Az OSZK Webarchívum 2021 augusztusi hírei

Archiválási munkák

A nyári szabadságok ellenére augusztusban is folyamatos volt az OSZK webarchívumának bővítése. Öt tematikus részgyűjtemény negyedéves mentése futott le a hétvégeken, továbbá újra learattuk az elektronikus időszaki kiadványok weboldalait, mert a júliusi ELPERI aratás statisztikája alapján módosítanunk kellett a robot működésén ahhoz, hogy arányosabb legyen a letöltött tartalom mennyisége és hogy a *seed*-ként megadott URL-eken mindenképpen végigmenjen. Két folyamatban levő esemény-alapú gyűjteménynél is történtek változások. A Nemzetközi Eucharisztikus Kongresszussal kapcsolatos oldalak mentésénél áttértünk a kéthetenkénti ütemezésre, mert az esemény közeledtével megszaporodtak a róla szóló hírek, továbbá az ArchiveWeb.page böngészőkiegészítővel lementettünk 28 egyedi cikket és közösségi média fiókot. A tokiói olimpia témájú NYAROL2020 nevű archívumunk seed listáját pedig kb. 30 URL címmel bővítettük, melyek főként a Paralimpiáról szóló hírekhez és egyéb forrásokhoz vezetnek. Közülük nyolc webkettes oldalt és Wikipédia szócikket szintén az ArchiveWeb.page segítségével egyenként töltöttünk le. Lementettük továbbá Heritrix-szel és HTTRack-kel is a sakk témájú Chess fórum archívumát a levelezőcsoport gazdájának kérésére, mivel a listát rövidesen megszünteti.

Címlisták aktualizálása

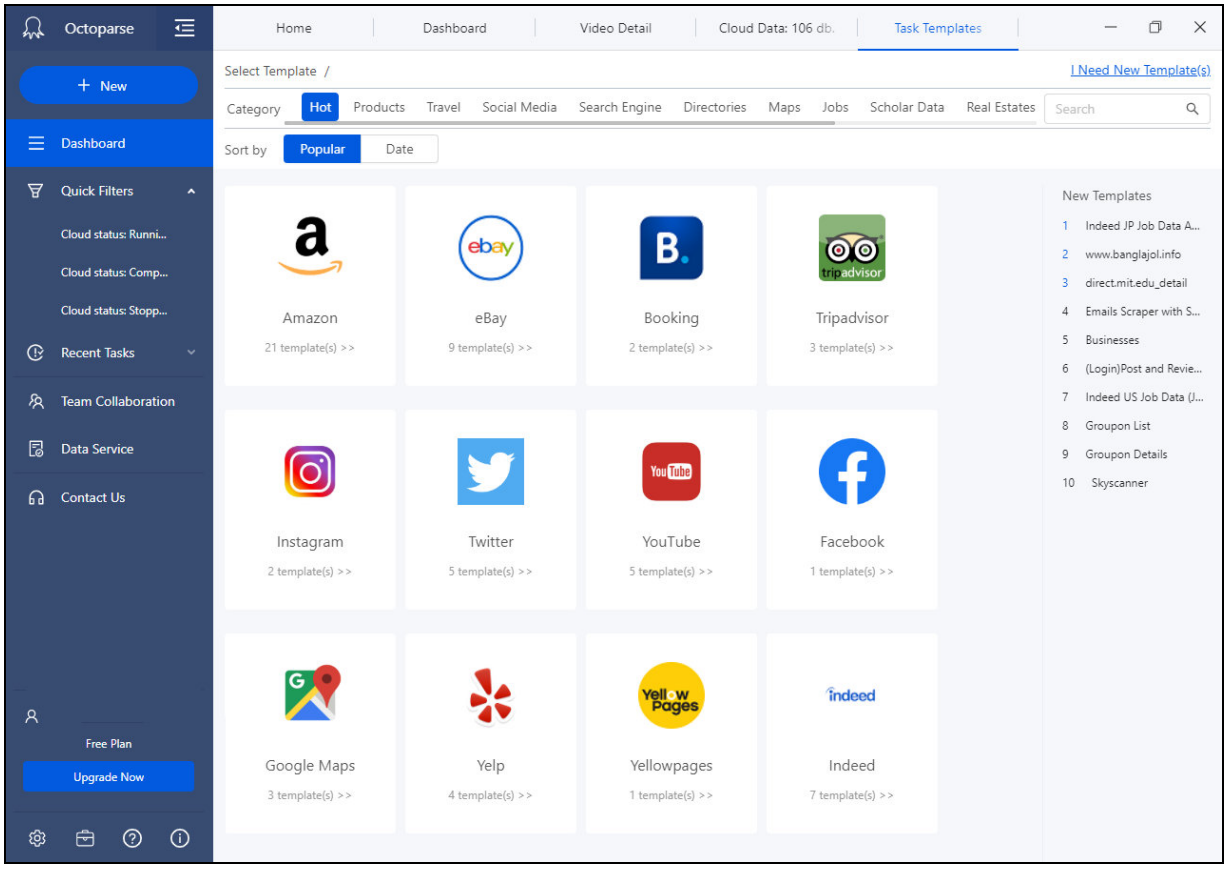
Befejeztük a webtér szintű aratásokhoz használt URL-ek esetében a ** NINCS CÍM ** jelzésű webhelyek júliusban elkezdett ellenőrzését. Eredetileg 38.394 olyan tétel volt a kb. 447 ezres listában, amelyhez nem sikerült automatikusan a *title* metaadatot begyűjteni. Ezeket az oldalakat böngészőben egyenként megnyitva ellenőriztük (augusztus folyamán 16.240-et) és 18.738 olyan URL-t találtunk, ami mögött van valami érdemi tartalom. A többi vagy már nem működő, illetve üres webhely, vagy parkoló domén volt, így ezeket törölve jelenleg 427.247 tételből áll a magyar webtér nyilvántartás, amelyben URL és/vagy *title* alapján lehet keresni mind a nyilvános, mind pedig a zárt archívum honlapján. Ezen két metaadat mellett böngészhetők a kezdőlapokról készült oldalképek (de a nyilvános felületen csak kis méretben), melyek összmérete már 2,65 terabájt, valamint be vannak linkelve az Internet Archive mentései is, amennyiben a Wayback Machine-ben megtalálható az adott webhely. A *title* nélküli URL-ek átnézése során 1.760 olyan webhelyet találtunk, amelyeket érdemes felvenni valamelyik már meglévő tematikus gyűjteményünkbe is, mert komolyabb mennyiségű tartalom van rajtuk, így célszerű őket nagyobb mélységben és negyedévente rendszeresen learatni. Ezeket szükség esetén az eredetinél informatívabb névvel láttuk el és besoroltuk a megfelelő témakörbe, majd az azon belüli alkategóriába is. Az ELPERI gyűjteményünk az átlagosnál többel, 62 időszaki kiadvánnyal bővült augusztusban, nagyrészt az újonnan talált webhelyek alapján.

