt válaszolták, hogy a webarchiválással, a digitális megőrzéssel kapcsolatos szabványok és technológiák terén szeretnék tudásukat elmélyíteni, ill. az egyes szoftvereszközök használatának oktatását tartanák fontosnak. A képzési formák közül pedig a legnépszerűbbnek a webináriumok, illetve a személyes jelenlétre épülő tanfolyamok bizonyultak (Drótos & Németh, 2018a).

Az IIPC tréning munkacsoportja a kérdőív tapasztalatait felhasználva először felmérte azokat a fajta oktatási tevékenységeket, melyekkel a leghatékonyabb segítséget tudják majd nyújtani tagintézményeik, illetve

a lehető legszélesebb körű szakmai célcsoportok számára. Fokozottan szeretnék építeni az egyes országokban eddig felhalmozott tapasztalatokra. Most először mérték fel azt is, hogy az egyes országokban milyen oktatási és képzési tevékenységek zajlanak, milyen keretek között s hányan vesznek azokban részt. A munkacsoport online értekezletein lezajlott eszmecserék során kiderült, hogy a legtöbb oktatási tevékenység igen szűk körű, az egyes webarchiválási tevékenységek ellátásához feltétlenül szükséges készségek és képességek átadására fókuszáló oktatási tevékenység nyert teret. Olyan széleskörű általános tananyagok kidolgozása lett tehát az elsődleges cél, melyek rugalmasan beilleszthetők bármiféle szervezett fizikai jelenlétre épülő, illetve távoktatási programjának a keretei közé, megbízható szakmai tartalmat kínálva fel a webarchiválással kapcsolatos különböző kompetenciák megalapozásához. A későbbiekben kerülhet sor haladóbb szintű webkuratori, illetve informatikai ismeretekhez kötődő témakörökhöz illeszkedő tananyagfejlesztés tervezésére is (IIPC Training Working Group, 2020).

Az IIPC a képzési anyagok fejlesztése során partnerségre lépett a Digital Preservation Coalition szervezettel, akik széleskörű tapasztalatokkal bírnak a hosszú távú digitális megőrzéshez kötődő különféle oktatási segédanyagok fejlesztésében. A közös munka eredményeként elkészült az első ilyen oktatási csomag (IIPC Training Working Group, 2020). Alapvető cél bevezetésként a webarchiválás fontosságának tudatosítása általános társadalmi szemszögből. A legfontosabb készségek és képességek megalapozása a cél meghatározó munkafolyamatokról és a hozzájuk kötődő alapfogalmakról. Terítékre kerülnek az általánosan használt

szoftverek. Ezen a téren elkülönült kategóriákat alkotnak a különféle webes tartalmakat rögzítő, az archivált anyagok megjelenítését segítő, illetve a hosszú távú megőrzés szempontjából releváns programok, programrendszerek. Fontos vetülete ennek az oktatási portfóliónak a webarchiválás vizsgálata a felhasználói igények szemszögéből. Bemutatásra kerülnek a potenciális felhasználói csoportok szakmai jellegzetességei, az általuk támasztott igények, a szolgáltatások tervezésének alapelemei. Ehhez szorosan kötődik a webarchívumokhoz kötődő különféle szolgáltatói felületek kialakítása. Ennek kapcsán egymáshoz szorosan kötődően kerülnek tárgyalásra a szakmai követelmények, valamint a felhasználóközpontú design megtervezése és kivitelezése is. Áttekintést kapunk a különféle webarchiválási és webarchiválás szolgáltatási politikák, irányelvek megtervezéséről, az eltérő szerkezetű politikai és jogi környezetekhez történő illeszkedésről. Felmerülnek a webarchiválást érintő kommunikációs stratégiák felépítésének alapelemei, illetve a szakmai, politikai lobbizás művészetének alaplépéseinek bemutatása is. A tananyag pedig olyan konkrét esettanulmányok rövid bemutatásával zárul, melyek komplex módon ábrázolják a korábban vázolt komponensek egységes szerkezetbe rendeződését és gyakorlatban történő megvalósulását (IIPC Training Working Group, 2020; Németh & Drótos, 2019a).

Alapvető cél a tananyagok tervezése során a moduláris alapszerkezet. Minden egyes fejezetnek önállóan is értelmezhető, használható alapszerkezettel kell bírnia. Könnyen áttekinthető módszertani szerkezet kialakítása a cél, melynek keretében először bevezetést kapunk az adott

fejezet tartalmába, bemutatásra kerülnek a legfontosabb összefüggések és alapfogalmak. A téma részletes kifejtését bőséges fotó, ábra, animáció és beágyazott videótartalom kíséri, az élményszerű tanulás megalapozásának céljából. Az egyes fejezeteket pedig minden alkalommal ellenőrző kérdések, illetve különféle kompetenciák elsajátításának mélységét felmérő feladatok is segítik. Az oktatási tartalom, illetve az ellenőrző kérdések, feladatok is rugalmas beilleszthetőséggel kell, hogy bírjanak különféle terjedelmű és mélységű oktatási program keretei közé. Éppen ezért a rugalmas modularitásnak nem csupán a teljes tananyag, illetve a fejezeti egységek szintjén kell működnie. A tananyag egyes belső fejezeteinek tagolását is úgy kell felépíteni, hogy az egyes altémakörökhöz kötődő anyagok is önállóan felhasználhatóak legyenek. Minden egyes szervezeti elem letisztult módon közvetítse önállóan is azt az ismerethalmazt (kompetenciakészletet), melyet átadni kíván (IIPC Training Working Group, 2020; *IIPC -awesome-web-archiving*, 2020; Németh & Drótos, 2019a).

A szakmai tartalmak meghatározása mellett tehát a módszertani felépítés is egyenrangú fontossággal bír a tananyagok fejlesztése során. A moduláris konzisztencia megteremtése különféle szakmai csoportokban, szinteken válik ellenőrizhetővé. A csoport általános álláspontjaként lett elfogadva, hogy habár fontos a fejlesztésre fordított időtényező optimalizálása is, de a megfelelő minőségű oktatási anyag rendelkezésre bocsátása élvez elsődlegességet (IIPC Training Working Group, 2020). A tananyagok mögött álló koncepció megvilágítására a zágrábi IIPC konferencia alatt videó interjúk is készültek a webarchiválás oktatásának

különféle aspektusairól. A munkacsoportban összeállítottunk egy mintegy száz tételből álló kérdéslistát. Az interjúalanyok számára a felvétel készítői (előzetes konzultációt követően) ezekből szemezgettek különféle témakörökhöz illeszkedve. Az interjúkat a tananyagok közrebocsátásával egyidőben publikálták („IIPC Training Video Case Studies”, 2020).

Az elkészült oktatási anyag az IIPC különféle fórumain keresztül kerül terjesztésre. Önálló GitHub csatorna létrehozását, az alapvetően informatikusok által szoftverek közzétételére használt felületen használatát több tényező is indokolja. A verziókövetéssel jól nyomon követhetők az oktatási anyagban bekövetkező változások. A kiegészítő vizuális és szoftveres elemek is könnyedén beilleszthetők a tananyag mellé. A felhasználói visszajelzések, korrekciós javaslatok precízen rögzíthetők és dokumentálhatók (*IIPC -awesome-web-archiving*, 2020).

Az első nemzetközi kurzus a Digital Preservation Coalition és az IIPC Training Working Group szervezésében 2020 októberében került megszervezésre, a webarchiválásban még nem jártas IIPC tagintézmények munkatársai részére. Itt kerültek először kipróbálásra a kifejlesztett tananyagok konkrét oktatási tevékenység közben is. A tanfolyam a korábban említett modulok anyagát kínálja, válaszokat keresve arra, hogy miért is kell a webet archiválni, hogyan néz ki egy webarchívum, milyen koncepciók szerint s milyen technológiai eszközökkel lehet webarchiválást szervezni és lebonyolítani, s hogy milyen kutatói és egyéb felhasználói igények kielégítésére kell megoldásokat találni.

A NETLAB projektben szerzett tapasztalataim mellett az IIPC tréning munkacsoportjában végzett tevékenységem teremtette meg a

nemzetközi háttérét a saját oktatási tevékenységeinknek. Saját elképzeléseim formálásában is nagy segítséget jelentettek a munkacsoport tagjai által érkező visszajelzések. A saját koncepciónk munkatársaimmal együtt végzett formálásakor szerzett tapasztalataink átadásával pedig én is jótékonyan tudtam befolyásolni a nemzetközi tananyagfejlesztési munkákat.

4.2 A Könyvtári Intézettel közösen kidolgozott blended learning koncepció

A magyarországi webarchiválási kísérleti projekt elsődleges céljaként szerepelt a kezdetektől fogva a szervezett webarchiválás alapjainak megteremtése mellett az oktatási ismeretterjesztő tevékenységek megalapozása is. Az előző fejezetben vázolt nemzetközi munkálatokkal folyamatos kölcsönhatásban elkezdtük kidolgozni saját oktatási tevékenységeink tartalmi és szervezeti kereteit is. Ennek kapcsán eredményes partnerséget alakítottunk ki a Könyvtári Intézet Oktatási Osztályával, illetve a saját kísérleti tevékenységeinknek is keretet adó OKR projekt e-learning munkacsoportjának munkatársaival is (Drótos & Németh, 2018a).

Első lépésben az oktatási tevékenységek potenciális igényeket és célközönséget vettük számba. A webarchiváláshoz rólunkról elengedhetetlenül szükséges az egyes földrajzi, illetve szakmai területekhez kötődő webes anyag gyűjtésének segítése, illetve a webkurátori munkában való közreműködés (szerzői jogtulajdonosok felderítése az archivált anyag nyilvános közzétételének céljára, az archivált anyag leíró metaadatainak elkészítésében való közreműködés

stb.). Ehhez kapcsolódva jelöltük meg oktatási célként olyan közgyűjteményi (elsősorban könyvtári) szakemberek szakmai továbbképzését, akik elsajátítják a webarchiválás munkafolyamataihoz kötődő alapvető készségeket és képességeket, hogy aztán szerepet tudjanak vállalni a fentebb vázolt együttműködési tevékenységekben is (Németh & Drótos, 2019a).

Az oktatási keretek számbavételekor partnereinkkel konzultálva felmerült, hogy kombinált, ún. blended learning típusú oktatási környezetet alakítsunk ki. Ennek részét képezi a Könyvtári Intézet által akkreditált, személyes részvételre épülő 30 órás tanfolyam. Másik fő komponense pedig egy olyan e-learning tananyag létrehozása, mely bármely szabványos oktatási keretrendszerbe betöltve segíteni tudja egyrészt a tanfolyami ismeretek elsajátítását, másrészt pedig egy csupán korlátozott személyes részvételre épülő távoktatási képzési forma megalapozását is (Németh, 2018c).

A tananyagfejlesztési koncepciókat, illetve annak megvalósítását élénk figyelem követte nemzetközi partnereink részéről, ezért az oktatási modell felépítését és gyakorlati formába öntését angol nyelvű szakcikkekben vázoltuk fel (Németh & Drótos, 2019a).

Elsődleges feladatunk az akkreditációs követelményeknek is megfelelő képzési célok megfogalmazása volt. Eszerint mindkét típusú (a személyes részvételre épülő, illetve a távoktatási tevékenységen alapuló) tanfolyam fő céljaként az lett megjelölve, hogy a résztvevők áttekintést szerezzenek az internetes tartalmak hosszú távú megőrzésére alkalmas eljárásokról, szoftveres technikákról. Képet kapjanak az alapvető webarchiválási

munkafolyamatokról. Szerezenek ismereteket a webarchiválás nemzetközi dimenziójáról, az ezzel a területtel foglalkozó legfontosabb külföldi projekteken, szolgáltatásokon keresztül. Betekintést kell szerezniük az OSZK-ban folyó webarchiválási tevékenységekbe is. Kimeneti követelményként fogalmazódott meg, hogy a résztvevők a kurzus elvégzése után képesek legyenek akár magáncélra, akár intézményük számára weboldalakot és webhelyeket tartalmazó gyűjteményeket létrehozni, illetve részt venni a magyar internet archívummal kapcsolatos, korábban már röviden vázolt együttműködési munkákban (Drótos & Németh, 2018a).

Az oktatási tevékenységek fő célközönségét a könyvtárosok, a múzeumi és levéltári dolgozók, az archivátorok, webkurátorok és az informatikusok jelentik. A képzés oktatói bázisát az OSZK Webarchiválási Osztályának munkatársai biztosítják.

A Könyvtári Intézet közreműködésével zajló tanfolyam mintegy évente kétszer kerül megrendezésre (az igényektől és a pénzügyi háttértől függően). Helyszíne az OSZK, illetve felmerült vidéki kihelyezett képzések szervezésének igénye is a jövőt tekintve. A koronavírus-járvány miatt a 2020. őszi tanfolyam elmaradt, a 2021. februári tanfolyam online a Microsoft Teams platformon kerül megrendezésre. A tanfolyam módszertani kerete a konstruktivizmusra épül. A részletes tantervet a mellékelt táblázat tartalmazza. A bemutatott témakörökben való mélyebb elmélyülést bőséges segédanyagok használatával tesszük elérhetővé. Az elméletibb jellegű frontális előadásokhoz is ellenőrző kérdések, gyakorlati feladatok kapcsolódnak. A tanfolyam anyagának magját pedig a személyes

részvételre épülő gyakorlati kurzusok adják. Az oktatók facilitátori szerepet töltenek be, az alapismeretek átadásán túl alapvetően törekednek arra, hogy minden résztvevő élményszerű módon, önmaga sajátíthassa el a szükséges készségeket és képességeket. A szükséges minimumszint teljesítésén túl pedig a résztvevőkön múlik, hogy az egyes anyagrészekben milyen mélységben mélyülnek el. Ebben a tananyag e-learning felülete is támogatást nyújt, melynek használata szervesen beépül a kontaktórákon alapuló tanfolyam keretei közé is. Az oktatási anyagokat pedig igyekszünk nyilvánosan, bármely érdeklődő számára elérhetővé tenni az egyéni ismeretszerzés előmozdításának céljából is (Németh & Drótos, 2019a).

A személyes jelenlétre épülő tanfolyam és a távoktatási anyag is öt modulból áll:

1. A digitális hosszú távú megőrzésről szóló alapismeretek. A webarchiválás a digitális közgyűjteményi szolgáltatások rendszerében. A webarchiválás alapvető terminológiájának, eljárásainak és munkafolyamatainak bemutatása
2. A webarchiváláshoz szükséges online szolgáltatások és Windows-alapú szoftverek bemutatása és alapvető felhasználói készségek megszerzése. A modul célja a webarchiváláshoz alapvetően szükséges munkafolyamatok szervezésének és működtetésének elsajátítása a megfelelő informatikai háttér segítségével.
3. Felhasználói szintű ismeretek megszerzése a Linux alapú webarchiváláshoz szükséges szoftverek használatához. A

webkuratori, különösen a metaadat gazdagítási munkafolyamatok elvégzéséhez szükséges készségek és képességek elsajátítása.

4. A webarchívumnak mint a tudományos kutatás tárgyának bemutatása. A tudományos hasznosítás különféle példáinak ábrázolása. A webarchívumok üzemeltetéséhez, a hosszú távú digitális megőrzés céljainak eléréséhez szükséges alapvető készségek és képességek elsajátítása
5. A tanfolyamon megszerzett ismeretek számonkérése két fő részből áll. A teszt az elméleti háttér értelmezésének képességét kéri számon, a tanfolyamon rendelkezésre bocsátott összes segédanyag felhasználható a megoldáshoz. A gyakorlati feladat pedig egy Windows-alapú szoftver (HTTrack) segítségével végrehajtott egyszerű archiválási feladat elvégzéséből áll. Általában nyolcvan és kilencven százalék közötti eredménnyel töltik ki a hallgatók.

(Németh & Drótos, 2019a)

A személyes jelenlétre épülő tanfolyam végeztével elégedettségi kérdőív, illetve az e-learning anyagok használatára vonatkozó kérdéssor kitöltésére nyílik önkéntesen lehetősége a résztvevőknek. A visszajelzések alapján a hallgatók több személyes konzultációt igényelnének (erre teret adunk a tanfolyami időkereteken kívül is). A tanfolyam belső szerkezetében pedig a visszajelzések kapcsán kisebb átcsoportosítást hajtottunk végre az elméleti időkeret rovására, hogy több idő jusson a különféle szoftverek kipróbálására, használatuk elsajátítására.

A távoktatási forma a járványhelyzetén túlnyúló későbbi intenzív használatára az OSZK e-learninges tartalomkezelő rendszerének beüzemelése után kerülhet majd sor. Az elsődleges célcsoportot azok a szakemberek jelentik majd, akik nem tudnak részt venni a Könyvtári Intézet tanfolyamain személyesen. A kurzus szerkezeti felépítése azonos a személyes jelenlétre építő tanfolyamával. Szerkezetileg természetesen kissé merevebb annál. Az online konzultációs eszközökkel ebben az esetben is erőteljesen építenénk az interaktív kommunikációra, az ellenőrző kérdéseknek, s a tananyag többszintű elsajátítását lehetővé tévő feladatoknak még nagyobb szerep jut.



ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR



KÖNYVTÁRI INTÉZET



Kattintson a képekre, és ismerje meg a kapcsolódó tartalmat!

? 🏠 📖

ÚJABB SÖTÉT KÖZÉPKOR FENYEGET?

A DIGITÁLISAN SZÜLETŐ ÉS AZ INTERNETEN TERJEDŐ KULTÚRA MEGŐRZÉSÉNEK FONTOSSÁGA





Digitális világunk az egyén és a társadalom életének minden szegmensét áthatja. **Az interneten tárolt és azon keresztül közvetített kultúra sokkal változékonyabb és tünékenyebb, mint a korábbi papíralapú.** [↓](#)

A megszűnt weboldalat jelző hibaüzenet  a legnépszerűbb online tartalom és sajátos népművészeti ággá vált. [↓](#)

Ha nem teszünk valamit digitális kultúránk legalább egy részének a hosszú távú fennmaradásáért, akkor egy új *dark age* jöhet: a jövőből megismerhetetlenek, értelmezhetetlenek lesznek korunk eseményei és folyamatai, amelyek elsősorban vagy már kizárólag a világhálón zajlanak.




