Az OSZK Webarchívum 2021 júniusi hírei

Webtér aratás

Elindult egy újabb webtér szintű aratás, melyet félévente futtatunk az összes olyan címre kiterjedően, amiről tudomásunk van. Legutóbb, 2020 december végén 251 ezer URL-ről indítottuk el a robotot. Ezt a listát most kibővítettük az időközben általunk összeválogatott webhelyekkel, valamint azokkal a .hu végű domén és aldomén nevekkel, melyeket a 2020 óta lementett weboldalakban levő linkekből gyűjtöttünk ki. A különböző szűrések után végül 446 ezer tételes lett az aktuális seed lista. Ezek közt is lehetnek még elég nagy számban duplumok, lényegében üres honlapok, parkoló domének, illetve olyan webszerverek, amelyek robottal nem archiválhatók, mert automatikusan nem deríthetők fel ezek teljes biztonsággal, de így is igen jelentősen sikerült megnöveli azt a kört, amelyről egy „pillanatfelvételt" tudunk készíteni.

Nyilvános gyűjtemény

Júniusban 45 újabb webhellyel – főként közgyűjtemények honlapjaival – bővült a nyilvános webarchívum: https://webarchivum.oszk.hu/demo-kezdolap/ A mentések a WCT keretrendszerben futó Heritrix robottal készültek, melyeket kiegészítettünk a HTTRack programmal készített másolatokkal is, amennyiben a robottal nem sikerült jól learatni valamelyik site-ot. A 2021 januárjában életbe lépett 626/2020. sz. kormányrendelet 6. § (3) pontja szerint már nem szükséges az állami és önkormányzati, illetve a közpénz felhasználásával készült webhelyek esetében egyedi szerződést kötni a Nemzeti Könyvtárral. Ennek köszönhetően a közeljövőben a publikus gyűjtemény további jelentős bővítése várható. Hogy a határeseteknél mi a követendő elv, arról jogászi véleményt kérünk majd. Ehhez egy munkaanyagot állítottunk össze az elmúlt hetekben az eddig felmerült konkrét példákból. Érdekes egybeesés, hogy jelzést kaptunk arról, a régi, még működő, de már nem frissülő megyei levéltári honlapok már nem sokáig lesznek elérhetőek. Három honlapra korábbról már volt szolgáltatási engedélyünk, a rendeletben kapott felhatalmazással most még hat webhely, valamint négy tematikus oldal archivált változatát tudtuk közzétenni és így biztosítani további elérhetőségüket. A Somogyi Könyvtár honlapja is hamarosan megújul, az eddig változat szintén megőrződik és elérhető a nyilvános archívumban is.