Technikai fejlesztések

A rendszergazdánk az OSZK informatikusának segítségével egy új virtuális szerverre költöztette át a Heritrix aratószoftvert, mert az egyéb munkafolyamatok (pl. a teljes szövegű indexelés és az oldalképkészítés) miatt a fő szerver már kezdett kifogyni az erőforrásokból. A biztonságosabb működés érdekében szigorítottunk a tűzfal beállításain, a jelszóval való bejelentkezésről áttértünk a kulcs-alapú azonosításra, valamint rendszeres backup-ok készülnek ezentúl a konfigurációs és más fontosabb fájlokról. A dán fejlesztésű SolrWayback kereső- és megjelenítő-rendszernek megkaptuk a legújabb béta verzióját, melyet a nyilvános archívumunkon teszteltünk (http://webadmin.oszk.hu/solrwayback/). A legfontosabb újdonság, hogy az ún. *service worker* megoldásnak köszönhetően sikerült csökkenteni annak az esélyét, hogy egy archivált weboldalon az élő webről jelenjenek meg beágyazott elemek. A SolrWayback megjelenítő képességei így már megközelítik a PyWb programét, bár a közösségi média oldalakkal még nem boldogul. A Kaptafa nevű saját segédprogramunkat, amivel a zárt archívumba készülő aratásokat paraméterezzük, kiegészítettük néhány további opcióval (URL címrészletek kizárása, cookie-k ignorálása, egy domain-ről letölthető URL-k számának korlátozása). Erre a minőségellenőrzések során tapasztalt problémák miatt volt szükség.

