Az OSZK Webarchívum 2024. áprilisi hírei

Archiválás és nyilvántartás

A legújabb részgyűjteményünk a kereskedelem, a szolgáltatás, a szállítás és a közlekedés tematikájába sorolható webhelyeket tartalmazza, beleértve egyebek mellett a webáruházakat, a szakemberkereső és az apróhirdetési oldalakat, az ingatlanközvetítőket, a taxitársaságokat és a futárszolgálatokat, továbbá az autósiskolákat is. Két hónap alatt összesen 9075 URL címet sikerült összeszedni és altémákba sorolni, melyek 8766 különböző honlaphoz vagy bloghoz tartoznak. A SZOLGKER kódnevet kapott válogatás első aratására április végén került sor. A Heritrix szoftver által letöltött tartalom mennyisége kevesebb mint egy nap alatt elérte a maximális mérethatárnak megadott fél terabájtot, ami érhető egy ilyen nagy méretű és korábban legfeljebb csak a webtér szintű aratás részeként kisebb mélységben mentett webhelyeket tartalmazó seed listánál.

A honlapunkról nyilvánosan elérhető demó archívumba a hónap folyamán 63 új tételt tettük ki, főként kormányhivatalok honlapjait, illetve választási és népszavazási oldalak mentéseit. Megint akadt a beválogatottak között néhány olyan webhely, amelyeket semmilyen eszközzel nem tudtunk jól archiválni a nem „robot-barát" technológiák használata miatt. Sajnos ilyenek az idei európai parlamenti és önkormányzati képviselőválasztás aloldalai is a Nemzeti Választási Iroda honlapján. Jó viszont, hogy a választásokra regisztrált szervezetek listája letölthető a honlapról, így ez alapján közel 80 tétellel tudtuk bővíteni a pártok és civil szervezetek weboldalait tartalmazó nyilvántartásunkat.

A Webrecorder projekt keretében fejlesztett Browsertrix Crawler és Browsertrix Cloud tesztelése céljából – a többi IIPC taggal együtt – több mint egy évig ingyenesen használhattunk egy távoli szervert és a hozzá tartozó tárhelyet. Ez a lehetőség május végével megszűnik, mert bár maguk a szoftverek nyílt forráskódúak és ingyenesek maradnak, a felhőszolgáltatást előfizetéshez kötik. Mivel a korábban sokszor akadozó tesztrendszer most már egészen megbízhatóan működik és a böngészőn keresztül archiváló robot is jól paraméterezhető lett, ezért a még ingyenes időszakot kihasználva április 17-én 11 részletben megadtunk neki egy 3264 URL címet tartalmazó listát, amit még tavaly ősszel állítottunk össze a Karikó Katalinnal kapcsolatos hazai és külföldi hírekből, intézményi aloldalakból, Wikipédia szócikkekből, podcast adásokból, YouTube és Vimeo videókból. Ezek többsége olyan megoldásokat használ, amelyekkel a Heritrix robot nem boldogult volna, viszont a Browsertrix néhány óra alatt betöltötte, majd végiggörgette az oldalakat, és ahol tudta, ott a beágyazott lejátszót is elindította. Problémák persze így is akadtak mind az archiválásnál (pl. a YouTube egy idő után egy captcha-t dobott fel), mind pedig a visszanézésnél (pl. a PyWb megjelenítő sokszor nem tudja összerakni a Browsertrix által generált WARC fájlokból a weboldalakat), de összességében több mint 24 gigabájtnyi, korábban még nem mentett tartalmat sikerült betölteni a Szegedi Egyetem könyvtárával közösen létrehozott Karikó Katalin webarchívumba.

Konferenciák

Április 4-én Kalcsó Gyula, a webarchiválási csoport vezetője „Képek és metaadataik gyűjteményezése scrapingtechnológiával közösségi képmegosztó oldalról" címmel tartott előadást Egerben az idei Networkshop konferencián. A prezentáció letölthető a honlapunkról.

