Az OSZK Webarchívum 2023 novemberi hírei

Archiválás és nyilvántartás

Elkészült a nyári technikai problémák miatt őszre tolódott webtér szintű aratás statisztikája és megtekinthető a honlapunkon. A két sikeresen lefutott job aratása együtt több mint 21 napig tartott és ez idő alatt a megadott 992 ezer címről kiindulva közel 200 millió URL-t talált a robot. Ezek közül 127 milliót sikerült letöltenie, majd a duplumszűrés után több mint 81 milliót el is tárolt 6,45 terabájt összméretben. A webhelyek kezdőlapjait ábrázoló oldalképek készítése több hétig tartott. Végül 842 ezernél is több PNG fájl készült el, melyek összesen 1,26 terabájt tárhelyet foglalnak el.

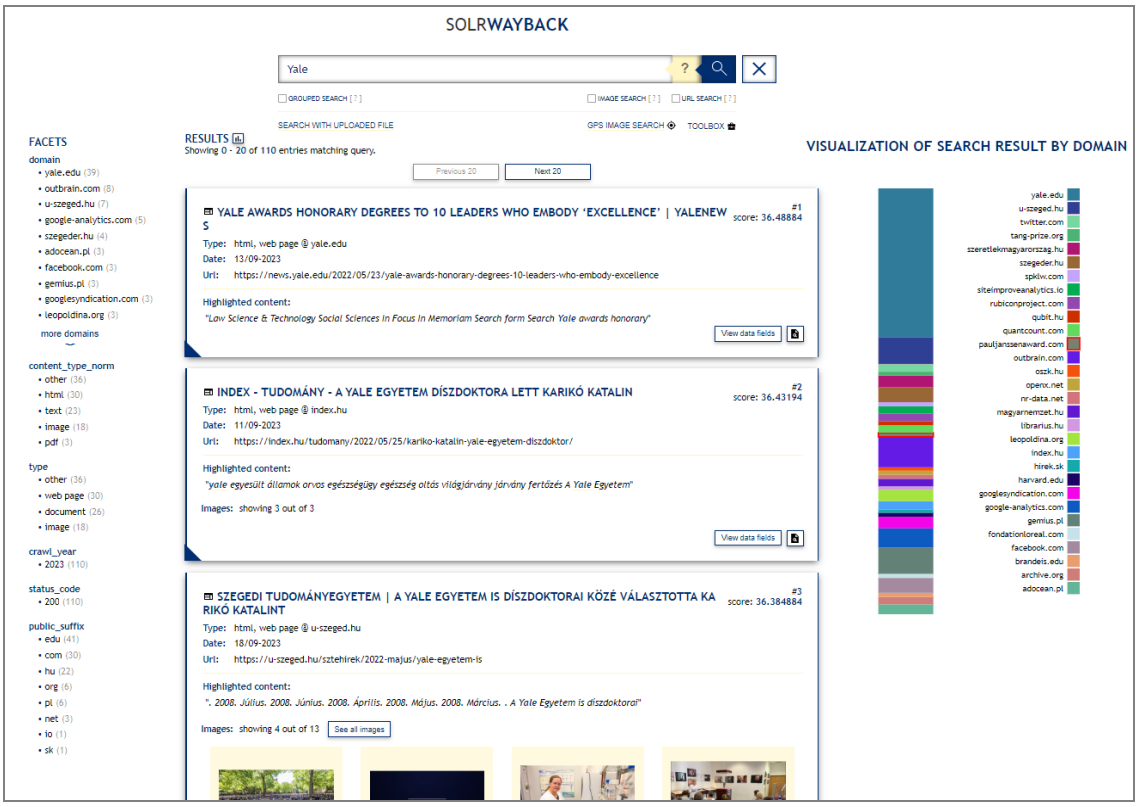
A webtér aratás után felgyorsított ütemezéssel indítottuk el 18 tematikus és 2 műfaji részgyűjtemény negyedik negyedéves mentését, melyek novemberben mind el is készültek, majd pedig ezek adatai és az azokból készült diagramok is felkerültek a honlapra. Ezeket az aratásokat a korábbiakhoz képest „udvariasabb" beállításokkal futtattuk, ami azt jelenti, hogy megnöveltük a webszervereknek küldött kérések közötti várakozási időt, hogy elkerüljük a robotunk esetleges kitiltását. A tárhelyköltözés és a fent említett nyári „válságos" informatikai helyzet miatt a részgyűjteményeknél egy vagy két aratás kimaradt az év folyamán.

