Az OSZK Webarchívum 2021 októberi hírei

Címlisták bővítése

Folytatódott az új TERMUSZ, vagyis a természet- és műszaki tudományos weboldalak címlistájának gyűjtése, októberben főként a földtan, a fizika, a kémia, a biológia és a matematika témákkal. A lista már közel ezer nevet és URL címet tartalmaz. E mellett a rendszeresen aratott egyéb tematikus gyűjteményeink is bővültek néhány száz címmel, az elektronikus periodikák nyilvántartásába pedig 59 új tételt vettünk fel e hónapban.

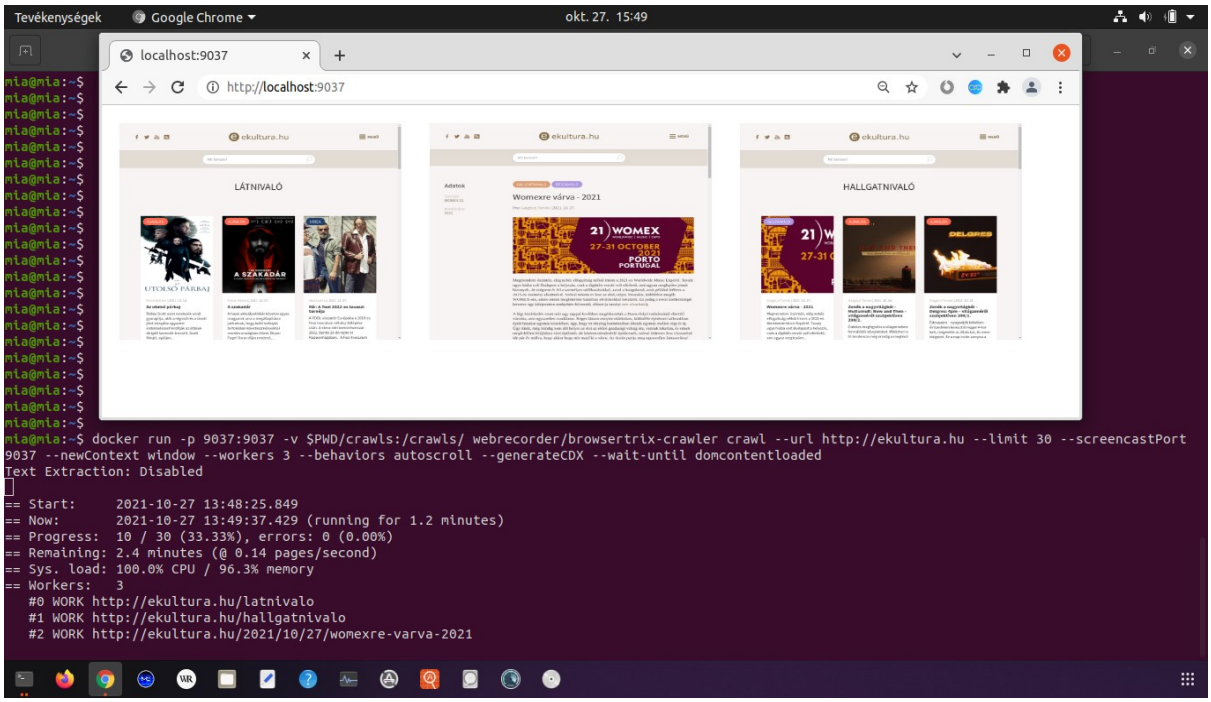
Teljes szövegű kereső

Több hónapnyi feldolgozás után kereshetővé vált a zárt webarchívumba 2017 óta begyűjtött dokumentumok teljes szövege. A Solr 31 ezer WARC fájlt indexelt le (további 773 esetben pedig hibát jelzett), az index fájl 966 millió URL-t tartalmaz és a mérete 2,8 terabájt. A visszakereső és megjelenítő eszköz ugyanaz a SolrWayback, amit a nyilvános gyűjteményben is használunk. Az első tesztek azt mutatták, hogy ekkora méretű állomány esetén nagyon lelassult a keresés és a találatok megnyitása, ezért segítséget kértünk a Dán Királyi Könyvtár informatikusától, a SolrWayback egyik fejlesztőjétől. A dán webarchívumnál 5 szerveren összesen 125 Solr példányt futtatnak, egyenként 1 terabájtos SSD háttértárral, ezeken van szétdarabolva az index állomány. Mivel a mi szerverünkhöz még nincs SSD tároló, ezért más módon próbáltuk meg optimalizálni a visszakeresést: az eredetileg 5 gigabájtos index részeket összefésültük, így érezhetően javult a sebesség, de még nagyon messze van az ideálistól. További probléma a webarchívumokra jellemző nagy redundancia, vagyis a sok duplum, amelyek nehezen szűrhetők ki, és természetesen a találati listáknál a relevancia szerinti rendezés messze nem olyan hatékony, mint amilyet a felhasználók a Google keresőnél megszoktak. A nagy index mellett két kisebb részgyűjteményben külön is lehet teljes szövegű keresést végezni: a válogatott híroldalak anyaga, illetve a koronavírus járvánnyal kapcsolatos források. Ezek a keresőfunkciók az OSZK-ban kialakított terminálszerveren keresztül lesznek elérhetők olvasók és kutatók számára.

Szoftverteszt

