Az OSZK Webarchívum 2021 júliusi hírei

Webtér aratás

A működő magyar webszerverek felderítéséhez, illetve a Heritrix aratórobot finomhangoláshoz szükséges néhány előzetes teszt után július 7. és 12. között elkészült egy újabb "pillanatfelvétel" jellegű, a kezdőoldaltól legfeljebb 2 linkmélységig terjedő mentés az összes általunk eddig összegyűjtött webhelyekről. A tavaly december végi 251 ezer tételes *seed* listához képest most jelentősen több, csaknem 434 ezer URL címről indítottuk el a robotot, majd egy újabb menetben még 12,5 ezer olyan URL-t is megadtunk neki, amelyekhez nem tartozik *robots.txt* fájl. A nagyobbik aratás során 6 nap és 9 óra alatt közel 72 millió fájlt töltött le a Heritrix, ennek több mint a fele volt az új tartalom, 3,2 terabájt összméretben. Részletes statisztika: https://webarchivum.oszk.hu/webarchivum/webter-szintu-aratasok/ A következő webtér aratásra ez év végén kerül sor, de már most elkezdtük a címlista további bővítését az interneten nyilvánosan elérhető, újonnan felfedezett adatforrások alapján.

Webtér seed kereső

A honlapunk Webarchívum/Keresés menüpontja alatt elérhetővé tettünk egy újabb űrlapot, amivel a webtér szintű aratásnál kiinduló címként használt URL-ek, valamint az ezekhez tartozó weboldalakról nagyrészt automatikus módszerekkel begyűjtött *title* metaadatok között lehet keresni. Az adatbázis most közel 441 ezer tételt tartalmaz és bár már többféle tisztítási fázison átment, még eléggé „szemetes", sok benne a nem működő vagy duplum URL, a hiányzó vagy semmitmondó név. A ** NINCS CÍM ** jelzésű, hiányzó *title* adatok pótlása emberi munkával folyamatban van, az elmúlt hetekben a 38 ezer tételnek körülbelül a felét néztük át. Ugyancsak folyik a kis méretű oldalképek gyártása és az Internet Archive-ban levő mentésekre mutató linkek ellenőrzése, így ezek fokozatosan jelennek majd meg a találati listákban. https://webarchivum.oszk.hu/kereses-a-webter-cimlistaban/

Zárt gyűjtemény

A webtér és a tematikus vagy műfaji részgyűjteményekben levő webhelyeknek csak a metaadatai és kis méretű oldalképei nézhetők meg a nyilvános honlapon, maguk a lementett fájlok egy, elsősorban kutatási célra szánt, nem publikus archívumba kerülnek. Az informatikus kollégák segítségével az elmúlt napokban sikerült kialakítani egy olyan terminál szervert az OSZK belső hálózatán belül, amin keresztül – mentési vagy továbbítási lehetőség nélkül – lehet böngészni és keresni ebben a gyűjteményben. A keresés egyelőre a metaadatokra vonatkozik, a teljes szövegű keresési lehetőség csak néhány hírportál naponta mentett anyagára terjed ki.

