

A Webarchiválási Osztály 2020. évi beszámolója

Készült: 2021-01-20

Munkaerő

Az OSZK 2020-ban lezajlott szervezeti megújulását követően a 2017-ben indult webarchiválási projekt egy önálló, 3 fős osztály keretében folytatódott tovább. A Digitális Tartalomfejlesztési és -szolgáltatási Osztállyal megmaradt a szoros együttműködés, például az EPA csapatban dolgozó egyik munkatárs részmunkaidőben a webarchívumba is besegít. Rajta kívül kivételesen más osztályokról is bekapcsolódtak OSZK-s kollégák az ellenőrzési feladatokba, akiknek a járványhelyzet miatt otthonról végezhető munkára volt szükségük.

Informatika

A szerverek és a tárolók felügyeletét az Informatikai Szolgáltató Osztály látja el, a napi szintű rendszergazdai feladatokat és programfejlesztéseket pedig a Vitéz Bt. végzi.

A webarchívum anyagát kétszer kellett költöztetni az év során. Tavasszal azért, mert a KIFÜ-től 2017-ben kapott tesztszerver tárhelye betelt, ősszel pedig az időközben beüzemelt új informatikai infrastruktúrára lett áttelepítve mindhárom szerverünk: *webadmin.oszk.hu* (nyilvános gyűjtemény), *webharvest.oszk.hu* (zárt gyűjtemény), *webarchivum.oszk.hu* (honlap). 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.

Az archiváláshoz, az archivált anyag visszanevezéséhez, a metaadatok között való kereséshez használt szoftverek köre jelentősen bővült tavaly. Ezekhez a munkafolyamatokhoz részben ingyenes és többnyire open source eszközöket teszteltünk és kezdtünk el használni, részben pedig saját szkripteket készítettünk. Előbbiek közül érdemes kiemelni a dán SolrWayback rendszert, amit ugyan már korábban beépítettünk a nyilvános gyűjteménybe a teljes szövegben való kereséshez, de 2020-ban egy jelentősen átalakított, több új funkciót is tartalmazó verzió készült belőle, melyet a jó szakmai kapcsolatnak köszönhetően a világon mi teszteltünk először.

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. Szintén új eszköz az a virtuális gép, melyet tesztelési és oktatási célokhoz állítottunk össze az év második felében. Ezzel egy átlagos PC-n Windows alatt tudunk futtatni linuxos archiváló és megjelenítő programokat.

Jogszabályok

2020-ban több olyan [jogszabály](#) is született, illetve módosult, amelyek az OSZK webarchiválási tevékenységét szabályozzák, vagy pedig azt is érintik. A legfontosabb ezek közül a kormány 626/2020. (XII. 22.) számú rendelete, amelyből végül kimaradt a tartalomszolgáltatók bejelentési kötelezettsége, viszont belekerült, hogy a „kormányzati és önkormányzati tartalomszolgáltató és az állami vagy önkormányzati fenntartású intézmény tartalomszolgáltató” archivált webtartalmait külön engedélyeztetés nélkül is nyilvánosan szolgáltatathatja az OSZK. A rendelet most január 1-től lépett hatályba. A 1358/2020. (VII. 1.) számú kormányhatározat pedig 2021-re 56.317.500 Ft-ot biztosított a központi költségvetésből a nemzeti könyvtár számára a webarchiválási feladatok ellátására.

Honlapok

2020 februárjában két honlapot tettünk nyilvánossá. A webarchivum.oszk.hu címen a projekttel, illetve általában az internetes tartalmak megőrzésével kapcsolatos dokumentumok találhatóak, továbbá kereshető és böngészhető a nyilvános gyűjtemény, és metaadat-szinten kereshető a zárt gyűjtemény is. Megnézhetők az egyes aratásokról készített [statisztikák](#), valamint a honlap látogatottsági adatai. A fontosabb információk angolul is elérhetők. A másik honlap a KDS-K pályázat részeként készült és a rakoczi2019.webarchivum.oszk.hu címen érhető el, hat egyenrangú nyelvi változatban. Ennek egyik célja a Rákóczi Emlékévvvel kapcsolatos digitális és digitalizált dokumentumok szolgáltatása egyfajta oktatási segédletként, a másik pedig annak demonstrálása, hogy hogyan egészítheti ki az archivált webtartalom a könyvtárak „hagyományos” digitális gyűjteményét.

