Németh Márton

A webarchiválás nemzetközi környezete

Mozaikok az IIPC 2019 kongresszusról

Elsőként áttekintést adunk a webarchiválás nemzetközi hátterét adó *International Internet Preservation Consortium* (IIPC) szervezetéről[1], az OSZK szerepéről a konzorcium munkájában, valamint szemelgetünk a konzorcium idei évi, Új-Zélandon megrendezett kongresszusán felmerült témakörök közül.

Az OSZK webarchiválással foglalkozó kísérleti projektje (http://mekosztaly.oszk.hu/mia) nemzetközi kapcsolatának gerincét adja az ezen a területen szerveződött nemzetközi konzorciumban (IIPC) kifejtett tevékenységünk. Az OSZK 2018. január elsejétől vált a konzorcium teljes jogú tagjává. A konzorcium 2013-ban alakult meg a Francia Nemzeti Könyvtár kezdeményezésére, 12 alapító taggal. Jelenleg már 45 országból vesznek részt könyvtárak, levéltárak, múzeumok, illetve a webarchiválás területén érdekelt civil és piaci szereplők is a tevékenységében. A konzorcium alapvető küldetése az interneten tárolt tartalmak archiválásához és biztonságos hosszú távú megőrzéséhez szükséges tevékenységek összefogására, támogatására irányul. A konzorcium munkáját az egyes tagintézmények képviselőiből megválasztott végrehajtó bizottság fogja össze, a titkárság segítségével. A közös munka nagy része tematikus munkacsoportokban zajlik, ezek tartalomfejlesztéssel, hosszú távú megőrzéssel, illetve oktatási kérdésekkel foglalkoznak. Az OSZK képviseletében ez utóbbi, 2017. év végén alakult munkacsoport tevékenységében veszünk intenzíven részt, melyről a későbbiekben szólunk.

A tagok számára nagyon nagy segítséget jelent az aktuális információk gyors megosztását lehetővé tevő elektronikus levelezőlista, illetve Twitter hírcsatorna, valamint a Slack nevű információmegosztó alkalmazás segítségével zajló kommunikáció. Ezek révén gyors áttekintést nyerhetünk első kézből a webarchiválást meghatározó nemzetközi trendekről, az egyes intézményi gyakorlatokról, projektekről, eredményekről.

Az IIPC tevékenységét a British Library bázisán koordinálják. Olga Holowniaval, az adminisztráció vezetőjével gyümölcsöző együttműködést sikerült kialakítani. 2018. nyár folyamán online találkozót szervezett számunkra mint új tagok számára az Internet Archive vezetőségi tagjával, illetve az IIPC végrehajtó bizottságának tagjával, Jefferson

Baileyvel. Alkalmunk nyílt részt venni az IFLA 2017. évi kongresszusán a lengyelországi Wrocławban[2], ahol önálló szekció foglalkozott az IIPC szervezésében a webarchiválással. Itt történt meg 2017 augusztusában az első személyes kapcsolatfelvétel az IIPC tagjaival, vezetőségével. Első kézből kaptunk tanácsot a webarchiválási kísérleti projektünk megalapozásához, még mielőtt hivatalosan taggá váltunk volna.

2018 januárjától váltak számunkra elérhetővé a fentebb jelzett információs csatornák. Bemutatkozó anyagot készítettünk terveinket taglalva a konzorcium honlapjára. A már említett konzorciumi titkár mellett a holland nemzeti könyvtár webarchiválásért felelős, részben magyar származású, az ELTE-n doktorált vezetője, Kees Teszelszky nyújtott nélkülözhetetlen segítséget tagságunk kezdetekor a szükséges információk összegyűjtéséhez. Ő irányította rá a figyelmünket az oktatási munkacsoport tevékenységére, melynek munkájába már 2018. januártól bekapcsolódtunk. A saját oktatási, képzési tevékenységeink megtervezése kapcsán igen jól jött, hogy fel tudtuk mérni a munkacsoport keretében az összes, a tagok körében már rendelkezésre álló oktatási, képzési tevékenységi formát. Ezen felül a konzorciumi tagok számára készült egy kérdőív is, amiben igényeiket, elvárásaikat mérték fel a webarchiválás oktatásának kérdései kapcsán. A kérdőív kiértékelésének előzetes eredményeit is be tudtuk építeni a munkánkba, illetve nyilvános előadásainkba is, melyeket 2018 áprilisában az egri Networkshop konferencián[3], illetve a Pozsonyi Egyetemi Könyvtár CDA 2018 digitális könyvtári konferenciáján tartottunk.

