Az OSZK Webarchívum 2023 augusztusi hírei

Archiválás, nyilvántartás

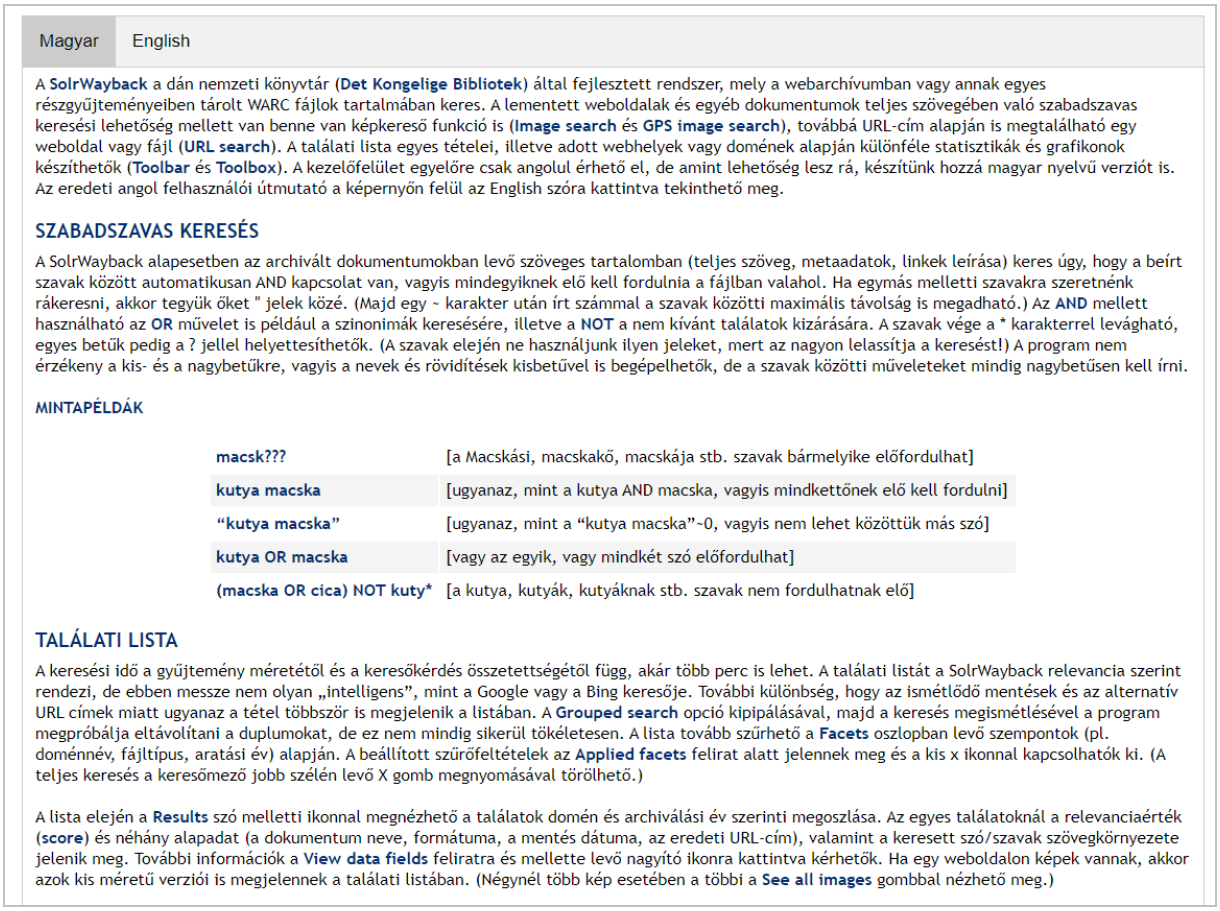
Július végén befejeződött a webarchívum anyagának még májusban elkezdett átköltöztetése egy 300 terabájtos, a korábbinál kétszer nagyobb tárhelyre, így ebben a hónapban már újból tudtunk aratásokat indítani a webharvest és a webharvest2 szervereken. A korábbi üzemszerű munka sajnos még nem állt teljesen helyre. A böngészőn keresztül archiváló Brozzler programmal naponta mentett hírportálok esetében jóval kevesebb az eltárolt tartalom, mint az első negyedévben volt, és az idei első webtér aratási kísérlet esetében is csak töredékét sikerült letölteni a tavaly decemberi mennyiségnek, a kiindulópontként megadott URL címek legnagyobb részénél el sem indult a Heritrix robot. A problémák okait az informatikusok vizsgálják, az egyik lehetséges magyarázat, hogy a webharvest2 szerverhez nem volt PTR rekord vagy más néven reverse DNS bejegyzés, ami a spam és a hacker támadások elleni védelemre szolgál. További gond, hogy az aratások log és report fájljai, valamint az oldalképek hibás dátumú mappákba kerülnek és ez a későbbiekben kavarodásokhoz vezethet.