Ugyancsak ő képviselte az OSZK-t az internetes tartalmak megőrzésével foglalkozó intézményeket tömörítő nemzetközi konzorcium, az IIPC idei közgyűlésén és konferenciáján Párizsban április 23. és 25. között. A rendezvény egyik kapcsolódó eseménye a fiatal kutatók számára megszervezett „Early Scholars Spring School on Web Archives" című szeminárium volt, melynek második napján ő is részt vesz az egyik kerekasztal beszélgetésen. Az ehhez készült „Web Archiving and its Research Use at the National Széchényi Library (Hungary)" című prezentációja szintén elérhető a honlapunkon. A konferenciát megelőző workshopokon, majd magukon az előadásokon és a pár perces rövid beszámolókon, a panelbeszélgetéseken és a poszter szekcióban is számos olyan téma merült fel, melyekkel már mi is találkoz-

tunk, vagy még jövőbeli feladatként előttünk állnak. A szünetek és a kísérő események pedig természetesen kiváló lehetőséget adtak a kapcsolatépítésre és a tapasztalatcserére. A rendezvényről részletes beszámoló jelenik meg majd az OSZK blogjában, a bemutatott újdonságokról pedig szócikkeket írunk a MIA Wikibe.

„Networking” az IIPC konferencián a francia nemzeti könyvtár ovális olvasótermében

A párizsi rendezvény előtt, április 11-én egy másfél órás videokonferenciát is szerveztek az IIPC tagoknak, részben azért, hogy akik nem tudnak személyesen részt venni a konzorcium közgyűlésén, azok is meghallgathassák az új elnökség és az egyes munkacsoportok beszámolóit, illetve a tavalyi kérdőíves felmérésre érkezett válaszok összefoglalóját, részben pedig hogy a már rendszeressé vált „member updates” keretében öt közgyűjtemény (Harvard University Libraries, Smithsonian Institution, University of North Texas Libraries, British Library, Royal Danish Library) munkatársai röviden bemutathassák a náluk folyó aktuális webarchiválási munkálatokat.

Együttműködés

Áprilisban két újabb megbeszélést tartottunk a székesfehérvári Vörösmarty Mihály Könyvtár munkatársaival a Web Curator Tool és a PyWb programokról. Megkaptuk tőlük azt a részletesen kidolgozott, az OSZK webarchívumában használt témakörökre, valamint a Köztauruszra alapozott kategória- és tárgyszójegyzéket, amit a Fejér vármegyei regionális webarchívum anyagának metaadatolására fognak használni. Több körös egyeztetés után rövidesen véglegesíteni fogjuk az együttműködés kereteit, a VMK és az OSZK vállalásait rögzítő szerződés szövegét is.

Statisztikák

Az osztályunk munkájának dokumentálásához leadtuk az első negyedévben lezajlott aratások összesített adatait. Frissítettük továbbá a honlapunkon található statisztikát és grafikont, mely így most a 2023 év végi állapotot mutatja. A honlapba beépített Google Analytics mérőkód alapján készült látogatói statisztika is elavult néhány héttel ezelőtt, mert GA4 verzió API-ja már nem támogatja azt módszert, amivel le tudtuk tölteni a forgalmi adatokat. Az új API lekérdezéséhez az Analytify nevű WordPress kiegészítőt telepítettük fel. Sajnos ennek az ingyenes változatában nem érhető el az a funkció, amivel az adatok és a diagramok a nyilvános felületre is kitehetők, ezért jelenleg csak a honlap adminisztrátorai tudják ezeket megnézni.

Közösségi média

Elraktároztuk az OSZK podcast csatornáján megjelent újabb adásokat (29 db hangfájl, 3 GB összméretben), valamint belinkeltük a Széchényi Ferenc archívum honlapjára a nemzeti könyvtár blogjában megjelent, a könyvtáralapító gróf naplóbejegyzéseiről és utolsó bécsi éveiről szóló kétrészes cikket, amit dr. Deák Eszter, a Régi Nyomtatványok Tárának munkatársa írt. (A blogban felhasznált képek forrásának beazonosításában is segítettünk.) Képernyőfotókat és rövid videókat készítettünk a webarchívummal kapcsolatos hírek illusztrálásához, melyek az OSZK Facebook oldalán és egyéb közösségi média felületein jelennek meg.

Az elmúlt hetekben lefutott tematikus aratások

Szolgáltatás, kereskedelem, szállítás, közlekedés (9075 db seed URL)
Életmód, szabadidő, hobbi (8830 db seed URL)
Történelem, hely- és családtörténet (1470 db seed URL)
Média, sajtó, műsorszórás (963 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1488 db seed URL)
Podkasztok (4730 db seed URL)
Elektronikus periodikák (10637 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (7221 db seed URL)

Az egyes 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: webarchivum@oszk.hu