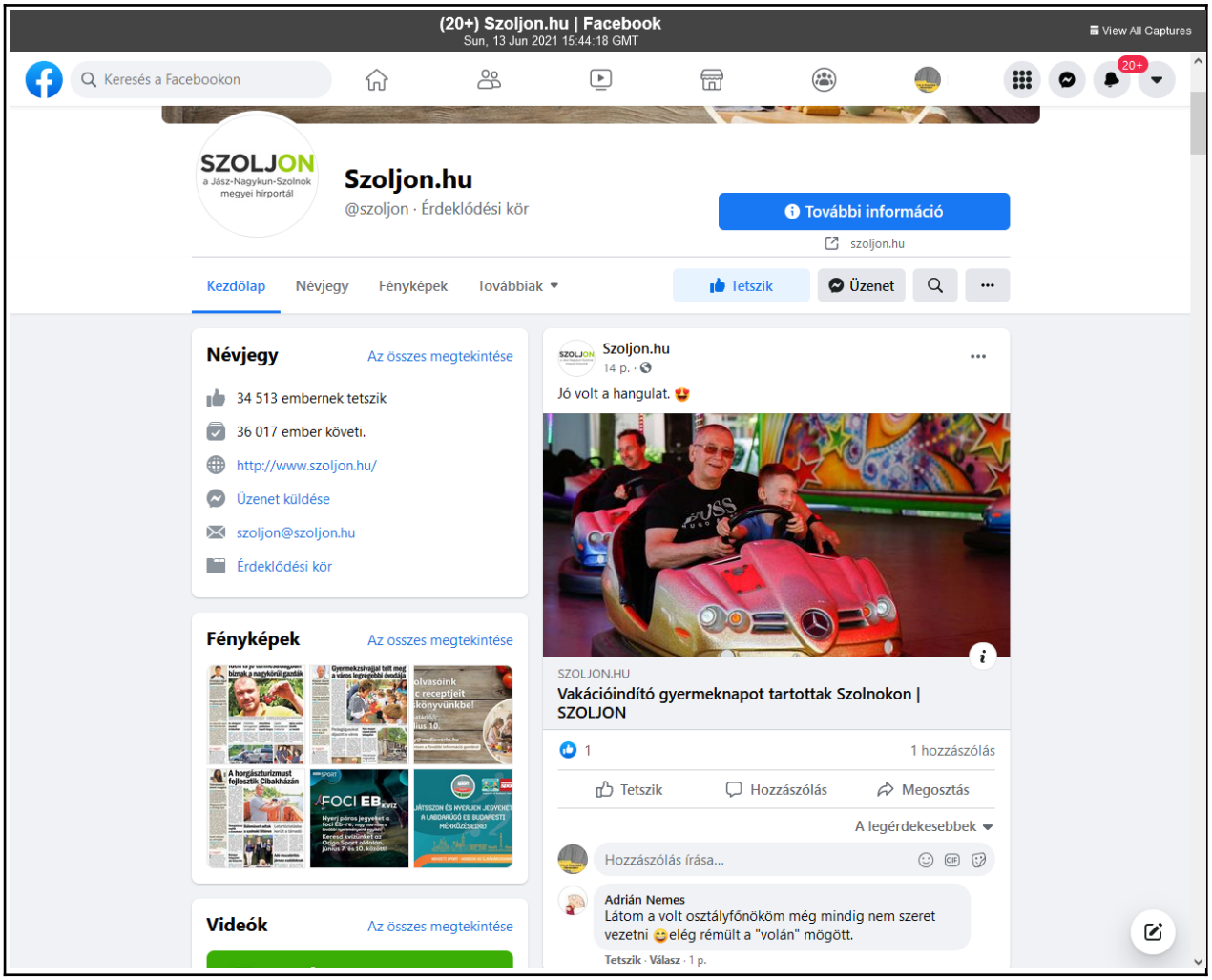
Közösségi média

A webkettes felületek archiválhatóságának tesztelése során az elmúlt hónapban 20 hazai és határon túli magyar híroldal **Instagram** posztjait mentettük le. Összesen 11.700 bejegyzést (fotót, képsorozatot és rövid videót) töltöttünk le 15,5 GB méretben, a legrégebbi 2013 júliusi. Ezekkel együtt már 745 Instagram fiókról van mentésünk: https://webarchivum.oszk.hu/bongeszes-instagram-oldalak/

Június elején lezárult a **Twitter** csatornák archiválhatóságának tesztelése. Több mint ezer Twitter fiókot mentettünk le, ebből 594 esetben sikerült teljes egészében letölteni a tweet listát és az abba beágyazott videókat és képeket (utóbbiakat csak kis méretben). A többinél csupán az utolsó kb. 750 tweetet lehetett visszagörgetni a rendszer korlátai miatt. A tapasztalatok szerint sok intézmény és közszereplő csak a Facebook és/vagy Instagram posztjainak linkjeit teszi ki a Twitterre, vagyis érdemi tartalom nincs a csatornájukon. A Twitter amúgy is kevéssé népszerű nálunk, a 2010-es években létrehozott fiókok elég nagy hányada már nem aktív. De azért vannak olyanok is, ahol naponta több, médiatartalommal kiegészített rövid szöveges bejegyzés jelenik meg, vagyis teljes értékű hírforrásként funkcionál a csatorna. Az archivált címek listája: https://webarchivum.oszk.hu/bongeszes-twitter-csatornak/

