Az OSZK Webarchívum 2022 júniusi hírei

Archiválás

A júniusi archiválási munkálatok két fő területre összpontosultak. Egyrészt a májusban létrehozott önálló podkaszt gyűjteményt bővítettük jelentősen. A korábban a MEDIA listában nyilvántartott 185 magyar csatorna helyett most már 1030 tételt tartalmaz ez az új archívumrész, melyek közül május-június folyamán 169-nél már az egyes adásokat is letöltöttük, mintegy 17 ezer fájlt, 1,1 terabájt összméretben. Az egyes podkasztokhoz tartozó MP3 vagy M4A formátumú fájlok önálló alkönyvtárakban böngészhetők a helyben használható archívum részeként.

A másik nagy feladat a hónap végén induló webtér szintű aratás seed listájának összeállítása és tesztek futtatása volt. A különböző forrásokból származó – köztük az Internet Archive-tól megvásárolt – közel 5 millió magyar domén és aldomén esetében megpróbáltuk begyűjteni a szerver által visszaadott státusz kódot és a webhely kezdőlapjának *title* adatát. A működő webszervereknél szétválasztottuk a tömegesen generált aldoméneket, a maradékból pedig töröltük a parkoló vagy más okból „üres" webhelyeket, valamint az egyértelműen nem magyar vonatkozásúakat. Utóbbiak beazonosítását nagyon megnehezíti, hogy vannak olyan nemzetközi platformok, amelyek a .com végződésen kívül országdomének alá is bejegyzik a felhasználóik által létrehozott webhelyek aldoménjeit, így jelenik meg például egy vietnámi művészeti iskola 1965-1970 között végzett diákjainak szóló blog a http://bantroi5.blogspot.hu címen is. Végül 380 ezer – általában aloldalakra vagy egy közös kezdőlapra átirányított – tömegesen generált aldomén címre és valamivel több, mint 1 millió egyedi doménre/aldoménre terjed ki az idei első webtér aratás, amely szám háromszorosa a tavaly decemberinek. Előbbi, vagyis a 380 ezres lista két szint mélységű mentése már le is futott június 24. és 27. között, de a letöltött tartalom mennyiségével nem vagyunk elégedettek, ennek okát most vizsgáljuk.

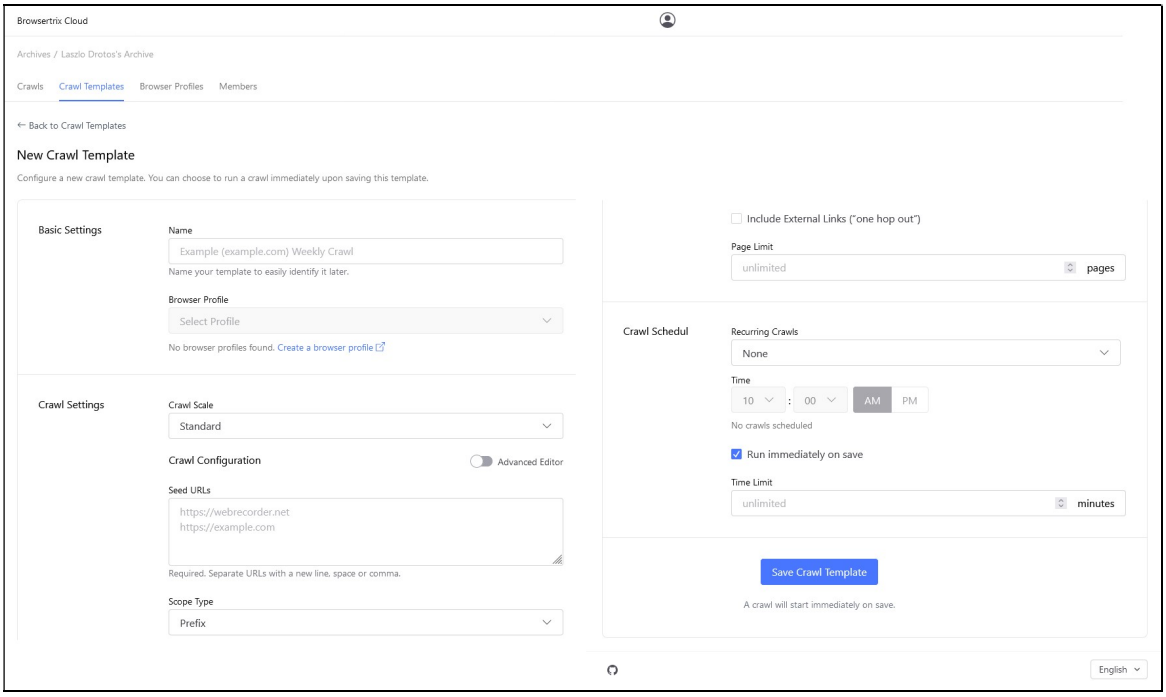
Az egyedi archiválási munkák közül két esetet érdemes külön is kiemelni:

