Az OSZK Webarchívum 2022 decemberi hírei

Archiválás, metaadatolás

December 2. és 21. között három menetben lezajlott az idei második webtér szintű aratás, a nyárival megegyező kiinduló URL címekkel és paraméterekkel. A több mint 1,37 milliós seed lista három részre osztását részben a szerver terhelésének csökkentése indokolta, részben pedig az, hogy az egyes jobokat eltérő beállításokkal futtattuk. Először a tömegesen (pl. webshopok termékkategóriáihoz) generált aldoméneket arattuk 3 napig és maximum 2 szint mélységig *(a letöltött URL-ek száma 17 millió, a fájlok összmérete 253 GB, ebből új vagy megváltozott tartalom 132 GB)*. Ezt követően a robots.txt-vel nem rendelkező webhelyeket mentettük 5 napig és 3 szintig *(a letöltött URL-ek száma 37 millió, a fájlok összmérete 1,8 TB, ebből új vagy megváltozott tartalom 1,4 TB)*. Végül pedig a maradék félmillió seed címről közel 10 napon keresztül töltöttünk le weboldalakat a kezdőlapról számítva legfeljebb 3 ugrásig követve a linkeket *(a letöltött URL-ek száma 104 millió, a fájlok összmérete 6,2 TB, ebből új vagy megváltozott tartalom 5 TB)*. Összességében tehát 6,53 terabájtnyi webtartalmat sikerült most archiválni, ami több, mint a június-júliusban eltárolt 4,29 terabájt. Mivel az oldalképek gyártása még zajlik, ezért a részletes statisztikák elkészítésével várnunk kell néhány hetet.

A karácsonyi szünetben lefutott a „Bölcsészet- és társadalomtudományok, szakterületek" nevű új részgyűjteményünk első aratása, amelybe addigra 4759 webhelyet válogattunk be. Sok tudományos oldalt már eddig is archiváltunk az EGYETEM és a KUTINT gyűjteményekben, így a TARSTUD elsősorban magán és céges honlapokat és blogokat tartalmaz, beleértve olyan üzleti szolgáltatásokat is, mint például az ügyvédi irodák, a pénzintézetek, a marketingcégek, vagy a fordítóirodák, mert sok esetben ezeken is található szakmai tartalom. Mivel ezeket a webhelyeket korábban még nem mentettük 5 szint mélységben, ezért mindössze 21 óra alatt elérte a robot által letöltött új tartalom az 500 gigabájtos mérethatárt. A TARSTUD gyűjtemény aratását a többi tematikus válogatáshoz hasonlóan negyedévente megismételjük és közben természetesen bővítjük a címlistát is, amely a honlapunkon böngészhető. Az ügyvédi irodák honlapjainak összeválogatásánál nagy segítséget jelentett egy, a „404-es" konferencia szünetében történt beszélgetést követően dr. Csernus Gábortól megkapott, 2578 URL-t tartalmazó Excel tábla. Ezeknek az ellenőrzése után 413 működő, de a nyilvántartásunkban még nem szereplő cím maradt, melyeket szintén felvettünk a seed listába.

A humán- és társadalomtudományi webhelyek gyűjtése közben 15 eddig még nem ismert podcast csatornát is sikerült nyilvántartásba venni és ezek hanganyagát is letöltöttük a hónap első hetében. Részben ugyanezen okból, részben pedig az ISSN Irodától kapott értesítéseknek köszönhetően 20 újabb e-periodika is a látókörünkbe került, ezeket januárban aratjuk majd első alkalommal. A tematikus részgyűjtemények közül a természettudományos és műszaki témájú bővült a legjelentősebben: ebben a hónapban 110 tételt vettünk fel a címlistába, főként informatikai cégek honlapjait és blogjait.

Mivel 2023. január 1-én megszűnik a koronavirus.gov.hu kormányzati tájékoztató oldal, ezért december utolsó napjaiban készítettünk róla egy-egy teljes mentést a zárt és a nyilvános gyűjteményben. A webhelyet korábban is arattuk heti, majd pedig havi rendszerességgel a KORONAVIRUS2020 nevű esemény-alapú gyűjteményünk részeként, de csak kis mélységben, az aktuális adatok és hírek begyűjtése céljából.