Nyilvános gyűjtemény

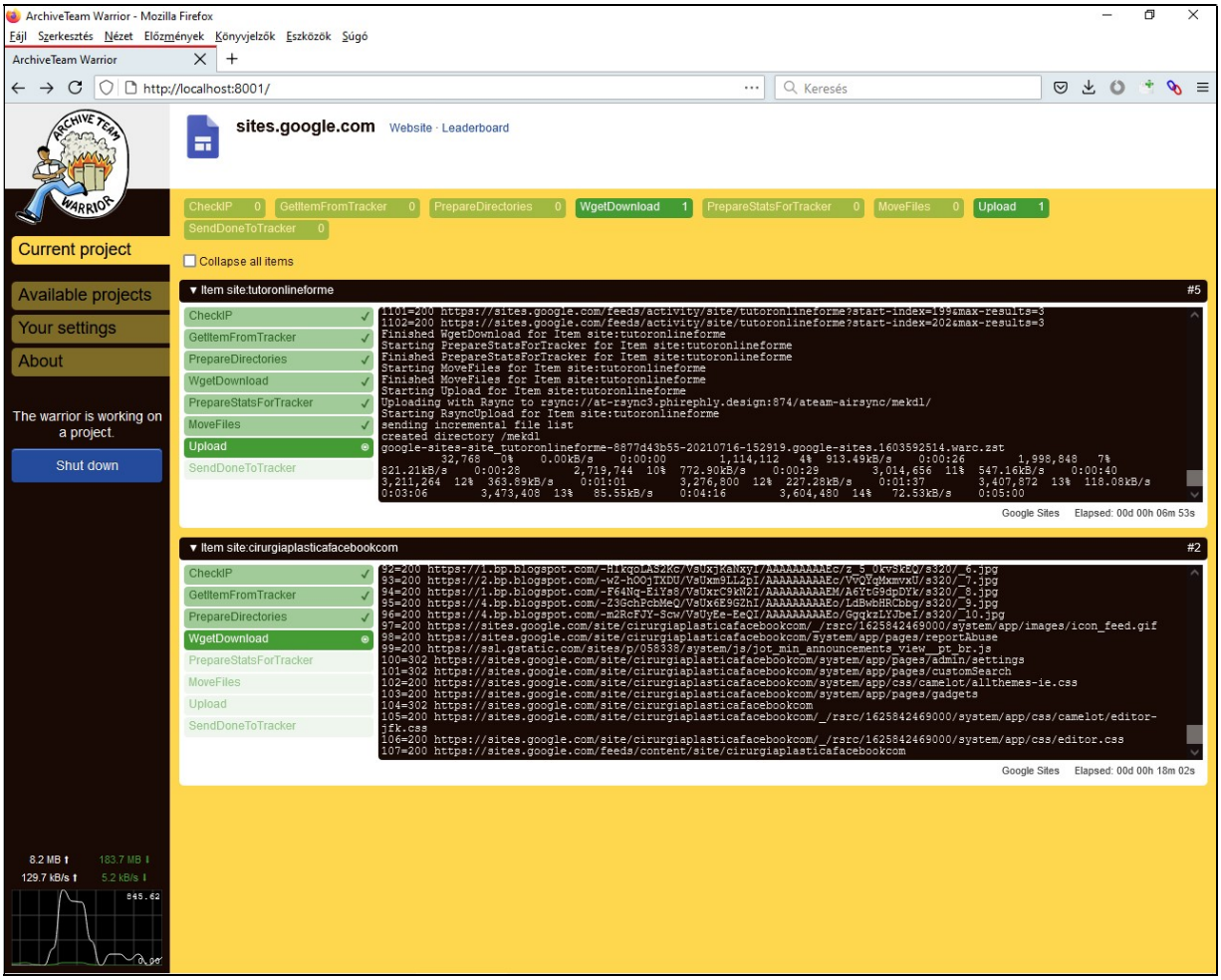
Június végén és július elején a Somogyi Károly Városi és Megyei Könyvtár 8 webhelyével bővült a nyilvános webarchívum: https://webarchivum.oszk.hu/demo-kezdolap/ A mentések a WCT keretrendszerrel futtatott Heritrix robottal készültek, ezeket két esetben kiegészítettünk a HTTRack programmal készített másolatokkal is. Egy OSZK-s honlap, a Mikes Program weboldala szintén júliusban került a nyilvános archívumba, illetve a https://webarchivum.oszk.hu/oszk-s-archivum-kezdolap/ címen elérhető OSZK-s válogatásba is. Ezek a tételek is a kormányrendeletben kapott felhatalmazás alapján kerültek be a nyilvános gyűjteménybe, miszerint külön szerződés nélkül szolgáltathatjuk a közpénzből készült webtartalmak archivált változatait. A június-júliusi első kör után a továbbiakban is folyamatosan tervezzük ez alapján bővíteni ezt a gyűjteményünket.

MIA Wiki

Júliusban 46 új szócikkel fejlesztettük tovább a wikinket: https://webarchivum.oszk.hu/mediawiki/ Igyekeztük feltárni minél több olyan szoftver vagy szolgáltatás elérhetőségét és dokumentációját, melyeket az utóbbi 15-20 évben internetes tartalmak lementése céljából fejlesztettek ki és vagy jelenleg is használatban vannak, vagy legalább dokumentált nyomaik fellelhetők még a világhálón. Június folyamán egy közel 150 URL-t tartalmazó listát válogattunk össze, azóta pedig az ezeken a weboldalakon levő adatok alapján szócikkeket írunk, összegezve az egyes programok és online szolgáltatások legfontosabb funkcióit, sajátosságait. Ezt a munkát augusztusban is folytatjuk, még kb. 70 tétel van hátra, ebből 48-nál már össze vannak gyűjtve az információk.