Augusztus 4-én még sikeresen lefutott a tavasszal kialakított „Életmód, szabadidő, hobbi" elnevezésű részgyűjtemény első aratása. A kb. 8 ezer seed címről elindulva kevesebb, mint egy nap alatt közel 5 millió URL-t töltött le a robot, 512 gigabájt összméretben.

Két új, esemény-alapú gyűjtést is elkezdtünk ebben a hónapban. Az egyik az „Európa Kulturális Fővárosa" nevet viseli, mivel 2023-ban Veszprém és a Bakony-Balaton Régió kapta meg ezt a címet, melyhez számos rendezvény és fejlesztési projekt is kapcsolódik. Az ezekről szóló weboldalakat igyekszünk ezentúl havi rendszerességgel összegyűjteni a hivatalos honlapokról, 26 hírportálról, a közösségi médiából és néhány további magyar és külföldi online forrásból. A másik részgyűjtemény az augusztusban megrendezett budapesti atlétikai világbajnoksághoz kapcsolódó webhelyeket, híreket, Facebook, Instagram és Twitter posztokat tartalmazza, összesen több mint száz információforrást. Az archiválás augusztus 18-31. között napi rendszerességgel történt, és szeptember 8-én még csinálunk egy kiegészítő mentést az utólag megjelent hírekről. Ehhez az anyaghoz egy nyilvános teljes szövegű kereső is tartozik, aminél maguk a mentett weboldalak ugyan nem nézhetők meg a szerzői jogok miatt, de a találati listák különféle szempontok alapján szűrhetők, számos metaadatot tartalmaznak, valamint adatvizualizációk és adatkészletek is előállíthatóak belőlük (lásd: https://atletikavb2023.webharvest.oszk.hu/solrwayback). A SolrWayback keresőhöz magyar nyelvű leírást készítettünk, amely a kérdőjel ikonra kattintva jelenik meg és közérthetőbb formában ismerteti a rendszer főbb funkcióit, mint az eredeti angol útmutató.

Folytatódott a tematikus címlisták ellenőrzése és az elérhetetlenné vált oldalak felkutatása vagy megszűntté nyilvánítása. A tudományos témájú gyűjtemények átnézése után augusztus első felében az „irodalom" és a „kiadók, nyomdák" témakörökhöz tartozó honlapokat és blogokat néztük végig. Előbbi lista 1475, utóbbi pedig 1438 db webhelyet tartalmazott az ellenőrzés előtt, utána pedig ezek a számok 1323-ra, illetve 1336-ra csökkentek, de ezek között is van több tucat, melyekről nem lehetett még eldönteni, hogy csak ideiglenesen nem működnek, vagy végleg bezártak.

Júliusban 161 újonnan felfedezett podcast adásait mentettük le és raktároztuk el önálló hangfájlok formájában, most pedig elkezdtük a tavaly nyáron archivált csatornák azóta megjelent részeit letöltögetni a „Podcasts" nevű Chrome kiegészítővel. Az elmúlt két hétben 131 csatornáról 6270 új adást archiváltunk 460 gigabájt összméretben. Pár podcast esetében régebbi, a tavalyi gyűjtésből hiányzó adásokat is sikerült megtalálni és lementeni.



Magyar English

A **SolrWayback** a dán nemzeti könyvtár (**Det Kongelige Bibliotek**) által fejlesztett rendszer, mely a webarchívumban vagy annak egyes részgyűjteményeiben tárolt WARC fájlok tartalmában keres. A lementett weboldalak és egyéb dokumentumok teljes szövegében való szabadszavas keresési lehetőség mellett van benne van képkereső funkció is (**Image search** és **GPS image search**), továbbá URL-cím alapján is megtalálható egy weboldal vagy fájl (**URL search**). A találati lista egyes tételei, illetve adott webhelyek vagy domének alapján különféle statisztikák és grafikonok készíthetők (**Toolbar** és **Toolbox**). A kezelőfelület egyelőre csak angolul érhető el, de amint lehetőség lesz rá, készítünk hozzá magyar nyelvű verziót is. Az eredeti angol felhasználói útmutató a képernyőn felül az English szóra kattintva tekinthető meg.

SZABADSZAVAS KERESÉS

A SolrWayback alapesetben az archivált dokumentumokban levő szöveges tartalomban (teljes szöveg, metaadatok, linkek leírása) keres úgy, hogy a beírt szavak között automatikusan AND kapcsolat van, vagyis mindegyiknek elő kell fordulnia a fájlban valahol. Ha egymás melletti szavakra szeretnénk rákeresni, akkor tegyük őket " jelek közé. (Majd egy ~ karakter után írt számmal a szavak közötti maximális távolság is megadható.) Az **AND** mellett használható az **OR** művelet is például a szinonimák keresésére, illetve a **NOT** a nem kívánt találatok kizárására. A szavak vége a * karakterrel levágható, egyes betűk pedig a ? jellel helyettesíthetők. (A szavak elején ne használjunk ilyen jeleket, mert az nagyon lelassítja a keresést!) A program nem érzékeny a kis- és a nagybetűkre, vagyis a nevek és rövidítések kisbetűvel is begépelhetők, de a szavak közötti műveleteket mindig nagybetűsen kell írni.