Szoftverteszt és szoftverbemutatók

A speciális feladatokra használható programokkal való ismerkedés során augusztus elején az Octoparse nevű web scraper szoftver ingyenes Windows-os változatát teszteltük, ami egy nagyon fejlett eszköz bizonyos adatok vagy oldalelemek tömeges kigyűjtéséhez a weboldalakról. A leghasznosabb funkciók (pl. a különféle webkettes szolgáltatásokhoz készített sablonok és a fejlesztő cég felhőszolgáltatásai) sajnos csak előfizetőknek érhetők el (https://www.octoparse.com). A hónap utolsó hetében az IIPC konzorcium egy újabb Zoom-alapú webináriumot tartott, ahol ezúttal a Dark & Stormy Archives nevű rendszert mutatták be a fejlesztői, mellyel a webarchívumok anyagából lehet különféle összeállításokat készíteni, valamint volt egy másik prezentáció is arról, hogy hogyan lehetne a Bloom Filters-nek nevezett, hash-alapú technológiát felhasználni arra, hogy a Memento protokollnál jelenleg alkalmazott módszernél lényegesen gyorsabban meg lehessen állapítani azt, hogy egy URL cím benne van-e valamelyik nyilvános archívumban (https://netpreserve.org/events/iipc-tss-webinar-dsa-and-bloom-filters/). Részt vettünk még egy másik előadáson is, ahol a korábban az OSZK-ban is bemutatkozott Arkivum cég szakembere beszélt a különböző fájlformátumok (főleg az Office dokumentumok és az elektronikus levelek) hosszú távú megőrzéséről (https://arkivum.com/webinar-recording-digital-preservation-of-valuable-files-and-documents/).



Az Oktoparse legnépszerűbb sablonjai a különböző platformokról való adatgyűjtéshez

Ismeretterjesztés és oktatás

A hónap folyamán 30 újabb szócikket írtunk a MIA Wikibe (https://webarchivum.oszk.hu/mediawiki/), valamint Zotero-ból többféle formátumba konvertálva kikerült a honlapra a nyári folyamán különböző forrásokból 217 új rekorddal bővített és így már több mint 700 tételből álló szakirodalmi bibliográfia (https://webarchivum.oszk.hu/szakirodalmi-bibliografia/). A Könyvtári Intézet meghirdette az őszre tervezett „Az internet archiválása mint közgyűjteményi feladat” című négy napos tanfolyamunkat. A tervek szerint a nagy érdeklődésre való tekintettel szeptember végén lesz egy jelenléti képzés az OSZK-ban (https://ki.oszk.hu/tanfolyamok/az-internet-archivalasa-mint-kozgyujtemenyi-feladat) és egy Teams-

alapú online kurzus november elején. Ha a járványhelyzet megengedi, akkor október 11. és 14. között egy kihelyezett tanfolyamot is tartunk a tatabányai József Attila Megyei és Városi Könyvtárban. A webtartalmak megőrzésével kapcsolatos ismeretek iránti növekvő érdeklődést jelzi az is, hogy néhány napja felkérést kaptunk a közgyűjtemények számára összeállított módszertani útmutató, a „Fehér könyv" új kiadásába egy webarchiválás témájú fejezet megírására.

Könyvtári együttműködés

Július utolsó hetében azoknak a megyei könyvtáraknak a vezetőinek is elküldtük az együttműködésre felkérő levelünket, akikkel nem beszéltünk a májusi első körben. Augusztus végéig 10 pozitív visszajelzést kaptunk azzal, hogy az érdemi egyeztetéseket majd csak szeptembertől tudjuk megkezdeni a nyári szabadságok miatt, de a szombathelyi Berzsenyi Dániel Megyei és Városi Könyvtár munkatársaival sikerült egy első, nagyon konstruktív megbeszélést lefolytatnunk még augusztusban. A könyvtár informatikusa néhány nap múlva már meg is keresett minket azzal kapcsolatban, hogy hogyan lehetne BDMK nemrég megújult honlapját archívum-barátabbá tenni és a tesztmentés eredményét továbbítva a fejlesztők felé már módosításokat is végeztek rajta, mi pedig kigyűjtöttük az általunk nyilvántartott, Vas megyei településekkel kapcsolatos webhelyek címlistáját, ami majd a regionális gyűjteményhez lesz hasznos kiinduló alap. Levelezésben folytatódott az egyeztetés az FSZEK és a KJMK képviselőivel a megállapodástervezet szövegéről. A többi 7, eddig még nem reagáló könyvtárnak szeptemberben emlékeztetőt fogunk küldeni.

Az elmúlt hetekben lefutott tematikus és műfaji aratások

Elektronikus periodikák (7.266 db seed URL)
Közoktatás és egyéb képzések (5.991 db seed URL)
Könyvtárak, levéltárak, múzeumok és galériák (1.875 db seed URL)
Képző-, előadó-, zene- és filmművészet (7.576 db seed URL)
Irodalom, irodalomtudomány és -történet (1.275 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (5.995 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