Az OSZK Webarchívum 2022 májusi hírei

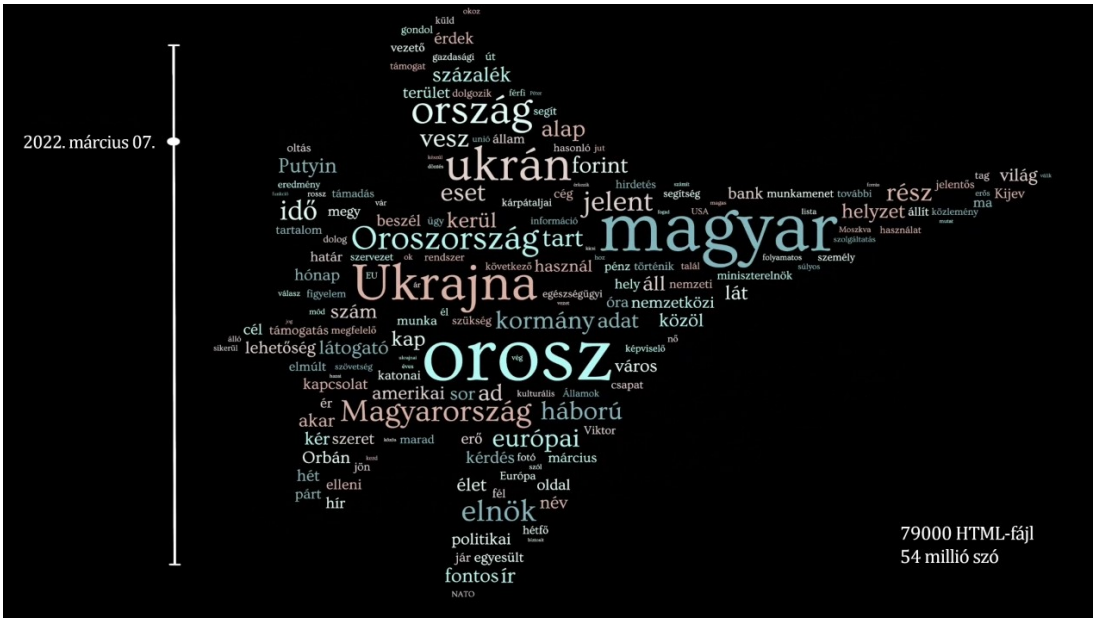
Archiválás

A megyei könyvtárakkal történő együttműködés kapcsán merült fel, hogy a honlapjaik kerüljenek be a nyilvános webarchívumba, amit megkönnyít az, hogy a 2020-as kormányrendeletnek köszönhetően már nem kell külön felhasználási szerződést kötni rájuk, mivel közpénzből fenntartott oldalak. 15 ilyen könyvtár honlapja még nem volt benne a nyilvános gyűjteményben – a többire korábban már kértünk engedélyt –, így most ezekről is elkészítettük a szükséges mentéseket, és egyedi paraméterezéssel próbáltuk javítani azok eredményességét.

A webtér szintű aratáshoz szükséges seed lista előkészítésének újabb munkafázisát végeztük el: az eredetileg 5,5 millió címből álló listából kiválogattuk azokat, amelyek tömegesen generált aldomének (pl. cég- és terméknyilvántartó adatbázisok alapján), mert ezeket nincs értelme részletesebben archiválni, mivel valójában a fő doménen található webhely aloldalai. Az így kapott 3,3 millió címből még kivettük a korridor.hu és a booked.hu alá bejegyzett aldoméneket, mert ezek együtt több mint 2 millió címet jelentettek, a többinél pedig begyűjtöttük a szerver által visszaadott fejlécet, benne a webhely elérhetőségét vagy átirányítását jelző státuszt és (jó esetben) a webhely nevét tartalmazó *title* adatot. A következő lépés ezek alapján annak eldöntése lesz, hogy ezek közül a tömeges aldomének közül melyeket érdemes legalább kisebb mélységben lementeni.