Újra elkezdtük nyilvános **Facebook** fiókok idővonalát menteni, amivel tavaly kénytelenek voltunk leállni a cég által bevezetett technikai változtatások miatt. Most az ArchiveWeb.page nevű Chrome kiegészítővel archiválunk, de az egyes posztokat külön nem mentjük, mert az nagyon időigényes. Első lépésben 14 történelmi témájú oldal idővonalát próbáltuk letölteni, ami az esetek felében az első bejegyzésig visszamenőleg sikerült is. A következő fázisban 64 hírportál és egyéb időszaki kiadvány Facebook oldaláról

készítettünk mentéseket, de ezeken olyan mennyiségű tartalom (kép és videó) van, hogy néha még az utolsó egy hónap anyagának letöltése is nehézséget jelent és a megjelenítéssel is vannak problémák. Jelenleg kormányzati és önkormányzati intézmények, továbbá politikai és civil szervezetek oldalait mentjük. Eddig 2483 Facebook oldalról készült legalább egy alkalommal teljes vagy részleges mentés. A címlista itt nézhető meg: https://webarchivum.oszk.hu/bongeszes-facebook-oldalak/



Időszaki kiadványok

Az elektronikus periodikák weboldalainak nyilvántartása 51 tétellel bővül az elmúlt hónapban, főként az Elektronikus Periodika Archívum és Adatbázis gyarapodása és a megyei könyvtárak honlapjain talált linkek alapján, így a teljes címlista mérete rövidesen eléri majd a 6000 tételt. Az ELPERI részgyűjtemény következő aratására júliusban kerül sor. https://webarchivum.oszk.hu/bongeszes-elektronikus-periodikak/

MIA Wiki

Közel 40 új vagy frissített szócikkel, főként közösségi oldalak archiválásra is alkalmas eszközök leírásával fejlesztettük tovább az internetes tartalmak megőrzésével foglalkozó wikinket, amely így már több mint 700 bejegyzésből áll: https://webarchivum.oszk.hu/mediawiki/ (Az éppen 4 évvel ezelőtt indított tudástár http://mekosztaly.oszk.hu/mediawiki/ címen levő korábbi változata a szerver meghibásodása miatt jelenleg nem elérhető, ezért a rá mutató régi linkek nem irányítódnak át automatikusan az új felületre!)

ASIS&T konferencia

Az ASIS&T, az információtechnológiával és információtudománnyal foglalkozó kutatókat és intézményeket tömörítő nemzetközi szervezet európai tagozata (European Chapter) “Information Science Trends 2021: Information Science Research During COVID-19 & Post-Pandemic Opportunities” címmel rendezett konferenciát a koronavírus járvánnyal összefüggő, az információtudományt sokféle vonatkozásban érintő tapasztalatokról, 2021. június 9-11. között az online térben. Ennek keretében június 11-én húszperces előadást tartottunk “Establish a COVID-19 Web Archiving Collection at the National Széchényi Library” címmel a koronavírus témájú gyűjteményünk gondozásáról, illetve a járvánnyal kapcsolatos nemzetközi archiválási tevékenységekről, melyekben mi is részt vállaltunk. Az előadás prezentációja itt letölthető: https://webarchivum.oszk.hu/wp-content/uploads/2021/06/asist-eu-2021-NM-v3-1.pptx

IIPC GA + WAC és RESAW

Lezajlott az International Internet Preservation Consortium (IIPC) éves közgyűlése és konferenciája, amit ezúttal is online rendeztek meg, összekapcsolva az archivált webtartalom kutatási célú hasznosításával foglalkozó 4. RESAW konferenciával: https://netpreserve.org/ga2021/ és https://www.resaw2021.net/ Mi most csak egy 5 perces összefoglalást tartottunk az elmúlt egy évről: az új jogszabályi környezetről és az önálló osztályról, a közösségi média archiválási kísérleteinkről és az együttműködési terveinkről. https://webarchivum.oszk.hu/wp-content/uploads/2021/06/A-quick-update-about-the-Hungarian-web-archive.pptx
A rendezvényhez használt videókonferencia rendszer speciális funkcióinak köszönhetően kötetlen beszélgetésekre is sor kerülhetett, így tovább tudtuk építeni a kapcsolatainkat az Internet Archive és a környező országok webarchívumainak munkatársaival.