Archiválás

A **szelektív aratásokat** tavaly már viszonylag rendszeresen, negyedéves ismétlődéssel, a hétvégeken futtattuk. Az év során három új témában (média, oktatás és sport) válogattunk össze több ezer webcímet, három korábbi kisebb részgyűjteményt (könyvtárak, levéltárak, múzeumok/galériák) pedig összevontunk. Jelenleg 12 tematikus gyűjteményünk van, melyekben összesen több mint 33 ezer webhely címe található. Elkezdtünk 8 további címlistát összeválogatni, közülük az idegenforgalom/vendéglátás témájú aratása 2021 elején már elindulhat. Bár ezeknél webhely szinten csak néhány alapadatot tartunk nyilván, de részgyűjtemény szinten részletes metaadat rekordokat készítünk [XML formátumban](#).

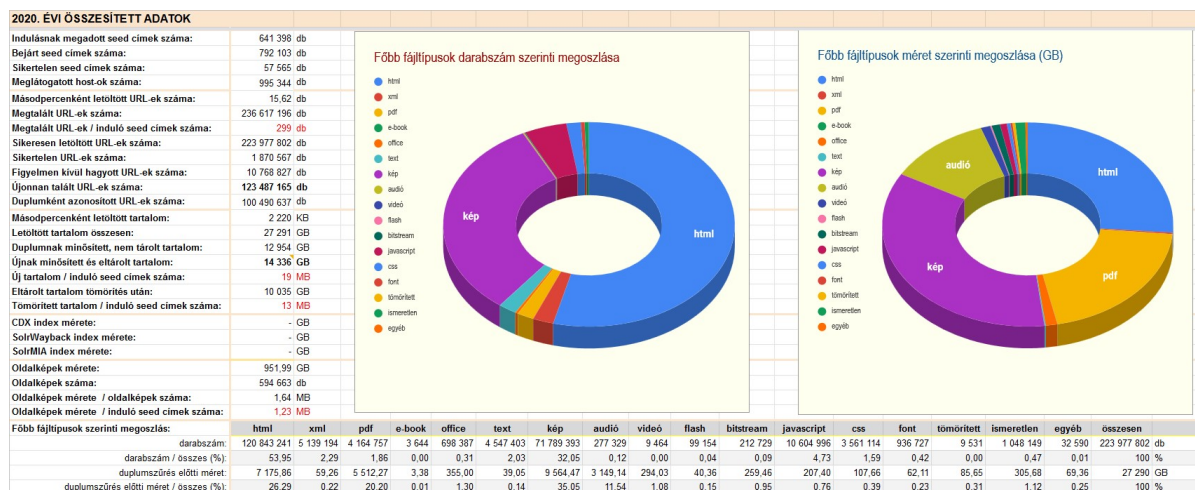
A műfaji részgyűjtemények közül kiemelten fontos az **elektronikus periodikák** webhelyeinek/weboldalainak archiválása, kiegészítve ezzel az EPA tevékenységét, amely a letölthető kiadványszámok megőrzésével foglalkozik. A két archívum nyilvántartása az EPA és MIA azonosítókra keresztül van összekapcsolva. A webarchívum által használt MIA azonosító már a 905341-es számnál tart, ami azt jelenti, hogy 5341 webhelyet/weboldalt tartunk nyilván ennél a részgyűjteménynél, ebből 5051-et aratunk jelenleg. (A többi vagy időközben megszűnt cím, vagy eleve egy digitális archívumban van.) 2020-ban 585 tétellel bővült ez a lista, jelentős arányban települési vagy regionális hírlapokkal, valamint határon túli magyar szervezetek kiadványaival. Az ELPERI nyilvántartásunk a tematikus gyűjteményeknél valamivel részletesebb metaadatokat tartalmaz, a negyedéves aratásokat általában hosszabb ideig futtatjuk és a kezdőoldaltól nagyobb mélységig követjük a linkeket a robottal. (2020-ban a szerverköltözések miatt sajnos két aratás is kimaradt.)

