Az OSZK Webarchívum 2022 júliusi hírei

Archiválás

A júliusi archiválási munkálatok nagy részét a fél évente végzett webtér szintű aratás tette ki, melynél az oldalképek készítése még jelenleg is folyamatban van. Maga az archiválás három menetben történt. Még június végén lefutott egy 2 szint mélységű és csak 3 napig tartó aratás 380 ezer tömegesen generált .hu végű aldoménre. Ezt követően július 4-én indítottunk egy 5 napos és 3 szint mélységű aratást azokra a webhelyekre, melyeknél az előzetes teszt szerint nincsen robots.txt fájl, ezek száma közel 478 ezer. Végül pedig szintén 3 szintig mentettük a maradék 514 ezer webhelyet július 10-től 10 napon keresztül. Mindhárom job az időkorlát elérése miatt állt le, vagyis még bőven lett volna letölthető tartalom. Összességében a 1,37 millió seed címről elindulva mintegy 180 millió URL talált a Heritrix robot és több mint 174 milliót sikeresen le is töltött, melyből 90 millió volt a korábban nem archivált vagy időközben megváltozott fájl. Méretileg a letöltött tartalom 8,89 terabájt, a ténylegesen eltárolt anyag pedig 6,13 terabájt lett. Az előző, 2021 december végén indított webtér aratáshoz képest darabszámban 3-szor, méretben pedig 2,5-szer több tartalmat mentettünk le, ami a megnövelt aratási mélységnek és futási időnek, valamint a seed lista megháromszorozásának köszönhető. Ugyanakkor a másodpercenként letöltött kilobájtok átlaga csaknem a felére csökkent a decemberi aratáshoz képest, miközben az URL/sec érték gyakorlatilag ugyanaz volt. Ennek a jelenségnek két oka is lehet: az Internet Archive-tól kapott doménlistában sok az inaktív vagy csak egyetlen oldalból álló webhely, melyek egy részét a több szintű előszűrésekkel sem sikerült eltávolítani; illetve az is közrejátszhatott, hogy ezekkel a nagy méretű aratásokkal már elértük a szerver erőforrásainak határait. A WEBTER gyűjteménnyel kapcsolatos munkák az aratások alatt és után is tovább folytak. Egyrészt készítettünk egy részletes statisztikát a letöltött tartalomról, másrészt Windows-os programokkal és emberi munkával megpróbáltuk begyűjteni a *title* metaadatot arról a 277 ezer webszerverről, melyekről a Linux szerveren ez nem sikerült, különböző technikai okok miatt. Végül 90 ezer „névtelen" webhely maradt, a többiről van legalább valamilyen – sajnos gyakran semmitmondó – névadatunk. A következő napokban a teljes 1,37 milliós listát nyilvánosan is kereshetővé tesszük URL cím és *title* alapján.

Miközben rendszeresen mentjük a magyar vonatkozású webhelyeket, nem szabad megfeledkezni a nemzeti könyvtár saját online tartalmairól sem (nehogy lyukas maradjon a suszter cipője...). Ez azért is érdekes kérdés, mert az OSZK Webarchívum a nyilvános webhelyek archiválásával foglalkozik, egy nem nyilvános digitális tartalom, mint például egy intézmény belső anyagának a megőrzése a tulajdonos felelőssége és lehetősége elsősorban. Ennek jegyében készítettünk egy egyedi mentést az OSZK intranetjéről a HTTrack nevű programmal, mivel az általánosan használt, de külső szerveren futó Heritrix szoftver nem is érné el azt. Összességében 3,29 GB-nyi tömörített anyagot sikerült lementeni erről a fontos intézménytörténeti dokumentumról a 2013-as indulásától kezdve. Korábban nem ilyen egységes szerkezetben léteztek a belső tartalmak, de érdekes véletlenként annak egyik töredéke, az akkori Digitalizálási Bizottság aloldala is előkerült egy munkatárs anyagai közül, amit beadási csomagként raktároztunk el.

Júliusban jelzést kaptunk a Neumann János Egyetem Könyvtár és Információs Központ igazgatójáról, hogy hamarosan megújul a honlapjuk és szeretnék megőrizni a régin összegyűlt sok anyagot, ami már nem kerül át az újra. Mivel problémás archiválhatóságú honlapról van szó, tesztmentésekkel próbáljuk kialakítani a megfelelő paramétereket. A későbbi könnyű hozzáférés érdekében az archivált oldalt érdemes lenne a nyilvános gyűjteményben is elhelyezni.