Szoftvertesztek és szoftverbemutató

A wiki bővítése során talált szoftverek közül párat ki is próbálunk, hogy hasznosítani tudjuk-e valamelyiket akár a napi munkánk, akár az oktatási tevékenységünk során. Júniusban ilyen volt a Twitterről való adatgyűjtést segítő TAGS (https://tags.hawksey.info), ebben a hónapban pedig a böngészőkiegészítőként is telepíthető Web Scraper (https://webscraper.io), amivel weboldalakról lehet adatokat vagy egyes tartalmi elemeket összegyűjteni. Egy virtuális gépen teszteltünk az ArchiveTeam Warrior nevű rendszerét (https://wiki.archiveteam.org/index.php/ArchiveTeam_Warrior), mellyel bárki bekapcsolódhat azokba az archiválási projektekbe, amiken ez az önkéntesekből álló csoport éppen dolgozik. (Az ArchiveTeam magyar honlapja ez év elején indul a http://archiveteam.hu címen.) Július 28-án pedig részt vettünk az IIPC konzorcium Zoom-alapú webináriumán, ahol az egyiptomi és új-zélandi fejlesztők a LinkGate rendszert mutatták be (https://netpreserve.org/projects/LinkGate/). Ezzel a webarchívumokban levő WARC fájlokból lehet kigyűjteni, majd vizualizálni a linkeket és azok időbeli változását is nyomon lehet követni.



Publikálás és oktatás

A Tudományos és Műszaki Tájékoztatás idei 7. számában megjelent Drótos László cikke "Az idő fogságában – Ki őrzi meg a közösségi médiát?" címmel. A tanulmány a Facebook, az Instagram és a Twitter bejegyzések archiválhatóságára vonatkozó OSZK-s tesztek eredményét ismerteti, bemutatja a szóba jöhető módszereket és szoftvereket, valamint egy rövid nemzetközi kitekintést is ad erről a speciális szakterületről: https://tmt.omikk.bme.hu/tmt/article/view/13062 A Könyvtári Figyelő számára Németh Márton és Drótos László önálló publikációban foglalta össze a webarchiválással kapcsolatos oktatási tapasztalatokat, különös tekintettel a COVID19 járvány alatt az online formátumra történt átállás módszertani és szervezési problémákra. A tanulmány várhatóan a folyóirat 2021/3. számában jelenik meg. A Könyvtári Intézet illetékes osztályának munkatársaival az elmúlt hetekben már egyeztettünk az őszre tervezett továbbképzési tanfolyamok ütemezéséről, melyek közt lesz jelenléti, online és kihelyezett is. Szintén júliusban került fel a honlapunkra egy középiskolásoknak szánt oktatási segédlet vázlata „Mentsük le az internetet! – Internetes tartalmak megőrzése intézményi és személyes archiválással" címmel: https://webarchivum.oszk.hu/tanfolyam-es-e-learning/mentsuk-le-az-internetet/ Ez a szöveg 2019 végén készült a KDS pályázat keretében egy, még megjelenés előtt álló multimédiás tananyaghoz. A benne levő linkek és adatok frissítve lettek.

Szakirodalmi bibliográfia

Az elmúlt hónap folyamán – az OSZK-ban elérhető EBSCO és Proquest adatbázisok szokásos havi, rutinszerű ellenőrzésein túlmutatva – a webarchiválás bibliográfia alaposabb frissítését végeztük el. Az utóbbi két év vonatkozásában rákerestünk többek között a Google Scholarban a *web archiving* és a *web scraping* kifejezésekre, ez utóbbiak révén tudtuk a legtöbb új releváns tételt feltárni és a bibliográfiába felvenni. Külső egyetemi hozzáféréssel, melyre Németh Márton PhD tanulmányai miatt volt lehetőségünk, a Proquest Central adatbázis csomag segítségével gyűjtöttünk a témakörhöz kapcsolódó szakdolgozatokat és doktori disszertációkat is. Mód nyílt az ACM informatikai szakadatbázisban található cikkek és konferenciaelőadások bibliográfiai adatainak rögzítésére is. Az ily módon feltárt anyaghoz illesztjük hozzá a saját publikációink adatait. A bibliográfiát a Zotero programmal gyarapítjuk, deduplikáljuk, s ezzel állítjuk elő a megfelelő formátumú kimeneteket, melyek a webarchívum honlapján kerülnek majd közzétételre várhatóan augusztus elején: https://webarchivum.oszk.hu/szakirodalmi-bibliografia/ Jelenleg még a 2020 februári utolsó nagyobb frissítés utáni állapot érhető itt el. A jövőben évente tervezzük az aktualizálását.