Bár a hónap folyamán nem gyűjtöttünk aktívan újabb címeket a válogatott részgyűjteményeinkbe, de különböző forrásokból legalább félszáz címmel bővültek a tematikus címlistáink és ugyanennyivel az elektronikus periodikákat tartalmazó nyilvántartásunk is. Csökkent viszont a MEDIA nevű seed lista mérete, mert kivettük belőle a podkasztokat. Ezekből egy külön műfaji gyűjteményt alakítottunk ki és elkezdtük a weboldalak mellett maguknak az adásoknak a hangfájljait is lementeni, mivel ezeket a legtöbb esetben nem tudta archiválni a Heritrix robot. Egy Chrome kiegészítő, illetve különböző online letöltő/konvertáló szolgáltatások segítségével egyedi mentéseket készítünk az MP3 vagy M4A formátumú állományokról, egységesítjük a fájlneveket (azonosító + sorszám + eredeti adásnév) és egy külön alkönyvtárban jelenítjük meg őket a zárt archívumon belül, a podkaszt archivált weblapja mellett. Az elmúlt két hétben az eddig nyilvántartott 185 podkasztból 110-nél már elkészült a mentés, nagyrészt az első adásig visszamenőleg, ami közel 10 ezer hangfájlt jelent 660 gigabájt összméretben. Mivel becslések szerint már több mint ezer magyar podkaszt létezik, ezért a következő hónapokban további jelentős bővítést tervezünk ennél az új részgyűjteménynél.

Az orosz-ukrán háború miatt kiemelten fontosnak számító és heti rendszerességgel mentett kárpátaljai webhelyek listája néhány irodalmi honlappal és bloggal bővült, melyeket azok tulajdonosa ajánlott archiválásra. Új Facebook oldalakat most nem gyűjtöttünk, viszont összeválogattunk 26 Instagram fiókot és egyedi mentéseket készítettünk róluk böngészőn keresztül (mivel ezek robottal nem arathatók). Ugyanígy lementettük a *#Kárpátalja*, a *#kárpátaljaimagyarok* és az *#oroszukránkonfliktus* hashtag-ekkel ellátott képeket is az Instagramról. A hazai és határon túli portálokról gyűjtött ukrajnai hírek szövegének feldolgozását a webarchívum rendszergazdája és a Digitális Bölcsészeti Központ (DBK) munkatársai együttműködésének köszönhetően nagyrészt sikerült teljesen automatizálni. A feldolgozott adatokról készült vizualizációk a https://dhupla.hu/page/kreativ/ oldalon nézhetők meg, a teljes szövegű kereső pedig itt érhető el: https://ukrajnapublic.webharvest.oszk.hu/solrwayback/. Az UKRAJNA2022 és a KARPATALJA nevű aratások hetente lefutnak, mint ahogy még a KORONAVIRUS2020 is, leállítottuk viszont az új kormány megalakulása után az országgyűlési választással kapcsolatos híreket begyűjtő OGYVAL2022 seed lista aratását, de csináltunk két soron kívüli mentést a KORMONKOR gyűjteményben levő kormányzati webhelyekről.



Az ukrajnai hírek szövegéből generált szófelhő

Ismeretterjesztés

Május folyamán több rendezvényen és egyéb fórumon is sikerült hírt adnunk a munkánkról. A debreceni BOBCATSSS 2022 konferencia első napján Németh Márton „Web Archiving in Higher Education" címmel tartott előadást a webarchiválás és a webarchívumok fontosságáról a felsőoktatásban, valamint egy szekciót is levezetett, az informális beszélgetéseken pedig igyekezett ráirányítani a figyelmet a tevékenységeinkre.

A hónap közepén volt az IIPC szervezet közgyűlése és konferenciája. Idén is mindkettő online zajlott és így valamennyi workgroup ülésén részt tudtunk venni, köztük a képzési munkacsoportén is, amiben Németh Márton képviseli a nemzeti könyvtárat. A beszámolók és a fórumbeszélgetések során tájékoztatást kaptunk a közös webarchiválási projektekről, valamint az archiváló és visszakereső/megjelenítő, nagyrészt nyílt forráskódú szoftverek fejlesztéseiről is. A konferencián Németh Márton és a DBK munkatársa, Kalcsó Gyula közös előadást tartott „Data extraction & visualization of harvested WARC files at National Széchényi Library" címmel az ukrajnai hírek archiválásáról, illetve azok természetes nyelvi elemző eszközökkel történő feldolgozásáról, s az ehhez kapcsolódó adatvizualizációkról. Az előadás előzetesen videón is rögzítésre került, majd ahhoz kapcsolódva élő fórumbeszélgetésen volt alkalmunk részt venni a British Library munkatársaival, illetve PhD hallgatókkal, akik a webarchívumok tudományos célú felhasználásával foglalkoznak.

A WARCnet projekt keretében megjelenő könyvnél bekerült a tervezetbe a két általunk javasolt fejezet az ukrajnai projektről, illetve a webarchiválás és a felsőoktatás kapcsolatáról.