Informatikai ügyek

A nyári szabadságok és az informatikusunk munkaszerződésével kapcsolatos adminisztrációs nehézségek miatt a technikai fejlesztések lelassultak az elmúlt hónapban. A Kubernetes platform kialakítása és az azon futtatandó Browsertrix Cloud keretrendszer beüzemelése csak minimálisan haladt előre.

Sokat fejlődött viszont a SolrWayback kereső, melynek 4.3.0-ás verzióját az elsők között kaptuk meg a dán programozóktól, akik a hibajavításoknál és az új konfigurációs lehetőségeknél a mi kéréseinket is figyelembe vették. A https://ukrajnapublic.webharvest.oszk.hu/solrwayback/ oldalon már nyilvánosan kipróbálható az új SolrWayback, ahol az ukrajnai háborúról szóló hírek közt lehet keresni vele.

Egy YouTube podkaszt szöveggé alakítása a Voice Notebook programmal

Utánanéztünk a magyar nyelvű hangfájlok szöveggé alakításának és csináltunk pár tesztet a Voice Notebook (más néven Speechpad) rendszerrel, ami a https://voicenotebook.com oldalon érhető el. A cél a webarchívumban levő podkasztok kereshetővé tétele lenne, melyek aratással és egyedi mentésekkel való archiválása júliusban is tovább folyt. A Voice Notebook a Google Cloud Speech API-t használja (https://cloud.google.com/speech-to-text/), ezért azokat a nyelveket ismeri fel és olyan

minőségben, amire a Google beszédfelismerő aktuálisan képes. Amikor jó minőségű a hangfelvétel és a résztvevők nem vágnak egymás szavába éppen, akkor a magyar szöveget szinte hibátlanul írja át. Központozás persze nincs, kérhető viszont időbélyeg és ha batch módban konvertálunk, akkor a fájlneveket is beleteszi a letölthető szövegbe (XML formátumú letöltés sajnos nincs, csak TXT). Az ingyenes verzió csak 15 percnyi hangot ír át egyszerre, a teljes változatnál $3.30 az egyszeri regisztrációs díj. Windows alatt a Vezérlőpulton a hangbeállításoknál a Felvétel fülön a Sztereó keverő-t kell alapértelmezett eszköznek megadni, majd a HTML5 audio opcióval lehet a korábban letöltött hangfájlokat megnyitni. Nagyon sok funkció van benne és beépíthető Chrome-ba, integrálható Windows és Ubuntu alatt is úgy, hogy bármely szövegszerkesztővel és hanglejátszóval működjön, van iOS és Android alkalmazás is hozzá, de alapvetően személyes használatra készült, nem szerveren parancsmódban futtatható tömeges konvertálásra. Vagy is valakinek meg kell nyitni a fájlokat, megvárni, amíg végigmegy a hangfelvétel, letölteni a szöveget és feltenni a TXT fájlt a szerverre, ahol a Solr leindexeli. További akadály egyelőre a költségkeret hiánya, mert a Google API használata is pénzbe kerül nagyobb mennyiségű hanganyag konvertálása esetén.

Július 1-től a webarchívum eddig zárt része elérhetővé vált az OSZK általános olvasótermében elhelyezett négy dedikált számítógépről a beiratkozott olvasók és a látogatók számára egyaránt. Az erről szóló hír megjelent a nemzeti könyvtár honlapján is: http://www.oszk.hu/hirek/hozzaferheto-az-oszk-webarchivumanak-nem-nyilvanos-gyujtemenye

Rendezvények

Mivel a „404 Not Found” nevű workshopunk megszokott novemberi idejére már több OSZK-s és DBK-s rendezvény is be van tervezve, ezért idén várhatóan októberben, annak is az utolsó hetében kerül megrendezésre ez a konferencia, reményeink szerint újra jelenléti formában (de tervezünk online részvételi lehetőséget is). Emiatt már most elkezdtük a szervezést: összeállítottuk az előzetes programot és felkértük a lehetséges előadókat. Mivel a nyári időszakban más érdemi előkészítő munka nem nagyon végezhető, szeptembertől feszített tempóban kell folytatni azt, hogy a hónap végén meghirdethető legyen a rendezvény.

