Az MNMKK OSZK Webarchívum 2024. júliusi hírei

Archiválás és nyilvántartás

A hónap legfontosabb a feladata a nyári webtér-szintű aratás lebonyolítása volt. A 2022 óta használt címlista kiegészítése az azóta emberi munkával vagy automatikusan gyűjtött URL-ekkel, valamint ezek aktuális HTTP státuszának és *title* metaadatának lekérdezése még júniusban megtörtént. Június utolsó napjaiban lefutott a mintegy 268 ezer, tömegesen generált aldomén aratása is. Ezt követően kezdtük el a többi, a webszerver által visszaadott státuszkód szerint működőnek tűnő oldal szétválogatását a *title* címkében levő szöveg alapján. Külön listába kerültek a parkoló (tehát valójában nem aktív) vagy bejelentkezést igénylő domének, illetve a gyűjtőkörön kívüli webhelyek. A következő lépés annak eldöntése volt, hogy a maradék címek esetében van-e *robots.txt* fájl a gyökérkönyvtárban? Mivel a Heritrix robotja alapértelmezésben nem kezdi el aratni azokat a szervereket, amelyeknél hiányzik ez a fájl, ezért az ilyeneket külön *job*-ként, más paraméterezéssel kell menteni. A „normál" *job* július 12-én indult el 452.422 címről, a *robots.txt* nélküli 145.337 darab cím aratása pedig három nappal később. Összesen 66,4 millió új vagy megváltozott fájlt mentett le a Heritrix, 5,4 terabájt összméretben. A naplófájlok elemzése során derült ki, hogy a gondos előkészítés ellenére is több mint 54 ezer olyan seed cím volt, amelyeknél a robot semmit nem tudott lementeni („-2"-es hibakód), pedig ezeknek legalább a felénél biztosan van valamennyi tartalom. Az egyik lehetséges magyarázat a túlterheléses támadások elleni védelmi technikák egyre elterjedtebb használata, melyek sajnos az archiválási célú robotok munkáját is akadályozzák.

Fontos előrelépés volt, hogy sikerült függetleníteni az oldalképek készítését az aratásokat indító scripttől, így egyrészt párhuzamosan futott a két címlistánál a kezdőoldalak „lefényképezése", másrészt nem kellett megvárni ezek elkészültét a szokásos negyedéves tematikus aratások esetében. Végül 545.236 db PNG képfájl jött létre, melyek összmérete közel 1 terabájt. A Redmine projektkezelő rendszer wikijében részletesen dokumentáltuk a webtér seed-lista karbantartásához készített scripteket és a teljes munkafolyamatot, magát az aktuális címlistát – a *title* adatokkal együtt – pedig betöltöttük egy Google Docs táblába és ezt leindexelve frissítettük a honlapunkon található seed-keresőt.

Befejeződött az idén létrejött WARC fájlokból kigyűjtött, korábban még nem nyilvántartott *.hu* végű domének és aldomének emberi munkával történő ellenőrzése és a fontosabbak besorolása valamelyik tematikus vagy műfaji gyűjteménybe. Július első két hetében több mint 1100 új tétellel bővültek ezek a gondozott címlisták, az elektronikus periodikák nyilvántartása például 78 kiadvánnyal. A „Történelem, hely- és családtörténet" elnevezésű táblázatban szintén besoroltuk a megfelelő altémák alá az elmúlt hónapokban összegyűlt több mint 40 webhelyet.

Mivel a Google Podcast szolgáltatás néhány hete megszűnt, ezért a PODCAST kódnevű gyűjteményünkben közel 1500 db podcasts.google.com kezdetű URL-t kellett törölni és ahol lehetett, ott helyettesíteni valamilyen más, lehetőleg jól aratható seed URL címmel. A másfél ezer közül végül 670 csatornát sikerült megtalálni a podcasters.spotify.com szerveren, ami a Spotify által 2023-ban magába olvasztott anchor.fm platform új neve. A címlista karbantartása közben 68, eddig még számunkra ismeretlen podcast adataival bővítettük a nyilvántartásunkat.

