Az OSZK Webarchívum 2023 októberi hírei

Személyi változás

A webarchívum munkafolyamatainak informatikai támogatását új kolléga, Bukovics János vette át, aki a napi teendők mellett a régóta esedékes fejlesztéseket is elkezdi majd. Az elmúlt napok a szervereken futó különféle szoftverek és scriptek megismerésével, valamint az aktuális technikai problémák okainak felderítésével teltek. A feladatok átvétele és a hibajelenségek elhárítása a következő hetekben is folytatódik még.

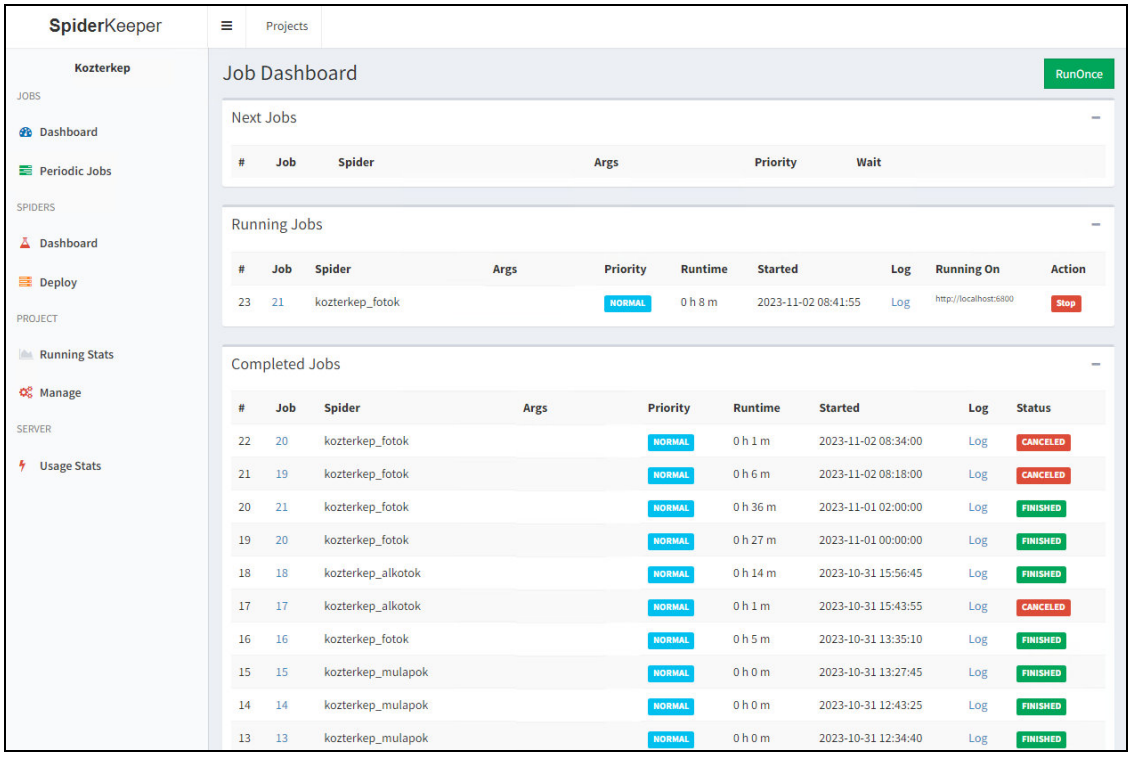
Archiválás

Az egyik technikai hibát a szerver IP címének megváltozása okozta, ami miatt hetekig nem működött az aratások paraméterezésére és indítására használt Kaptafa űrlap mögött futó program. Ezt végül sikerült megjavítani, de még vannak további scriptek is, amelyeket érint az IP címváltás és kihatással vannak az előző években már bejáratott munkafolyamatokra.

A korábbi problémák miatt szeptember 21-én újraindított webtér aratáshoz tartozó első jobot 16 nap és 5 óra futásidő után leállítottuk. Ez alatt az idő alatt 823 ezer szerveren több mint 101 millió URL-t talált és töltött le sikeresen a Heritrix, melyek közül 67,6 millió volt új vagy megváltozott tartalom. Így az összesen learatott 6,3 terabájtból a korábbi mentésekhez képest duplumnak számító fájlok eltávolítása után végül 5,3 terabájttal nőtt a webarchívum mérete. A WARC fájlok mellett 452.358 darab PNG oldalkép is készült 909 gigabájt összméretben a kiindulásként megadott 514.346 seed URL címről. A második job, ami a robots.txt fájl nélküli webhelyeket aratta, 5 napig futott október végén és 25,4 millió URL-t töltött le sikeresen, az eltárolt új tartalom pedig 1,1 terabájt. A tömegesen generált aldoménekre kiterjedő harmadik feladatot kétszer is megpróbáltuk elindítani, de mindkét esetben kevesebb, mint 1 gigabájt lementése után már csak hibaüzeneteket küldött a robot. Ennek a jelenségnek az okát még vizsgáljuk.

