Az OSZK Webarchívum 2022 novemberi hírei

Archiválás, metaadatolás

A jövő hónap első felére tervezett webtér aratás miatt a negyedéves ütemezés szerint decemberre eső részgyűjtemények mentését előre hoztuk novemberre. Mivel a KIFÜ C4I felhőjében futó virtuális szervereinken előre bejelentett karbantartás volt november közepén, ezért előző nap le kellett állítanunk négy, még futó aratást. Ezek közül a két kárpátaljai job-ot később újraindítottuk, a KUTINT és a MUVHAZ esetében elegendőnek ítéltük a 2 nap alatt letöltött tartalom mennyiségét.

Tovább folyt a bölcsész- és társadalomtudományok, illetve az ezekhez a szakterületekhez kapcsolódó cégek és szolgáltatások webhelyeinek válogatása. Ebben a hónapban főként néprajzi, nyelvészeti, szociológiai és politológiai témájú oldalakkal bővült ez az új részgyűjtemény, ami az október végi kb. 2500-hoz képest már több mint 4100 URL címet tartalmaz. A seed lista első aratására december elején kerül majd sor. A címgyűjtés során több száz olyan eddig nem ismert, vagy időközben megváltozott webhelyet is találtunk, melyek valamelyik másik tematikus gyűjteménybe vagy az elektronikus periodikák közé illenek, így ezek a listák is jelentősen gyarapodtak az elmúlt hetekben. (Az ELPERI nyilvántartás novemberben 104, októberben pedig 34 kiadvány weboldalával bővült.)

Informatikai feladatok

A szerverek karbantartása miatti leállás és újraindítás, valamint az üzemeltetési feladatok OSZK-s informatikusok általi átvétele miatt elég sok technikai problémát kellett megoldani a hónap folyamán. Ezeket most már a Redmine rendszeren keresztül jelezzük az informatikus kollégáknak, akik általában azonnal reagálnak és találnak is megoldást.

Szeptember végén kaptunk egy kérést Yves Maurertől, a luxemburgi nemzeti könyvtár informatikai főosztályvezetőjétől, aki a WARCNet projekt keretében készülő könyv egyik fejezetében a nemzeti web doméneket hasonlítja össze. Ehhez az archivált WARC fájlokból generált CDX(J) indexek elemzésével készít összesítéseket egy Python program segítségével. A saját archívumuk mellett az Internet Archive és a Common Crawl anyagát, valamint a dán, a francia és a magyar webarchívum állományát tervezi feldolgozni és ha az érintettek hozzájárulnak, „open data" formájában is közzétenni. Az első teszt után kiderült, hogy a Github-ról letöltött Python script az általunk használt OutbackCDX indexelővel készült fájlokkal nem boldogul, de november elejére a fejlesztő módosította a programot, így végül határidőre mi is el tudtuk küldeni az 1,6 millió aldomén MIME-típus adatait tartalmazó szövegfájlt.

Személyi ügyek

Németh Márton kollégánk, aki 2017 óta – szinte a projekt kezdetétől – dolgozott az OSZK-ban webkönyvtárosi munkakörben, november közepétől az Open Society Archives digitális gyűjteményét fejleszti tovább. Munkakörének átadása előtt készített egy részletes összefoglalást az által végzett feladatokról és a nemzetközi kapcsolatokról, melyeknek szintén ő volt a felelőse. November 23-án még tartott egy előadást az ELTE-n megrendezett DH_Budapest_2022 & DARIAH DAYS konferencián, már az OSA munkatársaként, de még az OSZK Webarchívumáról „The theoretical and practical fundamental elements of web archiving – The first steps of institutional web archiving in Hungary" címmel, valamint részt vett a webarchiválás jövőjéről szóló kerekasztal beszélgetésen. Elkészítette továbbá a beszámolókat azokról a külföldi rendezvényekről, melyeken októberben vett részt Prágában, Aarhusban és Pozsonyban.

Németh Márton tudományos és nemzetközi tevékenységét Kalcsó Gyula, a Digitális Bölcsészeti Központ munkatársa veszi át. Rajta kívül részmunkaidőben ismét bekapcsolódik a webarchiválásba Marcin Mirski, aki korábban már dolgozott velünk a Digitálistartalom-fejlesztési és -szolgáltatási Osztályon és az elmúlt napokban elkezdett újabb munkafolyamatokat is megtanulni.

Rendezvények

December 1-én az OSZK alapításának 220. évfordulójával kapcsolatos események részeként egy belső intézményi workshopra került sor, melyen Drótos László „Digitálisan született tartalom megőrzése a nemzeti könyvtárban" címmel tartott előadást. A prezentáció a „born digital" tartalom sajátosságait, archiválásának kihívásait, valamint az OSZK-ban ezen a téren eddig történt tevékenységet tekintette át, az egyedi dokumentumok gyűjtésétől és szolgáltatásától kezdve a webarchiválásig.