20. ábra Az e-learning tananyag egyik fejezetének részlete

Jelenleg kérdéses a Linux alapú alkalmazások bemutatása és a programok felhasználói szintű kipróbálása során be kell-e majd iktatni személyes jelenlétre épülő konzultációkat. Néhány program saját kipróbálásához mindenképpen megfelelő informatikai hardverkörnyezet, illetve

felhasználói ismeretek szükségesek. Ennek a résztvevők egy része nem biztos, hogy meg tudna felelni. A kurzus tehát alapvetően virtuális részvételre fog épülni, de valószínű szükséges lesz személyes konzultációkat is beépíteni annak keretei közé (Drótos & Németh, 2018a).

4.3 Jövőbeni oktatási tervek

A Közgűjteményi Digitalizálási Stratégia (KDS) e-learning ágának keretében elkészült egy új, a személyes webarchiválásról szóló tananyag. Ennek célközönségét a 9-12. osztályos gimnazista korosztály képezi. A cél az, hogy felvértezzük őket azokkal a készségekkel és képességekkel, amelyekkel saját maguk is el tudják kezdeni az érdeklődési körükbe eső webes tartalmak archiválását. A tananyag szöveges részének nyelvezete kissé informálisabb oldottabb a szakmai továbbképzésre szánt oktatási tartalomnál. A másik jelentős különbség az, hogy a tartalom fejlesztése során nekünk is alkalmazkodnunk kell ahhoz a mediatizált világhoz, melyben ez a korosztály felnő, illetve életének alapvető kereteit adja. A szöveges elemek csak mintegy felerészben ábrázolják a tudásanyagot. Rengeteg fotó, animáció, video részletek segítik majd felkelteni és folyamatosan fenntartani a célközönség figyelmét. Jelenleg a tanfolyam szöveges nyersanyaga, illetve a kiegészítő képi-média jellegű anyagok beillesztésének forgatókönyve áll rendelkezésre. Hátra van még az a fejlesztési szakasz, melynek révén megfelelő e-learning tartalomkezelő rendszerbe illeszkedve elnyeri majd végső formáját a tanfolyam. Éppen ezért a tananyag csupán nem nyilvános munkaanyag formájában érhető egyelőre el. A végleges formába öntéshez, a szükséges grafikai munkák megrendeléséhez a KDS projekt gazdájának, az EMMI-nek a hozzájárulása

is szükséges. A teljes tananyag elkészültét követően az OSZK E-learning munkacsoportja kér fel majd középiskolai tanárokat metodikai háttéranyagok elkészítésére, illetve a konkrét tantervekbe illeszkedő felhasználási lehetőségek megszervezésére.

A webarchiválás oktatása kapcsán a felsőoktatás egyes szereplőivel is együttműködési lehetőségeket szeretnénk megalapozni. A kísérleti projekt időszakában az ELTE-ről több kommunikáció és informatikus könyvtáros szakos egyetemista csoport is tett nálunk szakmai látogatást oktatóik kíséretében. Ezeken a szakterületeken látnánk lehetőséget további együttműködési formák kialakítására is. A jövőt tekintve szeretnénk, ha a szakmai látogatások mellett bekerülhetne a webarchiválás az egyetemi oktatási keretek közé akár az alapszakos, akár a mesterszakos képzési formákban. A közgyűjteményi szakmai továbbképzési anyagok felhasználásával, a képzőhelyekkel együttműködve szeretnénk az ő igényeikhez illeszkedő oktatási segédanyagokat összeállítani, s közreműködni az ismeretek átadásában is. Ezzel lehetne biztosítani azt, hogy a diplomát szerzett szakemberek a tudásukat a webarchiválás terén is tudják kamatoztatni, partnereinkké válva egy későbbi országos együttműködési háló kialakításában.

5 A webarchívumok tudományos hasznosításának keretei

A hagyományos tartalomfejlesztési feladatokon túl egyre fontosabb bemutatni, hogy a webarchívum miként jelenhet meg a tudományos kutatások tárgyaként. Nagyon fontos felvillantani, hogy társadalmi szinten

mi lehet az a hozzáadott érték, melyet a webarchívum kutatása kapcsán a felszínre kerülhet. A közgyűjtemények szempontjából nagyon fontos új perspektívákat kínál az új kutatási irányok megalapozása, külső partnerségi formák feltárása. A tudományos és társadalmi presztízs emelkedését kínálja intézményi szinten is egy-egy jól megalapozott projektben történő közreműködés. A felsőoktatási intézmények számára pedig fontos új szinergiákat tárulhatnak fel a digitális bölcsészetek, az információtudomány, illetve az informatika oktatásának-kutatásának kapcsán.

A fejezetben először kitérek a kutatási célú felhasználás körül formálódó új nemzetközi projektekre, majd ezt követően egy szinte napjainkban született új tudományágat a webtörténetírást mutatom be. A második témakörben a webarchívumról mint a digitális bölcsészeti kutatások tárgyáról teszek említést, a webarchívumban tárolt nagy mennyiségű adatkészletek tudományos kutatási célú felhasználásáról, illetve az archivált adatok vizuális megjelenítéséről lesz szó adatbányászati és adatelemzési megközelítésben. A webarchiválás, illetve az archivált tartalom visszakeresése hatékonyságának növelésére vonatkozó kutatási tervek felvázolása szemantikus webes eszközök segítségével a hatodik fejezetben kap helyet. Ez a mikroadatokkal foglalkozó fejezet olyan további kutatásokat segítő témafelvetéseket vezet elő, melyek hosszabb távon talán az egész dolgozat legfontosabb hozzáadott értékkel bíró tartalmi elemeiként jelenhetnek meg.

5.1 A WARCnet projekt

A webarchívumokra mint gyűjteményekre irányuló tudományos kutatásokat, illetve a webarchívumok gyűjteményeinek tudományos kutatási célú hasznosítását európai szinten a WARCnet projekt keretében fogják össze. Ennek finanszírozási kereteit dán forrásokból biztosítják 2022 végéig. Az Országos Széchényi Könyvtár (OSZK) részéről 2020 őszének végén, a második online konferenciát követően tudtunk csatlakozni a projekthez. Résztvevői európai nemzeti könyvtárak webarchiválást végző munkatársaiból, tudományos kutatóiból, egyetemi oktatást-kutatást végző személyekből (elsősorban kommunikációkutatók és történészek), illetve informatikus fejlesztők közül kerülnek ki. Négy munkacsoport keretében folyik a tevékenység. Az első munkacsoport a különböző intézményi webarchívumok gyűjteményeit érintő összehasonlító kutatásokra fókuszál, a második munkacsoport a több webarchívum által közösen épített nemzetközi gyűjteményekben rejlő kutatási, elemzési lehetőségeket tárja fel, a harmadik munka-csoport a webarchiváláshoz szükséges információtechnológiai fejlesztések területén mozog, a negyedik pedig a nyílt adatok menedzsmentjének webarchiválási vonatkozásait tartja szem előtt, beleértve a téma tágabb digitális közgyűjteményi vonatkozásait is (például webarchívumok integrációja integrált könyvtári rendszerekbe, teljesszövegű keresési funkciók fejlesztése) (Németh, 2020). Egyre fontosabb bemutatnunk a hagyományos tartalomfejlesztési feladatokon túlnyúlva, hogy a webarchívum miként jelenhet meg a tudományos kutatások tárgyaként. Nagyon fontos felvillantani, hogy társadalmi szinten mi lehet az a hozzáadott érték, melyet a webarchívum kutatása kapcsán a felszínre

kerülhet. A közgyűjtemények szempontjából nagyon fontos új perspektívákat kínál az új kutatási irányok megalapozása, külső partnerségi formák feltárása. A tudományos és társadalmi presztízs emelkedését kínálja intézményi szinten is egy-egy jól megalapozott projektben történő közreműködés. A felsőoktatási intézmények számára pedig fontos új szinergiákat tárulhatnak fel a digitális bölcsészetek, az információtudomány, illetve az informatika oktatásának-kutatásának kapcsán. A projekt fő koordinátora az aarhusi egyetem professzora, Niels Brügger, az ő munkáját segítik a munkacsoport-vezetők, akik egyben az irányító bizottság tagjai is. Sok egyéb mellett értékes információkat kaptunk például a Bajor Állami Könyvtár webarchívumát érintő gyűjteményfejlesztési irányelvekről, melyek jó mintául szolgálhatnak az OSZK saját hungarika alapú gyarapítási elveinek megfogalmazásához is. De szóba kerültek még a jogi szabályozás európai példái, a kutatási célú gyűjteményi hozzáférés európai szabályozási környezeteinek összehasonlításával. A koronavírus-járvány múltával lehetőség lesz pályázni rövid, 3-5 napos kutatási célú szakmai tanulmányutakra, a projektben résztvevő partnerintézményekbe. A projekt során a Belga Nemzeti Könyvtár munkatársa, Friedel Geeraert készített e sorok szerzőjével interjút az OSZK tematikus COVID-webarchívum gyűjteménye kialakításának tapasztalatairól (Németh, 2020b). Ezt érdemes lehet majd összevetni a későbbiekben a többi elkészülés alatt álló interjúval, melyek közül például a Dán Nemzeti Könyvtár, illetve a British Library munkatársaival is közzétételre kerülnek e témakörben.

5.2 Webtörténetírás

A webtörténetírás egy nagyon fiatal, igazából a 2010-es években kibontakozó tudományág. Önálló tudományos folyóirattal is rendelkezik már Internet Histories címmel, első számának bemutatkozó tanulmánya átfogó bemutatással szolgál e tudományos területről Niels Brügger és nemzetközi kutatócsoportjának segítségével (Brügger, Goggin, Milligan, & Schafer, 2017). A kutatások tárgya nagyon széleskörű témaköröket foglal magában. A webarchiválás és történeti kutatások viszonyával itthon is önálló tanulmány foglalkozik Kokas Károly és Drótos László révén, mely a Digitális Bölcsészeti című folyóirat debütáló számában jelent meg (Kokas & Drótos, 2018). Önálló rövid tanulmányban nemzetközi kitekintést is tettem a webtörténetírásról, a webarchiválás egyéb kutatási célú hasznosítási lehetőségeinek felvillantásával együtt (Németh, 2019d). A web első 25 éves történeti kontextusának felvázolását Niels Brügger végezte el. (Brügger, 2016b). Ugyanő mutat rá arra is, hogy a digitális anyagok kutatási célú felhasználása, a digitális bölcsészetek előtérbe kerülése mekkora hajtóerőt jelent a teljes bölcsész-társadalomtudományi terület vonatkozásában is. (Brügger, 2016a). A web múltjának tanulmányozása kulcsfontosságú a jelen fejlődési tendenciáinak értelmezéséhez is. Ez nem csupán a szűken vett világháló múltjára érvényes, hiszen az internet fejlődésének politikai, gazdasági, társadalomtörténeti aspektusai is kulcsfontosságúak a jelen trendjeinek elemzése szempontjából (Brügger, 2013). Az alábbiakban a terjedelmi korlátok miatt csupán rövid áttekintéssel szolgálok a legfontosabb vonatkozási pontok felvillantásával.

A történészeknek ugyanúgy fontos szerepet kellene játszaniuk a webarchiválás kutatási célú felhasználása mögött álló archiválási intézményrendszer kereteinek meghatározásakor, mint ahogyan az a hagyományos levéltárak szervezeti rendje és munkafolyamatai esetében a 19. században történt. Ezt a fontos nézőpontot is önálló tanulmány tárja fel Susanne Belovari révén (Belovari, 2017).

Fontos feladata a történészeknek a webarchívumok biztonságának vizsgálata is abból a nézőpontból, hogy a múlt archivált tényeinek manipulálását, aktuális politikai célokra történő újraírását is meg kellene akadályozni. Itt fonódik össze egymással az IT-biztonság és a történelmi hitelesség szempontrendszer (Lerner, Kohno, & Roesner, 2017).

A számítógépes világhálónak mint technikai infrastruktúrának a történetének, a web mint kommunikációs és publikációs platform történetének tanulmányozása is hangsúlyosan előtérbe kerül. Emellett egy adott személy, esemény, témakör, intézmény webes lenyomatának nyomon követésére is lehetőség nyílik. Érdekes példát kínál erre az első amerikai weboldal történetének rekonstruálása (AlSum, 2015) vagy a brit egyetemek weboldalainak története (Hale és mtsai., 2014).

A webes információ nagyon gyorsan avul, a digitális műveltség, illetve a személyes információk kezelésének fontos eleme lenne a személyes archiválás módszereinek széles körű oktatása és alkalmazása (Gomes & Costa, 2014). Az oktatási fejezetben említettem már, hogy ebben az irányban itthon is tervezünk lépéseket, az ifjúsági korosztálynak szóló digitális tananyagfejlesztéssel. Emellett a Networkshop 2020 című digitális

könyvtári és informatikai konferencián 2020 őszén önálló tutoriált szerveztem a témakör bemutatására.

A fentebb tárgyalt témaköröket egészíti ki az archivált szöveges, illetve vizuális webes tartalmak, vagy akár adott webszerverek naplófájljainak tanulmányozása a gépi tanulás (machine learning) eszköztárával, illetve a nagymennyiségű adatok elemzésének módszereivel (Brügger, 2009).

A kutatás szintjeként megjelenhet egy egyedi webes fájl vagy weboldal, illetve egy adott webhely, doméntartomány is. Legtágabb értelemben pedig a webes univerzumnak mint olyannak a történetét is vizsgálni lehet. A Memento protokoll segítségével több webarchívum archivált anyagai is összevethetővé válnak egy adott weboldallról, illetve témakörről, illetve ennek szükségszerű korlátait is feltárták már (Brunelle, Kelly, Salaheldeen, Weigle, & Nelson, 2015; Klein, Shankar, Balakireva, & Van de Sompel, 2019).

Egy másik nézőpont az, amikor azt vizsgáljuk, hogy ki és milyen céllal, hatókörrel folytat webtörténeti kutatásokat. Amikor kis mennyiségű forrásanyagot, akár csupán egy meghatározott honlap történetét tanulmányozzuk egy adott szoftverkörnyezettel, speciális kutatási céllal, azt nevezzük Niels Brügger terminológiájával élve mikroarchiválásnak, melynek felhasználási módjait gazdag szakirodalmi háttér villantja fel (Brügger, 2009; Jatowt, Kawai, Nakamura, Kidawara, & Tanaka, 2006; Jatowt, Kawai, Ohshima, & Tanaka, 2008; Jatowt, Kawai, & Tanaka, 2007). Persze dolgozhatunk webhelyek egy adott gyűjteményével, pl. valamilyen speciális szakterület webes lenyomatát tanulmányozva. Ilyenkor már a makro szintről beszélünk. Arra is voltak kísérletek, hogy egy teljes nemzeti

webtartományt tanulmányozzunk. Erre a legkorábbi 2000-es évek elejei francia példától napjainkig rengeteg esettanulmányt találunk például Dániából, Hollandiából és Szlovéniából is (Brügger & Laursen, 2019; Klasinc, 2019; Masanès, 2005; Teszelszky, 2019a, 2019b). A nemzeti domén tanulmányozásának speciális esetei közé tartozik a volt Jugoszláviának kiosztott .yu domén, ami mára már teljes mértékben el is tűnt az élő webről, s már csak webtörténeti módszerekkel vizsgálható, melyre Anat Ben-David tett, munkatársaival együtt, kísérletet (Ben-David, 2016; Ben-David, Amram, & Bekkerman, 2018). Különféle részletes módszertani összefoglalók is napvilágot láttak már a webarchiválás vizsgálatának különféle módszereiről és szintjeiről (Brügger, 2009, 2013; Brügger & Schroeder, 2017a).

Jelentős kihívásként jelenik meg a töredékes mementók megjelenése, a hibásan archivált objektumok, illetve a webarchívum visszanezése során felmerülő megjelenési problémák. A történeti hitelesség kérdését veti fel az, hogy az OpenWayback megjelenítő program különféle idősíkokban készült mentési elemeket csúsztat a visszanezés során egymásra, illetve az is előfordulhat, hogy néhány internaktív elem az élő webről szűrődik be a mentett anyag megjelenítésébe. Az archivált állomány autentikus volta tehát a visszanezés során is sérülhet. Egy adott webhely akár új helyre is költözhet, s az eredeti címén adott esetben már teljesen más tartalom kap helyet (Brügger, 2009; Locatelli, 2017).

Fontos megemlítenem ebben az összefüggésben azt is, hogy miután maga a webarchiválás eredendő módon egy összefüggő egészből csak töredékeket tud kiragadni, ez óhatatlanul hatással van a webtörténészek

tevékenységére is. Azzal kell dolgozniuk, ami a rendelkezésükre áll, s adott esetben megpróbálni pótolni a múlt hiányzó darabkait. Ebben persze semmiféle nívum sincs, hiszen a hagyományos történeti források használata is általában hasonló dilemmákat vet fel. Ami mégis különlegessé teszi a webtörténetírói munkát a hagyományos történészi tevékenységhez képest, hogy a vizsgálódásaink módszertani hátterének megtervezésének van egy speciális összetevője. Egyaránt tisztában kell lenni a webarchívumban tárolt források jellegzetességeivel, a webarchiválás munkafolyamatának főbb jellemzőivel, illetve annak a hardver- és szoftverkörnyezetnek a sajátosságaival, illetve korlátaival, melyek révén vizsgálódásainkat folytatni tudjuk.

