Az OSZK Webarchívum 2021 decemberi hírei

Aratások és címlisták bővítése

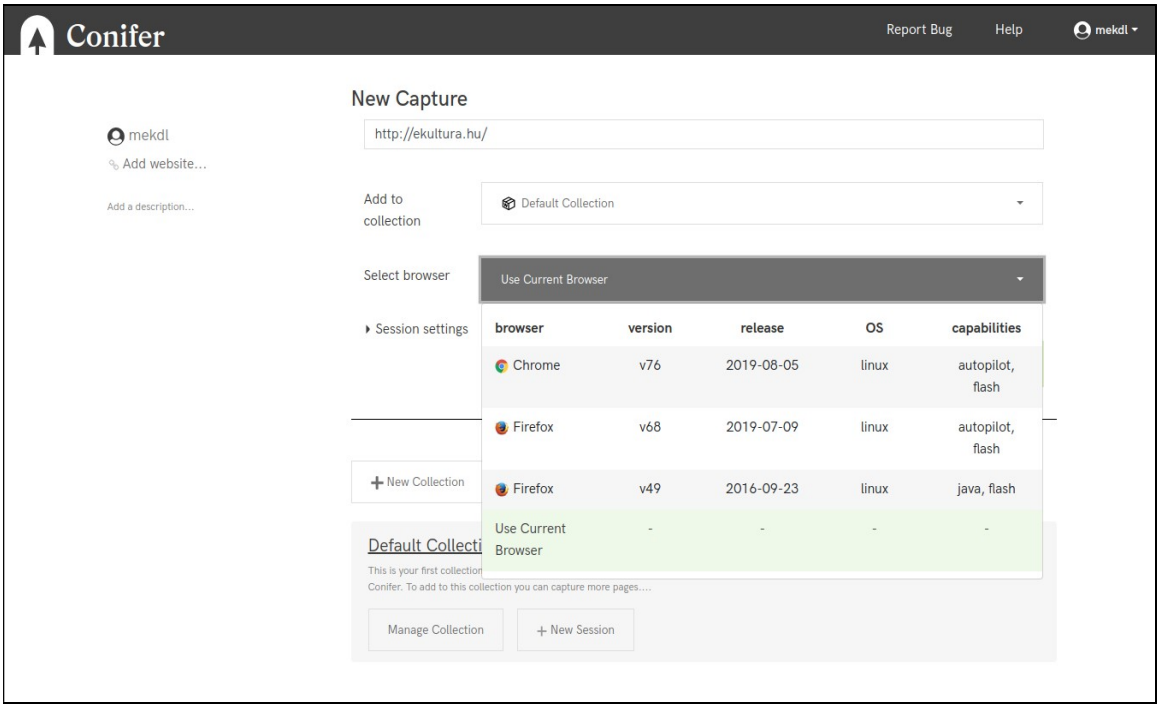
December közepén lefutott a természettudományos és műszaki témájú webhelyek részgyűjteményének első aratása, 1456 webcímről indítva. A seed lista intenzív bővítése a hónap elején is zajlott és jövőre is folytatódik még. A robot két nap alatt közel 2,4 millió URL címet talált és töltött le 275 GB összméretben. Ennél jóval nagyobb tömegű lesz a december 27-én indított webtér szintű aratás, ami az összes eddig általunk nyilvántartott magyar webhelyre, közel 450 ezer címre terjed ki és a tervek szerint tíz napig fut majd. Szintén decemberben elkezdett és januárra is áthúzó munkafolyamat az elektronikus periodikák címlistájának összevetése a nemzetközi ISSN portál adatbázisával. Az abban található 1377 magyar tételből eddig 895-öt ellenőriztünk, melyek közül sok volt a már nem létező vagy máshová költözött weboldal, de némi nyomozás után eddig 362 új címmel sikerült bővíteni a saját ELPERI nyilvántartásunkat, ami már közel 6000 tételes.