Kollégánkon keresztül jelzést kaptunk arról, hogy a Pálos Rend honlapja hamarosan megújul és örülnének annak, ha a régi változat megőrződne a webarchívumban. A honlapról több mentést is készítettünk, és végül sikerült majdnem teljességében archiválnunk, ahogy az új változatot is. A folyamatról részletesen tájékoztattuk képviselőjüket, aki érdeklődve követte a történteket és támogatta javaslatukat, hogy a nyilvános archívumba is bekerülhessen az anyag. (Az engedélyeztetés folyamatban van.)

Még a szolgáltatási szerződések megújítása kapcsán vettük észre, hogy a Ráday Gyűjtemény honlapjáról kitiltásra került a mentést végző robotot futtató szerverünk, így nem tudjuk archiválni azt. A problémát jeleztük a gyűjtemény vezetőjének, aki az informatikusukat bízta meg a helyzet rendezésével. Telefonon többször egyeztettünk vele, végül a szolgáltató IP-cím alapján beengedte az archiváló szervereinket. Ezt több próbamentéssel is ellenőriztünk, majd ezek eredményéről tájékoztattuk őket.

Informatikai ügyek

Az IIPC szervezet tagjaként hozzáférést kaptunk a – fejlesztés alatt lévő – Browsertrix Cloud nevű, felhőalapú webarchiváló rendszer teszt szerveréhez, melyen indítottunk is néhány próbamentést. A Browsertrix robotja egy böngészőn keresztül menti le a oldalakat, így bár lassabban, de sokkal jobb minőségben tudja archiválni a mai, dinamikusan generált weblapokat, mint a web korábbi generációjára kifejlesztett Heritrix. A felhasználók egy webes űrlap segítségével paraméterezhetik és ütemezhetik az aratási feladatokat, melyekhez előre definiált „böngésző profilok" is hozzárendelhetők, így lehet például bejelentkezést igénylő előfizetéses vagy közösségi média oldalakat is letölteni. A lementett tartalom azonnal visszanézhető a beágyazott ReplayWeb.page megjelenítővel, illetve kiexportálható WACZ formátumban, ami a WARC fájlok mellett a visszakeresést segítő indexállományokat és különféle metaadatokat tartalmazó csomag.



A Browsertrix Cloud aratások paraméterezésére szolgáló űrlap

Az elmúlt hetekben az informatikusunk, OSZK-s kollégája segítségével, összerakott egy Kubernetes platformot, egyelőre szintén csak teszt szervereken. Ezzel a megoldással a jelenleginél rugalmasabban és remélhetőleg kevesebb technikai problémába ütközve előre elkészített *image*-ek formájában lehet majd telepíteni/frissíteni az archiváláshoz és a megjelenítéshez szükséges különféle szoftvereket. Elsőként éppen a Browsertrix Cloud keretrendszert szeretnénk beüzemelni, később pedig egy SolrCloud-ot is létrehozni a teljes szövegű kereséshez.

