Az OSZK Webarchívum 2023 szeptemberi hírei

Archiválás, nyilvántartás

A hónap első felében tovább folyt a fő aratószerveren a hibakeresés, mert a Heritrix robot jóval kevesebb tartalmat mentett le a nyári webtér szintű aratási kísérlet során, mint korábban. A reverse DNS bejegyzés, az IP cím cseréje, a konfigurációs paraméterek módosítgatása, az egyéb archiválási feladatok leállítása és a Heritrix újraindítása után végül szeptember 21-én sikerült egy olyan aratási jobot elindítani, ami ugyan a korábbinál lassabban, de nagy mennyiségben tölti le a fájlokat a kiindulásként megadott több mint fél millió URL címen talált linkeket követve. Eddig hat nap alatt 2,1 terabájtot gyűjtött be a robot, ebből 1,5 volt az új vagy a megváltozott tartalom. A lassabb működésnek az az oka, hogy magasabbra állítottuk a várakozási időt a webszervereknek küldött kérések között, hogy ne terhelje túl őket a robot és ne kerüljön kitiltásra. A következő hetekben a webtér aratáshoz tartozó másik két jobot is lefuttatjuk, utána pedig folytatódhat a tematikus és egyéb részgyűjtemények ismétlődő mentése.

A nyilvános gyűjteményt szolgáltató szerveren is volt egy hiba, ami miatt leálltak a mentések. A Heritrix elvesztette a kapcsolatot az adatbázissal, így több job beragadt, a beütemezett feladatok pedig nem indultak el. Itt is a Heritrix és Web Curator Tool keretrendszer újraindítása segített.

Az új podcast adások letöltése folytatódott. Ebben a hónapban 131 csatornáról 3776 hangfájlt mentettünk le, 212 gigabájt összméretben. Egyedi mentéseket készítettünk továbbá a SZTE Klebelsberg Könyvtártól kapott URL címekről, melyeket a Karikó Katalin életét és munkásságát bemutató digitális gyűjteményhez válogattak össze a szegedi könyvtárosok. A webcímek között vannak tudományos publikációk és előadások, hírek és cikkek, hangfelvételek és videók – hazai és külföldi forrásokból egyaránt. Az 1175 tételes címlistából a duplumok és a már elérhetetlen oldalak kiszűrése után végül 1120-at sikerült lementeni az ArchiveWeb.page programmal WARC formátumban, összesen több mint 20 gigabájtot. (A 30 percnél hosszabb videókat mi nem archiváltuk.) A feladat részét képezte a mentett verziók ellenőrzése és a mentési vagy megjelenítési hibák és hiányok feljegyzése, továbbá a címlista bővítése. Utóbbihoz 3264 darab URL-t tudtunk eddig összeszedni, főként a hírportálokon használt címkék segítségével, a Wikipédia szócikkekben található linkek alapján, illetve speciális Google- és Bing-keresésekkel. A lista jól használható lesz majd az ArchiveWeb.page kiegészítőhöz hasonlóan a Chrome böngészőn keresztül archiváló, de automatizálható és betanítható Browsertrix robot teszteléséhez, melyet a közeljövőben szeretnénk mi is beépíteni az archiváló eszközparkunkba. Ugyancsak hasznos melléktermé volt ennek a munkának az, hogy a hírek keresgélése közben tucatnyi olyan hazai és határon túli magyar sajtóterméket találtunk, melyek eddig nem szerepeltek az e-periodika nyilvántartásunkban.

Jól halad a Scrapy nevű webscraper programmal való ismerkedés, ami szintén egy újabb eszköz lesz olyan feladatokhoz, amikor nem akarjuk vagy nem tudjuk az eredeti webhelyet megőrizni, hanem csak a rajta található tartalom bizonyos részeinek begyűjtése a cél. Elsőként a köztéri alkotásokról készült fotókat megosztó Köztérkép oldalhoz készítettük el azokat a spider kódokat, amelyekkel Scrapy a képek és az alkotók weblapjairól le tudja menteni a számunkra fontos metaadatokat és a képfájlokat. Mivel a kozterkep.hu szerverre sokan valamilyen Creative Commons licenc alatt töltik fel a fényképeket, ezért ezekkel a Digitális Képarchívum gyűjteményét is bővíteni tudjuk majd, amellett, hogy a teljes anyagot eltesszük a webarchívum zárt raktárába. Az elmúlt napokban felvettük a kapcsolatot a Köztérkép gazdájával és egyeztettünk a képekkel foglalkozó kollégákkal is arról, hogy hogyan feleltessük meg a begyűjtött metaadatokat a DKA-ban használt adatmezőknek és kötött szótáraknak. A több mint 42 ezer műlapon található 483 ezer fotó és a hozzájuk tartozó információk első „összegereblyézését” októberben tervezzük.