5.2.1 A webarchívumok történeti kutatási célú elemzésének támogatása

A webtörténészek munkájának támogatására egyre inkább előtérbe kerül speciális munkakörnyezetek kialakítása. Egyre súlyosabb kihívásként jelent meg, hogy a történettudósok közül sokan nem szándékoznak mélyreható informatikai ismeretekre szert tenni, miközben számukra is biztosítani kellene a webarchívumokban tárolt anyagok kutatási célú elemzését („Asking questions with web archives – introductory notebooks for historians—IIPC”, 2020). Az ausztrál, az új-zélandi és a brit nemzeti könyvtárak webarchívumai és az Internet Archive az IIPC-hez benyújtott nyertes pályázatában az ehhez szükséges munkakörnyezet kialakítását tűzte ki célul. Nem voltak előzmény nélküliek ezek a munkálatok, mivel az Archives Unleashed projekt keretében már a később tárgyalt projektben is részt vállaló Internet Archive elkezdett kifejleszteni online munkafüzeteket, illetve felhő alapú gyakorlókörnyezetet az Archive-IT

által gondozott gyűjteményekre irányuló kutatások segítésére elsősorban könyvtárosok, levéltárosok, illetve digitális bölcsészeti kutatásokkal foglalkozó szakemberek számára (Deschamps, 2019; Fritz & Milligan, 2018). A következőkben ismertetett projekt ezen túlnyúlik, mert olyan keretrendszert kínál mely bármely IIPC tagintézmény saját webarchívumához szabványosan illeszthető lesz (Németh, 2020c).

A történészek webarchívumokat érintő kutatásait segítő projektről részletesebb ismertető is készült (Németh, 2020c), itt csak egy szűkebb áttekintésre szorítkozunk. A könyvtárak, levéltárak, múzeumok digitális gyűjteményeinek használatához a GLAM Workbench szolgáltatási környezet eddig is biztosított eszközöket, gyakorlati témákat és leírásokat. Ehhez kellett hozzáilleszteni a webarchiválás témakörét (Németh, 2020c). A cél, tehát annak bemutatása lett, hogy a történeti kutatások során felmerülő témákat a webarchívumokból nyert adatok elemzésének segítségével miként lehet újszerű nézőpontokból feltárni („Jupyter notebooks for web archives”, 2020). Ehhez úgynevezett Jupyter digitális munkafüzeteket (notebook) hoztak létre. Az elméleti háttérrel tartalmazó szöveges útmutatók mellett részletes leírások is találhatók gyakorlati példákkal a különféle elemzési módszerek használatáról („Asking questions with web archives – introductory notebooks for historians – IIPC”, 2020). Az igazán újszerű megoldást azonban egy webböngészőbe integrált gyakorlófelület jelenti, melynek segítségével hozzáférhetünk egy adott webarchívumhoz és a gyakorlatban is kipróbálhatjuk, hogy miként lehet adatbányászati, illetve adatelemzési tevékenységeket folytatni, konkrét példákon keresztül lehet körbejárni különféle felhasználási

lehetőségeket. Sőt arra is van lehetőség, hogy ezeken a gyakorlófelületeken különféle előregyártott alkalmazásokat futtassunk le, s mérjük fel azok kimenetének felhasználási lehetőségeit („Jupyter notebooks for web archives”, 2020). A fejlesztők szándéka szerint a Memento protokoll, valamint a webarchívumok által használt korábban ismertetett eszközök (Heritrix, Brozzler, OpenWayback, PyWB) használatával bármelyik nemzeti könyvtári webarchívum összekapcsolható a tananyagokkal („Jupyter notebooks for web archives”, 2020). Így akár arra is lesz lehetőség, hogy össze lehessen hasonlítani a különféle webarchívumok archiválási módszereit, a metaadatmodelleket, illetve az azokra épülő szolgáltatásokat. Lehetőség nyílik az időbeli dimenzió tanulmányozására, hogy miként fejlődött egy-egy webhely, vagy egy-egy témakör mikor jelent meg a weben s hogyan gyarapodtak az ahhoz kötődő honlapok számszerűleg, illetve tartalmilag egyaránt („Jupyter notebooks for web archives”, 2020).

A fejlesztők szándékai szerint ez az oktatási környezet a zárt hozzáférésű archív gyűjteményekhez is hozzáilleszthető lesz a már említett szoftveres háttér biztosítása esetén. A prototípus 2020 őszre készült el. Ezt követően az IIPC keretében tanfolyamokat terveznek a különféle archívumokhoz történő illesztés, illetve a tananyagok használatának elsajátítására könyvtárosok számára. Felmérésre kerülnek a továbbfejlesztési lehetőségek is (*Final report. Asking questions with web archives—Introductory notebooks for historians*, 2020).

Remélhetőleg Magyarországon is sikerül majd bevonnunk ezt az újszerű oktató-munkakörnyezetet a saját magunk oktatási portfóliójába. Így az

akkreditált közgyűjteményi tanfolyamokon is újszerűen be tudnánk mutatni a webarchívumunk felhasználási lehetőségeit, valamint a kutatói közösség felé is új kapcsolatokat lehetne építeni az újszerű kutatási lehetőségek bemutatásának révén.

5.3 A webarchívum mint a nagy mennyiségű adatok forrása, az adattudományi kutatások tárgya

A webarchívumok nagy szövegtörzsek tárházaként adattudományi projektek középpontjában is állhatnak. Számos ilyen projekt felmerül már a szakirodalomban az utóbbi évekből (Farag, Nakate, & Fox, 2016; Holzmann, Goel, & Anand, 2016; Holzmann, Goel, & Gustainis, 2017; Maemura, Becker, & Milligan, 2016). Az értékes adatok gyors feldolgozása, illetve visszakeresésének biztosítása egyre inkább előkerül a webarchívumok használata során. Ilyen adatok lehetnek a naplófájlok adatai, speciális tranzakciós (pl. geolokációs adatok) vagy különféle az adott archív gyűjteményben tárolt szöveghez kötődő adattípusok is (Lnenicka, Hovad, & Komarkova, 2015). A webarchívumok a disztributív adatfeldolgozáshoz is segítséget nyújthatnak Apache Hadoop segítségével egy megadott alkalmazáskészlettel, adott platformon. Lnenicka és munkatársai (Lnenicka és mtsai., 2015) egy teljes munkafolyamatot vázolnak fel egy webes tartalombányászati alkalmazás fejlesztésére, s egy big data alapú archívum létrehozására, mely modern alkalmazáskörnyezetet használ (Python, PHP, JavaScript, MySQL, és felhőszolgáltatások). Felvázolják az architektúrát, a módszereket, az adatstruktúrát a weboldalak adatainak bányászására, disztributív feldolgozására, és big data alapú elemzésére. Új típusú együttműködés

jöhetne létre ennek alapján a közgyűjtemények, a webarchiválással foglalkozó szakemberek, illetve az adattudósok között. A szerzők arra is felhívják a figyelmet, hogy big data alapú alkalmazások az adattárolás, feldolgozás, elemzés kapcsán kiegészíthetik a hagyományosan használt programok tudását, de semmiféleképpen sem helyettesíthetik azokat! A részlegesen strukturált, illetve teljesen strukturálatlan adatkészletek számos kutatási célból vizsgálhatók. Fókuszálhatunk a webes tartalmakra, a tartalomhasználati adatok kinyerésére, illetve a webes szerkezeti elemek feltárására is. A tartalomfeltárás egyes webhelyek, illetve webhelycsoportok által közölt információk visszakeresésében segíthet. A fő cél az, hogy strukturált adatokat nyerjünk ki ezekből a tartalmi erőforrásokból. Ezek az adatforrások aztán integrálhatók szemantikailag hasonló adatelemekkel, valamifajta tartalmi hierarchia vagy tartalomintegráció alkotható meg a segítségükkel. (Lnenicka és mtsai., 2015).

A különféle strukturált, illetve részben strukturált adatkészletek kvantitatív alapú történeti elemzések tárgyául is szolgálhatnak. Ebben az esetben az adattudós kinyeri az adatokat a webarchívum gyűjteményéből, s segítséget nyújt a webtörténésznek azok elemzésében is. A Niels Brügger és Ralph Schroeder által szerkesztett munka, mely első ízben nyújt reprezentatív képet a webtörténetírás különféle alkalmazási példáiról, számos ilyen projektet sorol fel (Brügger & Schroeder, 2017a; Németh, 2018a). Ilyen például a brit országhomén bölcsész és társadalomtudományi szempontokból vizsgáló BUDDAH projekt (COWLS, 2015; Winters, 2015, 2017), melyet 2014-15-ben bonyolítottak le. 65TB-

nyi anyag került 1996 és 2013 között begyűjtésre, a projekt célja az volt, hogy különféle hasznosítási formákat találjanak ezen begyűjtött hatalmas adatkészlet kapcsán. Ez az anyag nem tükrözi teljes egészében a .uk domén tartalmát. A begyűjtött adatelemeket az aratás dátumával rögzített időbélyegekké látták el az archiválási folyamat során (Brügger & Schroeder, 2017b; Németh, 2018a). A fejlesztők és a kutatók közös munkájának eredményeként megszületett a SHINE névre hallgató keresőfelület, mely a begyűjtött anyagban történő teljesszövegű keresést tette lehetővé. A visszakeresést segítette az anyag különféle témakörökre bontása is, mely a szabadszavas keresés mellett szintén a visszakeresés alapjául szolgálhatott. A webarchívumban tárolt anyag koncepcionális elemekké szervezése, a kutatási stratégiák felállítása, illetve a visszakereső eszköz, illetve annak navigációs felülete tervezése közben új együttműködési területek tárultak fel a különféle tudományágak képviselői között (WEB Archive UK & JICS, 2015). Számos kihívás persze továbbra is fennállt a projekt lezárását követően is. Hogyan kezeljük a nem teljeskörűen, illetve zavaros tartalommal archivált adatelemeket, a webarciválásra fókuszáló kutatási irányok hogyan illeszthetők be az egyes hagyományos tudományterületek keretei közé, illetve a kutatás során felmerülő kérdéseket hogyan lehet közérthető módon bemutatni. A történeti típusú kutatásokra szolgáló keresőmotor prototípusát nyilvánosan is elérhetővé tették (WEB Archive UK & JICS, 2015).

Egy újabb érdekes alkalmazási területet villant fel a 2020 végén zárult LinkGate nevű projekt, mely a webarchívumokban tárolt nagy mennyiségű adatok vizualizációjával foglalkozik („LinkGate: Core Functionality and

Future Use Cases—IIPC”, 2020). A részletek önálló publikáció keretei között is kifejtésre kerültek (Németh, 2020c), itt most csupán a komponensek rövid ismertetésére szorítkozunk. A projekt gazdái az egyiptomi Biblioteca Alexandrina és az Új-Zélandi Nemzeti Könyvtár. Előbbi a technikai fejlesztésért, az utóbbi a kutatói, felhasználói igények becsatornázásáért felel elsősorban. A projekt három alap komponensből áll. Az első egy Link-indexer névre hallgató indexelő eszköz („IIPC RSS LinkGate Webinar”, 2020). Ez kinyeri a szükséges metaadatokat a webarchívumokban tárolt WARC fájllokból (WARC fájl URI címe, WARC fájl dátuma, az adott webhelyről kifelé mutató linkek listája), s egy önálló WAT névre hallgató fájltypusban tárolja azokat. A Link-serv komponens a Link-indexer által kinyert adatokat szemantikus adattárban (data store) tárolja el, s gráf alapú adatsémát rendel hozzá (Németh, 2020c). A gráf alapú adatbázist a Neo4j nevű No-SQL adatbáziskezelő rendszer menedzseli („IIPC RSS LinkGate Webinar”, 2020). A Link-Viz nevű harmadik komponens pedig egy megjelenítési felületet biztosít, ahol az adatbázisban tárolt adatok webböngészőn keresztül vizuálisan gráf-adatszerkezetben megjeleníthetők („IIPC RSS LinkGate Webinar”, 2020). Az egyes webhelyek egy-egy csomópontot alkotnak a hálózatban a közöttük lévő kapcsolatok pedig térben és időben is tanulmányozhatóvá válnak.

Remélhetőleg a jövőben egyre több hasonló projektről fogunk hallani, illetve a webarchiválás szolgáltatási környezetének megszilárdulása után, én is szeretnék Magyarországon kutatókkal együtt dolgozni big data alapú projekteken.

5.4 Hiteles webarchívum

A nemzeti könyvtárak hatókörén általában kívül eső komponens a webarchívumokból történő hitelesített adatok szolgáltatása. Az archivált hiteles jogi dokumentumok gyűjtése és felhasználása az üzleti és a közigazgatási szféra szintjén jelenik meg. Számos országban (például Nagy-Britanniában vagy Ausztráliában) törvény írja elő, a cégek teljes online tevékenységét (ideértve a közönségkapcsolati csatornák forgalmát a közösségi média felületekkel együtt) archiválni kell, a hiteles archivált anyagot jogi eljárásokban felhasználhatóvá kell tenni. Széles szakirodalma van a jogi hitelesség biztosításával kapcsolatos webarchiválási tevékenységnek, ebből idézünk néhány tételt (Alam, Weigle, Nelson, & Klein, 2019; Cunningham & Phillips, 2005; Flanagan, 2010; Strodl, Becker, Neumayer, & Rauber, 2007; Togan & Florea, 2015). Egyes cégek erre építik fel üzleti modelljüket, hogy hiteles módon megőrizzék a céges webes kommunikáció mindenféle lenyomatát, legyen szó weboldalról, vagy akár a közösségi médiáról is. A brit MirrorWeb („Website Archiving and Monitoring Solutions | MirrorWeb”, 2020) cég például a tartalmi elemekben bekövetkező változásokat is rögzíti, naplózza. Így vissza lehet keresni, hogy egy jogi vita esetén adott konkrét időpontban milyen információkat bocsátott az adott cég ügyfeleinek rendelkezésére. A webhelyek mögött álló adatbázisok rekordjait napi szinten archiválják, s elérhetővé teszik audit, egyéb rendszeres ellenőrzési tevékenység, illetve ügyfélpanaszok kivizsgálásának céljából. Az államigazgatásban is egyre inkább előtérbe kerül néhány területen ez a kérdéskör, ahol a hiteles archivált anyagok szolgáltatása az államigazgatás átláthatósága, illetve a jogi viták eldöntése szempontjából jelenik meg. Itt most egy ausztrál

példát említünk meg (Flanagan, 2010). Természetesen ezt a hiteles másolatok begyűjtésére és kezelésére létrehozott teljes rendszerkörnyezetben lehet a leginkább megoldani. Ennek felépítése, a csatlakozó közigazgatási-pénzügyi szolgáltatások némelyikének áttekintése is megjelenik a szakirodalomban (Thornton, 2012; Togan & Florea, 2015). Új begyűjtési módszerek is előtérbe kerülnek ennek kapcsán (Alam és mtsai., 2019). Az üzleti élet és a közigazgatási szolgáltatások tisztességes és zavartalan működésének garantálásához ez a webarchiválást érintő terület várhatóan még jobban fel fog fog a jövőben értékelődni.

6 Mikroadatok használata a webarchiválás és a visszakeresés hatékonyságának növelésére

Ennek a fejezetnek a legfőbb célja a további kutatómunka megalapozása a webarchiválás hatékonyságának növelése terén, olyan adatkészletek kidolgozásával melyek szabványos szemantikus jelölési tulajdonságokkal felvértezve beilleszthetők lennének bármely webhely forráskódjába és értelmezhetők lennének a webarchiválást végző szoftverek által. Elsőként definiáljuk azt, hogy milyen típusú adatkészletekről beszélünk, majd a későbbiekben rátérünk azok használatának jelentőségére a webarchiválás során.

6.1 A mikroadatok a dokumentumalapú web és a szemantikus web metszéspontjában

A HTML szabvány születésétől fogva néhány nagyon alapvető metaadat elem (például a cím feltüntetése) beilleszthető lett egy honlap forráskódjának fejlécébe. Az adott webhelyhez csatolt robots.txt fájl segít

lehatározni annak tartalmát, miután utasításokat tartalmaz, melyek meghatározzák, hogy merre mehet az aratórobot, s mely tartalmi elemek tiltottak a számára (Németh, 2019d). Ezek a szolgáltatások a dokumentumalapú webhez kapcsolódnak, ahol csak nagyon hozzávetőleges elképzelésekkel rendelkezhetünk egy webhely tartalmi jellemzőiről. Ezek a szolgáltatások nem segítik elő a webhelyekre vonatkozó tartalmi információk visszakeresését, s különböző webhelyek metaadatait sem tudjuk összekötni egymással (Németh, 2019c).

A szemantikus web megjelenése a következő fejlődési fok, mely nem egyedi dokumentum egységeket (webhelyeket) tekint alapegységnek, hanem azok tartalmára fókuszál, hogy miképpen lehet a tartalmi jellemzőket azonosítani, visszakeresni, katalogizálni, s egy globális információs univerzum részévé tenni egyben azokat. A webalapú erőforrások ezáltal egységes felületen kezelhetővé válnak, függetlenül attól, hogy könyvtári katalógusokból vagy akármely, a nyílt kapcsolt adatok szabványai szerint közzétett információforrásból származik (Németh, 2019c). A mikroadatok voltaképpen hidat képeznek a dokumentumalapú web és a szemantikus web között. Lehetővé teszik szemantikus állítások megfogalmazását metaadat formájában, majd beilleszteni azokat egy webhely forráskódjába, s így közzé lehet tenni annak tartalmát szemantikus platformokon is (Horváth, 2016). A mikroadatok a webes szabványok és a szemantikus webes modellek alapján kerültek meghatározásra. Az RDFa (Resource Description Statement in Attributes) a W3C konzorcium által közzétett ajánlás („XHTML+RDFa 1.1—Third Edition”, 2015), mely attribútum szintű

kiterjesztéseket ad hozzá a HTML, XHTML és különféle egyéb XML alapú dokumentumtípusokhoz, a webes dokumentumok ellátására gazdagon metaadatokkal.