A zárt archívumhoz való hozzáférést is sikerült megoldani ebben a hónapban. Kizárólag erre a feladatra dedikált gépek lettek beállítva az OSZK „Általános olvasó” nevű részén, melyekről csak megtekinteni lehet a weboldalakat, letöltési lehetőség szerzői és személyiségi jogok miatt nincsen. A gyűjteményben való böngészéshez és kereséshez elkészítettünk egy egyszerű felületet és egy tájékoztató szöveget.

Gyűjtőkör

Az OSZK megújítás alatt lévő gyűjtőköri szabályzatához a Digitális Bölcsészeti Központ munkatársaival közösen összeállítottunk egy szövegjavaslatot, ami a digitális született dokumentumok megőrzésére és ennek részeként a webarchiválásra vonatkozik. A digitális univerzum gyors változása miatt ebben a szövegben csak az általános elveket fogalmaztuk meg, a webarchívum működésére egy külön, gyakrabban frissíthető belső szabályzatot készítünk a közeljövőben.

WARCnet

2022. június 13. és 15. között került sor a WARCnet kutatási hálózat újabb értekezletére a University of London patinás főépületében a Senate House-ban, melyen Németh Márton képviselte az OSZK-t. A projekt hat munkacsoportjából háromban vagyunk érdekeltek: az 1. számúban, mely a webarchívumok összehasonlító kutatásával, visszakereshetőségével foglalkozik; az 5.-ben, ami a kutatók igényeinek megfelelő adatformátumok és adatsémák meghatározásán dolgozik; és a 6. munkacsoportban, melyben minden projekttag részt vesz és amelyben a fenntarthatóság a cél új pályázati lehetőségek és projektek felkutatásával 2023 tavasza utánra, amikorra a WARCnet projekt kifut. A találkozón a fő hangsúly a különféle kutatási irányokat, projekteket és eredményeket összegző tanulmánykötet

struktúrájának egyeztetésén és a főbb témakörök bemutatásán volt. Mi két fejezet összeállításában veszünk részt. Az egyikben Németh Márton és Kalcsó Gyula az orosz-ukrán háború kapcsán létrehozott gyűjtemények korpuszain folytatott szövegelemzési-adatvizualizációs tevékenységekről számolnak majd be (ennek alapjai egy rövid előadás keretében az értekezleten is ismertetésre kerültek). A másik téma a webarchiválás megjelenése különféle felsőoktatási képzési programok keretei között. Ezzel kapcsolatban a WARCnet hálózat tagjainak segítségével szeretnénk felmérni a meglévő jó gyakorlatokat és a jövőbeni lehetőségeket.

A találkozón – a személyes kommunikáció előnyeit kihasználva – betekintést lehetett nyerni mélyebben is egyes országok (Dánia, Hollandia, Franciaország, Nagy-Britannia) munkafolyamataiba, valamint ismereteket szerezni az egyetemi oktatói és kutatói igényekről, valamint a súlyos szerzői, jogi, adatvédelemi kihívásokról ezek kielégítése kapcsán. Új, még fejlesztés alatt álló szoftvereket lehetett kipróbálni és konzultálni a fejlesztőkkel. Mi is meg tudtuk mutatni a SolrWayback-nek azt az új verzióját, amely bár a nyilvános felületen nem jeleníti meg az Ukrajnával kapcsolatos archivált hírek teljes szövegét, viszont visszakereshetővé teszi őket kutatási célból. Bemutattuk továbbá az Internet Archive által számunkra fejlesztett portált, ahol a .hu doménről 1996 óta napjainkig archivált magyar anyag teljes szövegében lehet keresni.

Az értekezletet követően megbeszélésre került sor a British Library webarchívumában a gyűjtemény PR felelősével, illetve a könyvtár magyar állományának fő katalogizálójával, aki egyben a webarchívum klímaváltozás tematikájú részgyűjteményének gazdája is. Témaként a webarchiválást segítő tágabb közgyűjteményi, kutatói együttműködési hálózatok szervezése, működtetésük előnyei, a BL-ben használt archiválási munkafolyamatok és stratégiai alapelvek áttekintése szerepelt.