Ismeretterjesztés és ismeretszerzés

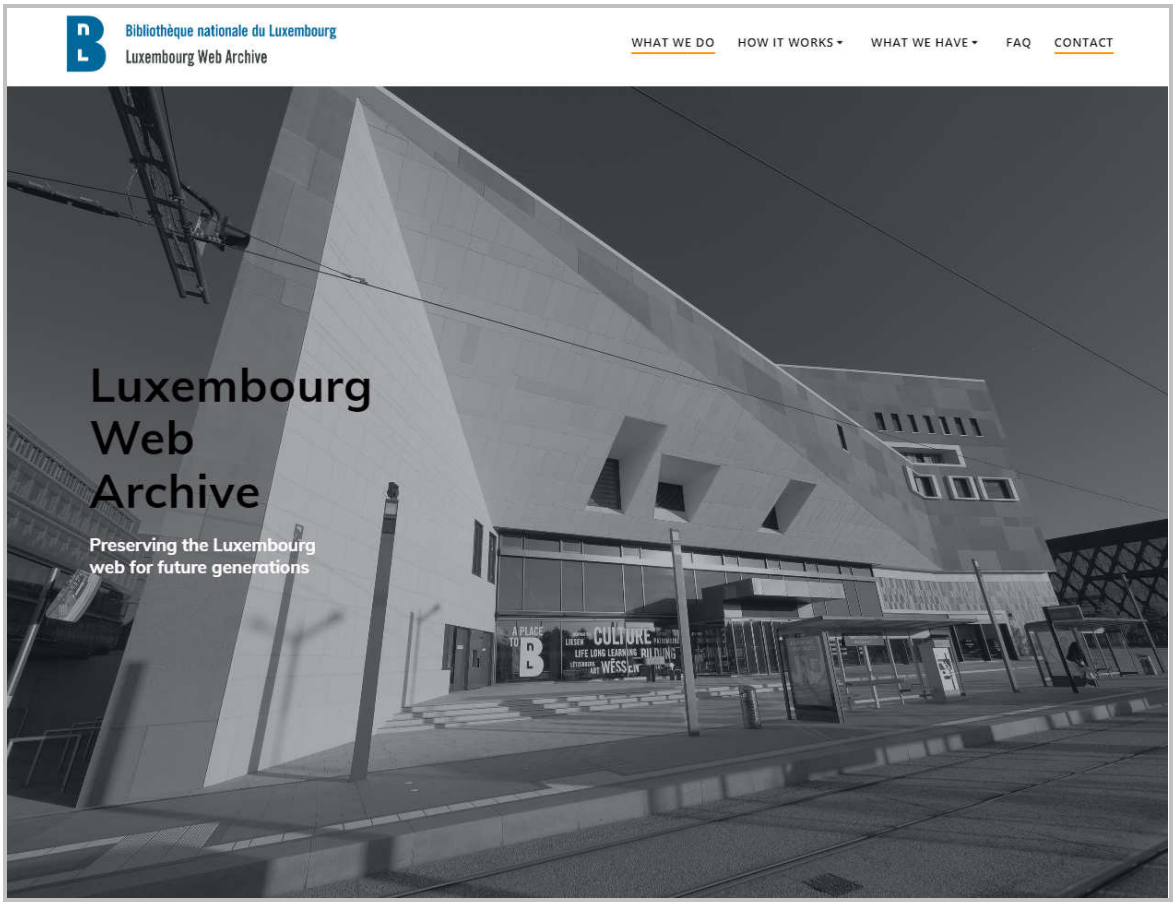
Leadtuk az International Internet Preservation Consortium jövő évi konferenciájára szánt, „Archiving of a community based image sharing website by scraping in the Hungarian National Library” című előadásunk absztraktját, melyben a Köztérkép projekt tapasztalatait szeretnénk ismertetni. Mivel az IIPC rendezvényeken főként az aratásról (harvesting) esik szó, ezért a külföldi kollégák számára érdekes lehet egy ilyen scraping témájú esettanulmány.