Az RDFa az RDF adatmodell része, melyet a szemantikus web fő sarokpontjának tekinthetünk. A megfeleltetési eljárás lehetővé teszi RDF alapú alany-állítmány-tárgy típusú kijelentések beillesztését XHTML dokumentumokba. A hatalmas előnye ennek, hogy az RDF kifejezések állítmányi része bármely szabványos adatszótárból megadható. Így tehát különböző szótárak vagy sémák egységes felületen használhatók (Németh, 2019c). Az egyik legfontosabb ilyen szótár a schema.org („Bib.schema.org-1.0—Schema Bib Extend Community Group”, 2020; „Bib.schema.org-1.0—Schema Bib Extend Community Group”, 2020; „Getting Started—Schema.org”, 2020). Ez a Bing, a Google és a Yahoo közös fejlesztéseként jött létre, s nagyon hasonló egy általános tárgykörű teauruszhoz. A két legfontosabb hozzáadott érték, melyet ez képvisel, hogy lehetővé teszi bármiféle, a bennünket körülvevő világ egyes elemeiről alkotott kijelentésünk szabványos osztályozási környezetbe helyezését, s ez a teauruszhoz hasonló rendszer felhasználható bármely weboldal, illetve webhely tartalmi leírására. Hatalmas előnye tehát a felhasznált kijelentések szabványos pontossága, illetve egymáshoz illeszkedése. A schema.org jelölőkkel rendelkező HTML oldalak sokkal értelmezhetőbbek a keresőmotorok számára, mint az azokat nélkülöző társaik, s így tartalmilag sokkal relevánsabb, pontosabb információkhoz lehet hozzájutni általuk (Németh, 2019c). Ezek a jelölők voltaképpen szabványos mikroadatok, melyek az adott oldal forráskódjában jelennek

meg (Adamich, 2015; „Getting started with schema.org using Microdata”, 2016). S mivel a keresőgépek számára a releváns tartalmi jelölőkkel ellátott oldalakat sokkal könnyebb indexelni, így előnyben is fogják részesíteni azokat. Az egyik első és legnagyobb felhasználója a schema.org rendszerének az OCLC (Online Computer Library Center) lett. A WorldCat nevű globális katalógusrendszerükben minden tétel leírása schema.org kijelentéseket is tartalmaz. Ennek az eredménye, hogy az OCLC rekordjai előkelő helyeken szerepelnek a keresőrendszerek találati listáiban (Adamich, 2013). Egyre több nyílt forráskódú integrált könyvtári rendszer is elkezdte hasonló okból alkalmazni tehát a schema.org használatát.

A mikroadatok ugyanakkor nem csupán az indexelés pontosságának növelése érdekében hasznosak. A kapcsolt adatok univerzumában a szabványos szemantikus formában leírt (címkézett) adatkészletek összekapcsolhatók egymással, olyan hatalmas tartalmi bázist kínálva az információk visszakeresése számára, mellyel eddig az emberiség történetében nem találkozhattunk (Németh, 2019c). A mikroadatok által az adott weboldalba, webhelybe beépíthető egy olyan információkészlet is, mely azok hosszútávú digitális megőrzéséhez járul hozzá (Enis, 2015; „Getting started with schema.org using Microdata”, 2016).

Az üzleti életben, ahol valós üzleti érdekek állnak a különféle szemantikus sémák fejlesztése és használata mögött, a mikroadatok használata már elterjedtnek mondható. Nagyon jó példa erre a különféle árukereső szolgáltatások működése, ahol az árucikkek rengeteg adata be van címkézve a kereskedelmi oldalak forráskódjába, s így válnak azok feldolgozhatóvá és összehasonlíthatóvá (Németh, 2019c). Esetünkben

ugyan nem kereskedelmi, üzleti háttér indokolná a mikroadatok használatát, de az érdekelt intézmények, tartalomipari szereplők megítélésem szerint igen könnyen ráébredhetők lennének használatuk előnyeire egy megfelelő koncepció felvázolása esetén. Éppen ezért tekintjük most át előbb a legközvetlenebb előnyöket, majd a főbb különféle potenciális hasznosítási területeket.

6.2 A mikroadatok potenciális felhasználási lehetőségei a webarchiválás során

Az OCLC közrebocsátott egy útmutatót, mely a webarchiváláshoz szükséges leíró jellegű metaadatok alapkészletét teszi közzé (Jackie & Bowers, 2018). Ebben hozzáillesztették a megfelelő schema.org megfeleltetéseket a főleg Dublin Core metaadat rendszerből kiinduló alapelemekhez. Ez az ajánlás azonban csak a lehető legszükségesebb leíró elemeket tartalmazza. További kihívás, hogy a schema.org vagy akármely egyéb szemantikus szótár nem tartalmaz olyan speciális adatelemeket, melyek a webarchiválási tevékenységekhez kapcsolódnak. Egy új séma létrehozása, illetve valamely meglévőnek a kiegészítése jelentős további kutatómunkát igényel (Németh, 2019c). Ebben a fejezetben arra vállalkozom, hogy felsoroljuk azokat a legfontosabb területeket és adatkészlet típusokat, melyek kidolgozása és egy szemantikus sémába történő beillesztése rengeteget segíthetne a webarchiválás hatékonyságának növelésében.

6.2.1 A webarchiválás hatékonysága növelésének általános kihívásai

Mielőtt a figyelmünket a mikroadatok webarchiválási környezetben történő felhasználására fordítanánk, néhány, azok felhasználását is érintő webarchiválási kihívásra világítunk rá.

Az első ilyen kihívás, hogy a robotok számos esetben nem tudják feltérképezni egy webhely teljes struktúráját, mivel az nincs kellően definiálva. További kihívást jelent a tartalomformátumok dinamikus változása, hogy a régebben általánosan használt formátumok megjelenítésének némelyikét a mostani böngészők már nem támogatják (elég, ha csak a Flash formátumra, illetve technológiára utalunk). Számos esetben napjainkban már azt sem egyszerű megállapítani, hogy egy adott webhely egyáltalán milyen formátumokra, technológiai elemekre támaszkodik. Csupán annak vagyunk tudatában, hogy a leartott webhelyek anyagából egyes komponensek hiányoznak, de a hiány okai rendkívül összetettek lehetnek, s a web, mint médium rendkívül komplex volta miatt nagyon nehezen is állapítható meg az orvoslás módja (Németh, 2019c). Másik jelentős kihívásként jelentkezik a nagytömegű adatok kezelése az azokat tartalmazó adatkészletekben. Az adatkészletek tartalma a tudományos kutatások tárgyaként szolgálhatna, de valóban kutathatóvá tenni azt rendkívül összetett feladatot jelent. A különféle szabványos mikroadat sémák (szótárak) használata jelentős segítséget jelenthet az imént jelzett kihívások kezelésében (Németh, 2019c). A továbbiakban tehát a különféle használati módok összefoglalására térünk rá.

6.2.2 A robotokkal végzett webaratás szabályozása mikroadatok segítségével

Néhány típusú mikroadat segítséget nyújthat számunkra abban, hogy az arató robotok számára pontosan és teljesen feltérképezhető legyen egy adott webhely szerkezete. A webhely térképe elkészíthető XML formátumban. Ezt a leírást egy robots.txt nevű fájlban tárolható el, melybe speciális utasítások illeszthetők melyek a webarchiválást végző robotok tevékenységét befolyásolják. A robots.txt fájlban tárolt információk megjelenése ki is bővíthető, hogy ne csupán egyetlen fájl jelenjen meg az adott webhely gyökérkönyvtárában, hanem egyes információk a webhely egyes oldalainak fejlécében legyenek rögzítve, pontosan leírva az adott oldalon található adatokhoz történő hozzáférés módjait. Ezek a robots.txt fájlban, illetve az egyes lapok fejléceiben rögzített járulékos információk jelenleg nem szabványos kiegészítések formájában érhetők el, ilyen például a Robots Exclusion Standard („Robots exclusion standard—Wikipedia”, 2020). Ebben vannak leírva például a webhely egyes oldalainak, illetve elemeinek indexelését kizáró noindex metaadat címke, vagy noindex http választ nyújtó fejléc elem (html response header). De egyéb utasításokat is adhatunk a robotok viselkedésének szabályozására, mint például az allow a sitemap host, vagy a universal match. A lényeges problémát itt az jelenti, hogy ez a szabvány, illetve annak elemei nem részei a hivatalos html szabvány leírásának. A webarchiválást végző szakmai közösségnek arra kellene törekednie, hogy bekerüljenek abba, s elérhetőek legyenek minden robot számára az összes platformon, a lehető legtöbb webhelyen. Ennek segítségével sokkal pontosabb honlaptérképek készülhetnének el a mostaniaknál. Az

öröknaptár alkalmazások, a robotok számára nem bejárható adatbázisokra mutató linkek például mind külön címkézhetők lennének, hogy kizárásra kerüljenek az aratásból, megelőzve, hogy a robot olyan csapdába kerüljön, melyből nem tud kijutni (Németh, 2019c). A noindex http fejléc elem pedig adott esetben egy webhely egész oldalát is kizárhatja az aratásból, amennyiben annak tartalma nem aratható le a robotok számára. A robotok működését szabályozó szabvány további kiterjesztése lehetne olyan szabályok kidolgozása, melyek az adott webhely különféle platformokhoz illeszkedő felületeit határoznák meg (mobilbarát felület, akadálymentes felület, robotbarát felület stb.). Így meg lehet adni a robotok számára, hogy csupán egy bizonyos felületet járjanak be, vagy külön-külön akár az összeset is, ahogy az archiválási szabályok megalkotásakor rendelkezünk az adott webhely szerkezeti, formai és tartalmi jellemzőinek ismeretében (Németh, 2019c).

Mikroadatok segítségével egyes tartalmi elemek is könnyebben elhatárolhatók lennének az aratási tevékenységet segítő. Ilyenek például a hirdetési elemek, amelyek megfelelő címkézésére támaszkodva az archivátor beállíthatná, hogy reklámokkal vagy reklámok nélkül kérje egy adott webhely mentését. De ugyanígy címkézhetők lennének a külön kis ablakokban felbukkanó tartalmi elemek (pop-up contents), melyek arathatóságát szintén opcionálisan lehetne rögzíteni az archiváló intézmény irányelvei szerint (Németh, 2019c).

Különleges szabályokat állíthatnánk be a robotok számára, amikor a webhely tulajdonosa azt szeretné, hogy csak egy bizonyos felület legyen aratható. Ugyanígy válogathatna a webhely tulajdonosa a különféle nyelvi

verziók arathatóságának beállítása kapcsán is. Külön kihívást jelent, amikor egy webhely kezdőoldala csupán valamely egyszerű lekérdező elemet tartalmaz, melyről tovább haladva érhetőek el különféle műveletekkel az egyéb tartalmak. A robot azonban automatikusan nem tudja értelmezni az itt felkínált műveleteket, ezért meg kell adni számára azt a kezdőpontot, esetleges alternatív felületet, ahonnan kiindulva el tudja kezdeni egy adott honlap aratását. Ezt is be lehetne állítani megfelelő mikroadat címke segítségével, így a robotok a kezdőlapról automatikusan átirányításra kerülnének arra az oldalra, ahonnan az aratás elkezdődhet. Ennek révén, amennyiben nem is archiválható a maga teljességében egy webhely teljes keresési funkcionalitása, illetve információ-visszakeresési szolgáltatásai azok eredeti formájában, a felület mögött lévő adatok arathatók és megőrizhetők lennének (Németh, 2019c).

A fentebb felsorolt információk mikroadatok segítségével történő szabványos megadása a webaratás hatékonyságát még a jelenlegi szoftveres háttérre támaszkodva is jelentősen megnövelné, anélkül, hogy fejlesztői oldalon jelentős befektetéseket kellene eszközölni az archiváló szoftverek továbbfejlesztésének irányában. A webhely tulajdonosoknak, fejlesztőknek pedig a szabványos ajánlások érvényesítése szintén nem jelentene jelentős többlet ráfordítást sem időben, sem pénzben (Németh, 2019c).

6.2.3 Az archivált állomány megjelenítésének segítése mikroadatok révén

Jelentős kihívásként jelentkezik a webarchiválás terén, hogy amennyiben az archiválni kívánt anyag begyűjtése sikeres volt, az archivált állomány a

megjelenítő szoftveres háttér korlátai miatt nem jeleníthető meg helyesen. Különféle mikroadatokban elhelyezett utasítások segíthetnék a leírt tartalom szerkezetének és grafikai formátumának helyes megjelenítését. Például számos esetben speciális JavaScript vagy pdf bővítményeket használ egy adott webhely az azokhoz hozzárendelt beágyazott tartalmak megjelenítéséhez, vagy ugyancsak gyakran lehulló menük tárják fel egy webhely tartalomkínálatát, szolgáltatásait. Amennyiben mikroadatok segítségével a webhely forráskódjába illesztve alternatív módokat állítunk be a fotók, dokumentumok, vagy egy egyszerűsített formájú menüstruktúra megjelenítéséhez, akkor az archivált tartalom tökéletesen megjeleníthető lesz (Németh, 2019c).

6.2.4 A hosszú távú digitális megőrzés támogatása mikroadatok segítségével

Speciális mikroadattípusok nyújthatnak segítséget abban, hogy egy adott webhely forráskódjába illesztve feltárják a webhelyen használt külféle tartalom formátumok típusait, verzióit és a kapcsolódó szoftverek jellemzőit. Különösen problémás elemek lehetnek a webarchiválás kapcsán a külféle java appletek, beágyazott Flash alkalmazások és tartalmak. Ezeket igen fontos lenne külön mikroadat címkékkel ellátni, meghatározva azok pontos verziószámát és funkcióit. Ezek a pontos leírások különösen fontossá válhatnak majd a jövőben, amikor az archivált webhelyek tartalmának konverziója, illetve emulációja válik majd szükségessé új megőrzési, illetve szolgáltatási platformokon (Németh, 2019c).

6.2.5 A kutatástámogatási szolgáltatások segítése mikroadatok segítségével

A webarchívumok kutatási célú felhasználásának támogatása nagyon fontos részét képezi a szolgáltatási kínálatnak bármely archiválást végző intézmény esetében. A szemantikus adatok szolgáltatása fontos segítséget jelent a kutatók számára az archívumból történő hatékony információkeresésben, amikor számos nagyméretű adatkészlettel, egyéb tartalmi erőforrásokkal kell dolgozniuk. A különféle adatkészletek közötti kapcsolatok megteremtése a nyílt kapcsolt adatok modelljének segítségével szintén igen fontos kutatástámogató tényező. Így ugyanis kiléphet egy adott kutató az archívum virtuális határai közül, s együtt kezelheti az onnan származó adatkészleteket másfajta szabványos adatforrásokkal is. A legfontosabb eszközök e tekintetben a hivatkozást teremtő relációk, amelyek leíró attribútumként rendelhetők hozzá egy adott hiperhivatkozáshoz, meghatározva annak típusát és/vagy annak kapcsolatát a hivatkozás forrása felől annak célpontja felé (Németh, 2019c).

A szemantikus weben az RDF típusú hivatkozások alapvető jelentőségűek a nyílt kapcsolat adatokra épülő adatkészletekben, meghatározva a kapcsolatot (állítás formájában) az RDF tripletekben, lehetővé téve a gépi úton is értelmezhető állítások automatikus feldolgozását egy gigantikus globális gráf keretei között a szemantikus weben. A nyílt kapcsolat adatokra épülő adatkészletekben az RDF nyelven rögzített linkek adják meg az értékeket az `rdf:type` property elem segítségével, illetve megadják azok kapcsolatait a már jól ismert szemantikus sémákban, szótárakban

található kifejezések, definíciók segítségével. Például személynevek, illetve földrajzi nevek is címkézhetők, illetve szűrhetők különféle névtér sémák s az egyes elemekhez rendelt azonosítók segítségével. A szemantikus wikipédia, illetve szemantikus könyvtári katalógusok, illetve bármely más szemantikusan publikált adatkészletek névtér elemei is hozzáadhatók kapcsolatként (Németh, 2019c).

Egy másik fontos nézőpont egy adott webhely létrehozatala időpontjának, illetve a legutóbbi változtatások dátumának pontos rögzítése. Nagyon fontos rögzíteni egy adott archivált anyag, mint történeti forrás hitelességét is. Sajnos néhány jelenleg is elterjedten használt visszanevező szoftver különböző idősíkokban készült mentések elemeit képes egységes felületen megjeleníteni, csorbítva így az archivált anyagok történetileg hiteles megjelenítését. A webhelyek, weboldalak, vagy annak akár egyes részei létrehozatali, módosítási dátum adatainak rögzítése nagyban segíthet abban, hogy az azonos időpontban mentett tartalmi elemek hiteles módon legyenek megjeleníthetők az archivált állományból (Németh, 2019c).

A mikroadatok néhány digitális bölcsészeti alapú kutatási tevékenység támogatásában is segítséget nyújthatnak. A digitális filológusok nagyszámú XML alapú elemzőeszközt, illetve eljárást használnak. Az általuk használt eszközöket, módszereket megfelelő tartalmú mikroadatokkal szemantikus sémák, szótárak elemei használata révén is támogatni lehet. Például egy angol nyelvű szövegben megbújó ógörög nyelvű idézet, egy adott személy vagy földrajzi név (mint névtér elem)

különbéle megnevezései, a többnyelvű névalakok, a történeti írásmódok használata is feltárható ily módon (Németh, 2019c).

6.3 Mikroadatok előállítása és alkalmazása

Mint a fentiekben láthattuk a webarchiválási tevékenységet és a webarchívumok használatát rendkívül változatos módon lehet megfelelő mikroadatforrások előállításával támogatni. A központi kérdés ott bújik meg, hogy ki állítja össze az adatok rögzítését és használatát előíró ajánlásokat, illeszti hozzá azokat a meglévő szabványkörnyezetekhez, sémákhoz, illetve vállalja fel azok folyamatos gondozását. Világosnak tűnik, hogy elsősorban a jelentős webarchívumokat üzemeltető, digitális dokumentumállománnyal bíró közgyűjtemények, illetve tartalomipari oldalról a nagy tartalomszolgáltató platformok részvétele tűnik ebben elkerülhetetlennek (ideértve a jelentős blogszolgáltatókat, illetve közösségi média felületeket is). Ezek a szereplők elegendő erőforrással rendelkeznek megfelelő ajánlások, illetve szótárak kidolgozásának segítésére, illetve használatuk elterjesztésére. A kritikus felhasználói tömeg elérése a tartalom előállító oldalon kulcskérdés, hiszen így válnak igazán értelmezhetővé a mikroadatok használatának előnyei. A webarchiválás területén az e területen érdekelt IIPC konzorcium vállalhatja vezető szerepet a megfelelő ajánlások és szótárak kidolgozásában. Partnerségi kapcsolatok építésével pedig hozzájárulhatna, hogy a tartalomipar legfőbb szereplői is felismerjék a mikroadatok implementálásának hasznosságát a saját szolgáltatási környezeteik keretei közé (Németh, 2019c).