Október végén befejeződött a magyar podcast csatornákon az elmúlt egy éves időszak alatt megjelent adások mentése. A több mint 1500 nyilvántartott feedben kb. 1000 esetben voltak új részek, a többi műsor már valószínűleg befejeződött, sőt már esetleg teljesen el is tűnt az internetről, illetve van néhány, amelyek még működnek ugyan, de olyan platformon, ahol nincs letöltési lehetőség. A frissítés során ebben a hónapban 545 podcast 17.112 adását töltöttük le MP3, MP4 vagy M4A fájlok formájában, melyek mérete összesen csaknem 1 terabájt. Időközben több mint félszáz újabb csatornát is sikerült találni, ezek nyilvántartása és archiválása elkezdődött.

Október 10-én beszéltünk a kozterkep.hu képmegosztó oldal gazdájával a webhely anyagának scraping módszerrel való archiválásáról, a hónap végén pedig már néhány „éles” tesztet le is futtattunk a szerverünkön. Ezekhez négy önálló Scrapy spidert, vagyis Python nyelvű scriptet írtunk, melyek az alkotók, a köztéri művek, illetve az egyes fotók metaadatait töltik le, továbbá magukat a képfájlokat. A próbák során előjött egy-két eddig nem tapasztalt hiba, ezeket a következő napokban meg kell még oldani. Biztató jel, hogy a teszt során begyűjtött 236 fotóadatlap közt mindössze 17 olyan volt, amin nincs valamilyen Creative Commons licenc, vagyis a fényképek nagy része bekerülhet majd a Digitális Képarchívum nyilvános gyűjteményébe is.



A Scrapy vezérlésére használt SpiderKeeper program „műszerfala”

Rendezvények

Az LA Europe (a Special Libraries Association európai tagozata) október 19-én tartott webináriumán a brit webarchívum mutatkozott be. A British Library vezető munkatársa, az archívum kurátora a gyűjtemény történetének, összetételének és szolgáltatásainak ismertetése mellett érdekes adatokat is megosztott a web eróziójáról, a webtér aratásaik méretéről (5-10 millió webhely, évi 70-100 terabájt növekedés), az emberi munkával gondozott részgyűjteményeikről (több mint 100 téma/esemény), valamint a belső és külső együttműködésekről. Számunkra a legnagyobb újdonság a „Document Harvester” nevű, saját fejlesztésű programjuk volt, amivel a kormányzati honlapokról készült mentésekből kiválogatják a PDF fájlokat és az azokban található metaadatokat. Utóbbiakat a katalogizáló könyvtárosok még átnézik és kiegészítik, majd az így létrejött rekord megjelenik a könyvtár katalógusában, a hozzá tartozó link pedig a webarchívumban található fájlra mutat. Mivel a kormányzati tartalmak esetében náluk sem kell külön engedély a nyilvános szolgáltatásra, ezért ezzel a módszerrel a szabad hozzáférésű digitális dokumentum gyűjteményüket is bővíteni tudják.

Összeállítottuk a november 29-re tervezett „404 Not Found – Ki őrzi meg az internetet?” című rendezvényünk programját, készülnek a grafikai anyagok és rövidesen elérhető lesz a regisztrációs űrlap is. Október 25-én volt egy újabb megbeszélés a konferenciára meghívott luxemburgi kollégákkal, akik a nyitóelőadások mellett délután két workshopot is tartanak majd.

November 9-én és 10-én kerül megrendezésre a „Pécs–Baranya évszázadai” című országos helyismereti konferencia a Csorba Győző Könyvtár, a Pécs Története Alapítvány és az MTA PAB Város- és Helytörténeti Munkabizottsága közös szervezésében. A közgyűjteményi szekcióban Visky Ákos László, az OSZK webarchívumának kurátora tart majd előadást „Archivált webtartalom a könyvtári gyűjteményekben – A közgyűjtemények szerepe az internetes tartalmak megőrzésében” címmel, melynek kivonata már olvasható a rendezvény programfüzetében és elkészült az előadás szövegének írott verziója is.

A projekt hírei a https://webarchivum.oszk.hu/a-projektrol/hirek-esemenyek/ oldalon kísérhetők figyelemmel. Kapcsolati cím: webarchivum@oszk.hu