A periodikák közül az **online hírportálok** speciális műfajt jelentenek, egyrészt azért, mert ezeknél olyan gyorsan változik a címlapon megjelenő új tartalom, hogy a negyedéves aratási gyakoriság nem elegendő, másrészt mert olyan bonyolult, dinamikusan generált weboldalakból állnak, amelyekkel a „buta” robotok nem boldogulnak. Ezért 2020 őszén elkezdtünk kísérletezni a Brozzler nevű szoftverrel, ami egy böngészőn keresztül archivál, így bár jóval lassabban, de jobb minőségben menthetők vele weboldalak. Egyelőre 18, eltérő típusú magyar hírportál címlapját és előző napi híreit mentjük vele napi gyakorisággal, több-kevesebb sikerrel.

Tavaly elkezdtünk foglalkozni a **közösségi média** megőrzésével is, egyelőre szintén csak kísérleti jelleggel. Ezek a szolgáltatások nagy mértékben a felhasználó interaktivitására épülnek és sokszor kifejezetten gátolják a tartalom automatikus módszerekkel való letöltését. Az archivált oldalak visszanevezése szintén problémás az ezeken a platformokon használt, ráadásul gyakran változtatott technológiai megoldások miatt. 2020 elején a viszonylag egyszerűbb Instagrammal kezdtünk el kísérletezni, majd nyár végén a Facebook oldalakkal, decemberben pedig a Twitter csatornákkal. A Conifer nevű online szolgáltatást, illetve a Webrecorder Desktop programot használtuk ezekhez, melyekkel videómagnó-szerűen lehet „felvenni”, vagyis eltárolni egy böngészési folyamatot. Bár mindkét rendszerben van egy Autopilot funkció, amellyel szimulálható egy emberi felhasználó (görgetés, lapozás, kattintgatás), de a tapasztalataink szerint ez nem, vagy csak rövid ideig működik, így valódi emberi közreműködéssel, illetve részben billentyűmakrókkal tudtunk végül lementeni tartalmakat ezekről a helyekről. Tavasszal 717 Instagram fiókból töltöttünk le 110 ezer posztot: fotókat és rövid videókat. Szeptember-októberben pedig 850 Facebook oldal nyilvános idővonalát mentettük, melyek közt 147 volt könyvtári témájú és ezeknél az egyes bejegyzéseket is letöltöttük, összesen kb. 30 ezret. Az idővonalakat igyekszünk legalább egy évre visszamenőleg legörgetni, de az esetek felénél sikerült egészen az oldal létrehozási dátumáig visszamenni anélkül, hogy lefagyott volna a böngésző. A Twitter esetében még kevés tapasztalatunk gyűlt össze, de az Instagram és a Facebook mellett erre is igaz, hogy számos technikai probléma van mind a mentés, mind a lementett fájlok indexelése, mind pedig azok megjelenítése során, nem beszélve a magas élőmunka igényről. Úgyhogy ezeknek a tartalmaknak a tömeges mentése a jelenlegi lehetőségek mellett nem tűnik fenntarthatónak.

A tematikus és műfaji részgyűjteményeken túl négy új **esemény-alapú** webarchívumot is létrehoztunk. Ezek az események a már említett Rákóczi emlékév, a trianoni békeszerződés 100. évfordulója, a Tokióban tervezett nyári olimpia, valamint a koronavírus járvány. Mindegyiknél mentettünk vagy még jelenleg is mentünk teljes webhelyeket, webhely-részleteket (pl. híroldalak specifikus rovatait), illetve egyedi weboldalakat (pl. egyesével beválogatott, vagy releváns címkék alapján kiszűrt híreket), sőt esetleg hang- vagy videófájlokat és közösségi médiatartalmakat is. Ezeket az aratásokat heti vagy havi rendszerességgel futtatjuk/futtattuk addig, amíg az adott esemény aktuális.