A hatodik „404 Not Found – Ki őrzi meg az internetet?" című konferencia és workshop december 8-án kerül megrendezésre. A hónap folyamán nagyrészt ennek az előkészítésével foglalkoztunk. A rendezvény november 28-án lett nyilvánosan meghirdetve az OSZK honlapján, Facebook oldalán és egyéb kommunikációs csatornáin. Személyesen és online is részt lehet venni rajta, előzetes regisztráció után.

A tervezett program a következő:

10.00 Köszöntő
Rózsa Dávid főigazgató, Országos Széchényi Könyvtár

10.10 Megnyitó
Szóllás Péter a Könyvtári és Levéltári Főosztály vezetője, Kulturális és Innovációs Minisztérium

10.20 Az OSZK Webarchívum 2022. évi eredményei
Drótos László könyvtáros, Országos Széchényi Könyvtár

Az előadás az elmúlt egy évben elvégzett munkát foglalja össze, beleértve a webarchívum bővítését, a szoftver-teszteket, a tudományos és ismeretterjesztési tevékenységet, valamint a hazai és nemzetközi kapcsolatok építését. Szó lesz az újonnan létrehozott téma-, műfaj- és földrajzihely-alapú részgyűjteményekről, a fontosabb eseményekkel kapcsolatos hírek archiválásáról, a közösségi média és a podcastok mentéséről, a nyári nagy webtéraratásról, a nyilvános gyűjtemény bővítéséről és a nem publikus archívumhoz való hozzáférésről. 2022. június 1-jétől a webarchiválás a Digitális Bölcsészeti Központ szervezeti keretei között folyik tovább, így

kiemelten fontossá vált a webtartalom kutatási célú hasznosítása és a tudományos munka. Az előadásban ezekről is beszámolunk, részletesebben ismertetve az IIPC-konzorcium és az idén lezáruló WARCnet projektben elért eredményeket.

10.45 Az apokalipszis archiválása: a válságok eseménytermése webarchívumokban kutatók számára
Dr. Teszelszky Kees kurátor, Holland Királyi Könyvtár

Ha a világméretű válságokat a digitális bölcsészettudományok szemszögéből akarjuk tanulmányozni, széles körű nemzetközi webgyűjteményekre van szükségünk, amelyek gondosan kurált és megőrzött internetes tartalmakat fűznek össze. Az International Internet Preservation Coalition (IIPC) fontos nemzetközi eseményekről született digitális webes tartalmak közös gyűjtését végzi. Ezeket az erőfeszítéseket az IIPC tartalomfejlesztési csoportjának tagjai koordinálják, akik több mint 35 ország – köztük nemzeti, egyetemi és regionális könyvtárak és archívumok – kurátorai, gyűjteményi szakemberei és webarchivátorai. A kurált tartalmakat az Archive-It-tel, az Internet Archive részével együttműködve archiválják. Az IIPC közvetlenül kutatókkal és kutatóhálózatokkal dolgozik együtt az archivált internetes tartalmak használatának és elemzésének előmozdításán. Előadásomban két kurátori együttműködésben létrehozott gyűjteményt szeretnék bemutatni: az Éghajlatváltozás gyűjteményt (2019-ben épült) és az Ukrajnai háború gyűjteményt (idén indult). Szeretném ismertetni, hogyan épült fel ez a gyűjtemény, mit tartalmaz, és milyen haszna lehet a digitális bölcsészettudományi kutatásokban.

11.10 Adatbányászati módszer kidolgozása archivált webes tartalmakon. Az Ukrajna-projekt tanulságai
Dr. Kalcsó Gyula digitálistartalom-fejlesztő, Országos Széchényi Könyvtár

Az OSZK a webtérszintű és a tematikus aratások mellett eseményalapú gyűjteséket is készít a jelentősebb kulturális, politikai és sporteseményekről. 2022. február 21. után elkezdtük gyűjteni az orosz–ukrán konfliktussal, majd később háborúval kapcsolatos híreket 75 magyarországi és határon túli portálról. A gyűjtés alapvetően a portálokon használt címkék vagy kategóriák alapján történik (ez 445 seed URL-t jelent). A mentések hetente egyszer futnak. A learatott WARC-fájlokból kibányászható adatok számos statisztikai kimutatáshoz, adatvizualizációhoz használhatók. A WARC-okból a Digitális Bölcsészeti Központ természetesnyelv-feldolgozásra alkalmas szöveget nyer ki, majd nyelvi elemzéseket végez. A kapott adatokat összesíti és rendezi. A különböző időpontokban végrehajtott aratások anyagából szófelhők készíthetők, amelyeket megfelelő módon animálva látványosan ábrázolható a magyar hírportálok szókészletének alakulása. Az adatok idővonalon ábrázolhatók, a különböző beállításokkal számos megjelenítési lehetőség elérhető. A kidolgozott módszer alkalmas arra, hogy a gyűjtemény bármely részéből nyelvileg elemezhető szövegkorpuszokat építsünk, és azokból szövegbányászati módszerekkel adatokat nyerjünk.