Nagyon hasznos lenne, ha fellendülnének a mikroadatok potenciális alkalmazásával foglalkozó kutatások a webarchiválás területén, s nemzetközi összefogással a közgyűjteményekre támaszkodva, egyre több tartalomipari szereplőt bevonva elterjedne a kidolgozott ajánlások, szótárak, sémák használata, valahogy hasonlóan, mint ahogy az akadálymentes webes felületek kialakításának fontossága is a közgondolkodás részévé vált. A webes kulturális örökség megőrzése ugyanis közös ügye kellene, hogy legyen a tartalomipari szereplőknek s közgyűjteményeknek egyaránt.

7 Epilógus - A kutatás értékelése

A zárófejezetben először a hipotézisekben foglaltakkal kapcsolatos kutatás eredményekre térek ki.

Az első hipotézisben foglalt célkitűzések véleményem szerint teljesültek. Többféle megközelítésből próbáltam alátámasztani a webarchiválás mint tudományos téma relevanciáját, különös gondot fordítottam annak alapvető közgyűjteményi kereteinek felvázolására. Több fejezetben is szóba kerültek azok a dilemmák, hogy milyen szakmai, technológiai, illetve elméleti módszertani határai vannak a webarchiválással kapcsolatos tevékenységeknek, milyen értelmezési tartományban határozhatjuk meg a webarchiválás fogalomkörét. Önálló témakörként került tárgyalásra ennek kapcsán az archivált állomány minőségének ellenőrzésével, illetve hosszú távú megőrzésével kapcsolatos kihívások halmaza is.

Vizsgálódásaim során különös gondot kívántam fordítani az itthoni webarchiválással kapcsolatos gyakorlati munka, illetve elméleti kutatási háttér megalapozása kapcsán a vonatkozó nemzetközi szakirodalom széleskörű feltárására az egyes, a dolgozatban tárgyalt résztémák kapcsán. A dolgozat fontos összetevőjét alkotja a webarchiválásra irányuló jelenlegi tudományos kutatási irányok összegzése. A webarchiválás mint átfogó interdiszciplináris szakmai kutatási terület egyre jobban intézményesül a nemzetközi tudományos életben. Ugyanez magyar viszonyok között még nem mondható el. Miután a webarchiválás mint állandó szakmai feladat törvénymódosítás révén az Országos Széchényi Könyvtár alaptevékenységei között kap helyet, remélhetőleg a kialakuló gyűjtemény, mint a tudományos vizsgálatok tárgya is szélesebb körben teret nyer majd a későbbiekben. Ezzel a disszertációval voltaképpen a kezdőlökést szerettem volna megadni e folyamat kibontakozásához. Egyrészt szükséges volt egy általános áttekintéssel szolgálni a webarchiválással kapcsolatos általános tudományos vonatkozási pontokról. A legfontosabb célja mindezek mellett e disszertációnak mégis az, hogy áttekintést nyújtson a webarchiválást érintő különböző kutatási, illetve kutatás-fejlesztési tevékenységekről. A szerző, amellett, hogy meg kívánta jelölni a személyes érdeklődése homlokterében álló kutatási területeket, egyben inspirációval is kíván szolgálni ahhoz, hogy minél többen, minél több nézőpontból válasszák a webarchiválást, a webarchívumot mint gyűjteményt a tudományos kutatásaik tárgyául.

Az elméleti és kutatásmódszertani keretek mellett a második hipotézisben foglaltak révén a dolgozat másik nagyon hangsúlyos összetevője a formálódó hazai gyakorlat elemeinek felvázolása. A nemzetközi gyakorlat itthon leginkább hasznosítható vonatkozásainak tárgyalása után sort kellett keríteni a tevékenységet jellemző gyakorlati jellegű munkafolyamatok bemutatására. Sokszínű elemzések révén adtam választ arra a hipotézisekben feltett kérdésre, hogy sikerrel alapoztuk meg munkatársaimmal együtt a webarchiválást mint önálló nemzeti könyvtári szolgáltatást. A dolgozatban felvázoltam a kiindulási munkatervünket, s azt, hogy ehhez képest meddig sikerült eljutnunk az egyes célkitűzések, munkafolyamatok, szakmai tevékenységek formálása kapcsán. Előkerült az összes olyan főbb tevékenységi irány, melyekkel foglalkozni kellett az állandó szolgáltatás megteremtése kapcsán.

A hipotézisekben foglalt harmadik lényeges témakörként merült fel még a webkönyvtárosi, webkurátori feladatkörök tisztázása, hogy egyáltalán mennyire határozható meg általában a webarchiválás könyvtári feladatként. Ennek kapcsán egyrészt bemutattam a webarchiválás leglényegesebb könyvtári vonatkozásait tágabb közgyűjteményi értelmezési keretek között, másrészt a dolgozatban tárgyalt gyakorlati munkafolyamatok is megfelelő példákat nyújtanak arra, hogy a hagyományos könyvtári tevékenységek, mint például a gyarapítás, feldolgozás, formai és tartalmi leírás, tárolás, megőrzés és a szolgáltatási dimenzió milyen módon értelmezhetők a webarchívum mint állományrész vonatkozásában.

A hipotézisekben foglalt negyedik témakör kapcsán nagyon lényeges elemeként került bemutatásra a szabályozási, jogi háttér kialakításának segítése is. Az infrastrukturális, jogi, és a humán erőforrások szűkösségéből adódó, az állandó könyvtári szolgáltatás megalapozása kapcsán felvetődött kihívások időről időre nem kis feladatot jelentettek. A nyitott és többnyire segítőkész nemzeti könyvtári vezetői és a feltétlenül pozitív fenntartói, jogszabályalkotói hozzáállásnak köszönhetően a hipotézisekben felvetett kihívások jó részét sikerült leküzdeni. A webarchiválási tevékenység fenntartható ellátásához azonban feltétlenül szükséges a humán erőforrás bővítése a kezdeti időszakban jelenleg garantált finanszírozási háttér folyamatos fenntartása, az információtechnológiai feltételek folyamatos biztosítása, az intézményes partnerekkel való együttműködés.

Az elméleti háttér és a webarchiválási gyakorlati tevékenység megalapozásának egyfajta közös metszeteként kerültek hangsúlyosan tárgyalásra a hipotézisekben foglalt ötödik témakörben felvetett oktatási kérdések is. A webarchiválás témaköre sikeresen bekerült az akkreditált könyvtáros továbbképzés keretei közé. Lényeges jövőbe mutató feladat e téren a felsőoktatási partnerekkel való kapcsolatok kiépítése, hogy az elméleti és gyakorlati webarchiválási tevékenységeket a graduális képzés oldaláról is támogatni lehessen.

A hipotézisekben foglaltakra adott válaszok összegzése után a publikációs tevékenységem rövid értékelésére keríték még sort. Minden fő fejezethez természetesen hozzárendelhetők önálló publikációk, konferencia előadások. A kísérleti projekt munkafolyamatait kapcsán kiemeltem

foglalkoztam előadások, illetve publikációk szintjén a nemzetközi gyakorlattal, a webarchiválás közgyűjteményi kereteinek bemutatásával, a webarchívumok metaadat ellátásával, a webarchiválás oktatásának, tudományos célú felhasználásának kérdéskörével, a szemantikus technológiák használatával, a kísérleti projekt eredményeinek bemutatásával, az állandó nemzeti könyvtári szolgáltatás háttéréül létrehozott új honlap tartalomkínálatával (mely egyben keresztmetszetet nyújt az összes lényeges gyakorlati webarchiválási és azzal összefüggő kommunikációs tevékenységről), valamint a bárki által személyes webarchiváláshoz használható szoftverek áttekintésével.

Végezetül arról szeretnék szót ejteni, hogy az internet fejlődése számos nyitott kérdést rejt, melyek megválaszolása alapvető hatással lehet arra, hogy a weben megjelenő tartalmakat miként lehetséges a jövőben archiválni majd. Megfigyelhető egyfajta egyre gyorsuló széttöredezettség, egyrészt a különféle platformok szintjén, másrészt a nagyhatalmi rivalizálás virtuális kivetüléseként. Amennyiben a számítógépes világháló egységessége az eddigieknél még jobban háttérbe szorulna, az az archiválás szemszögéből is új helyzetet teremthet.

A webarchiválásnak mint üzleti tevékenységek a tárgya eddig csak nagyon korlátozottan nyert teret. Amennyiben e terület üzletileg esetleg felértékelődne a jövőben, s nagy üzleti súlyú szereplők is megjelennek a szolgáltatásaikkal, az pedig az eddig a döntően a szintéren feltűnő közgyűjtemények, illetve egyéb nonprofit szereplők tevékenységi körének átértékelésével járhat majd.

Halvány elképzelésekkel rendelkezünk tehát arról, hogy mit hozhat a jövő. Egy azonban biztos. A 2017-ben elindult webarchiválási gyűjteményépítési tevékenység hamarosan szintet fog lépni, s a webarchívum mint gyűjtemény gyorsan bővülő anyaga remélhetőleg egyre több kutató érdeklődését kelti majd fel és sokszínű kutatási együttműködési lehetőségek kialakításának esélyét rejtheti magában.

Summary

The Phd thesis is focusing on web-archiving issues. By offering a broad overview of theoretical and practical elements of this discipline mainly in a Hungarian context, however in international perspective, the main aim is to facilitate further research and development activities in this field.

The thesis offers an introduction to some basic conceptions and major challenges of web-archiving. A contextual analysis is being offered about web archiving related to the library, archive, and museum fields. A brief description of the international context is also being described.

The core of the thesis has built on a detailed analysis about the initial conception, professional frameworks, and workflows of web-archiving in the National Széchényi Library, Hungary as the first organized project in that field in this country. The description of major workflow elements, the challenges in collection development and metadata management fields are being described together with the international context of the project. A brief outlook to the main elements of the software and hardware infrastructure, an introduction the challenges in long-term digital preservation context and a brief description of the legal framework

together with an outlook to the communication conception of the project are also essential elements of this analysis.

A whole chapter is focusing on the educational context of web archiving by international and Hungarian perspectives. An essential part of the thesis is describing the research perceptions in web archiving field by introducing various new sub-disciplines in this context and offering an overview to theoretical and practical research activities in the future related to the use of semantic microformats.

Bibliográfia

1358/2020. (VII. 1.) Korm. Határozat (2020). Letöltés helye:

<https://net.jogtar.hu/jogszabaly?docid=A20H1358.KOR&txtreferer=00000001.txt>

2020. Évi XXXII. törvény, Pub. L. No. 32. (2020). Letöltés helye:

http://njt.hu/cgi_bin/njt_doc.cgi?docid=219746.383606

A Kormány 195/2019. (VIII. 1.) Korm. Rendelete (2020). Letöltés helye:

<https://magyarkozlony.hu/dokumentumok/b880213d783e1a39ea4b0a3c86290da2ec9c5cac/megtekintes>

About IIPC. (2019). Letöltés 2020. augusztus 17., helye:

<http://netpreserve.org/about-us/>

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*. <https://doi.org/10.1007/s10502-019-09325-9>

Adamich, T. (2013). BIBFRAME Take 3: SchemaBibEx, OCLC WorldCat, and Search Optimization. *Technicalities*, 33(5), 9–12. Letöltés helye:

<http://search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=97868634&site=ehost-live>

Adamich, T. (2015). Schema BibEx Extensions: Another 21st Century Paradigm Shift. *Technicalities*, 35(4), 15–17. Letöltés helye:

<http://search.proquest.com/docview/1726721148?accountid=26452>

Adobe Flash. (2020). In *Wikipedia*. Letöltés helye:

https://en.wikipedia.org/w/index.php?title=Adobe_Flash&oldid=972195466

- Alam, S., Weigle, M. C., Nelson, M. L., & Klein, M.; Van de Sompel, H. (2019). Supporting Web Archiving via Web Packaging, 3. arXiv. 1906.07104. Letöltés helye: <https://pure.know.nl/portal/files/12677017/1906.07104v1.pdf>
- Albanese, A. (2020). Internet Archive Answers Publishers' Copyright Lawsuit. *Publishers Weekly (Online)*,. Letöltés helye: <https://www.publishersweekly.com/pw/by-topic/digital/copyright/article/83971-internet-archive-answers-publishers-copyright-lawsuit.html>
- AlSum, A. (2015). Reconstruction of the US First Website. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries—JCDL '15* (pp. 285–286). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2756406.2756954>
- Anderson, K. E. (2020). Getting acquainted with social networks and apps: Capturing and archiving social media content. *Library Hi Tech News*, 37(2), 18–22. <https://doi.org/10.1108/LHTN-03-2019-0011>
- Androvič, A., Bizík, A., Hausleitner, P., Katrincová, B., Lacková, I., & Matúšková, J. (2017). Digitálne pramene—Národný projekt zberu a archivácie v roku 1. *Knihovna PLUS*, (1), 1–14.
- Androvič, A., & Prókai, M. (2007). Web-archívum made in Slovakia: Kísérleti projekt az elektronikus információforrások gyűjtésére és archiválására A web mint kulturális örökség. *Tudományos és műszaki tájékoztatás*, 54(10). Letöltés helye: <https://tmt.omikk.bme.hu/tmt/article/view/6656/7660>
- Archive-friendly website – MIA WIKI. (2020). Letöltés 2020. augusztus 17., helye: https://webarchivum.oszk.hu/mediawiki/index.php?title=Archive-friendly_website

Archive-It—Web Archiving Services for Libraries and Archives. (2020). Letöltés 2020. augusztus 17., helye: <https://archive-it.org/>

Asking questions with web archives – introductory notebooks for historians—
IIPC. (2020). Letöltés 2020. augusztus 12., helye:
<http://netpreserve.org/projects/jupyter-notebooks-for-historians/>

Az Országos Széchényi Könyvtár gyűjtőköri szabályzata, Pub. L. No. F18/2015
(2015). Letöltés ideje 2020. augusztus 12., helye:
http://www.oszk.hu/sites/default/files/f_18_2015_gyujtokori_szabalyzat.pdf

Bailey, S., Thomson, D., Szalóki, G. (2006). Az első nyilvános webarchívum az
Egyesült Királyságban. *Tudományos és műszaki tájékoztatás*, 53(10). Letöltés
helye: <https://tmt.omikk.bme.hu/tmt/article/view/5761/6771>

Belovari, S. (2017). Historians and Web Archives. *Archivaria*, (83), 59–79.
Letöltés helye:
<https://archivaria.ca/index.php/archivaria/article/view/13600/14985>

Ben-David, A. (2016). What does the Web remember of its deleted past? An
archival reconstruction of the former Yugoslav top-level domain. *New Media
& Society*, 18(7), 1103-1119. 1103-1119. doi:10.1177/1461444816643790.
Letöltés helye: <https://doi.org/10.1177/1461444816643790>

Ben-David, A. (2020). Counter-archiving Facebook. *European Journal of
Communication*, 35(3), 249–264. doi: 10.1177/0267323120922069. Letöltés
helye: <https://doi.org/10.1177/0267323120922069>

Ben-David, A., Amram, A., & Bekkerman, R. (2018). The colors of the national
Web: Visual data analysis of the historical Yugoslav Web domain.

International Journal on Digital Libraries, 19(1), 95–106.

<https://doi.org/10.1007/s00799-016-0202-6>

Beresford, P. (2007). Web Curator Tool. *Ariadne*, (50). Letöltés 2020. augusztus 14., helye: <http://www.ariadne.ac.uk/issue/50/beresford/>

Berlin, J. (2017). *Wail/Electron*. JavaScript. Letöltés helye: <https://github.com/N0taN3rd/wail> (Original work published 2016)

Bib.schema.org-1.0—Schema Bib Extend Community Group. (2020). Letöltés 2020. augusztus 14., helye: <https://www.w3.org/community/schemabibex/wiki/Bib.schema.org-1.0>

Bingham, N. (2014). Quality Assurance Paradigms in Web Archiving Pre and Post Legal Deposit. *Alexandria: The Journal of National and International Library and Information Issues*, 25(1–2), 51–68. doi: 10.7227/ALX.0020. Letöltés helye: <https://doi.org/10.7227/ALX.0020>

Bingham, N. (2019, július 6). British Library Access Policies: Challenges and Approaches [Presentation]. Letöltés 2020. augusztus 20., helye: <https://digital.library.unt.edu/ark:/67531/metadc1609016/>

Bratislava University Library. (2016). Webdepozit. Letöltés 2020. január 8., helye: <http://www.webdepozit.sk>

British Library. (2019). UK Web Archive. Letöltés 2020. január 8., helye: <http://webarchive.org.uk>

British Library. (2020a). Save a UK website. Letöltés 2020. augusztus 17., helye: <https://www.webarchive.org.uk/en/ukwa/info/nominate>

British Library. (2020b). UK Web Archive—Technical information. Letöltés 2020. augusztus 17., helye:

<https://www.webarchive.org.uk/en/ukwa/info/technical>

Brozzler – MIA WIKI. (2020). Letöltés 2020. augusztus 18., helye:

<https://webarchivum.oszk.hu/mediawiki/index.php/Brozzler>

Brunelle, J. F., Kelly, M., Salaheldeen, H., Weigle, M. C., & Nelson, M. L. (2015).