Az IIPC két webináriumot is tartott ebben a hónapban. Szeptember 6-án a finn, a dán és a norvég webarchívum munkatársai mutatták be a gyűjteményeiket és a náluk folyó munkát, 27-én pedig a holland nemzeti könyvtárból és a Harvard University könyvtár- és levéltárából hallgathattunk meg két előadást a webarchívumok metaadatolásáról.

Felkérést kaptunk egy, a webarchiválást bemutató előadás tartására a november 9–10-én megrendezésre kerülő „Pécs–Baranya évszázadai” című országos helyismereti konferencia közgyűjteményi szekciójában, amelynek idei témája a digitalizálás, a digitális örökségvédelem és az online adatbázisok jelentősége a társadalomtudományi kutatásokban.

Elkezdtük a saját „404 Not Found – Ki őrzi meg az internetet?” című konferenciánk és workshopunk szervezését, amire a tervek szerint november 29-én kerül sor.

Az október első hetére meghirdetett tanfolyamunk sajnos elmarad, mert nem gyűlt össze a kellő számú jelentkező. Mivel ennek a tananyagnak az akkreditációja az idén lejár, jövő tavasszal már egy megújított tematikájú kurzust szeretnénk indítani.



A luxemburgi webarchívum honlapja: https://www.webarchive.lu

Együttműködések

Szeptemberben több alkalommal is egyeztettünk a szegedi egyetemi könyvtárban dolgozó kollégákkal a Karikó-archívummal kapcsolatos részfeladatokról és a tervezett fejlesztésekről (teljes szövegű kereső, szövegkorpusz, vizualizálás), valamint a https://mediateka.ek.szte.hu/exhibits/show/kariko_katalin_szte/ címen található gyűjtemény felvételéről a nyilvános webarchívumba.

Szeptember 11. és 15. között az OSZK Digitális Bölcsészeti Központjának három munkatársa a Conference of European National Librarians pályázatán nyert támogatásnak köszönhetően ellátogatott a luxemburgi nemzeti könyvtárba és az ottani egyetemre. Az első napon a két ország webarchívumainak részletes bemutatására került sor, melyhez angolra is lefordítottuk a belső szabályzatunkat. A második napon a luxemburgi fél egész napos workshop keretében ismertette a Browsertrix használatát, valamint az ún. paywall mögötti tartalmak mentésének módszertanát és ezzel összefüggésben a saját deduplikációs eljárásukat. A harmadik napon délelőtt a luxemburgi kollégák a SolrWayback megjelenítővel és keresővel kapcsolatos terveiket vázolták fel, délután pedig mi osztottuk meg velük az eddigi SolrWayback-es tapasztalatainkat. Ezután látogatást tettünk a Luxemburgi Egyetemen, ahol Valerie Schäfer, a webarchívumok kutatásának egyik legjelentősebb szakembere doktoranduszával együtt fogadott minket. Tapasztalatot cseréltünk velük a témában, illetve számos új ötlet merült fel a webarchivált tartalmak kutathatóságával kapcsolatban. A negyedik napon délelőtt bemutattuk az OSZK webarchívumának megújítására vonatkozó elképzeléseinket, majd délután a pályázat teljesítésével kapcsolatos további terveket beszéltük meg. Az utolsó nap délelőttjén még a luxemburgi fél látogatásának előkészítésével töltöttük az időt, illetve rögzítettük a további szakmai teendőket, összeállítottuk a megosztandó anyagok, kódok stb. listáját.

A projekt hírei a https://webarchivum.oszk.hu/a-projektrol/hirek-esemenyek/ oldalon kísérhetők figyelemmel. Kapcsolati cím: webarchivum@oszk.hu