Július 15-én az ArchiveWeb.page nevű Chrome kiegészítővel lementettünk tömörítve kb. 4 gigabájtnyi tartalmat a Foci EB-vel kapcsolatos Facebook, Instagram, X (korábban Twitter) és YouTube oldalakról. A hónap második hetében pedig összeválogattunk több mint 160 forrást (239 URL-t) a párizsi olimpiával kapcsolatban. Ezek többsége hazai és határon túli hírportál, a többi pedig önálló honlap vagy aloldal, illetve Wikipédia szócikk vagy közösségi média tartalom. A robottal aratható címek mentése július 16-án indult és az olimpia ideje alatt napi rendszerességgel történik, azt követően pedig egy ideig hetente egyszer még letöltjük híreket, illetve csinálunk majd egy mentést a közösségi platformokon közzétett bejegyzésekről is. Ehhez a NYAROL2024 gyűjteményhez beüzemeltünk egy teljes szövegű SolrWayback

keresőt, ami nyilvánosan elérhető a https://olimpia2024.webharvest.oszk.hu/solrwayback/ címen. (Jogi okokból itt maguk a mentett weboldalak nem nézhetők meg, csak a metaadatok.) Készült három rövid videó is, melyek bemutatják ezt a részgyűjteményt és annak keresőfelületét, továbbá az OSZK Facebook oldalán is lesz erről egy animáció és egy kísérőposzt. Ehhez a témához tartozik még, hogy az IIPC levelezőlistáján és blogjában megjelent egy felhívás arról, hogy a konzorcium tagjai ajánljanak helyi információforrásokat a „Paris 2024 – Summer Olympics and Paralympics" webarchívumhoz, aminek első aratására július 22-én került sor, az utolsót pedig szeptember 11-re tervezik. Mi 17 magyar címet javasoltunk ebbe a nemzetközi gyűjteménybe, és mivel Románia még nem tagja az IIPC-nek, ezért a román olimpiai és sport bizottság honlapját, néhány ottani sporthír portált, valamint román nyelvű Wikipédia szócikkeket is ajánlottunk az erre a célra létrehozott Google űrlapon keresztül.

További sport témájú feladat ezekben a hetekben a SPORT jelű részgyűjtemény címlistájának aktualizálása, ami a 2020 szeptemberi létrehozása óta nem lett ellenőrizve. Július elején két közösségi szolgálatot teljesítő diák átnézte a kb. 3800 URL több mint kétharmadát: böngészőben megnyitották a weboldalakat és jelezték egy táblázatban, ha megváltozott valamelyik oldal címe, neve vagy tartalma, illetve ha teljesen eltűnt. A problémás tételeket javítjuk vagy megszűntnek minősítjük az éles nyilvántartásunkban, és megnézzük természetesen az általuk már nem ellenőrzött URL-eket is. Tanulságos, hogy az eddig átnézett 2 ezer honlap közül több mint 300 vált elérhetetlenné nem egészen négy év alatt. Egy részük ugyan megújult formában más domén alatt megjelent vagy a Facebookra költözött, de közel 10 százalékuk nyom nélkül tűnt el az élő webről.

Újabb, a korábbiaknál teljesebb mentéseket próbáltunk készíteni L'Harmattan kiadó egyik, megszűnés előtt álló webhelyéről és elraktároztuk az általuk csomagként beadott elektronikus kiadványokat. Érkezett három újabb archiválási javaslat, ezekről is készítettünk tesztmentéseket és kettőnél elindítottuk az engedélyeztetési folyamatot is. A nyilvános webarchívum 7 webhellyel, Magyarország külföldi képviseleteinek honlapjaival bővült az elmúlt napokban.

A megújuló tanfolyamunk 5.1-es moduljához készült prezentáció címlapja

Egyéb feladatok

Elkezdtük összeállítani annak az adatbázisnak a tábláit, melyben a jelenleg különböző helyeken tárolt bibliográfiai, adminisztratív és technikai metaadatokat fogjuk nyilvántartani, és ami az egyes munkafolyamatok automatizálásának alapjául is szolgál majd. Az adatbázis a DBeaver Community nevű ingyenes eszközzel készül, egyelőre a tesztelési célokra használt szerverünkön.

A megújított tematikájú, „Internetes tartalmak archiválása" című tanfolyamunkhoz elkészült egy mintaként használható prezentáció a Browsertrix rendszerről. A böngészőn keresztül való archiválás jelenlegi legfejlettebb eszköze mellett annak alternatíváját, a Brozzlert is ismerteti néhány dia, továbbá a prezentáció végén vannak linkek a modulhoz kapcsolódó ajánlott forrásokra és a fontosabb szakkifejezéseket magyarázó wiki szócikkekre.

A Magyar Nemzeti Múzeum Közgyűjteményi Központ (MNMKK) létrejöttével módosítani kellett a nyilvános szolgáltatáshoz szükséges felhasználási engedély sablonjait. A javított PDF, DOCX és ODT fájlok már letölthetők a honlapunkról.

Az elmúlt hetekben lefutott aratások

Életmód, szabadidő, hobbi (9155 db seed URL)
Szolgáltatás, kereskedelem, szállítás, közlekedés (11579 db seed URL)
Webtér szintű aratás (865982 db seed URL)

Az egyes 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: webarchivum@oszk.hu