„60 terabájtnyi weboldalt őriz a nemzeti könyvtár!" címmel készült egy rövid videó az OSZK YouTube csatornájára, melyben Németh Márton az áprilisi Networkshopon elhangzott előadása témájáról, a webarchívum nemzetközi kapcsolatairól beszél. Az OSZK blogján pedig szintén egy általa írt beszámoló olvasható a Networkshop 2022 konferenciáról.

Május 18-án az OSZK-ban tartotta a Magyar Könyvtárosok Egyesülete Helyismereti Könyvtárosok Szervezete a taggyűlését, melyen Visky Ákos László „ad hoc" jelleggel egy rövid, bemutatkozó előadást rögtönzött a webarchiválásról és a megyei könyvtárakkal tervezett együttműködésről. Erre egy véletlen folytán került sor: a taggyűlés előadója, Sipos Júlia bemutatta munkatársunkat Mennyeiné Várszegi Judit szekció elnöknek, aki érdeklődve hallgatta annak az együttműködésnek a történetét, ami pont helyismereti szempontból fontos a résztvevő könyvtárak számára, ezért történt a soron kívüli felkérés.

Szepesi Judit a Magyar Művészeti Akadémia könyvtárosa meghívta az osztály munkatársait az őszi Tudomány Napjára szervezett beszélgetésre, mely a webarchiválás és a művészetek kérdését szeretné körbejárni „Big Data / Internet of Things, avagy hogyan archiváljuk művészetet?" címmel.

Együttműködések

Májusban újra elkezdtük a webarchiválás és egyedi *born digital* dokumentumok gyűjtéséhez kapcsolódó együttműködés kialakításának szervezését, és megpróbáltuk ismét felvenni a kapcsolatot azokkal a megyei hatókörű könyvtárakkal, melyekkel tavaly különböző okokból ez nem sikerült. Ezúttal gyors és érdemi válaszokat kaptunk, a kilenc megszólított könyvtár közül hattal már a megbeszélések is megtörténtek május folyamán, de sajnos három könyvtártól továbbra sem érkezett reagálás, nekik emlékeztető levelet küldtünk. A megbeszéléseken részt vevő kollégák pozitívan fogadták a kezdeményezésünket, egyetértettek a felvetéssel, jelenleg a formális intézményi válaszokra várunk. Addig is a településnevek alapján kigyűjtöttük az egyes megyékhez kapcsolódó, általunk már nyilvántartott webhelyek listáit, melyek a leendő virtuális regionális gyűjtemények alapjaiként szolgálhatnak. A hónap folyamán ezekkel a könyvtárakkal beszéltünk: Csorba Győző Könyvtár, Pécs; Eötvös Károly Megyei Könyvtár, Veszprém; Hamvas Béla Pest Megyei Könyvtár, Szentendre; Móricz Zsigmond Megyei és Városi Könyvtár, Nyíregyháza; Takáts Gyula Megyei és Városi Könyvtár, Kaposvár; Tolna Megyei Illyés Gyula Könyvtár, Szekszárd.

Felvettük a kapcsolatot Szőcs Endrével, a Székelyudvarhelyi Városi Könyvtár igazgatójával, miután kollégánkkal folytatott magánbeszélgetésen kifejezte érdeklődését a webarchiválás iránt. Formális megbeszélésre még nem került sor a lehetséges együttműködésről, mert előbb szeretne több közgyűjteményi szakemberrel is konzultálni a romániai magyar intézményekből, hogy minél tágabb kör tudjon bekapcsolódni a közös munkába. Ha sikerülne kialakítani egy élő kapcsolatot, az a helyi beágyazottság előnyénél fogva nagyban segítené a romániai magyar és magyar vonatkozású webhelyek összegyűjtését, és példaként szolgálhatna más szomszédos országok felé is.

A webarchiválási tevékenység 2022. június 1-től a Digitális Bölcsészeti Központ keretei között folyik tovább, ezt az átállást is elő kellett e hónapban készítenünk.

Az elmúlt hetekben lefutott tematikus aratások

Könyvtárak, levéltárak, múzeumok és galériák (1996 db seed URL)
Irodalom, irodalomtudomány és -történet (1410 db seed URL)
Természet- és műszaki tudományok (1486 db seed URL)
Közoktatás és egyéb képzések (6544 db seed URL)
Képző-, előadó-, zene- és filmművészet (8085 db seed URL)
Elektronikus periodikák (9428 db seed URL)

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