Not all mementos are created equal: Measuring the impact of missing resources. *International Journal on Digital Libraries*, 16(3–4), 283–301. doi: 10.1007/s00799-015-0150-6. Letöltés helye:

<http://dx.doi.org/10.1007/s00799-015-0150-6>

Brunelle, J. F., Kelly, M., Weigle, M. C., & Nelson, M. L. (2016). The impact of

JavaScript on archivability. *International Journal on Digital Libraries*, 17(2), 95–117. doi: 10.1007/s00799-015-0140-8. Letöltés helye:

<https://doi.org/10.1007/s00799-015-0140-8>

Brunelle, J. F., Weigle, M. C., & Nelson, M. L. (2017). Archival Crawlers and

JavaScript: Discover More Stuff but Crawl More Slowly. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 1–10). Piscataway, NJ, USA: IEEE Press. Letöltés helye:

<http://dl.acm.org/citation.cfm?id=3200334.3200336>

Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, 11(1–2), 115–132. doi: 10.1177/1461444808099574.

Letöltés helye: <https://doi.org/10.1177/1461444808099574>

Brügger, N. (2013). Web historiography and Internet Studies: Challenges and perspectives. *New Media & Society*, 15(5), 752–764. doi:

10.1177/1461444812462852. Letöltés helye:

<https://doi.org/10.1177/1461444812462852>

Brügger, N. (2016a). Digital Humanities in the 21st Century: Digital Material as a Driving Force. *Digital Humanities Quarterly*, 10(3), 17. Letöltés helye:

<http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html>

Brügger, N. (2016b). Introduction: The Web's first 25 years. *New Media & Society*, 18(7), 1059–1065. doi: 10.1177/1461444816643787. Letöltés helye:

<https://doi.org/10.1177/1461444816643787>

Brügger, N. (2018). *HOW TO DIG INTO THE HISTORY OF A NATION'S WEB? THE DEVELOPMENT OF THE DANISH WEB 2006-2015*. Előadás IIPC RWG webinar series, online webinar. Letöltés ideje 2019. december 15., helye:

<https://netpreserve.org/events/iipc-rss-webinar-the-development-of-the-danish-web-2006-2015-2/>

Brügger, N., Goggin, G., Milligan, I., & Schafer, V. (2017). Introduction: Internet histories. *Internet Histories*, 1(1–2), 1–7. doi:

10.1080/24701475.2017.1317128. Letöltés helye:

<https://doi.org/10.1080/24701475.2017.1317128>

Brügger, N., & Laursen, D. (2019, június 6). A National Web Trend Index [Presentation]. Letöltés 2020. augusztus 20., helye:

<https://digital.library.unt.edu/ark:/67531/metadc1608974/>

Brügger, N., Laursen, D., & Schafer, V. (2019, június 6). Opportunities and Challenges in Collecting and Studying National Webs [Presentation]. Letöltés 2020. augusztus 20., helye:

<https://digital.library.unt.edu/ark:/67531/metadc1609008/>

- Brügger, N., & Schroeder, R. (Szerk.). (2017). *The Web as History: Using Web Archives to Understand the Past and the Present* (1. kiad.). United States, North America: UCL Press.
- Cartledge, C., & Nelson, M. (2015). When should I make preservation copies of myself? *International Journal on Digital Libraries*, 16(3/4), 183–205. Letöltés helye: <http://10.0.3.239/s00799-015-0155-1>
- Chrislane Lupovici. (2006). Web Crawling: The Bibliothèque nationale de France's Experience. *International Cataloging & Bibliographic Control*, 35(4), 82–84.
- Clifton, G. (2005, március 2). Comments on IIPC web archiving metadata set (v 0.1) [Text]. Letöltés 2020. augusztus 20., helye: <https://digital.library.unt.edu/ark:/67531/metadc1457772/>
- Conifer. (2020). Letöltés 2020. augusztus 18., helye: <https://conifer.rhizome.org>
- Cowls, J. (2015, április 27). Research Using Big UK Domain Data [Presentation]. Letöltés 2020. augusztus 20., helye: <https://digital.library.unt.edu/ark:/67531/metadc1476399/>
- Crawler-friendly website – MIA WIKI. (2020). Letöltés 2020. augusztus 17., helye: https://webarchivum.oszk.hu/mediawiki/index.php?title=Crawler-friendly_website
- Cunningham, A., & Phillips, M. (2005). Accountability and accessibility: Ensuring the evidence of e-governance in Australia. *Aslib Proceedings*, 57(4), 301–317. doi: 10.1108/00012530510612059. Letöltés helye: <https://doi.org/10.1108/00012530510612059>

Czech National Library. (2015a). Comprehensive Harvest- Czech Webarchive.

Letöltés 2020. január 8., helye:

<http://www.webarchiv.cz/en/comprehensive-harvests>

Czech National Library. (2015b). Czech webarchive-topic collections. Letöltés

helye: <http://www.webarchiv.cz/en/topic-collections>

Czech Webarchive. (2011). Webarchiv.cz. Letöltés 2020. január 8., helye:

<http://www.webarchiv.cz>

Dancs, Sz. (2011). Webarchiválási politikák. *Könyv, könyvtár, könyvtáros*,

20(10), 14–20. Letöltés helye:

http://www.epa.hu/01300/01367/00248/pdf/EPA01367_3K_2011_10.pdf

Danmarks Kulturministeriet. (2015). Pliktaflevering lov. Letöltés 2020. január 8.,

helye: <http://pliktaflevering.dk/loven/index.htm>

Davis, R. C. (2016). The Future of Web Citation Practices. *Behavioral & Social Sciences Librarian*, 35(3), 128–134. doi: 10.1080/01639269.2016.1241122.

Letöltés helye: <http://dx.doi.org/10.1080/01639269.2016.1241122>

Decman, M. (2011). Problems of Long-Term Preservation of Web Pages TT -

Problematika Dolgorocne Hrambe Spletnih Strani. *Knjiznica*, 55(1), 193–208.

Letöltés helye:

<https://search.proquest.com/docview/1266143497?accountid=27464>

Deschamps, R. (2019, március 12). Exploring Web Archival Data through

Archives Unleashed Cloud Jupyter Notebooks. Letöltés 2020. augusztus 21.,

helye: <https://news.archivesunleashed.org/exploring-web-archival-data-through-archives-unleashed-cloud-jupyter-notebooks-7605c6ca2b33>

- Di Pretoro, E., & Geeraert, F. (2019). Behind the Scenes of Web Archiving: Metadata of Harvested Websites. In Depoortere, R., Gheldof, T., Styven, D., & Van, J.Eycken, D. (Szerk.), *Press, Trust and Understanding: The value of metadata in a digitally joined-up world* (pp. 63–74). Brussels: Archives et Bibliothèques de Belgique - Archief- en Bibliotheekwezen in België. Letöltés helye: <https://hal.archives-ouvertes.fr/hal-02124714>
- Drótos, L. (2006). Mi a MIA? Javaslat egy magyar Internet Archívum létrehozására. *Tudományos és műszaki tájékoztatás*, 53(6), 267–274.
- Drótos, L. (2017a). Az internet archiválása mint könyvtári feladat. *Tudományos és műszaki tájékoztatás*, 64(7–8), 361–371.
- Drótos, L. (2017b). OpenWayback. MIA WIKI. Letöltés 2020. február 23., helye: <http://mekosztaly.oszk.hu/mediawiki/index.php/Wayback>
- Drótos, L. (2017c). *OSZK webarchívum pilot koncepció -2017-04-17*.
- Drótos, L. (2018a). *Webarchívum „gépház”*. Prezentáció, Budapest. Letöltés ideje 2018. december 20; helye: http://mekosztaly.oszk.hu/mia/doc/Webarchivum_gephaz.pptx
- Drótos L. (2018b). Webes tartalmak digitális megőrzése. *Könyv, könyvtár, könyvtáros*, 27(10), 11–17. Letöltés helye: https://epa.oszk.hu/01300/01367/00307/pdf/EPA01367_3K_2018_10_011-017.pdf
- Drótos L., & Németh M. (2018a). A webarchiválás oktatása = The education of web-archiving. In *NETWORKSHOP 2018 konferenciakiadvány* (pp.31–37). Budapest: HUNGARNET Egyesület. doi: 10.31915/NWS.2018.4 Letöltés helye: <https://doi.org/10.31915/NWS.2018.4>

Drótos, L., & Németh, M. (2018b). Web museum, web library, web archive The responsibility of public collections to preserve digital culture. In Petrovska, L., Īvāne-Kronberga, B., & Meldere, Z. (Szerk.), *The Power of Reading: Proceedings of the XXVI Bobcatsss Symposium, Riga, Latvia, January 2018* (pp. 124–126). Riga: The University of Latvia Press. Letöltés helye: http://blogi.lu.lv/bobcatsss2018/files/2018/08/BOBCATSSS_2018_TheProceedings.pdf

Drótos, L., & Visky, Á. L. (2018a). *Útmutató a szelektíven archivált egyedi webhelyek XML-alapú metaadatolásához*. Letöltés ideje 2020. november 14; helye: https://docs.google.com/document/d/11nGzbRgFk9RnyK_05IKBcud01F-eli2mT03aqCOTIGk/

Drótos, L., & Visky, Á. L. (2018b). *Útmutató az archivált webhelyekből/weboldalakból álló részgyűjtemények XML-alapú metaadatolásához*. Letöltés ideje 2020. november 14; helye: https://docs.google.com/document/d/1NlgWfzHLS0AzcVEXweuSHBq9-AkvZFbQw_totLoEG7I/

Drótos, L., Németh M., Visky, Á.L. (2018). Alapozó ismeretek. Letöltés ideje, 2020. január 8; helye: http://mekosztaly.oszk.hu/mia/e-learning/internet-archivalas_1modul.html

Dynamic HTML. (2020). In *Wikipedia*. Letöltés helye: https://en.wikipedia.org/w/index.php?title=Dynamic_HTML&oldid=970834686

Dynamic web page. (2020). In *Wikipedia*. Letöltés helye: https://en.wikipedia.org/w/index.php?title=Dynamic_web_page&oldid=967203212

- Egense, T., & Klindt Myrvoll, A. (2018, november 15). SolrWayback: Demo of the SolrWayback search interface, tools and playback engine for WARCs [Presentation]. Letöltés 2020. augusztus 20., helye: <https://digital.library.unt.edu/ark:/67531/metadc1477121/>
- Elson, A. K. (2020). Getting acquainted with social networks and apps: Capturing and archiving social media content. *Library Hi Tech News*, 37(2), 18–22. doi: 10.1108/LHTN-03-2019-0011. Letöltés helye: <https://doi.org/10.1108/LHTN-03-2019-0011>
- Enis, M. (2015). Ending the Invisible Library | Linked Data. *Library Journal*. doi: 10.1016/S0022-3913(12)00047-9. Letöltés helye: [https://doi.org/10.1016/S0022-3913\(12\)00047-9](https://doi.org/10.1016/S0022-3913(12)00047-9)
- Espley, S., Carpentier, F., Pop, R., & Medjkoune, L. (2014). Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content. *Alexandria: The Journal of National and International Library and Information Issues*, 25(1–2), 31–50. doi: 10.7227/ALX.0019. Letöltés helye: <https://doi.org/10.7227/ALX.0019>
- Estonian National Library. (2008). About Estonian Web Archiving. Letöltés 2019. január 8., helye: <https://www.nlib.ee/en/content/estonian-web-archive>
- Fansler, C., Gilbertson, K., & Petersen, R. (2014). The Missing Link: Observations on the Evolution of a Web Archive. *Journal for the Society of North Carolina Archivists*, 11(1), 46–59. Letöltés helye: http://www.ncarchivists.org/wp-content/uploads/2017/08/JSNCA_2014_Vol11No1_Fansler-Gilbertson-Petersen.pdf
- Farag, M., Nakate, P., & Fox, E. A. (2016). Big Data Processing of School Shooting Archives. In *Proceedings of the 16th ACM/IEEE-CS on Joint*

- Conference on Digital Libraries—JCDL '16* (pp. 271–272). New York, New York, USA: ACM Press. doi: 10.1145/2910896.2925466. Letöltés helye: <https://doi.org/10.1145/2910896.2925466>
- Final report. Asking questions with web archives—Introductory notebooks for historians.* (2020). Letöltés 2020. december 10., helye: <http://netpreserve.org/projects/jupyter-notebooks-for-historians/>
- Finnemann, N. O. (2019). Web Archive. *KNOWLEDGE ORGANIZATION*, 46(1), 47–70. doi: 10.5771/0943-7444-2019-1-47. Letöltés helye: <https://doi.org/10.5771/0943-7444-2019-1-47>
- Flanagan, G. P. (2010). *Digital Preservation and Authentic Legal Information* (SSRN Scholarly Paper No. ID 2463288). Rochester, NY: Social Science Research Network. doi: 10.2139/ssrn.2463288. Letöltés helye: <https://doi.org/10.2139/ssrn.2463288>
- Fritz, S., & Milligan, I. (2018). Archive-It Blog – Analyze your Web Archives at Scale: The Archives Unleashed Cloud. Letöltés 2020. augusztus 21., helye: <https://archive-it.org/blog/post/analyze-your-web-archives-at-scale-the-archives-unleashed-cloud/>
- Getting started with schema.org using Microdata. (2016). Letöltés 2015. január 1., helye: <http://schema.org/docs/gs.html>
- Getting Started—Schema.org. (2020). Letöltés 2020. augusztus 14., helye: <https://schema.org/docs/gs.html>
- Gomes, D., & Costa, M. (2014). The Importance of Web Archives for Humanities. *International Journal of Humanities and Arts Computing*, 8(1), 106–123. doi: 10.3366/ijhac.2014.0122. Letöltés helye: <https://doi.org/10.3366/ijhac.2014.0122>

Government Web& Data Archive. (2017). Letöltés 2020. január 6., helye:

<https://digital2.library.unt.edu/nomination/GWDA/about/>

Grab Them All. (2017). Letöltés 2020. február 23., helye:

http://mekosztaly.oszk.hu/mediawiki/index.php/Grab_Them_All

Grimshaw, J. (2016). UK Official Publications: Managing the Transition to

Electronic Deposit at the British Library. *Legal Information Management*,

16(1), 3–9. doi: 10.1017/S1472669616000037. Letöltés helye:

<http://dx.doi.org/10.1017/S1472669616000037>

Gulyás, L., Jurányi, Z., Soós, S., & Kampis, G. (2014). Can web presence predict academic performance? In *Proceedings of the 23rd International Conference on World Wide Web—WWW '14 Companion* (pp. 1183–1188). New York,

New York, USA: ACM Press. doi: 10.1145/2567948.2579037. Letöltés helye:

<https://doi.org/10.1145/2567948.2579037>

Halbert, M., Skinner, K., Wilson, M., & Zarndt, F. (2016). Here Today, Gone

within a Month: The Fleeting Life of Digital News. In *IFLA WLIC 2016 –*

Columbus, OH – Connections. Collaboration. Community in Session S21—

Satellite Meeting: News Media. In: News, new roles & preservation

advocacy: Moving libraries into action, 10-12 August 2016, Lexington, KY,

USA. Lexington, KY, USA: IFLA -- International Federation of Library

Associations and Institutions. Letöltés helye:

<http://library.ifla.org/id/eprint/2077>

Hale, S. A., Yasseri, T., Cows, J., Meyer, E. T., Schroeder, R., & Margetts, H.

(2014). Mapping the UK Webspace: Fifteen Years of British Universities on the Web. In *Proceedings of the 2014 ACM Conference on Web Science* (pp. 62–70). New York, NY, USA: ACM. doi: doi.org/10.1145/2615569.2615691.