A hónap folyamán több tesztet is végeztünk a 0.4-es verziójú, nyílt forráskódú Browsertrix szoftverrel, ami idővel felválthatja a jelenleg a hírportálok napi szintű mentéséhez használt Brozzler programot, sőt talán a közösségi média oldalak automatizált letöltéséhez is alkalmas lesz a nagyon élőmunka-igényes ArchiveWeb.page helyett. A Browsertrix a nevét a „browser” és a „Heritrix” szavak összevonásából kapta, ami a működési módjára utal: böngészőt használ a megadott weboldal betöltéséhez, majd elmentéséhez, de utána ugyanúgy követi a benne levő linkeket, mint az Internet Archive Hertirix nevű robotja. A Chrome-ot egy Puppeteer nevű eszköz segítségével vezérli, a mentést pedig a PyWb végzi *capturing* üzemmódban. Vannak hozzá úgy nevezett *behavior* szkriptek (pl. autoscroll, video autoplay, valamint webhely-specifikus viselkedések), így a Heritrixhez képes sokkal jobban archiválhatók vele a dinamikusan generált, emberi interaktivitást igénylő weboldalak. A szabványos WARC mellett WACZ formátumba is tud menteni, ami a ReplayWeb.page megjelenítő számára szükséges indexeket és technikai adatokat is tartalmazza egyetlen zip csomagként. A böngészőn keresztül való archiválásnak egyetlen hátránya a sebessége, mert meg kell várni, amíg az oldalak betöltődnek és végiggörgeti őket a szkript, ezért ez a módszer lényegesen lassabb, mint a „buta” robottal való aratás. A tesztjeink során ezt vizsgáltuk különböző környezetekben: **1.** a könyvtár által biztosított Macbook Air (2017-es modell, 8 GB memória, 256 GB HDD) laptopon, **2.** egy 4 GB memóriával 128 GB Flash memóriával, ARM processzoron futó Raspberry Pi miniszámítógépen, **3.** egy 8 GB memóriájú Intel I7-es processzorral rendelkező Windows laptopon kialakított, 40 GB méretű háttértárral, 4 GB memóriával bíró VirtualBox-alapú virtuális gépen, **4.** egy egy hónapos ingyenes próbaidőre beüzemelt, felhőben futó virtuális szerveren (2 magos processzor, 4 GB RAM, 60 GB SSD). Tapasztalataink szerint ez utóbbi virtuális gép a többi konfigurációnál

legalább kétszer gyorsabban hajtotta végre a kijelölt feladatokat. Ez is arra utal, hogy a megfelelő processzorteljesítmény SSD alapú tárhellyel társítva számottevő mértékben képes megnövelni még a kisebb aratási műveletek végrehajtásának hatékonyságát is.



Az ekultura.hu honlap tesztmentése a Browsertrix segítségével egy virtuális gépen.
(A fekete terminálablakban az indítási parancs és az aratás előrehaladása látható, a Chrome böngészőben pedig az, ahogyan a program három szálon párhuzamosan nyitja meg az oldalakat.)