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 Egensével. Sikerült infrastruktúrát szolgáltatnunk a közös munkával kifejlesztett, a webarchiválás során született anyagok teljesszövegű anyagok indexelésére szolgáló SOLRWayback illetve SOLRMIA[4] eszközök folyamatos teszteléshez. A dán fejlesztők számos nemzetközi szakmai rendezvényen adtak számot a közös munka eredményeiről, ami az IIPC vezetőségének körében is komoly pozitív visszhangot keltett. Különösen kiemelkedik szakmai jelentőségben az IIPC 2018. évi közgyűlésén és konferencián a közös erőfeszítéseink eredményeiről szóló dán előadás, ahol bemutatták az OSZK informatikai platformján futó SOLRWayback szolgáltatást[5], mely a rendezvényhez kötődő közösségi média csatornákon is jelentős publicitást kapott.

A szintén friss IIPC-tag belga webarchiválási projekt munkatársaival az Aarhusi Egyetem Netlab[6] kutatócsoportja által szervezett angol nyelvű webarchiválással foglalkozó e-learning kurzus[7] keretében építettünk ki kapcsolatot. Az itt folyó szakmai diskurzus is jelentős segítséget jelentett pilot projektünk szakmai megalapozásában.

Személyes tapasztalatcserén alapuló sikeres szakmai kapcsolatokat sikerült kiépíteni a cseh[8], a szlovák[9] és az osztrák[10] webarchívumok munkatársaival is. A pozsonyi Egyetemi Könyvtárban több előadást is tartottunk, illetve eredményes szakmai véleménycserét folytattunk a házigazda intézmény, illetve a Cseh Nemzeti Könyvtár webarchiválásért felelős munkatársaival. Az Osztrák Nemzeti Könyvtárban pedig egynapos szakmai tanulmányúton vettünk részt.

Sajnos eddig nem nyílt lehetőségünk a konzorcium éves közgyűlésén illetve kapcsolódó konferenciáján történő részvételre, mivel idén arra Új-Zélandon került sor. Lehetőségünk nyílt viszont ötperces összefoglalót készíteni a rendezvény számára, bemutatva magunkat s eddigi eredményeinket. A konzorciumi titkár felajánlására videóhoz felhasznált prezentáció, illetve szöveges anyag alapján összefoglaló készül az IIPC honlapjára

is. Jövőre ez a legnagyobb éves konzorciumi rendezvény Zágrábban kerül majd megrendezésre 2019. június 5-7. között[11]. Németh Mártonnak lehetősége nyílt bekapcsolódni a konferencia programbizottságának munkájába, részt vállalva így a szakmai tartalom összeállításában. Két előadással is készülünk majd e rendezvényre, egy rövid 15 perces bemutatkozásra az OSZK-ban folyó webarchiválási munka kapcsán, illetve 30 percben bemutatnánk a már említett indexelő eszközök fejlesztési eredményeit, emellett felvázolnánk a további lehetséges kutatás-fejlesztési lépéseket a webarchívumokban begyűjtött anyagok metaadatokkal történő ellátása, s a visszakeresés biztosításának területén.

Reményeink szerint a konzorcium keretei között folyó folyamatos tapasztalatcsere és közös szakmai munka az ideinél is elmélyültebb formában teljesedhet ki 2019 folyamán, a webarchiválási munkafolyamatokban végzett egyes tevékenységeinkkel összhangban. Számos szakmai előadásban, publikációban tervezzük továbbra is felmutatni a közös eredményeinket.

2018. november 12.-én került sor az IIPC idei közgyűlésére, majd ezt követően november 13–15. között a szakmai konferenciára. A közgyűlésen az összes konzorciumi tagintézmény munkatársai számára ingyenes a részvétel, létszámtól függetlenül. A konferencián tagintézményenként egy munkatárs ingyenesen, a többiek kedvezményes részvételi díjjal vehetnek részt. Idén a helyszín az új-zélandi Wellington volt. A helyszín kiválasztását elsősorban az indokolta, hogy rá akarták irányítani a figyelmet az ausztráliai-óceániai térségben zajló jelentős webarchiválási tevékenységre, illetve tágabb értelemben az ázsiai térség számára is kedvező helyszínen kívántak a szakmai tapasztalatcserére lehetőséget adni. A minden évben megrendezésre kerülő rendezvény jó alkalmat ad a konzorciumon belül előző évben folytatott közös tevékenységek számbavételére, a jövőbeni célok kijelölésére.