A Magyar Könyvtárosok Egyesülete (MKE) 2022. évi vándorgyűlésének július 14–16. között a Magyar Agrár- és Élettudományi Egyetem (MATE) Kaposvári Campusa adott otthont, a szervezésben pedig oroszlánrészt vállaltak a Takács Gyula Megyei Hatókörű Városi Könyvtár munkatársai is. Mivel a KDS projekt keretében korábban elkészítettünk a 9-10. osztályos korcsoport számára egy tananyagot a számukra érdekes digitális tartalmak archiválásáról, ezért a rendezvényen a Könyvtárostanárok Egyesületének vezetőivel tárgyaltunk arról, hogyan tudnának segíteni abban, hogy megfelelő módszertani ajánlással ez az összeállítás eljuthasson az oktatási intézményekbe. A résztvevők közül – ahogy az együttműködés kapcsán is – többen jelezték az őszi, a Könyvtári Intézet által szervezett tanfolyamunkon való részvételi szándékukat. A Kecskeméti Neumann Egyetem Könyvtárától pedig kérést kaptunk, hogy mivel új weboldalra fognak áttérni, próbáljuk meg a jelenlegit archiválni a számukra.

Októberre meghívást kaptunk a Cseh Tudományos Akadémia Irodalomtudományi Intézetétől előadás tartására az archivált webtartalom metaadatolása, illetve a webarchiválás aktuális trendjei témájában. Előadásra kértek fel novemberre a hosszú távú digitális megőrzéssel foglalkozó – a pozsonyi Egyetemi Könyvtárban megrendezésre kerülő – konferenciára is.

Publikációk

Németh Márton az OSZK blogjára bejegyzéseket írt a júniusi WARCnet rendezvényről, illetve az MKE Vándorgyűlésről, ezek augusztus elején fognak majd megjelenni; valamint bekapcsolódott a Networkshop 2022 konferencia előadásainak publikált változatait tartalmazó kötet lektorálásába. A

korábban a Könyvtári Figyelő folyóiratnak leadott, a WARCnet projektről szóló cikk lektorálása során lehetőségünk nyílt kiegészíteni azt a 2022. év első felének történéseivel is.

Nemzetközi ügyek

Az International Internet Preservation Consortium (IIPC) orosz-ukrán háborúval foglakozó tematikus gyűjteményébe beadtuk az általunk relevánsnak ítélt magyar URL címek listáját, a kurátorok által kért alapvető metaadatokkal együtt.

Az IIPC blogja az ukrajnai háborúval kapcsolatban indult nemzetközi webarchiválási projektről szóló blogbejegyzésében (https://netpreserveblog.wordpress.com/2022/07/20/web-archiving-the-war-in-ukraine/) külön kiemeli az OSZK Digitális Bölcsészeti Központjának (benne a webarchiváló csoportnak) a munkáját, amelyről a szervezet júniusi konferenciáján számoltunk be.

Együttműködések

Júliusban három könyvtártól: a székesfehérvári Vörösmarty Mihály Könyvtártól, a győri Dr. Kovács Pál Megyei Könyvtár és Közösségi Tértől és a szentendrei Hamvas Béla Pest Megyei Könyvtártól kaptunk pozitív visszajelzést az együttműködési javaslatunkra, így már csak öt megyei könyvtártól nincs végső válaszunk, csak jelzésünk. Augusztusban szeretnénk elkészíteni az egyedi szerződéseket is, és szeptemberben minden megkeresett intézménnyel véglegesíteni ezeket.

Az elmúlt hetekben lefutott tematikus aratások

Elektronikus periodikák (9476 db seed URL)
Podkaszt csatornák (2945 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (6290 db seed URL)
Magyar webtér (1371617 db seed URL)

Az egyes aratások részletes statisztikai adatai a https://webarchivum.oszk.hu/szelektiv-aratasok/ és a https://webarchivum.oszk.hu/webter-szintu-aratasok/ aloldalakon nézhetők meg a honlapunkon. A projekt hírei a https://webarchivum.oszk.hu/a-projektrol/hirek-esemenyek/ lapon kísérhetők figyelemmel. Kapcsolati cím: mia@mek.oszk.hu