Letöltés helye: <https://doi.org/10.1145/2615569.2615691>

- Haskovcova, M. (2019). Webarchiv. Czech National Library. Letöltés ideje 2020. december 14. helye: https://webarchivum.oszk.hu/wp-content/uploads/2020/03/DH_Budapest_2019_Marie_Haskovcova.pdf
- Hegyközi, I. (2014). Hol tart ma a webarchiválás? Szakirodalmi áttekintés. *Könyvtári figyelő*, 60(4), 527–534. Letöltés helye: http://epa.oszk.hu/00100/00143/00095/pdf/EPA00143_konyvtari_figyelo_2014_4_527-534.pdf
- Heritrix—IA Webteam Confluence. (2020). Letöltés 2020. augusztus 18; helye: <https://webarchive.jira.com/wiki/spaces/Heritrix/overview>
- Holub, K., & Rudomino, I. (2015). A decade of web archiving in the National and University Library in Zagreb. In *Preservation and Conservation with Information Technology. IFLA 2015 South Africa*. Cape Town: IFLA -- International Federation of Library Associations and Institutions. Letöltés ideje: 2020 november 15; helye: <http://library.ifla.org/id/eprint/1092>
- Holzmann, H., Goel, V., & Anand, A. (2016). ArchiveSpark. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries—JCDL '16* (pp. 83–92). New York, New York, USA: ACM Press. doi: 10.1145/2910896.2910902 Letöltés helye: <https://doi.org/10.1145/2910896.2910902>
- Holzmann, H., Goel, V., & Gustainis, E. N. (2017). *Universal distant reading through metadata proxies with archivespark*. 2017 IEEE International Conference on Big Data (Big Data), Big Data (Big Data), 2017 IEEE International Conference on (p. 459). IEEE. doi: 10.1109/BigData.2017.8257958. Letöltés helye: <https://doi.org/10.1109/BigData.2017.8257958>

- Horváth, Á. (2016). *RDFa—Schema.org: Unity of document and semantic web*.
Előadás Networkshop 2016, Debrecen. Letöltés ideje: 2020 november 14;
helye:
<https://conference.niif.hu/event/5/session/10/contribution/27/material/slides/0.ppt>
- How and when to use Brozzler. (2020). Letöltés 2020. augusztus 18; helye:
<http://support.archive-it.org/hc/en-us/articles/360000351986>
- HTML 5.2. (2017). Letöltés 2020. augusztus 15; helye:
<https://www.w3.org/TR/html52/>
- HTTPS. (2020). In *Wikipedia*. Letöltés helye:
<https://en.wikipedia.org/w/index.php?title=HTTPS&oldid=971569450>
- HTTrack Website Copier—Free Software Offline Browser (GNU GPL). (2020).
Letöltés 2020. augusztus 18; helye: <https://www.httrack.com/>
- HTTRack—MIA WIKI. (2017). Letöltés 2020. február 23; helye:
<http://mekosztaly.oszk.hu/mediawiki/index.php/HTTrack>
- IFLA Library. (2017). Session 186—A memory bank for the digital society:
National libraries and web archiving—National Libraries with Information
Technology. Letöltés 2020. augusztus 17; helye:
<http://library.ifla.org/view/conferences/2017/2017-08-23/798.html>
- IIPC. (2017). OpenWayback. Letöltés 2020. február 23; helye:
<https://github.com/iipc/openwayback/wiki>
- IIPC. (2019). IIPC portal. Letöltés 2019. január 28; helye:
<https://netpreserve.org>

IIPC. (2020a). Legal Deposit. Letöltés 2019. január 28; helye:

<http://netpreserve.org/web-archiving/legal-deposit/>

IIPC - Collection development policies. (2018). Letöltés 2020. augusztus 17;

helye: <http://netpreserve.org/web-archiving/collection-development-policies/>

IIPC Training Working Group. (2020). Training materials. Letöltés 2020.

augusztus 17; helye: <http://netpreserve.org/web-archiving/training-materials/>

IIPC -awesome-web-archiving. (2020). IIPC. Letöltés 2020. augusztus 17; helye:

<https://github.com/iipc/awesome-web-archiving>

IIPC Content Development Working Group. (2019). Letöltés 2020. augusztus 17;

helye: <http://netpreserve.org/about-us/working-groups/content-development-working-group/>

IIPC -Országos Széchényi Könyvtár. (2018). Letöltés 2020. augusztus 17; helye:

<http://netpreserve.org/about-us/members/orszagos-szechenyi-konyvtar/>

IIPC, R. W. G. (2020b). Research Working Group—IIPC. Letöltés 2020. augusztus

12; helye: <https://netpreserve.org/about-us/working-groups/research-working-group/>

IIPC RSS LinkGate Webinar. (2020). Letöltés 2020. augusztus 13; helye:

https://docs.google.com/presentation/d/1mYSciOvbU9Hm3jsMSJgioSVr3ZGuVkvYr10HnHXJwXA/edit#slide=id.g8cec5e7f8a_0_67

IIPC Training Video Case Studies. (2020). Letöltés 2020. augusztus 18; helye:

<http://netpreserve.org/web-archiving/iipc-training-video-case-studies/>

IIPC TWG. (2017). IIPC Training Working Group survey. Letöltés 2018. június 12;
helye: <https://www.surveymonkey.com/r/V7MVXXW>

IIPC TWG. (2019). IIPC Training Working Group. Letöltés 2019. január 28; helye:
<http://netpreserve.org/about-us/working-groups/training-working-group/>

IIPC Web Archiving Metadata Set—UNT Digital Library. (2005). Letöltés 2020.
augusztus 20; helye:
<https://digital.library.unt.edu/ark:/67531/metadc1457778/>

IIPC Working groups. (2019). Letöltés 2020. augusztus 17; helye:
<http://netpreserve.org/about-us/working-groups/>

Ilácsa, Sz. (2020). A webarchiválás metaadatolásának problémái. *Könyvtári
figyelő*, 30. (66.)(1), 74–82. Letöltés helye:
https://epa.oszk.hu/00100/00143/00359/pdf/EPA00143_konyvtari_figyelo_2020_01_074-082.pdf

International Internet Preservation Consortium (IIPC) repository. (2020).
Letöltés 2020. augusztus 2; helye:
<https://digital.library.unt.edu/explore/partners/IIPC/>

Internet Archive. (2019). Internet Archive portal. Letöltés 2019. január 28;
helye: <http://www.archive.org>

Jackie, D., & Bowers, K. (2018). *Descriptive Metadata for Web Archiving:
Recommendations of the OCLC Research Library Partnership Web Archiving
Metadata Working Group*. Letöltés helye:
<https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata/recommendations.html>

- Jackson, A. (2015, április 27). Ten Years of the Uk Web Archive: What Have We Saved? [Presentation]. Letöltés 2020. augusztus 20; helye: <https://digital.library.unt.edu/ark:/67531/metadc1476398/>
- Jackson, A. (2018, november 13). Continuous, Incremental, Scalable, Higher-Quality Web Crawls with Heritrix [Presentation]. Letöltés 2020. augusztus 20; helye: <https://digital.library.unt.edu/ark:/67531/metadc1477144/>
- Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y., & Tanaka, K. (2006). Journey to the past. In *Proceedings of the seventeenth conference on Hypertext and hypermedia—HYPERTEXT '06* (p. 135). New York, New York, USA: ACM Press. doi: 10.1145/1149941.1149969. Letöltés helye: <https://doi.org/10.1145/1149941.1149969>
- Jatowt, A., Kawai, Y., Ohshima, H., & Tanaka, K. (2008). What can history tell us? In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia—HT '08* (p. 5). New York, New York, USA: ACM Press. <https://doi.org/10.1145/1379092.1379098>
- Jatowt, A., Kawai, Y., & Tanaka, K. (2007). Detecting Age of Page Content. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management* (pp. 137–144). New York, NY, USA: ACM. doi: 10.1145/1316902.1316925. Letöltés helye: <https://doi.org/10.1145/1316902.1316925>
- Java applet. (2020). In *Wikipedia*. Letöltés helye: https://en.wikipedia.org/w/index.php?title=Java_applet&oldid=972322018
- Johnson, P. (2016). Where Did All the Information Go? Well at Least the Important Stuff. *Refer*, 32(2), 8–12. Letöltés helye: <https://search.proquest.com/docview/1803538845?accountid=27464>

Jupyter notebooks for web archives. (2020). Letöltés 2020. augusztus 12; helye:
<https://slides.com/wragge/iipc-jupyter>

Jurányi, Zs. (2013). Heritrix 3 konfigurációs kézikönyv. Letöltés 2020. február 23;
helye: <https://juzraai.github.io/blog/2013/heritrix-3-konfig-receptkonyv/>

Kampis, G. (2017). *A MIA pilot rövid bemutatása*. Előadás 404 Not Found” workshop, Budapest. Letöltés ideje: 2020 március 14; helye:
https://webarchivum.oszk.hu/wp-content/uploads/2020/03/Kampis_Gyorgy_MIA_pilot_GK_ea.pptx

Kampis, G., & Gulyás, L. (2013). Big is small, and changes slowly in Hungary. In *Coginfo 2013 Conference*.

Kelly, M. (2020). *Wail—Github*. Python. Letöltés helye:
<https://github.com/machawk1/wail> (Original work published 2013)

Klasinc, J. (2019). Web Archiving Overview: National and University Library - Slovenia [Presentation]. Letöltés 2020. augusztus 20., helye:
<https://digital.library.unt.edu/ark:/67531/metadc1609023/>

Klein, M., Balakireva, L., & de Sompel, H. (2018). Focused Crawl of Web Archives to Build Event Collections. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 333–342). New York, NY, USA: ACM. doi. 10.1145/3201064.3201085. Letöltés helye:
<https://doi.org/10.1145/3201064.3201085>

Klein, M., Shankar, H., Balakireva, L., & Van de Sompel, H. (2019). The Memento Tracer Framework: Balancing Quality and Scalability for Web Archiving. In Doucet, A., Isaac, A., Golub, K., Aalberg, T., & Jatowt, A. (Szerk.), *Digital Libraries for Open Knowledge. TPDL 2019. Lecture Notes in Computer Science, vol 11799*. (pp. 163–176). Cham: Springer International Publishing.

doi: 978-3-030-30760-8_15. Letöltés helye: https://doi.org/10.1007/978-3-030-30760-8_15

Kokas, K., & Drótos, L. (2018). Webarchiválás és a történeti kutatások. *Digitális Bölcsészet*, 1(1), 35–55. doi: 10.31400/dh-hun.2018.1.129 Letöltés helye: <https://doi.org/10.31400/dh-hun.2018.1.129>

Kornhoffer, M. (2010a). *100 nap az élet: A világhálón található információk gyűjtésének és megőrzésének áttekintése hazánkban, valamint Kelet- és Közép-Európában*. Pécs: Pécsi Tudományegyetem. Letöltés helye: <https://redmine.oszk.hu/documents/82>

Kornhoffer, M. (2010b). *A világhálón található információk gyűjtésének és megőrzésének hazai és nemzetközi áttekintése* (BA szakdolgozat). Pécsi Tudományegyetem. Letöltés helye: <http://mek.oszk.hu/10400/10424>

Kornhoffer, M. (2011). Internet-archívumok hazánkban és Közép-Európában. *Felderítő Szemle*, 10(3–4), 63–78.

Kõuts, J. (2012). Estonian Web Archive: Preserving the Estonian Mind [Presentation]. Letöltés 2020. augusztus 20; helye: <https://digital.library.unt.edu/ark:/67531/metadc1638385/>

Kvasnica, J. (2016). Vědecké využití dat z webových archivů. *Knihovna: knihovnická revue*[online]., 27(2), 23–34. Letöltés helye: <https://knihovnarevue.nkp.cz/archiv/2016-2/recenzovane-prispevky/vedecke-vyuziti-dat-z-webovych-archivu>

Kvasnica, J., & Kreibich, R. (2013). Formátová analýza sklizených dat vrámci projektu webarchiv NK ČR. (Czech). *File Format Recognition of Data Harvested by Web Archiving Project of National Library of the Czech Republic. (English)*, (2), 1.

Leach-Murray, S. (2018). Archive-It. *Technical Services Quarterly*, 35(2), 214.

Letöltés helye:

<http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds>

Leadership. (2019). Letöltés 2020. augusztus 17; helye:

<http://netpreserve.org/about-us/leadership/>

Lee, J. (2017). Where should the culture of our lives and memory be preserved?

- Rethinking the role of the library. In *IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity. Society. In Session 189—Asia and Oceania*. Wrocław:

IFLA -- International Federation of Library Associations and Institutions.

Letöltés helye: <http://library.ifla.org/1691/>

Lerner, A., Kohno, T., & Roesner, F. (2017). Rewriting History. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security—CCS '17* (pp. 1741–1755). New York, New York, USA: ACM Press.

doi: 10.1145/3133956.3134042. Letöltés helye:

<https://doi.org/10.1145/3133956.3134042>

LinkGate: Core Functionality and Future Use Cases—IIPC. (2020). Letöltés 2020.

augusztus 12; helye: <http://netpreserve.org/projects/LinkGate/>

Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., ... Wrubel,

L. (2018). API-based social media collecting as a form of web archiving.

International Journal on Digital Libraries, 19(1), 21–38. doi: 10.1007/s00799-

016-0201-7. Letöltés helye: <http://dx.doi.org/10.1007/s00799-016-0201-7>

Liu, J. P. (2016). Fair Use, Notice Failure, and the Limits of Copyright as

Property. *Boston University Law Review*, 96(3), 833–856. Letöltés helye:

<http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=118226656&lang=hu&site=ehost-live>

Live web leakage – MIA Wiki. (2017). Letöltés 2020. augusztus 18; helye:
https://webarchivum.oszk.hu/mediawiki/index.php?title=Live_web_leakage

Lnenicka, M., Hovad, J., & Komarkova, J. (2015). A Proposal of a Big Web Data Application and Archive for the Distributed Data Processing with Apache Hadoop. In Núñez, M., Nguyen, N. T., Camacho, D., & Trawiński, B. (Szerk.), *Computational Collective Intelligence. Lecture Notes in Computer Science, vol 9330* (pp. 285–294). Cham: Springer International Publishing. doi: 10.1007/978-3-319-24306-1_28. Letöltés helye:
https://doi.org/10.1007/978-3-319-24306-1_28

Locatelli, E. (2017). The role of Internet Wayback Machine in a multi-method research project. In *“Researchers, practioners and their use of the archived web”, London, School of Advanced Study, University of London*. London.

Cerbová L. (2009). A cseh köteleespéldány rendelet és a web. *Könyvtári figyelő*, 55(3), 518–520. Letöltés helye:
https://epa.oszk.hu/00100/00143/00072/pdf/2009_3%20szam_referatumo k.pdf

Macnaught, B. (2018). Social Media Collecting at the National Library of New Zealand. In *IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies Session 93—National Libraries and Social Media—Meeting the Challenges of Acquiring, Preserving and Proving Long-Term Access—National Libraries*. Kuala Lumpur: IFLA. Letöltés helye:
<http://library.ifla.org/id/eprint/2274>

- Maeda, N., & Oyama, S. (2017). The Technologies of Web Archiving. *Joho No Kagaku to Gijutsu*, 67(2), 73. Letöltés helye:
<https://search.proquest.com/docview/2007445742?accountid=27464>
- Maemura, E., Becker, C., & Milligan, I. (2016). Understanding computational web archives research methods using research objects. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 3250–3259). IEEE. doi: 10.1109/BigData.2016.7840982. Letöltés helye:
<https://doi.org/10.1109/BigData.2016.7840982>
- Masanès, J. (2005). Web Archiving Methods and Approaches: A Comparative Study. *Library Trends*, 54(1), 72–90. Letöltés helye:
<https://search.proquest.com/docview/220467286?accountid=27464>
- Massis, B. (2016). Libraries and digital memory. *New Library World*, 117(9/10), 673–676. Letöltés helye:
<https://search.proquest.com/docview/1830312026?accountid=27464>
- Matúsková, J., & Hausleitner, Ing. P. (2019). Archiving of Websites and Born Digital Documents in Slovakia. In *8th Colloquium Library and Information Experts of the V4+ Countries*. Letöltés: 2020. január 14; helye:
https://kolokvium.ulib.sk/wp-content/uploads/2019/06/B32_P_Matuskova_SK.pdf
- Mayr, M., & Predikaka, A. (2016). Nationale Grenzen im World Wide Web – Erfahrungen bei der Webarchivierung in der Österreichischen Nationalbibliothek. *Bibliothek Forschung Und Praxis*, 40(1), 90–95. doi: 10.1515/bfp-2016-0007. Letöltés helye: <https://doi.org/10.1515/bfp-2016-0007>

- Michel, A. (2019, július 6). The legal framework for web archiving: Focussing on GDPR [Presentation]. Letöltés 2020. augusztus 20; helye: <https://digital.library.unt.edu/ark:/67531/metadc1608995/>
- Microsoft Silverlight. (2020). Letöltés 2020. augusztus 15; helye: <https://www.microsoft.com/silverlight/>
- Mobile operating system. (2020). In *Wikipedia*. Letöltés helye: https://en.wikipedia.org/w/index.php?title=Mobile_operating_system&oldid=973051284
- National Archives. (1996). UK Government Web Archive. Letöltés: 2020. Január 15; helye: <http://www.nationalarchives.gov.uk/webarchive/>
- National Library of Australia. (2019). TROVE DATABASE. Letöltés ideje: 2020. január 14; helye: <https://trove.nla.gov.au/general/about>
- Németh, M. (2017). Nemzetközi körkép a webarchiválás gyakorlatáról. *Könyvtári figyelő*, 63(4), 575–582. Letöltés helye: https://epa.oszk.hu/00100/00143/00349/pdf/EPA00143_konyvtari_figyelo_2017_04_575-582.pdf
- Németh, M. (2018a). A webarchiválásról történeti megközelítésben. *Könyv, könyvtár, könyvtáros*, 27(2), 48–52. Letöltés helye: <http://ki2.oszk.hu/3k/2018/06/a-webarchivalasrol-torteneti-megkozelitesben/>
- Németh, M. (2018b). *How to catalogue a web archive?* Előadás Information Interactions 2018 workshop, Bratislava. Letöltés ideje 2020. január 8; helye: http://mekosztaly.oszk.hu/mia/doc/How_to_catalogue_a_web_archive.ppt

- Németh, M. (2018c). *The education of web archiving*. Előadás CDA 2018 conference, Bratislava. Letöltés ideje 2020. január 8; helye: http://mekosztaly.oszk.hu/mia/doc/The_education_of_web-archiving_CDA_2018.pptx
- Németh, M. (2019a). *Metadata supported full text search in a web archive*. Előadás IIPC conference, Zagreb. Letöltés ideje 2020. január 8; helye: http://mekosztaly.oszk.hu/mia/doc/Metadata_supported_full-text_search_in_a_web_archive-IIPC-2019-Zagreb.pptx
- Németh, M. (2019b). Nemzetközi tanácskozás az internet megőrzéséről. Letöltés 2020. augusztus 17; helye: https://nemzetikonyvtar.blog.hu/2019/06/18/nemzetkozi_tanacskozas_az_internet_megorzeserol
- Németh, M. (2019c). Using semantic microformats for web archiving – an initial project conception. In Katarina Tomková (Szerk.), *Nové trendy a východiská pri budovaní LTP archívov: Zborník príspevkov zo 4. Medzinárodnej konferencie o dlhodobej archivácii Bratislava, 5. 11. 2019*. (pp. 31–38). Bratislava: Univerzitná knižnica v Bratislave. Letöltés ideje 2020. január 14; helye: <https://cloud.ulib.sk/index.php/s/JITVgF9TmRruFtC>
- Németh, M. (2019d). Web Archives as a Research Subject. In *Information and technology transforming lives: Connection, interaction, innovation proceedings* (pp. 471–478). Osijek: Faculty of humanities and social sciences, University of Osijek. Letöltés helye: http://bobcatsss2019.ffos.hr/docs/bobcatsss_proceedings.pdf
- Németh M. (2019e). Webarchiválás két szakmai rendezvény tükrében. *Könyv, könyvtár, könyvtáros*, 28(6), 26–29.

