Az OSZK Webarchívum 2022 márciusi hírei

Archiválás

Az elmúlt havi archiválási munkák nagy részét – a beütemezett tematikus aratások mellett – az orosz-ukrán háborúval kapcsolatos feladatok jelentették. Az ukrajnai híreket már február 21-én elkezdtük gyűjteni a hazai és a határon túli magyar hírportálokról. Akkor 74 forrásból 115 linket válogattunk össze, melyek címkék alapján szűrik ki a releváns cikkeket ezekről az oldalakról. A seed listát egy hónappal később az időközben bevezetett, a háborúval, a menekültekkel és a szankciókkal kapcsolatos új címkékkel bővítettük, így már több mint 400 tételes, az aratása pedig minden hét elején történik. (https://webarchivum.oszk.hu/bongeszes-orosz-ukran-konfliktus-2022/) Visszamenőleg is megpróbáltuk összegyűjteni az ukrajnai híreket, ezért pár napja csináltunk két további mentést is, melyek a januárban, illetve a február 1. és 20. közötti időszakban megjelent cikkekre terjedtek ki. Ezekhez a pótaratásokhoz 35, illetve 41 seed URL-t adtunk meg a robotnak.

Március elején főigazgatói kérésre egy újabb részgyűjteményt kezdtünk építeni a kárpátaljai magyar és magyar vonatkozású webhelyekből, melyek elérhetősége vagy akár fennmaradása is veszélyben forog. A valamelyik tematikus vagy műfaji gyűjtemény részeként eddig is nyilvántartott és archivált honlapok, blogok, periodikák és közösségi oldalak listáját különféle linkgyűjtemények (pl. szaknévsor, ittmagyarulis.hu), a keresőgépek találati listái, a régebben archivált magyar weboldalakról a .ua országdoménre mutató linkek, valamint a Wikimédia Magyarország Egyesület vezetőjének segítségével a Wikidata adatbázisból kigyűjtött ukrajnai doménnevek alapján kezdtük el bővíteni. Továbbá segítséget kértünk intézményektől, könyvtárosoktól, a térséggel foglalkozó vagy ott élő szakemberektől, valamint a közösségi médián keresztül is, hogy ajánljanak számunkra eddig még ismeretlen kárpátaljai oldalakat. Érkezett is jó néhány javaslat, melyek többségét – ellenőrzés után – felvettük a nyilvántartásunkba. Jelenleg valamivel több mint 900 webcímet tartalmaz a lista, ezek közül kb. 400 darab Facebook oldal. (https://webarchivum.oszk.hu/bongeszes-karpatalja/) Utóbbiak ugyan elérhetőség szempontjából nem veszélyeztetettek, de fontos kordokumentumok lehetnek és kétséges, hogy meddig tudják még frissíteni őket a fenntartóik. A Heritrix robottal aratható webhelyekről heti két mentést készítünk (a magyar nyelvűekről nagyobb mélységben), a közösségi oldalakat pedig egyelőre legalább egyszer, de lehetőség szerint az oldal létrejöttéig visszagörgetve töltjük le egyenként az ArchiveWeb.page programmal.

Ezeken kívül február végén és március elején több mint 300 olyan Facebook fiókot is lementettünk, melyeket magyarországi intézmények vagy szervezetek tartanak fenn. Elsősorban a hazai településekhez kapcsolódó önkormányzatok, polgárőrségek, önkéntes tűzoltóságok és mentőszolgálatok oldalai kerültek bele ebbe az archiválási ciklusba.

Érdekes feladat volt a szegedi Somogyi Károly Városi és Megyei Könyvtár március közepén megújult honlapjának archiválása. A régi honlapról – a könyvtár kérésére – a leállítás előtti napokban csináltunk mentéseket többféle módon is, majd az új felületet is teszteltük archiválhatóság szempontjából. Mindkét honlapverzió visszanézhető különböző megjelenítőkkel a nyilvános gyűjteményünkben is. (https://webarchivum.oszk.hu/demo-kezdolap/)

A nyilvános archívum bővítése sajnos nem haladt előre márciusban az archiváló szerver és a rajta futó WCT keretrendszer hibái miatt, sőt néhány honlap elérhetőségét meg is kellett szüntetnünk, mert nem tudtuk megújítani a tartalomgazdákkal korábban határozott időre kötött szolgáltatási szerződést. A technikai hibákat a hónap végére sikerült elhárítani, így áprilistól megjelennek majd újabb archivált webhelyek a nyilvános gyűjteményben a mostanában engedélyezett vagy szolgáltatási engedélyt nem igénylők közül.

A zárt archívumhoz való hozzáférhetőség is megoldódni látszik. A technikai tesztek lezajlottak és a lehetőségek függvényében, illetve a jogászi szakvélemény alapján egyelőre két, kizárólag erre a feladatra dedikált gép lesz beállítva az OSZK-ban, az „Általános olvasó” nevű részlegen.