Az emberi munkával vagy automatikus módszerrel (pl. az archivált weboldalakban található linkekből) eddig összegyűjtött valamennyi URL címre kiterjedő, kis mélységű, „pillanatfelvétel”-szerű **webtér aratások**nál a fél éves ütemezést igyekszünk most már tartani. A 2019 végén indított aratás 2020 januárjára is átnyúlt, a következő pedig nyár közepén zajlott le. Ekkor 270 ezer URL-t adtunk a robotnak kiindulásként, ami azután öt nap alatt a linkeket követve további 46 milliót talált, közülük 22 millió volt az új, korábban még nem letöltött fájl, 2,4 terabájt össz méretben. A 2020. december 30. és 2021. január 4. közötti év végi aratás ugyan csak 251 ezer URL-ről indult, de köztük jóval kevesebb volt a már megszűnt vagy duplum cím, mert előtte különböző félautomatikus megoldásokkal valamennyire kitisztítottuk a listát. Ez alkalommal 50 millió megtalált fájlból 24,5 millió volt az új, melyek 2,3 terabájt tárhelyet igényeltek.



A tematikus, az ELPERI és a webtér aratások 2020. évi összesített adatai

Együttműködések

A 2019-es KDS-K pályázaton nyertes könyvtáraknak a digitalizálás mellett az OSZK webarchíváló tevékenységének támogatását is vállalniuk kellett. A pályázat elhúzódó adminisztrációja miatt végül csak 2020 tavaszán tudtuk felvenni a könyvtárakkal a kapcsolatot és nagyon kevés idő maradt az érdemi munkára. Az eredetileg tervezett címggyűjtés, minőségellenőrzés, meta-adatolás és az engedélyeztetésben való közreműködés közül csak az első feladat valósult meg. Június végéig 11 partnerkönyvtártól összesen 3411 webhelycímet kaptunk, melyek többsége az adott megyéhez vagy annak tágabb környezetéhez kapcsolódik. Az ellenőrzések után 838 URL bizonyult általunk már ismertnek, 219 érdektelennek, és 2354 volt közöttük a hasznos, új cím. Ezekről csináltunk egy gyors mentést, majd besoroltuk őket a tematikus részgyűjteményeinkbe, így a továbbiakban már rendszeresen aratásra kerülnek. Bár a pályázat lezárultával ez az együttműködés abbamaradt, de néhány könyvtár jelezte, hogy szívesen folytatná ezt a munkát, illetve kipróbálna más munkafázisokat is, ha ennek a kereteit ki tudjuk alakítani.

A nemzetközi együttműködések fő katalizátora az International Internet Preservation Consortium, melynek 2018 óta az Országos Széchényi Könyvtár is tagja. Az IIPC blogjában 2020 júniusában jelent meg az OSZK webarchívumának és új honlapjának bemutatása. Az internetes tartalmak megőrzésével foglalkozó intézményeket és projekteket összefogó szervezet 2020-as közgyűlését online formában tartotta meg, ami számunkra szerencsés volt, mert az eredetileg tervezett helyszínre, Montrealba nemigen tudtunk volna eljutni, így viszont mi is figyelemmel kísérhettük a rendezvényt. A konzorcium munkacsoportjai közül az oktatási témájában veszünk részt, amely az év folyamán több angol nyelvű tananyagot és ismeretterjesztő prezentációt is közzétett. A tagintézményektől időről-időre érkeznek hozzánk körlevelek, ezekben vagy egy felméréshez kérnek adatokat, vagy valamilyen speciális témájú archívumhoz helyi információforrások címlistáit. Előbbire példaként a belga BESOCIAL kutatást emelem ki, ami azt mérte fel, hogy a nemzeti könyvtárak hogyan próbálják a közösségi médiatartalmakat megőrizni. Utóbbival kapcsolatban pedig a COVID járvánnyal foglalkozó fontosabb online források összegyűjtésében segítettünk magyar részről. Az általunk gondozott koronavírus témájú webarchívumról egy külön interjú is készült és jelent meg a WARCnet Papers sorozatban. A WARCnet a 2020 és 2022 közötti időszakra kiterjedő nemzetközi projekt, központja a dán Aarhus University, a célja pedig a nemzeti és nemzetközi domének történetének kutatása a webarchívumokban megőrzött digitális tartalomra alapozva. Reményeink szerint mi is be tudunk majd kerülni az egyik munkacsoportba és publikálni a kiadványaikban.