Gyűjtőköri leírás

Az IIPC levelezőcsoportban közzétett kérésünkre a külföldi kollégáktól kapott információk, valamint a saját jelenlegi gyakorlatunk alapján összeállítottunk egy előkészítő anyagot a webarchívum leendő gyűjtőköri szabályzatához.

Internet Archive

Levelet váltottunk az Internet Archive-val a magyar domének listájának átadására, illetve az általuk 1996 óta archivált magyar weboldalak kereshetővé tételére vonatkozóan. Megerősítették, hogy a 2018-ban kapott árajánlat lényegében továbbra is érvényben van, annak ellenére, hogy azóta jelentősen megnőtt a náluk található magyar webtartalom.

WCT 3.0

Elkezdtük tesztelni a nyilvános gyűjteményhez használt Web Curator Toolkit keretrendszer 3-as verzióját (https://webcuratortool.org/). Egyelőre még egyeztetünk az új-zélandi nemzeti könyvtár fejlesztőivel, mert beleütköztünk egy technikai problémába. Időközben már a 4-es változatról is vannak hírek: bemutatták az IIPC idei konferenciáján és a 2020-as Web Archiving and Digital Libraries Workshopon is: https://vtechworks.lib.vt.edu/handle/10919/99569 A 4.0-ás WCT főként a minőségbiztosításhoz fog új funkciókat nyújtani és magába integrálja a PyWb megjelenítőt is, amely egyben archiváló eszközként is használható lesz a Heritrix-es mentésből hiányzó weboldalak vagy egyéb fájlok utólagos pótlásához.

Regionális gyűjtemények

Folyik a megyei könyvtárakkal való együttműködést formalizáló szerződés előkészítése. Kialakítottunk egy sablont (https://docs.google.com/spreadsheets/d/1VbxOarc2sDzNrySEG2lM7-85f9Q-In9E5W1jqu1-ACs/) a regionális és tematikus virtuális gyűjtemények számára, és két megye esetében a településnevek alapján már leválogattuk az általunk eddig összegyűjtött webhelyeket. Új ötletként a Digitális Tartalomfejlesztési és -szolgáltatási Osztállyal együtt szeretnénk bevonni ezeket a partnereket egyéb online digitális

dokumentumtípusok válogatásába és engedélyeztetésébe is. Megbeszélést folytattunk az ELTE Állam- és Jogtudományi Kar könyvtárának vezetőjével is. Ők az együttműködés keretében a jogi vonatkozású webhelyek nyilvántartásában érdekeltek.

Az elmúlt hetekben lefutott tematikus aratások

Egyetemek, főiskolák (3598 db seed URL)
Kutatóintézetek, tudományos szervezetek (1000 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (881 db seed URL)
Vallások, hitrendszerek, egyházak (2620 db seed URL)
Idegenforgalom, vendéglátás (5206 db seed URL)
Közoktatás és egyéb képzések (5989 db seed URL)

Közösségi média mentések

Facebook: 185,5 GB (79 warc fájl)
Instagram: 16 GB (9 warc fájl)
Twitter: 10,3 GB (23 warc fájl) *(májusi és áprilisi adatokkal együtt)*

A tematikus aratások részletes statisztikai adatai a https://webarchivum.oszk.hu/szelektiv-aratasok/ weblapon nézhetők meg. A projekt hírei a https://webarchivum.oszk.hu/a-projektrol/hirek-esemenyek/ oldalon kísérhetők figyelemmel. Kapcsolati cím: mia@mek.oszk.hu