Szerződések megújítása

Még 2018-ban, a webarchiválási projekt pilot szakaszában alakítottuk ki az archivált tartalmak nyilvános szolgáltatásának első verzióját, azonban a kísérleti jelleg miatt akkor még csak határozott idejű felhasználási szerződéseket kötöttünk a partnereinkkel. Ezek a szerződések ez év végével járnak le, így fontos feladat volt megújításuk. Mivel azóta megteremtődött a webarchiválás jogi háttere és ez év elejétől már üzemszerűen végzi az OSZK ezt a tevékenységet, nem volt akadálya a felhasználási szerződések határozatlan idejűvé alakításának sem. Szerencsére a 626/2020. (XII. 22.) számú kormányrendelet lehetővé tette az állami és önkormányzati, valamit a közpénzből megvalósuló webhelyek lementett változatának egyedi engedély nélküli nyilvános szolgáltatását, így egy jelentős csoporttal nem szükséges megújítani a szerződésünket. Azonban így is mintegy 70 partnerrel kellett e-mail-ben felvenni a kapcsolatot, kérve a felhasználási engedélyük újrakötését. Ez a nyilvános szolgáltatás nemcsak klasszikus értelemben vett archívumi hozzáférést biztosít bizonyos tartalmakhoz, hanem egyúttal demonstrálja is a webarchiválás lehetőségeit, bemutatva annak előnyeit és hátrányait. A gyűjteményben jelenleg 307 db webhely nézhető vissza különböző témakörben, illetve műfajban, a ma már hagyományosnak tekintett honlapoktól kezdve a blogokon és periodikákon át a webkettes tartalmakig (https://webarchivum.oszk.hu/demo-kezdolap).

Technikai ügyek

A PyWb megjelenítő szoftver új verziójához frissíteni kellett a nyilvános gyűjteményt szolgáltató szerver több komponensét, ami miatt az ott futó, rendszeres mentéseket végző Web Curator Tool keretrendszer eddig használt változata működésképtelenné vált, így ennél is átálltunk az új, 3.0-ás verzióra. Ezt már néhány hónapja teszteltünk a másik szerverünkön, de mind az aratási job-ok elindításánál, mind pedig a letöltött tartalom tárolására szolgáló WARC fájlok indexelésénél problémák voltak, melyek a nyilvános gépen is előjöttek, így az azon beütemezett mentések december első felében szüneteltek. A hibák elhárításában a WCT fejlesztőitől kértünk segítséget és bár a rendszer saját indexelője továbbra sem működik, de remélhetőleg a PyWb-vel már visszanézhetők lesznek az azóta újraindult mentések. A WCT mellett saját, illetve felhőalapú virtuális gépeken, valamint a Windows 11 új Linux alrendszere alatt is teszteltük a Conifer és a Browsertrix Crawler nevű szoftvereket. Ezek böngészőn keresztül töltik le a weboldalakat és ezért a WCT által is használ Heritrix robot helyett sokkal alkalmasabbak például a hírportálok és a közösségi média archiválására. Érdekes tanulság volt, hogy a Conifer a Windowson belüli Linux alatt jobban működik, mint egy önálló Ubuntu rendszeren.



Egy saját virtuális gépen futó Conifer rendszer, amiben régebbi böngészőket is lehet választani a mostaniak által már nem támogatott Flash- vagy Java-alapú webhelyek mentéséhez