Ugyancsak kimaradtak alkalmak az először augusztusban aratott „Európa Kulturális Fővárosa - 2023" nevű hír- és webhely-válogatás esetében, részben a webtér aratás miatt, részben pedig azért, mert nem működött az automatikus ütemezés. Végül november 20-án „kézzel" indítottuk el a három, eltérően paraméterezett jobot, melyek ezúttal sikeresen lefutottak.

A hónap folyamán elkészült a kozterkep.hu képmegosztó oldal fotóanyagának és metaadatainak begyűjtése a Scrapy programmal, de a képfájlok lementése csak második nekifutásra sikerült, mert a tesztszerveren elfogyott a tárhely. Négy spidert (Python scriptet) írtunk és futtattunk le a SpiderKeeper keretrendszerrel vezérelve. Az egyik a köztéri alkotások műlapjait járta be 24 perc alatt és egy több mint 42 ezer sorból álló CSV fájlba írta az azokon található fontosabb adatokat és szövegeket. A másik spider megnyitotta az összes műlap fotógalériáját és elmentette a képeket (ennek a futási ideje 6,7 nap volt), egy további script pedig a több mint 475 ezer fotó mellett feltüntetett főbb metaadatokat írta egy szintén CSV formátumú szöveges állományba 33 perc alatt. Végül a negyedik spider az alkotók adatlapjait mentette le mindössze 2 perc alatt egy 9300 soros fájlba. Az eredmények ellenőrzése során találtunk kisebb eltéréseket a honlapon levő statisztikák és az általunk lementett darabszámok között, de ezeknek az lehet a magyarázata, hogy nem minden műlap és fénykép nyilvános. A képeket letöltő spider esetében csak pár darab hibaüzenet volt a naplófájlban, de később ezeket a fájlokat is sikerült lementenie. A jogi státusz alapján legalább 258 ezer fotó rendelkezik valamilyen Creative Commons licenccel, így ezek – a metaadatok megfeleltetése és adatbázisba töltése után – megjelenhetnek majd a Digitális Képarchívum gyűjteményében is. A teljes anyagot pedig eltettük a webarchívum raktárába hosszú távú megőrzés céljából. További feladat lesz még azoknak a scripteknek a megírása és beütemezése, amelyekkel frissíteni lehet az archív anyagot a Köztérképre az utolsó mentésünk óta felkerült képekkel és adatokkal.

Újrakezdtük a Karikó Katalinnal kapcsolatos hírek és videók egyedi mentéseit az ArchiveWeb.page böngészőkiegészítővel. A SZTE Klebelsberg Könyvtár munkatársától 80 új címet kaptunk, melyek közül már 30-at archiváltunk is november utolsó napjaiban. Elkészült továbbá egy kereső, amivel a még szeptemberben lementett 1120 weboldal teljes szövegében lehet keresni, valamint különféle statisztikai adatokat generálni (https://kariko.webharvest.oszk.hu/solrwayback). Maguk a mentések jogi okokból csak az OSZK olvasótermében tekinthetők meg, de a metaadatok között a nyilvános felületen is megtalálható az eredeti URL cím, melyet kimásolva megnézhető az élő weboldal, amennyiben még elérhető. Az együttműködés részét képezi a szegedi könyvtárosok által gondozott Karikó Katalin virtuális kiállítás rendszeres archiválása is (https://mediateka.ek.szte.hu/exhibits/show/kariko_katalin_szte).