Ismeretterjesztés

Az archiválás és a kutatás-fejlesztés mellett a téma népszerűsítését is fontosnak tartjuk, mert csak úgy lesz fenntartható a nemzeti könyvtár webarchívuma és majd minden további más, az internetes tartalmak megőrzésére indított hazai projekt, ha a memóriaintézményekben dolgozó szakemberek, az illetékes döntéshozók, a tartalomszolgáltatók és a szélesebb nagyközönség is megérti, hogy a digitálisan születő kultúránk érték, érdemes erőforrásokat fordítani a hosszú távú megőrzésére és kutathatóvá tételére. Ezért indítottuk el a Könyvtári Intézet segítségével azt az akkreditált tanfolyamot, melyet legutóbb 2020 elején tartottunk meg. Az őszi alkalom a korlátozások miatt elmaradt ugyan, de 2021 februárjában online formában ismét meghirdette a KI és a 25 fős keret néhány nap alatt betelt. Emellett elkészítettük egy másik, elsősorban a középiskolás korosztálynak szánt ismeretterjesztő anyag kéziratát is, ami a személyes és az intézményi internet archiválás fontosságáról és módszereiről szól. Ennél egyelőre a tananyag fejlesztő cégre várunk, hogy elkészítse a multimédia mellékleteket és e-learning formátumra konvertálja a kéziratot.

2020. december 2-án negyedik alkalommal tartottuk meg „404 Not Found - Ki őrzi meg az Internetet?” című rendezvényünket, melyet a járvány miatt online, a Teams alkalmazáson keresztül követhettek a résztvevők. Valószínűleg részben ennek is köszönhető, hogy minden korábbinál többen, kb. 160-an regisztráltak és az online csatornán egyszerre jelenlevők száma meghaladta a 130-at. A konferenciát Rózsa Dávid főigazgató és Dr. Latorcai Csaba közigazgatási államtitkár nyitotta meg, majd előadást tartott Szóllás Péter, az EMMI főosztály-vezetője. A mi részünkről Drótos László számolt be a projekt 2020-as eredményeiről, majd három előadás foglalkozott a fő témával választott együttműködésről: Maczelka Árpád a KDS projektet ismertette, Visky Ákos László az eddigi és a lehetséges intézményközi munkamegosztásokról beszélt, Szécsényi Orsolya, a Bródy Sándor Megyei és Városi Könyvtár könyvtárosa pedig az intézményük digitális helyismereti gyűjteményét mutatta be és kiemelte a lehetséges kapcsolódási pontokat az OSZK webarchívumával. A délutáni, workshop jellegű részben Drótos László és Németh Márton egyedi webarchiváláshoz használható szoftvereket mutatott be Windows, illetve a virtuális gépen futó Linux alatt. Az előadások és a workshop anyagát feltöltöttük a [honlapra](#) és a [Videotoriumba](#).

Az elmúlt évben is sikerült [több publikációt](#) megjelentetnünk hazai és külföldi szaklapokban, illetve előadást tartani különféle rendezvényeken. Az őszi halasztott és videókonferencia formájában megrendezett Networkshopon Moldován István a Rákóczi archívumot mutatta be, Németh Márton pedig a mikroadatok felhasználási lehetőségeiről beszélt a webarchiválás különböző fázisaiban és egy tutoriált is tartott. A szemantikus webes technológiák webarchívumokban való hasznosítása egy nemzetközileg is újdonságnak számító téma és bekerült Németh Márton PhD dolgozatába is, melynek elővédése néhány napja sikeresen megtörtént. Szakirodalmi téren említést érdemel még az „A blended learning-based curriculum on web archiving in the National Széchényi Library” című 2019-es tanulmányunk, mivel 2020-ban beválasztották az Emerald kiadó által kiemelten fontosnak ítélt publikációk közé és megkapta a „Literati Award” elismerést.

A szakmai kommunikáción túl tavaly több alkalommal sikerült a média érdeklődését is felkeltenünk az OSZK webarchívuma iránt; megjelentek rólunk sajtóhírek, szerepeltünk tévében és rádióban, így szélesebb körben is fel tudtuk hívni a figyelmet erre a területre.