A konferencia egyes szekcióit élően közvetítették a Youtube-on, illetve rövid ideig felvételről is elérhetők voltak a rendezvény után, de mára már sajnos eltűntek. Ebből is látszik, hogy még egy konferencia videoanyagának webes rögzítése és megőrzése is mekkora problémákat vet fel. A szakmai előadásokat elsősorban technikai jellegű workshopok egészítik ki. Az előadások két fő típusba sorolhatók be. Az egyik fajta az egyes tagintézmények projektjeit, fejlesztési eredményeit mutatja be. A másik típus, amikor különféle általános fejlesztésekre hívják fel a figyelmet, illetve olyan szoftvercsomagokat mutatnak be, melyek a teljes nemzetközi közönség érdeklődésére számot tarthatnak. Előbbi kihívások közül előkelő helyen szerepelt a jogi kérdések tisztázása, illetve az egyes nemzeti kötelespéldány szabályok és a webarchiválás viszonyának felmérése. A webarchiválás kapcsán súlyos veszélyként jelenik meg az Európai Unióban a felejtés jogához kapcsolódó rendelkezések érvényesítése, ami technikai okok, valamint az állomány integritásának megőrzése miatt szinte megoldhatatlan követelményeket támaszt az intézményekkel szemben. De elvi, történeti nézőpontból sem indokolható az állományból történő adattörlés. Olyan megoldások képzelhetők el, amikor az indexelési algoritmusokon módosítva az adott adatokhoz való hozzáférés kerül korlátozásra.

Több intézmény képviselője ráirányította a figyelmet a webarchiválás általánosan tapasztalható súlyos alulfinanszírozottságára. Intézményi költségvetési szinten aprópénznek tűnő összegekből lehetne jelentős eredményeket elérni a webes kulturális örökség egyes szeleteinek mentése terén, de ezzel az intézményi vezetők sokszor nincsenek tisztában. Arról, hogy mekkora a tét, szintén hangzottak el adatok. A kilencvenes évektől

a 2000-es évek közepéig a brit webarchiválási konzorcium által begyűjtött anyagok 60-80%-a már nem érhető el az élő weben! Aminek a mentéséről tehát ma pénzügyi okok, szűklátókörűség miatt lemondunk, az félő, hogy nagyon rövid időn belül örökre elvész a számunkra.

Az intézményi gyakorlatok ismertetésnek tengeréből kiemelkedik a British Library webarchiválásért felelős szakmai illetékesének előadása.[12] A 2013-as kötelespéldány törvénymódosítás miatt minden időszaki kiadványt folyamatosan menteniük kell, a hírportálokat naponta többször is. Ennek következtében alaposan át kellett tervezni a webarchiválási munkafolyamatot. Folyamatosan fut a Heritrix szoftver segítségével megvalósuló aratási művelet, ebbe ágyazzák bele a különféle aratási feladatokat, automatizált módon készülnek megadott időpontokban a mentések a beállított webes szeletekről a megadott szempontok szerint. Biztosítaniuk kellett, hogy a paramétereket közben dinamikusan lehessen módosítani. Felmerült, hogy a korábbiaknál nagyobb eredményességgel kellene menteni a Javascript kódokkal teli webes tartalmakat. Különös gondot kellett fordítani a folyamatosan, automatikusan futó archiválási tevékenység zavarmentességének biztosítására pl. áramellátási, adathálózati problémák fellépése esetén.