Elkezdtünk egy új tematikus részgyűjteményt építeni CSALAD kódnévvel, melyben ilyen altémák lesznek: „Család általában”, „Párkapcsolat, esküvő, házasság”, „Szexualitás”, „Születés, anyaság”, „Gyermekek, gyermeknevelés”, „Idősek”, „Halál”. Eddig kb. ezer honlap és blog címét vettük nyilvántartásba, a gyűjtés decemberben is folytatódik még.



A Karikó-webarchívum keresője és a találati lista megoszlása doménnév szerint

Rendezvények

November 11-én Visky Ákos László webkurátor „Archivált webtartalom a könyvtári gyűjteményekben – a közgyűjtemények szerepe az internetes tartalmak megőrzésében” címmel tartott előadást a Pécs-Baranya évszázadai helyismereti konferencián a webarchiválásról és az OSZK archívumáról. A prezentáció letölthető a honlapunkról.

November 29-én lezajlott a hetedik „404 Not Found – Ki őrzi meg az internetet?” konferencia és workshop. A résztvevők száma meghaladta a kilencvenet, ebből több mint ötvenen online követték az eseményt. Az előadások témái: az OSZK webarchívumának megújítása, célzott adatmentés scraping technológiával, a webaratásból származó szövegek automatikus feldolgozása, valamint a szegedi egyetemi könyvtárral együtt épített Karikó-webarchívum. A konferencia egyúttal záróeseménye volt a Luxembourgi Nemzeti Könyvtárral közösen elnyert nemzetközi pályázatnak. A három napos látogatásra Budapestre érkezett luxemburgi kollégák két előadásban mutatták be tevékenységüket, valamint egy kétrészes workshopot is tartottak az általuk használt technológiákról. A prezentációk és a helyszínen készült fotók megnézhetők a honlapunkon.

Internet Archive kereső

A webarchívum nem publikus része az OSZK olvasótermében elhelyezett két, erre a célra dedikált gépről böngészhető. Néhány napja ezekről a PC-kről már az Internet Archive szerverén kialakított portál is elérhető, melyen az amerikai webarchívum által 1996 és 2022 között a magyar .hu domén alól lementett weboldalak teljes szövegében lehet keresni, valamint az ezekbe belinkelt kép-, hang-, videó- és PDF-fájlok is megtalálhatók vele. A találatok a Wayback Machine szolgáltatással nézhetők meg, ahol az adott doménra vonatkozó különböző statisztikák és vizualizációk is lekérhetők. A keresőfelület angolra és magyarra is átváltható, és kétnyelvű súgót is készítettünk hozzá.

Az elmúlt hetekben lefutott aratások

Életmód, szabadidő, hobbi (8125 db seed URL)
Természet- és műszaki tudományok, szakterületek (2151 db seed URL)
Könyvtárak, levéltárak, múzeumok és galériák (2108 db seed URL)
Irodalom, irodalomtudomány és -történet (1391 db seed URL)
Képző-, előadó-, zene- és filmművészet (8563 db seed URL)
Történelem, hely- és családtörténet (1330 db seed URL)
Média, sajtó, műsorszórás (931 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1459 db seed URL)
Podkasztok (4694 db seed URL)
Elektronikus periodikák (10083 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (6777 db seed URL)
Vallások, hitrendszerek, egyházak (2847 db seed URL)
Idegenforgalom, vendéglátás (6850 db seed URL)
Sport, testkultúra (3707 db seed URL)
Kutatóintézetek, tudományos szervezetek (1261 db seed URL)
Egyetemek, főiskolák (4211 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (950 db seed URL)
Egészségügy, szociális szféra (8212 db seed URL)
Közoktatás és egyéb képzések (7012 db seed URL)
Bölcsészet- és társadalomtudományok, szakterületek (5903 db seed URL)

Az egyes 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: webarchivum@oszk.hu