- Németh M. (2020a). Az OSZK Webarchívum új honlapjának felépítése és szolgáltatásai. *Könyv, könyvtár, könyvtáros*, 29.(6.), 16–26. Letöltés helye: https://epa.oszk.hu/01300/01367/00329/pdf/EPA01367_3K_2020_06_016-026.pdf
- Németh, M. (2020b). Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary [E-paper]. Letöltés helye: https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_Hungary.pdf
- Németh, M. (2020). Webarchívum mint a tudományos kutatások tárgya. *Tudományos és Műszaki Tájékoztatás*, 67(12), 757–765. Letöltés helye: <https://tmt.omikk.bme.hu/tmt/article/view/12804/14543>
- Németh, M., & Drótos, L. (2019a). A blended learning-based curriculum on Web archiving in the National Széchényi Library. *Digital Library Perspectives*, 35(2), 97–114. doi: 10.1108/DLP-03-2019-0012. Letöltés helye: <https://doi.org/10.1108/DLP-03-2019-0012>
- Németh, M., & Drótos, L. (2019b). Metadata Management and Future Plans to Generate Linked Open Data in the Hungarian Web Archiving Pilot Project. *ITLIB*, 2019(2). Letöltés helye: <https://itlib.cvtisr.sk/buxus/docs/38-metadata.pdf>
- Németh, M., & Drótos, L. (2020). Egyedi mentésekre szolgáló webarchiváló szoftverek. *Könyv, könyvtár, könyvtáros*, 29(12), 3–11. Letöltés helye: https://epa.oszk.hu/01300/01367/00334/pdf/EPA01367_3K_2020_12_003-011.pdf

NetArchiveSuite. MIA WIKI (2017). Letöltés 2019. február 23; helye:

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

NetArchiveSuite documentation. (2018). Letöltés 2020. február 23; helye:

<https://sbforge.org/display/NAS/NetarchiveSuite>

Netarkivet. (2015a). Indsamling materialer til Netarkivet. Letöltés 2020. január

8; helye: [http://netarkivet.dk/wp-](http://netarkivet.dk/wp-content/uploads/2015/10/Politik_for_indsamling_af_materiale_til_Netarkivet.pdf)

[content/uploads/2015/10/Politik_for_indsamling_af_materiale_til_Netarkivet.pdf](http://netarkivet.dk/wp-content/uploads/2015/10/Politik_for_indsamling_af_materiale_til_Netarkivet.pdf)

Netarkivet. (2015b). Pligtaflevering til webstedejere. Letöltés 2020. január 8;

helye: <http://netarkivet.dk/til-webstedejere/pligtaflevering/>.

Netarkivet. (2018). Solrwayback. Letöltés 2020. február 23; helye:

<https://github.com/netarchivesuite/solrwayback#about-solrwayback>

Netarkivet. (2020). Netarkivet | Adgang til Netarkivet. Letöltés 2020. augusztus

17; helye: <http://netarkivet.dk/adgang/>

Netlab. (2018a). NETLAB Web Archiving Course Brochure. Letöltés 2019. január

28; helye: <http://www.netlab.dk/wp-content/uploads/2017/04/NetLab-Web-Archiving-Course-Brochure.pdf>

Netlab. (2018b). Netlab-courses. Letöltés 2019. január 28; helye:

<http://www.netlab.dk/services/courses/>

Nielsen, J. (2016). *Using web archives in research – an introduction* (1.). Aarhus:

Netlab. Letöltés helye: [http://www.netlab.dk/wp-](http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf)

[content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf](http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf)

Nimbus screen capture. (2018). Letöltés 2020. február 23; helye:

http://mekosztaly.oszk.hu/mediawiki/index.php/Nimbus_Screen_Capture

- Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, 18(3–4). doi: 10.1045/march2012-niu1. Letöltés helye: <https://doi.org/10.1045/march2012-niu1>
- Nuys, C. van, Albertsen, K., Pedersen, L., & Szalóki, G. (2005). A Paradigma projekt. *Tudományos és műszaki tájékoztatás*, 52(11–12). Letöltés helye: <https://tmt.omikk.bme.hu/tmt/article/view/1372/1838>
- Országos Széchényi Könyvtár. (2018). Pywb search interface -NSZL. Letöltés 2020. február 23; helye: <http://webadmin.oszk.hu/pywb/>
- Osborne, A. (2020). Heritrix wiki. Letöltés 2020. február 23; helye: <https://github.com/internetarchive/heritrix3/wiki>
- OSZK. (2007). Bibliotheca Nationalis Hungariae, „Az Országos Széchényi Könyvtár programja, 2007. Letöltés helye: http://www.oszk.hu/sites/default/files/szakmai_munkaterv_2007_0.pdf
- OSZK Webarchiválási munkacsoport. (2017). *Az OSZK webaratás pilot projektjének gyűjtőköri tervezete*. Letöltés: 2020. augusztus 17; helye: https://webarchivum.oszk.hu/wp-content/uploads/2020/02/Webarchivum_gyujtokor_tervezet.pdf
- OSZK Webarchiválási munkacsoport. (2018). *Részletes beszámoló a 2017-2018 évi webarchiválási pilot projektről*. Letöltés: 2020. augusztus 17; helye: https://webarchivum.oszk.hu/wp-content/uploads/2020/02/Webarchivalas_vezetoi_osszefoglalo_2017-2018.pdf
- Pandora Archive—NLA Selection Guidelines. (2009). Letöltés 2020. augusztus 17; helye: <http://pandora.nla.gov.au/selectionguidelines.html>

- Pandora Archive—Preserving and Accessing Networked Documentary Resources of Australia. (1999, május). Letöltés 2020. augusztus 17; helye: <http://pandora.nla.gov.au/>
- Paterson, J. (2015). *Snapshot of Web harvesting in Canada*. Canada, North America: SelectedWorks. Letöltés helye: <http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds>
- Paul-Choudhury, S. (2011). Your digital legacy. *New Scientist*, 210(2809), 40–43. Letöltés helye: <https://www.newscientist.com/article/mg21028091-400-digital-legacy-the-fate-of-your-online-soul/>
- Peternák, M. (2017). 404 és akiknek nem kell—Médiamúzeum Magyarországon: A „Magyar Tartalom” húsz éves jubileuma alkalmából. Budapest: c3 kulturális Központ. Letöltés: 2020. február 14; helye: http://mekosztaly.oszk.hu/mia/doc/workshop/Peternak_Miklos_404_OSZK_2017_10_13_ea.pptx
- Phillips, M. E., & Phillips, K. K. (2018). End of Term 2016 Presidential Web Archive. *Against the Grain*, 29(6), 27. Letöltés helye: <https://digital.library.unt.edu/ark:/67531/metadc1114874/>
- Puppeteer. (2020). Letöltés 2020. február 23; helye: <https://pptr.dev/>
- Pywb. (2019). Letöltés 2020. február 23; helye: <http://mekosztaly.oszk.hu/mediawiki/index.php/PyWb>
- RealAudio. (2020). In *Wikipedia*. Letöltés helye: <https://en.wikipedia.org/w/index.php?title=RealAudio&oldid=972363209>

- Rich web application. (2020). In *Wikipedia*. Letöltés helye:
https://en.wikipedia.org/w/index.php?title=Rich_web_application&oldid=970947425
- Riigikogu. Legal Deposit Copy Act—Estonia (2016). Estonia. Letöltés helye:
<https://www.riigiteataja.ee/en/eli/514092016001/consolide>
- Riva P., Boeuf P. L., Žumer M., & Szerkesztőbizottsága E. (2017). IFLA könyvtári referenciamodell, 99. Letöltés helye:
https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712-hu.pdf
- Robots exclusion standard—Wikipedia. (2020). Letöltés 2020. augusztus 14;
helye: https://en.wikipedia.org/wiki/Robots_exclusion_standard
- Rosenthal, David. S. (2017). Losing the Battle to Archive the Web. Letöltés 2020. január 7; helye: <http://dpconline.org/blog/idpd/losing-the-battle-to-archive-the-web?%20>
- Ruusalep, R. (2017). Preserving digital legal deposit—New challenges and opportunities. In *IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity. Society. In Session 210—Preservation and Conservation (PAC) Strategic Programme*. Wrocław: IFLA -- International Federation of Library Associations and Institutions. Letöltés helye:
<http://library.ifla.org/id/eprint/1677>
- Schostag, S. (2019a). Harvesting YouTube videos—A manual approach [Presentation]. Letöltés 2020. augusztus 20; helye:
<https://digital.library.unt.edu/ark:/67531/metadc1609005/>
- Schostag, S. (2019b, július 6). Why, what, when and how?: Curatorial decisions and the importance of their documentation [Presentation]. Letöltés 2020.

augusztus 20; helye:

<https://digital.library.unt.edu/ark:/67531/metadc1608959/>

Schostag, S., & Fonss-Jorgensen, E. (2012). Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective. *Microform & Digitization Review*, 41(3–4), 110–120. Letöltés helye:

<https://search.proquest.com/docview/1520327503?accountid=27464>

Sierman, B., & Teszelszky, K. (2017). How can we improve our web collection? An evaluation of webarchiving at the KB National Library of the Netherlands (2007-2017), 27(2), 94–106. doi: 10.1177/0955749017725930. Letöltés helye: <https://doi.org/10.1177/0955749017725930>

Signori, B. (2017). Preserving cultural heritage: Better together! In *IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity. Society. In Session S08— Satellite Meeting: Preservation and Conservation Section joint with the Association International Francophone des Bibliothécaires et Documentalistes (AIFBD) in collaborati*. Wrocław: IFLA -- International Federation of Library Associations and Institutions. Letöltés helye: <http://library.ifla.org/id/eprint/1801>

SolrMIA. (2018). Letöltés 2020. február 23; helye:

<http://193.6.201.202/solrmia/>

Solrwayback kereső. (2018). Letöltés 2020. február 23; helye:

<http://webadmin.oszk.hu/solrwayback/>

Stirling, P., Illien, G., Sanz and, P., & Sepetjan, S. (2012). The state of e-legal deposit in France: Looking back at five years of putting new legislation into practice and envisioning the future. *IFLA Journal*, 38(1), 5–24. Letöltés helye: <http://10.0.4.153/0340035211435323>

- Storarr, T., & Seaman, G. (2016). Researching User Needs: The UK Government Web Archive—UNT Digital Library. Letöltés 2020. augusztus 20; helye: <https://digital.library.unt.edu/ark:/67531/metadc1477155/>
- Strodl, S., Becker, C., Neumayer, R., & Rauber, A. (2007). How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 29–38). New York, NY, USA: ACM. doi: 10.1145/1255175.1255181
Letöltés helye: <https://doi.org/10.1145/1255175.1255181>
- Sylvie-Rollason-Cass. (2020). What is brozzler? Letöltés 2020. február 23; helye: <https://support.archive-it.org/hc/en-us/articles/360000343186-What-is-Brozzler->
- Szabo, S. (2018, november 15). Website Archiving Guide Dutch National Archives [Presentation]. Letöltés 2020. augusztus 20; helye: <https://digital.library.unt.edu/ark:/67531/metadc1477176/>
- SZTAKI. (2011). Living Web Archive—LIWA. Letöltés 2020. augusztus 18; helye: <https://www.sztaki.hu/en/projects/liwa>
- SZTAKI. (2013). LAWA - Longitudinal Analytics of Web Archive Data. Letöltés 2020. augusztus 18; helye: https://old.sztaki.hu/tudomany/projektek/projekt_informaciok/?uid=00303
- Temporal drift – MIA. (2017). Letöltés 2020. augusztus 18., helye: https://webarchivum.oszk.hu/mediawiki/index.php?title=Temporal_drift
- Teszelszky, K. (2017). The harvest of the Dutch digital fields: The landscape of webarchiving in The Netherlands. Letöltés ideje: 2020. január 8; helye: <https://webarchivum.oszk.hu/wp->

content/uploads/2020/03/Kees_Teszelszky_2017_Presentatie_webarchiveri
ng_KB_BUDAPEST_404.pptx

Teszelszky, K. (2019a). Distant reading: The Frisian Web Domain. Letöltés 2020.
január 8; helye:
[http://mekosztaly.oszk.hu/mia/doc/DH_2019/2019.09.11_Teszelszky_Friese
_web.pptx](http://mekosztaly.oszk.hu/mia/doc/DH_2019/2019.09.11_Teszelszky_Friese_web.pptx)

Teszelszky, K. (2019b). Web archaeology in The Netherlands: The selection and
harvest of the Dutch web incunables of provider Euronet (1994–2000).
Internet Histories, 3(2), 180–194. doi: 10.1080/24701475.2019.1603951.
Letöltés helye: <https://doi.org/10.1080/24701475.2019.1603951>

The dead web – The end | online kiállítás | Ludwig Múzeum. (2020). Letöltés
2020. augusztus 17., helye:
<https://www.ludwigmuseum.hu/valogatas/dead-web-end-online-kiallitas>

Thomson, S. D., & Kilbride, W. (2015). Preserving Social Media: The Problem of
Access. *New Review of Information Networking*, 20(1/2), 261–275. Letöltés
helye: <http://10.0.4.56/13614576.2015.1114842>

Thornton, J. B. (2012). Archiving of Comprehensive Annual Financial Reports
(CAFRs) on State Government Web Sites. *Behavioral & Social Sciences
Librarian*, 31(2), 87–95. doi: 10.1080/01639269.2012.686244. Letöltés
helye: <http://dx.doi.org/10.1080/01639269.2012.686244>

Tiiu, D. (2018, november 13). Becoming a Web Archivist: My 10 Year Journey in
the National Library of Estonia [Presentation]. Letöltés 2020. augusztus 20;
helye: <https://digital.library.unt.edu/ark:/67531/metadc1477143/>

Togan, M., & Florea, I. (2015). A Reference Model for a Trusted Service
Guaranteeing Web-content. In Reimer, H., Pohlmann, N., & Schneider, W.

- (Szerk.), *ISSE 2015* (pp. 216–224). Wiesbaden: Springer Fachmedien
Wiesbaden. doi: 10.1007/978-3-658-10934-9_18. Letöltés helye:
https://doi.org/10.1007/978-3-658-10934-9_18
- UNESCO. Charter on the Preservation of Digital Heritage (2003). Letöltés ideje
2020. augusztus 18; helye: http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html
- Veikkolainen, P., Lager, L., Lehtonen, J., & Kaunonen, K. (2016, december 4).
Long-Term Preservation of the Finnish Web Archive [Presentation]. Letöltés
2020. augusztus 20; helye:
<https://digital.library.unt.edu/ark:/67531/metadc1477130/>
- Visky, Á. L. (2020). Gyorsmérés – az OSZK Webarchívumának és néhány
könyvtárnak a KDS-K pályázat keretében történt együttműködéséről. *Könyv,
könyvtár, könyvtáros*, 29.(7-8.), 24-38. Letöltés helye:
https://epa.oszk.hu/01300/01367/00330/pdf/EPA01367_3K_2020_07-08_024-038.pdf
- Visky, Á. L., & Drótos, L. (2020). Rákóczi-archívum. *Könyv, könyvtár, könyvtáros*,
29(3), 35–48. Letöltés helye:
https://epa.oszk.hu/01300/01367/00326/pdf/EPA01367_3K_2020_03_035-048.pdf
- Vitéz, G. (2020). Az arató és a megjelenítő szoftverekről. Letöltés 2020. február
23; helye: <http://mekosztaly.oszk.hu/mia/e-learning/prezentaciok/VG-eloadas.pptx>
- Vozár, Z. (2019). CZ web archive development. Data processing and provision in
cluster. In *8th Colloquium Library and Information Experts of the V4+
Countries*. Letöltés ideje: 2020. február 15; helye:

https://kolokvium.ulib.sk/wp-content/uploads/2019/06/B34_P_Vozar_CZ.pdf

VRML Virtual Reality Modeling Language. (1995). Letöltés 2020. augusztus 15; helye: <https://www.w3.org/MarkUp/VRML/>

W3C Consortium. (2019). Introduction to Web Accessibility. Letöltés 2020. augusztus 17; helye: <https://www.w3.org/WAI/fundamentals/accessibility-intro/>

WAIL - MIA WIKI. (2017). Letöltés 2020. február 23; helye: <http://mekosztaly.oszk.hu/mediawiki/index.php/WAIL>

Wappalyzer—Identify technology of websites. (2020). Letöltés 2020. augusztus 15; helye: <https://www.wappalyzer.com/>

WARC – MIA. (2017). Letöltés 2020. augusztus 17; helye: <https://webarchivum.oszk.hu/mediawiki/index.php?title=WARC>

WCT manuals. (2019). Letöltés 2020. február 23; helye: <http://webcurator.sourceforge.net/manuals.shtml>

WEB Archive UK & JICS. (2015). Shine Project Historical Research Prototype. Letöltés ideje 2020. február 14; helye: <https://www.webarchive.org.uk/shine>

Web Curator Tool. (2018). Letöltés 2020. február 23; helye: <http://mekosztaly.oszk.hu/mediawiki/index.php/WCT>

Webrecorder—MIA WIKI. (2018). Letöltés 2020. február 23; helye: <http://mekosztaly.oszk.hu/mediawiki/index.php/Webrecorder>

Webrecorder/warcit. (2020). Python, Webrecorder. Letöltés helye:

<https://github.com/webrecorder/warcit>

Website Archiving and Monitoring Solutions | MirrorWeb. (2020). Letöltés

2020. augusztus 20., helye: <https://www.mirrorweb.com>

Wichtowski, R. (2017). Increasing Copyright Protection for Social Media Users

by Expanding Social Media Platforms' Rights. *Duke Law & Technology*

Review, 16(1), 253–268. Letöltés helye:

<https://scholarship.law.duke.edu/dltr/vol15/iss1/13/>

Winters, J. (2015, április 27). Big UK Domain Data for the Arts and Humanities

[Presentation]. Letöltés 2020. augusztus 20; helye:

<https://digital.library.unt.edu/ark:/67531/metadc1476406/>

Winters, J. (2017). Breaking in to the mainstream: Demonstrating the value of

internet (and web) histories. *Internet Histories*, 1(1–2), 173–179. doi:

10.1080/24701475.2017.1305713. Letöltés helye:

<https://doi.org/10.1080/24701475.2017.1305713>

XHTML+RDFa 1.1—Third Edition. (2015). Letöltés 2020. augusztus 14 helye:

<https://www.w3.org/TR/xhtml-rdfa/>

Zeng, M., & Quin, J. (2016). *Metadata*. Facet Publishing.