PhD védés

Németh Márton kollégánk október 15-én sikeresen, „summa cum laude” minősítéssel megvédte doktori értekezését, melynek címe: „A webarchiválás elméletének és gyakorlatának alapelemei. A szervezett keretek között zajló webarchiválás kezdetei Magyarországon.” A disszertáció a Debreceni Egyetem Természettudományi és Informatikai Doktori Tanácsa Informatikai Tudományok Doktori Iskolájában készült, Eszenyiné dr. Borbély Mária egyetemi adjunktus témavezetésével. A dolgozat áttekintő képet nyújt a webarchiválás elméleti alapjairól, a webarchívumok kutatási célú felhasználásáról, a webarchiválás oktatásáról, illetve az OSZK keretei között folyó webarchiválási tevékenység alapjainak megteremtéséről, az informatikai háttérről és a munkafolyamatokról. (https://dea.lib.unideb.hu/dea/handle/2437/310638)

Workshop előkészítés

Október legnagyobb feladata a november 23-án és 24-én megrendezésre kerülő „404 Not Found – Ki őrzi meg az internetet?” című videókonferencia és workshop előkészítése volt. A különböző részfeladatokhoz kapcsolódóan nagy segítségünkre voltak más osztályokon dolgozó OSZK-s kollégák is, például a szervezésben, a lektorálásban, valamint a grafikai anyag, a regisztrációs űrlap, a honlap hír, a PR-címlista és a sajtóanyag elkészítésében. Mindez sok megbeszéléssel, egyeztetéssel és levelezéssel járt, és mivel a korábbi évektől eltérően idén kettős, részben nemzetközi rendezvény lesz a „workshop”, külön figyelmet kellett fordítani arra, hogy minden anyag angolul is elkészüljön és a partnerek, valamint az érdeklődők rendelkezésére álljon. A konferencia első napjához kapcsolódó, a közép-európai webarchívumok együttműködését megcélzó CEWA memorandumot előzetesen elküldtük a külföldi partnereinknek véleményezésre. A végső egyeztetések után, várhatóan november elején tudjuk majd hivatalos formában is meghirdetni a rendezvényt.

Külföldi kapcsolatok

Megújul az IIPC szervezet végrehajtó bizottsága, a jelölési folyamat lezárult, a szavazatunkat október közepén adtuk le. Zajlik a szerződéskötés előkészítése az Internet Archive-val, melynek eredményeként egy általuk kialakított portál felületen teljes szöveggel kereshetővé válik a náluk levő magyar web-tartalom, az OSZK pedig megkapja az általuk nyilvántartott .hu domén alá tartozó címlistát. „Brief introduction to the Hungarian Web Archive at the National Széchényi Library" címmel összeállítottunk egy prezentációt a Lengyel Állami Levéltárban november közepére meghirdetett webarchiválási témájú rendezvényre.

Hazai együttműködések

A workshop előkészítése miatt újabb vagy emlékeztető megkereséseket nem küldtünk ki a célba vett megyei könyvtáraknak, csak a folyamatban lévő ügyeket vittük tovább. A hónap elején érkezett meg az egyik érintett intézmény válasza a együttműködés véglegesíthetőségéről, így velük együtt már öt megyei és egy egyetemi kari könyvtárral tudtunk elviekben megállapodni. Új partnerként a Békés Megyei Könyvtár (BMK) munkatársaival ismertettük az elképzeléseinket (velük technikai okokból nem tudtunk korábban beszélni), ami náluk is pozitív fogadtatásban részesült. Elkészítettük a leendő regionális gyűjtemény táblázatát és leválogattuk a nyilvántartásunkból a vonatkozó címeket. Reméljük, a hivatalos válaszuk is hamarosan megérkezik, ahogy a többi függőben lévő is. Október 11-én megbeszélést folytattunk az ELTE Digitális Bölcsészeti Központ munkatársaival, akik beszámoltak a webarchiválást érintő tevékenységeik előrehaladásáról. Abban állapodtunk meg, hogy december elején visszatérünk rá, hogy milyen területeken tudnánk szorosabban együttműködni. Indig Balázst felkértük egy előadás megtartására az idei workshopunkon, aki el is vállalta azt.

Az elmúlt hetekben lefutott tematikus és műfaji aratások

Történelem, hely- és családtörténet (1.097 db seed URL)
Média, sajtó, műsorszórás (999 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (6.150 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1.481 db seed URL)
Elektronikus periodikák (7.615 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