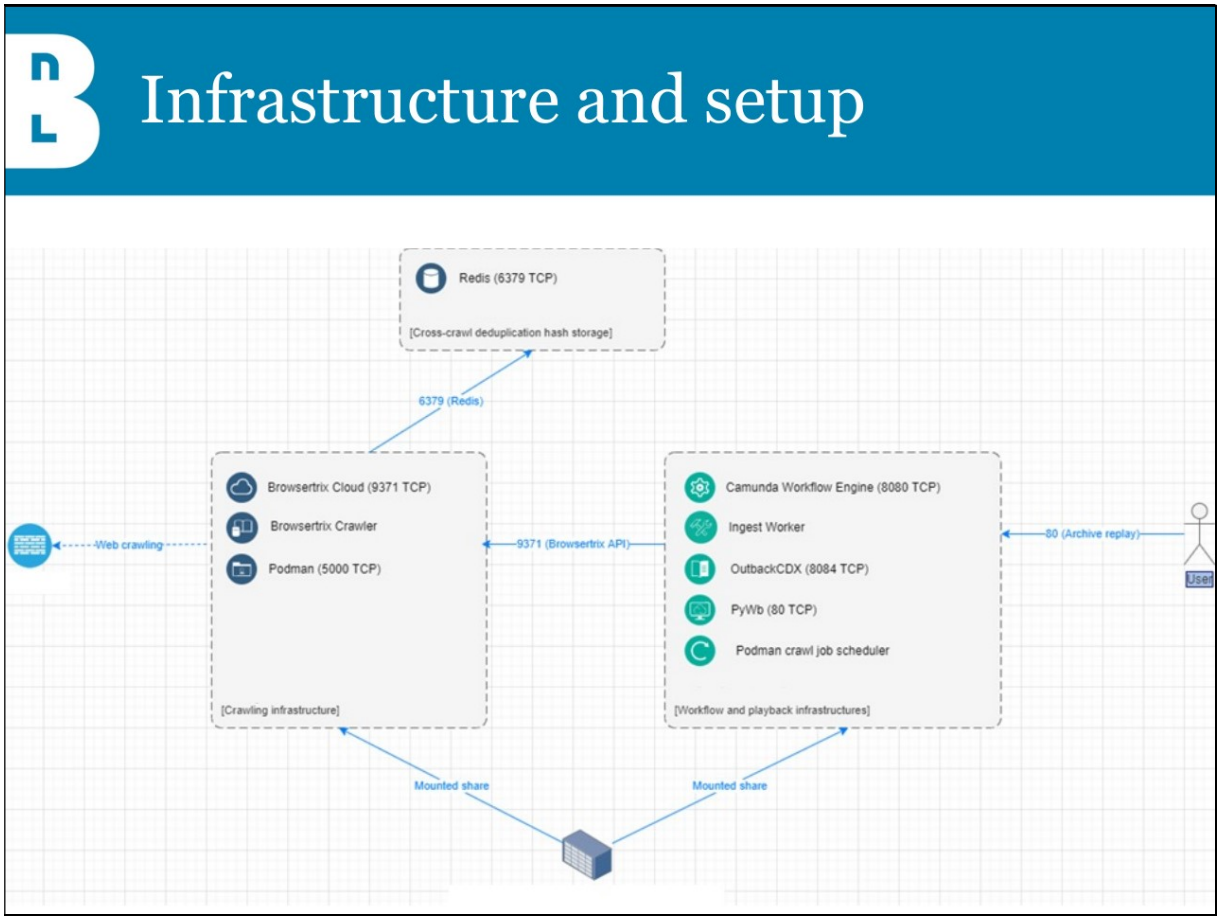
A nyilvános webarchívumban elkezdtük pótolni azokat a metaadat rekordokat, amelyek elkészítésére korábban nem jutott idő. Mivel továbbra sincs elegendő kapacitás erre az élőmunkaigényes feladatra, ezért most csak a legfontosabb adatokat rögzítjük, köztük a tárgyszavakat, melyek majd felhasználhatók lesznek egy automatikusan tárgyszavazó nyelvi modell tanításához. Az elmúlt 2

hétben 30 ilyen „egyszerűsített” XML fájl készült el, melyeket egy, a korábbinál halványabb barna gomb jelez a demó gyűjtemény oldalán.

Az OSZK digitális dokumentumgyűjteményeit gondozó kollégák bevonásával a leendő Széchényi Ferenc archívumhoz is készülnek a metaadatok és a címlapképek. Decemberben a MEK-ből beválogatott 18 könyvhöz/könyvrészlethez, az EPA-ból származó 74 cikkhez/tanulmányhoz és a DKA-ba felkerült 486 képi dokumentumhoz készültek el az XML és a PNG fájlok. A webarchívumba még októberben 71 oldalt mentettünk le, ezek metaadatolása is elkezdődött az elmúlt napokban.