Kutatás

Új feladatként az OSZK Digitális Bölcsészeti Központjának munkatársaival együtt elkezdtünk dolgozni a webarchívum anyagának kutathatóvá tételén. Ennek első fázisa az ukrajnai, illetve kárpátaljai aratások kielemzése és vizualizációja, amihez egy projekttervet állítottak össze a DBK-s kollégák és kialakítottak egy munkafolyamatot. Hogy milyen méretű szövegkorpusz feldolgozásáról lesz szó, ahhoz némi támpontot ad az, hogy csak a február 21-i első hírportál aratás során 11 darab WARC fájl keletkezett, melyekben 82 ezer HTML oldal volt és ezekből 380 millió karakternyi szöveg állt elő. Az egyik részfeladat olyan szófelhők készítése, melyeken nyomon lehet követni a hírekben előforduló jellemző szavakat és ezek időbeli változását. Egy további feladat lesz az elmúlt években begyűjtött weboldalakban levő linkekben eddig talált 65 ezer darab ukrajnai webhelyről egy nyelvi elemző-eszközzel eldönteni, hogy melyek a magyar nyelvűek vagy magyar vonatkozásúak.

Ismeretterjesztés

Március második hetében ismét megtartottuk „Az internet archiválása mint közgyűjteményi feladat” című négy napos tanfolyamunkat a Könyvtári Intézet szervezésében. Az oktatás ezúttal is online zajlott és a 14 jelentkezőből végül 11-en vettek rajta részt végig és vizsgáztak le sikeresen.

A Webarchiválási Osztály munkatársainak több cikke is publikálásra kész állapotba került az elmúlt hetekben, a lektori és szerkesztői javaslatokat követő átdolgozások után. Németh Mártonnak a Könyvtári Figyelőben fognak megjelenni a tavalyi IFLA konferenciáról, illetve a „404 Not Found” workshop nemzetközi napjáról írt beszámolói. A szintén tavaly megrendezett varsói webarchiválási konferencia eseményeit összefoglaló cikke pedig a Könyv, Könyvtár Könyvtáros c. lap számára került leadásra március végén. Ugyancsak a 3K folyóirat jelenteti meg a „Webhelyek archiválási problémái” című írást, ami eredetileg Visky Ákos László és a Berzsenyi Dániel Könyvtár rendszergazdája, Perger Ádám 2021 novemberi közös workshop előadásának cikkváltozatához készült, de végül önállóan jelenik majd meg.

A kárpátaljai webtartalmak archiválásával kapcsolatban három rádióinterjú is készült márciusban Drótos Lászlóval, így a szakmán kívül sikerült szélesebb körben is felhívni a figyelmet a nemzeti könyvtárban folyó webarchiválási tevékenységre.

Az Internet Archive készülő magyar keresőfelülete

Nemzetközi kapcsolatok

Az Internet Archive elkészítette annak a portálnak az első verzióját, amellyel az 1996 óta a .hu domén alól begyűjtött weboldalakon levő szavakra, valamint a belinkelt kép, hang, videó és PDF fájlokra lehet keresni. A találatok az IA Wayback Machine szolgáltatásában jelennek meg. A teljes szöveg mellett természetesen konkrét URL-re is rá lehet keresni és ha egy olyan, az élő weben még elérhető webcímet adunk meg, ami még nincs benne az amerikai archívumban, akkor a rendszer felajánlja annak lementését. A hónap során több hibalistát és módosítási javaslatot is eljuttattunk a fejlesztőknek, köztük a felület automatikus magyar fordításában talált problémákat.

Március hónapban is folytatódtak a nemzetközi WARCnet projekt munkálatai és elkezdődött a jövőre megjelenő WARCnet könyv témáinak összegyűjtése. Németh Márton az OSZK képviseletében különféle munkacsoportok online értekezletein vett részt.

Válaszoltunk a pozsonyi Comenius Egyetem alapszakos hallgatójának a közösségi média archiválást érintő kérdéseire. Interjút egyeztettünk április végére a University of London egyik PhD hallgatójával, aki a webarchiváláshoz kötődő témában – szintén a webkettes oldalak megőrzési problémáira koncentrálva – írja a disszertációját és ehhez kért segítséget tőlünk.

Kitöltöttünk egy kérdőívet, mellyel az International Internet Preservation Consortium azt mérte fel, hogy hogyan próbálják a memóriaintézmények megőrizni az orosz-ukrán konfliktus internetes lenyomatait, a híroldalakon és a közösségi médiában megjelenő információkat.

Március 24-én részt vettünk az Open Preservation Foundation, az IMPACT Centre of Competence és az IIPC közös szervezésű webináriumán. A „Preservation of Digitised and Born-Digital Collections – Interconnections, Policies and Workflows" című rendezvényen a francia, a brit és a holland nemzeti könyvtár szakemberei beszéltek arról, hogy a digitalizált és az eleve digitálisan születő dokumentumok kezelését miként oldották meg náluk, hogyan illeszkednek ezek a munkálatok a szervezeti struktúrába, milyen jogszabályok és belső szabályzatok vonatkoznak rájuk, és milyen kapcsolódási pontok vannak közöttük, illetve a könyvtárban zajló más (pl. kutatási) tevékenységek között.

Az elmúlt hetekben lefutott tematikus aratások

Idegenforgalom, vendéglátás (5.926 db seed URL)
Sport, testkultúra (3.490 seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (910 db seed URL)
Vallások, hitrendszerek, egyházak (2.733 db seed URL)
Kutatóintézetek, tudományos szervezetek (1.151 seed URL)
Egyetemek, főiskolák (3.952 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