MINTAPÉLDÁK

macsk???	[a Macskási, macskakő, macskája stb. szavak bármelyike előfordulhat]
kutya macska	[ugyanaz, mint a kutya AND macska, vagyis mindkettőnek elő kell fordulni]
“kutya macska”	[ugyanaz, mint a “kutya macska”~0, vagyis nem lehet közöttük más szó]
kutya OR macska	[vagy az egyik, vagy mindkét szó előfordulhat]
(macska OR cica) NOT kuty*	[a kutya, kutyák, kutyáknak stb. szavak nem fordulhatnak elő]

TALÁLATI LISTA

A keresési idő a gyűjtemény méretétől és a keresőkérdés összetettségétől függ, akár több perc is lehet. A találati listát a SolrWayback relevancia szerint rendezi, de ebben messze nem olyan „intelligens”, mint a Google vagy a Bing keresője. További különbség, hogy az ismétlődő mentések és az alternatív URL címek miatt ugyanaz a tétel többször is megjelenik a listában. A **Grouped search** opció kipipálásával, majd a keresés megismétlésével a program megpróbálja eltávolítani a duplumokat, de ez nem mindig sikerül tökéletesen. A lista tovább szűrhető a **Facets** oszlopban levő szempontok (pl. doménnév, fájltípus, aratási év) alapján. A beállított szűrőfeltételek az **Applied facets** felirat alatt jelennek meg és a kis x ikonnal kapcsolhatók ki. (A teljes keresés a keresőmező jobb szélén levő X gomb megnyomásával törölhető.)

A lista elején a **Results** szó melletti ikonnal megnézhető a találatok domén és archiválási év szerinti megoszlása. Az egyes találatoknál a relevanciaérték (**score**) és néhány alapadat (a dokumentum neve, formátuma, a mentés dátuma, az eredeti URL-cím), valamint a keresett szó/szavak szövegkörnyezete jelenik meg. További információk a **View data fields** feliratra és mellette levő nagyító ikonra kattintva kérhetők. Ha egy weboldalon képek vannak, akkor azok kis méretű verziói is megjelennek a találati listában. (Négynél több kép esetében a többi a **See all images** gombbal nézhető meg.)

A SolrWayback magyarított súgójának részlete

Néhány optikai- vagy mágneslemezen megőrzött régi magyar webes tartalmat is elraktároztunk augusztusban ZIP csomagok formájában a webarchívum erre a célra kialakított tárterületén. Ilyen beadási csomagként került be a gyűjteménybe például a 90-es évek magyar internetjének két népszerű honlapja: a Számítógép Gyűlölők Társasága és a Magyar Anime Útmutató, továbbá a HiX, vagyis a Hollósi Information eXchange 1990-1998 közötti, annak idején CD lemezen is terjesztett archívuma.

Ismeretterjesztés és ismeretszerzés

A webarchivum.oszk.hu/mediawiki címen levő wikihez 6 újabb szócikk készült el a hónap elején, főként az IIPC májusi konferenciájának videofelvételeiben bemutatott új szolgáltatásokról és szoftverekről.

Az International Internet Preservation Consortium felmérést készít a nagyobb webarchívumokban használt megoldásokról az archiválás, az indexelés és megjelenítés, a kutatás és elemzés, a szolgáltatás és a gyűjtemény gondozása, valamint a közösségi média megőrzése területén. Nemrég mi is kitöltöttük ezt a táblázatot az OSZK webarchívumában használt programok adataival.

Augusztus 30-án volt az IIPC tagok szokásos videobeszélgetése, melyen ezúttal nem voltak előre bejelentett beszámolók, de a szolgálati közlemények után egy-két új hírt azért megosztottak egymással a kollégák, például a SolrWayback-ről, a paywall mögötti weboldalak aratásáról és az egyik nagy francia tárhelyszolgáltató bezárása előtt megszervezett „vészhelyzeti” archiválásról.

Együttműködések

Ismét voltak megbeszélések a luxemburgi webarchívum munkatársaival az őszre tervezett tanulmányutakról. Az utazások részletei mellett a szeptemberi luxemburgi látogatás programját is egyeztettük velük.

Augusztus 17-én Váradi Tamással, a Nyelvtudományi Kutatóközpont főigazgató-helyettesével beszéltünk az OSZK és a Kutatóközpont közötti együttműködésnek a webarchívumot érintő részéről. Első lépésben a 2018-as első webtér szintű aratásig visszamenőleg átadjuk ezeknek az aratásoknak az indexeit és a fájlok ellenőrző összegeit, melyeket összevetnek majd a náluk levő, a Common Crawl projekt anyagából leszűrt magyar oldalak URL címeivel és checksum adataival, hogy kiderüljön, mennyi az átfedés a két állomány között.

Az elmúlt hetekben lefutott tematikus aratás

Életmód, szabadidő, hobbi (8018 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: webarchivum@oszk.hu