Megemlítendő még két fontos általánosan használt szoftvercsomag továbbfejlesztésének ismertetése. Az egyik a Web Curator Tool, melynek fejlesztése anno Új-Zélandról indult, majd megakadt, most pedig holland segítséggel folyik újra tovább. Ez azért nagyon praktikus eszköz, mert nem csupán vezérelni lehet általa a tartalmakat learató Heritirix robotszoftvert, hanem a learatott anyag indexelésére és metaadatokkal történő ellátásra is önálló munkakörnyezetet nyújt. Sajnos a fejlesztés elakadása éppen abban nyilvánult meg, hogy a legújabb Heritrix verziókat már nem lehetett vezérelni általa, a kompatibilitás megbomlása miatt. Ezt próbálják orvosolni a legújabb verzióval a fejlesztők, finomítva persze a metaadatkezelő funkciókon is.

A WebRecorder[13] teljesen másfajta módszerrel használható webarchiválásra. Itt arról van szó, hogy kiválasztjuk a saját böngészőnket, vagy a szolgáltatásba beépített különféle típusú, funkcionalitású böngészők valamelyikét, majd a világhálón böngészve a háttérben lementődik az a tartalom, amit megnézünk. A mentett tartalmakat aztán szabványos WARC konténerfájlokban elmenthetjük külső szerverre, vagy akár a saját szerverre is, a megjelenítésről pedig a Webrecorder Player[14] program gondoskodik, mely elérhető az összes elterjedt operációs rendszer alatt. A kihívás itt abban jelentkezik, hogy miként lehetne a böngészőhasználat emberi tevékenységét automatizálni, tehát megadni, hogy milyen tartalmakat járjon be a böngészőprogram, melyek aztán mentésre kerülnek. Az idei konferencián a fejlesztő csapat ennek a régen várt fejlesztésnek a próbaváltozatát mutatta be[15].

Vint Cerf, akit az internet egyik atyjának tekintenek, manapság pedig a Google technológiai evangélistájaként tevékenykedik, átfogó előadást tartott a hosszú távú digitális megőrzést övező technológiai, jogi illetve emberi tényezőből fakadó követelményekről és dilemmákról. Igazából a legérdekesebb pontja az előadásnak az volt, amit nem tartalmazott. Az internetes anyagok szolgáltatásában, visszakeresésében domináns szerepet játszó piaci szereplőknek jelenleg a webarchiválásához szinte semmi közük nincs. Nem látunk pl. a Google digitalizálási programjához hasonló partnerségi kezdeményezéseket, ami a webarchiválására irányulna. Pedig ezen a területen különösen félő, hogy ilyenfajta partnerségi formák nélkül az előrehaladás kizárólag közgyűjteményi energiákból merítve csak

részlegesen valósulhat meg. Több előadó felvetette egyébként a rendezvényen, hogy a közgyűjteményeknek sokkal határozottabb, ha úgy tetszik konfrontatívabb stílusba hajló módon kellene felvetniük a hosszú távú megőrzés fontosságát, mind belső fórumokon, mind a külső partnereik felé. Az udvarias, szinte félős visszafogottság, ami a terület kommunikációs kultúráját általában jellemzi, szintén erőteljes problémát jelent.

Zárszóként elmondhatjuk, hogy minden bizonnyal 2019 júniusában, Zágrábban a következő konferencián hatalmas élmény lesz végre személyesen is bekapcsolódni az élénk szakmai párbeszédbe, hozzájárulni a webarchiválás fejlődéséhez.

Jegyzetek

1. http://www.netpreserve.org
2. Németh Márton: Szubjektív beszámoló az IFLA wrocławi világkonferenciájáról. Könyvtári Figyelő 2018/2. pp. 273-278.
3. https://kifu.videotorium.hu/hu/recordings/24659/a-webarchivalas-oktatasa
4. http://193.6.201.202/solrmia/
5. http://193.6.201.202/solrwayback/
6. http://www.netlab.dk
7. http://www.netlab.dk/wp-content/uploads/2017/04/NetLab-Web-Archiving-Course-Brochure.pdf
8. https://www.webarchiv.cz/en/
9. https://www.webdepozit.sk/en/english-homepage/
10. https://webarchiv.onb.ac.at/
11. http://netpreserve.org/ga2019/
12. https://anjackson.net/2018/11/13/continuous-incremental-heritrix/
13. https://webrecorder.io/
14. https://github.com/webrecorder/webrecorder-player
15. A prezentáció elérhetősége: https://docs.google.com/presentation/d/1_AoCavSoZRFZp6KNpRcYfhJe9am4XCJIWN26mP4Haro/edit#slide=id.g39f46cab1b_0_161