Nemzetközi ügyek

A novemberi „404-es” workshop első, közép-európai napjának előadásait és az azokat követő kerekasztal beszélgetésen elhangzottakat Németh Márton önálló tanulmányban foglalta össze, ami a Könyvtári Figyelőben fog megjelenni, várhatóan 2022 első negyedévében. Erről a közép-európai webarchívumok közötti együttműködés tervezetről beszámoltunk röviden az International Internet Preservation Consortium már rendszeresnek tekinthető *update*-jén is, melyen rajtunk kívül a brit National Archives, az amerikai Harvard Library, a holland Koninklijke Bibliotheek és a dániai Det Kgl. Bibliotek - Aarhus munkatársai ismertették az aktuális tevékenységüket a webarchiválás területén. (A dán kolléga prezentációjában jó volt látni az OSZK nyilvános webarchívumáról készített képernyőfotókat, amivel az általuk fejlesztett SolrWayback szoftvert demózni szokták.) A december 15-i videóbeszélgetést egy nappal később egy másik webinárium is követte. Ezen az IIPC által támogatott két projekt állásáról számoltak be a fejlesztők. Az elsőn a Los Alamos National Laboratory informatikusai dolgoznak, felhasználva a horvát webarchívumtól kapott adatokat. A kutatás célja annak felmérése, hogy a WARC fájlokból generált CDX indexek helyett hogyan lehetne az ú.n. „Bloom filter” technológiát használni annak ellenőrzésére, hogy egy adott URL címről van-e mentés egy webarchívumban. A hasítófüggvény (hash) alapú, valószínűségi modellen alapuló adatszerkezet tárhelyigénye a CDX indexnél jóval kisebb, a visszakeresés sebessége pedig lényegesen nagyobb. Ideális megoldás lenne a tervezett közép-európai portálhoz és közös keresőhöz. A másik projekt szintén jól használható lesz majd ehhez a portálhoz, mivel a Dark and Stormy Archives szoftvercsomaggal könnyen lehet látványos összeállításokat készíteni annak demonstrálására, hogy mi található az egyes webarchívumokban, vagy azok egyes részgyűjteményeiben. A rendszert a Los Alamos National Laboratory és az Old Dominion University munkatársai fejlesztik és a National Library of Australia webarchívumának anyagából csináltak vele demonstrációkat az ottani könyvtárosok. A munka során szerzett tapasztalatokat azután felhasználták az egyes szoftverkomponensek továbbfejlesztéséhez.

Végül két további hír: Elkezdődött a WARCnet projektet bemutató könyv tervezése, a fejezetek címének és terjedelmének felvázolásával. A dán koordinálású nemzetközi projektben az OSZK is részt vesz, erről a novemberi beszámolónkban részletesebben is írtunk. Ugyancsak utaltunk rá korábban, hogy folyik az egyeztetés az Internet Archive és az Országos Széchényi Könyvtár között a Wayback Machine-ban található magyar tartalom teljes szövegének kereshetővé tételéről és a .hu alatti domén lista megvásárlásáról. A karácsonyi ünnepek előtti napokban mindkét fél részéről aláírásra került a szerződés, így januártól elkezdődhet a szolgáltatás kialakítása.

Hazai együttműködések

A „404 Not Found" workshop mellett novemberben már nem jutott idő arra, hogy a szolnoki Verseghy Ferenc Könyvtár és Közművelődési Intézmény munkatársaival lefolytassuk az együttműködés lehetőségeit körbejáró megbeszélésünket, így ezt végül december elején tudtuk megejteni. Bár hivatalos válasz még nem érkezett a kezdeményezésünkre, de örömmel láttuk, hogy nem feledkeztek meg levelünkről és nem hagyták feledésbe merülni a kérdést. Az év végéhez közeledvén nem is szándékoztunk már újabb lehetséges partnerekkel felvenni a kapcsolatot, de jövőre mindenképpen folytatni szeretnénk a projektet, megkeresni az eddig nem reagáló könyvtárakat is, illetve eljutni végre a szerződéskötésig és a tényleges munkáig.

Szintén még december elején került sor egy újabb megbeszélésre a székesfehérvári Vörösmarty Mihály Könyvtár munkatársaival, de már az együttműködés egy lehetséges megvalósulásáról. A könyvtáruk szeretne kapcsolódni a 2022-es Aranybulla Emlékévhez, melyhez egy tematikus gyűjteményt terveznek létrehozni különféle digitális és digitalizált dokumentumokból, köztük archivált webhelyekből, a mi Rákóczi Archívumunk (https://rakoczi2019.webarchivum.oszk.hu) mintájára. Ezt nagyon fontos kezdeményezésnek tartjuk, nem egy egyszerű webarchiválási projektnek, mivel önálló szándékból született és az általunk megosztott ismeretek felhasználására épül, ezért szeretnénk minden szakmai támogatást megadni hozzá. Első lépésként összeállítottunk és megosztottunk velük egy válogatást a Rákóczi-gyűjteményben levő metaadatokból, oldalképekből és WARC fájlokból.

Az elmúlt hetekben lefutott tematikus és műfaji aratások

Vallások, hitrendszerek, egyházak (2.716 db seed URL)
Természet- és műszaki tudományok (1.456 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (898 db seed URL)
Sport, testkultúra (3.478 db seed URL)
Egyetemek, főiskolák (3.931 db seed URL)
Kutatóintézetek, tudományos szervezetek (1.118 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