Szakmai gyakorlat

Első alkalommal tudunk gyakornokot fogadni az önálló Webarchiválási Osztályon egy, az ELTE Könyvtár és Információtudományi Intézetében tanuló, szeptembertől harmadéves BA szakos hallgató személyében. A gyakorlat 80 óra időkeretben online formában zajlik, folyamatos konzultációkkal, illetve az elvégzett tevékenységek munkanaplóban történő dokumentálásával. Először részletesen bemutattuk az osztály tevékenységét, majd ismertettük azokat a munkafolyamatokat, amelyek szóba jöhetnek a gyakorlat tárgyaként. Ezek közül gyakornokunk az egyedi webhelyek metaadatolásával történő megismerkedést, illetve a webcímek gyűjtését választotta a társadalom- és humántudományok témakörében. Mindkét munkafolyamat alapvető ismereteit betanítottuk neki júliusban, augusztusban pedig már önálló tevékenység formájában segíti majd az osztály munkáját gyakorlatának teljesítése során.

RDA alkalmazásprofil

Ilácsa Szabina, a Könyvtári Intézet szabványosításért felelős osztályának munkatársa készített egy új koncepciót arról, hogy egy alkalmazásprofil létrehozásával folytatódhatna a webarchívum RDA-alapú metaadat leírási lehetőségének modellezése, majd gyakorlati tesztelése. Erre az évre vonatkozóan abban állapodtunk meg, hogy külön dokumentumban rögzítjük a projekt vízióját: a kutatás-fejlesztési cél az RDA lehetőségeinek alkalmazása az élő és az archivált webre; távlatilag ez a séma lenne a webarchívum

metaadat leírásának az alapja; valamint a további cél az lenne, hogy az általunk elvégzett munka nemzetközileg is elérhetővé váljon és továbbfejleszthető legyen egyéb webarchívumok igényei szerint. A vízió megfogalmazása után elkészítjük a projekt munkatervét, egy értelmező szótárat a fontosabb fogalmakról, valamint egy funkcionális követelmény vázlatot. Erre és a vízióra épül majd rá a használati esetmodell, melyet a jövő évtől tervezünk megalkotni. Az eredményeket bemutatjuk és publikáljuk a szakmai közösség felé, a végleges dokumentumokat lefordítjuk angolra is és az IIPC-n keresztül megpróbáljuk elérhetővé tenni más intézmények számára is, tekintve, hogy a szemantikus technológiák és az RDA használata a webarchiválásban nemzetközi szinten is gyerekcipőben jár még.

Könyvtári együttműködés

Július utolsó hetében küldtük el a webarchiválásban történő együttműködésre felkérő leveleinket a megyei hatókörű könyvtárak vezetőinek. Május végén még csak azt a három könyvtárat szólítottuk meg, melyek a 2019-2020-as KDS projekt végén jelezték további együttműködési szándékukat, közülük kettővel, a kecskeméti Katona József Könyvtár, az egri Bródy Sándor Megyei és Városi Könyvtár, valamint az önként jelentkező Fővárosi Szabó Ervin Könyvtár és az ELTE Állam- és Jogtudományi Kar könyvtárának munkatársaival már részletekbe menően egyeztettünk a közös munkáról, ezekről májusi és júniusi híreinkben már beszámoltunk („Együttműködés kialakítása könyvtárakkal” és „Regionális gyűjtemények” című részek), és felvettük a kapcsolatot régi partnerünkkel a Szegedi Tudományegyetem Klebelsberg Kuno Könyvtárában is. Mivel ezek a tervezett együttműködések hálózatszerűen kapcsolódva az egész országot lefedő regionális gyűjtemények alapjai lennének, ezért fontos volt valamennyi lehetséges partner megkeresése is, ez történt meg most a felkérések elküldésével. A tervezett együttműködés már nemcsak a szűken vett webarchiválásra vonatkozna, hanem a tágabban értelmezett egyedi webtartalmakra, az online dokumentumok gyűjtésére is, melyekkel társosztályunk a Digitális Tartalomfejlesztési és -szolgáltatási Osztály foglalkozik és csatlakozik ezáltal a projekthez. A felkérések elküldése után szinte azonnal pozitív reakció érkezett öt könyvtártól, de bízunk abban, hogy a többiek esetében is csak a nyári szabadságok miatt késik az igenlő válasz. Sajnos a szabadságolás időszaka az érdemi egyeztetés megkezdését is nehezíti, így várhatóan ezekre majd csak szeptembertől fog sor kerülni. Addig azonban potenciális partnereinknél is folyhatnak előkészítő tervezések, a felkérés átgondolása, a kérdések megfogalmazása, így ez a technikai szünet sem válik kárba veszett idővé, inkább lehetőséget ad a felkészülésre. Külön érdekességnek ígérkezik, hogy Szegeden az egyetemi és a megyei-városi könyvtár között profil szintű munkamegosztás is várható.

Az elmúlt hetekben lefutott tematikus és műfaji aratások

Elektronikus periodikák (7266 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1462 db seed URL)
Történelem, hely- és családtörténet (1042 db seed URL)
Média, sajtó, műsorszórás (944 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