11.35 Tematikus digitális gyűjtemények elemzése az AVOBMAT többnyelvű kutatási eszközzel
Dr. Péter Róbert docens, Szegedi Tudományegyetem

Az előadás célja, hogy bemutassa az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz működéséhez kapcsolódó munkafolyamatot és a különböző elemzőfunkciókat. A webes alkalmazás segítségével nagy mennyiségű metaadatot és szöveget lehet kinyerni és kritikusan elemezni adatvezérelt mesterséges intelligenciával és természetesnyelv-feldolgozásos technológiákkal támogatott módszerekkel és eszközökkel. Az AVOBMAT szöveg- és adatbányászati eszköz újdonságai a következők: számos nyelven képes előfeldolgozni, (szemantikusan) gazdagítani és elemezni metaadatokat és szövegeket; a beépített funkciók lehetőséget adnak a szoros és távoli olvasásra egyaránt; egy felhasználóbarát, interaktív grafikus felületen integrál metaadat- és szövegelemzéssel kapcsolatos kutatási eszközöket. A platformfüggetlen alkalmazás elsősorban olyan felhasználók számára lett kifejlesztve, akiknek nincsenek programozási ismereteik. Az egyszerűen használható felület interaktív paraméter-beállítást és vezérlést biztosít a normalizálást is támogató előfeldolgozástól az analitikai szakaszokig. A felhasználók interaktív módon kísérletezhetnek az elemzések különböző beállításaival a munkafolyamat során. Ezáltal az AVOBMAT segít felismerni a számítógépes szöveg- és adatelemzés episztemológiai kihívásait, korlátait és erősségeit, valamint kritikus módon értelmezni az alkalmazott módszereket és eredményeket.

12.00 Ebédszünet

13.00 A virtuális kiállítástól az archiválásig: Karikó Katalin nyomában
Dr. Kokas Károly informatikai főigazgató-helyettes, Szegedi Tudományegyetem Klebelsberg Kuno Könyvtára

Az előadás azt az utat mutatja be, ahogy a korábban megpendített „szegedikumok" anyagából kiemeltük a Karikó Katalin-témát és virtuális kiállítást készítettünk belőle, melynek része lett egy nemzetközi kitekintésű Karikó-weblinkgyűjtemény is. Ennek kapcsán a webtérből elkezdtük gyűjteni a lementhető és efemernek tűnő anyagokat, hogy házilagosan file-szinten archiváljuk őket. Ezután a MIA projekttel együttműködve ajánlójegyzéket kezdtünk készíteni azokról a webhelyekről, amelyeket elmentésre javaslunk egy Karikó-archívumhoz. Az előadás érinti a tervezés, kivitelezés mozzanatait, a gyűjthetőség, lementés, archiválás nehézségeit, a szerzői jogi kérdéseket és az archiválás és/vagy szolgáltatás, redundancia és teljesség mindnyájunkat gyötrő dilemmáit is.

13.25 Hogy kerül egy oklevél a webarchívumba? Aranybulla-emlékév-archívum kialakítása a Vörösmarty Mihály Könyvtárban
Horváth Adrienn igazgatóhelyettes, Vörösmarty Mihály Könyvtár

13.50 Szünet

14.00–16.00 Workshop könyvtárak számára
Drótos László – **Visky Ákos László**, Országos Széchényi Könyvtár

Az elmúlt hetekben lefutott tematikus aratások

Sport, testkultúra (3514 db seed URL)
Vallások, hitrendszerek, egyházak (2788 db seed URL)
Egyetemek, főiskolák (4088 db seed URL)
Kutatóintézetek, tudományos szervezetek (1190 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (926 db seed URL)
Idegenforgalom, vendéglátás (6012 db seed URL)
Irodalom, irodalomtudomány és -történet (1440 db seed URL)
Könyvtárak, levéltárak, múzeumok és galériák (2029 db seed URL)
Természet- és műszaki tudományok, szakterületek (1633 db seed URL)
Képző-, előadó-, zene- és filmművészet (8143 db seed URL)
Közoktatás és egyéb képzések (6590 db seed URL)
Elektronikus periodikák (9513 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