A rendezvényről beszámoló készült, mely az OSZK blogján jelenik majd meg.

Ide kapcsolódik még, hogy megválaszoltuk egy, a WARCnet projekt munkatársai által összeállított online kérdőívet a COVID-19 járvány kapcsán végzett webarchiválási tevékenységekről.

Együttműködések

Júniusban folytattuk a megyei könyvtárakkal az együttműködést megcélzó megbeszéléseket. A Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér (Győr) és a Deák Ferenc Megyei és Városi Könyvtár (Zalaegerszeg) kollégái is nyitottan fogadták a kezdeményezést, ahogy a Pécsi Tudományegyetem Egyetemi Könyvtár és Tudásközpont (PTE EKTK) is. Utóbbi a megyei könyvtárral is együtt tudna működni feladatmegosztásban, ezért került megkeresésre „soron kívül”. Az elmúlt egy év alatt valamennyi megyei könyvtárral megtörtént a kapcsolatfelvétel, közülük egy tartalmát tekintve utasította el az együttműködést, míg egy másik a kezdeményezéssel egyetértve, de kapacitáshiány miatt. Jelenleg a beérkező válaszokat várjuk és az esetlegesen felmerülő kérdéseket válaszoljuk meg, hogy véglegesíteni lehessen a megállapodásokat és elkezdődhessen a közös munka. Addig is a Bács-Kiskun Megyei Katona József Könyvtár (Kecskemét) már nagy lendülettel belevágott a munkába, hiszen a megyei gyűjtemény kialakítása önmagában is hasznos helytörténeti vonatkozásban.

Június 28-án Holl Andrással, az MTA Könyvtár és Információs Központ főigazgató-helyettesével tartottunk megbeszélést arról, hogy hogyan lehetne megőrizni azokat az online forrásokat, melyek létrehozását az MTA pályázati pénzekkel támogatja, de nem olyan formátumúak, amilyeneket a REAL repozitórium be tud fogadni. Mivel ezek általában amúgy is az OSZK Webarchívumának gyűjtőkörébe tartoznak, ezért csak annak a folyamatát kell kidolgozni, hogy az archívum értesüljön róluk, a tartalmak előállítói pedig a támogatásért cserébe engedélyezzék a nyilvános hozzáférést az archivált verzióhoz – legalább az után, amikor az már lekerült az élő webről, és legalább ahhoz a részéhez, ami a pályázati pénzből valósult meg. A beszélgetés második felében arról volt szó, hogy hogyan lehetne kutatási célokra használni a webarchívum állományát (pl. nyelvi elemzéssel kigyűjteni belőle a tudományos publikációkat tartalmazó PDF fájlokat), valamint hogy érdemes lenne felhívni az MTA Könyvtárát használó kutatók figyelmét erre az OSZK-ban található *big data* korpuszra és az Internet Archive által biztosított keresőre és API-ra.

Közösségi szolgálat

Idén nem egyetemről érkezett hozzánk szakmai gyakorlatra hallgató, hanem egy közösségi szolgálatra jelentkező középiskolás diákot fogadhattunk, összesen 40 órányi foglalkoztatással. Az első nap általánosságban ismertettük meg a webarchiválással, majd pedig olyan feladatokkal látjuk el, amelyek ebben az életkorban már megbízhatóan elvégezhetők és érdekesek lehetnek (pl. webcímek és oldalképek ellenőrzése, podkasztok mentése).

Az elmúlt hetekben lefutott tematikus aratások

Sport, testkultúra (3490 db seed URL)
Vallások, hitrendszerek, egyházak (2740 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (910 db seed URL)
Kutatóintézetek, tudományos szervezetek (1157 db seed URL)
Egyetemek, főiskolák (3960 db seed URL)
Idegenforgalom, vendéglátás (5926 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