Informatikai ügyek

December 14-én Zoom megbeszélést tartottunk Tóth Lászlóval, a luxemburgi nemzeti könyvtár webarchívumának magyar informatikusával a Browsertrix Crawler és a Browsertrix Cloud szoftverekkel kapcsolatban. A Browsertrix egy böngészőn keresztül archiváló robot, mely parancssorból és webes felületről is vezérelhető. A modern, sok Javascriptet és egyéb programkódokat használó weboldalak esetében sokkal jobb eredményeket produkál, mint az általában használt Heritrix. A luxemburgi kollégák a *paywall* mögötti, vagyis előfizetést igénylő hírportálok archiválását kezdték el vele és egy komplett rendszert alakítottak ki, olyan komponenseket is beépítve, mint a Podman konténermotor és a Camunda munkafolyamat-vezérlő. Mivel jövőre mi is egy hasonló rendszert tervezünk a híroldalak és a közösségi média jó minőségű automatizált mentésére, ezért ez a beszélgetés nagyon hasznosnak bizonyult, mert olyan részletekre is rákérdezhettünk, amelyek a korábbi IIPC webináriumokon tartott prezentációkból nem derültek ki.



A luxemburgi nemzeti könyvtárban kialakított rendszer az előfizetéses hírportálok archiválására

Az elmúlt hetekben az OSZK-s informatikus kollégáktól is sok segítséget kaptunk a webtér-aratás alatt, valamint a „404 Not Found" konferencia és workshop lebonyolítása során, továbbá megoldottak több kisebb-nagyobb technikai problémát is. Utóbbiak közül néhány példa: a zárt archívumban az OWB megjelenítő nem találta meg a WARC fájlokat; a Heritrix által naponta e-mailben küldött státusz jelentések nem jöttek meg; az ukrajnai híreket tartalmazó WARC fájlok egy ideje nem másolódtak át a DBK tárhelyére; a nyilvános szerveren az indexelő script sosem állt le; a lefutott jobok adatlapja helyett egy hibaüzenet jelent meg a Heritrix adminisztrációs felületén; a publikus gyűjteménynél a böngésző és kereső funkciók nem működtek; a személyi változások miatt módosítani kellett a hozzáféréseken.

Tudományos munka és ismeretterjesztés

December 13-án megbeszélést tartottunk az új SZMSZ szerint már szintén a DBK-hoz tartozó Könyvtári Szabványosítási Iroda munkatársaival egy, az élő és az archivált webhelyek leírására egyaránt használható RDA alkalmazásprofil kidolgozásáról. Ez a munka már 2019-ben elkezdődött, de egyéb feladatok miatt abbamaradt. Jövőre szeretnénk rendszeressé tenni ezeket az egyeztetéseket és reményeink szerint nemzetközileg is érdeklődésre számot tartó eredményük lesz.

A Könyvtári Intézet Könyvtári Szak-és Továbbképzési Osztálya összeállította jövő évi képzési tervét, mely szerint 2023-ban március 6. és 9., valamint október 2. és 5. között kerül meghirdetésre „Az internet archiválása mint közgyűjteményi feladat" elnevezésű tanfolyamunk.

Rendezvények

December 1-én az OSZK-ban rendezett belső intézményi workshopon Drótos László „Digitálisan született tartalom megőrzése a nemzeti könyvtárban" címmel tartott előadást a *born digital* dokumentumok és webhelyek archiválhatóságáról és az Országos Széchényi Könyvtár ez irányú tevékenységéről. A prezentáció letölthető a honlapunkról: https://webarchivum.oszk.hu/wp-content/uploads/2022/11/Born_digital_DL.pptx.

2022. december 8-án tartottuk meg a hatodik „404 Not Found – Ki őrzi meg az internetet?" című konferenciát és workshopot, melyen összesen 130-an vettek részt (ebből kb. 80-an online). A prezentációk és a fotók már megnézhetők a rendezvény weboldalán, a videofelvételek pedig rövidesen felkerülnek a Videotoriumba. A https://webarchivum.oszk.hu/404-not-found-workshop-2022-december-8/ oldalról elérhető továbbá az OSZK blogjában megjelent összefoglaló és az a linkgyűjtemény is, ami a délutáni workshophoz készült, ahol is a regionális könyvtári gyűjtemények kialakításáról, a webhelyek válogatásáról, metaadatolásáról és archiválásáról volt szó.

Együttműködések

A konferencia után két könyvtártól is kaptunk olyan levelet, melyben már az együttműködés konkrétumai iránt érdeklődnek. A Pécsi Tudományegyetem Egyetemi Könyvtár és Tudásközpont a PTE és a PTE-hez kapcsolódó intézmények és cégek weboldalainak archiválását fogja felvállalni. A Bács-Kiskun Megyei Katona József Könyvtárban pedig már el is kezdtek dolgozni – egyelőre kísérleti jelleggel – egy megyei webarchívum kialakításán.

Az elmúlt hetekben lefutott tematikus és webtér aratások

Bölcsészet- és társadalomtudományok, szakterületek (4878 db seed URL)
Magyar webtér (1 371 617 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