

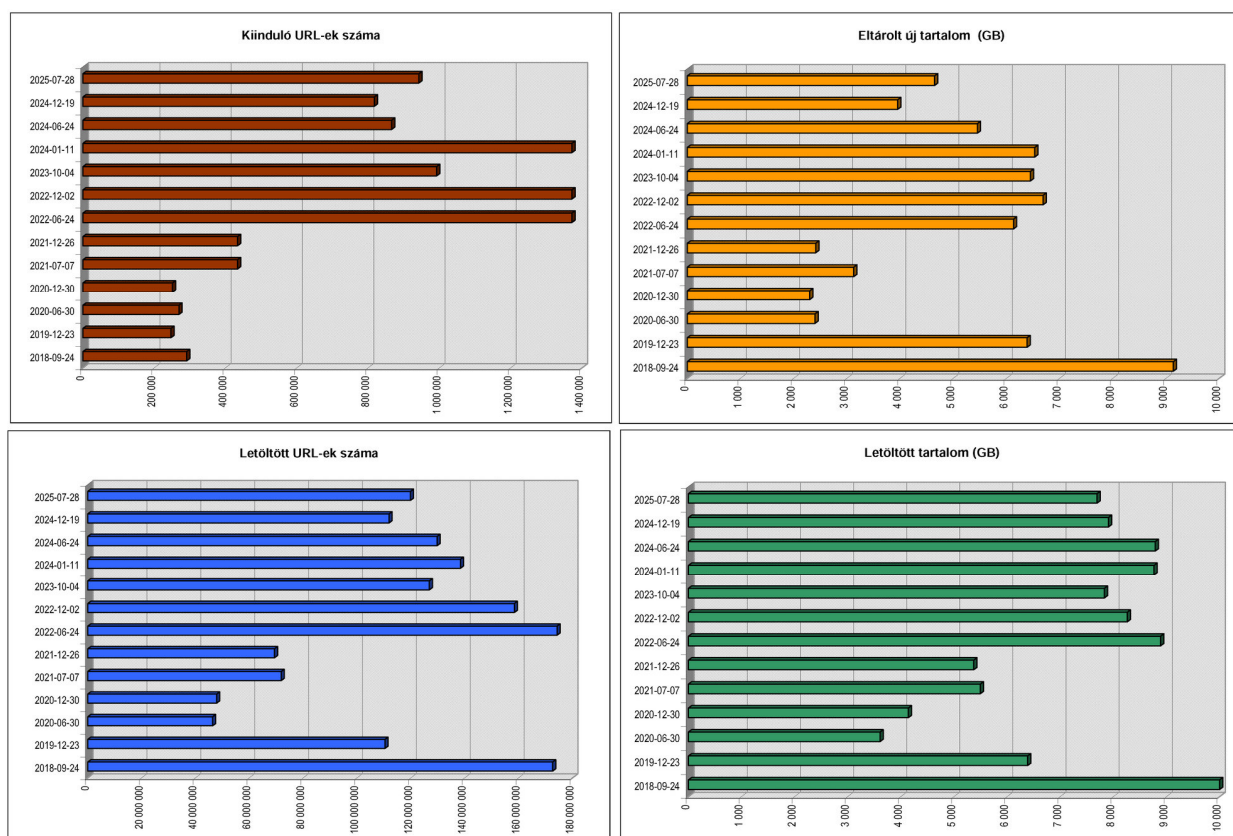
Az MNMKG OSZK Webarchívum 2025. júliusi és augusztusi hírei

Archiválás és nyilvántartás

Júliusban az idei első webtér-szintű aratáshoz készítettük elő a kiinduló címlistákat az archiváló szoftver számára. Mivel ezúttal egyesítettük az összes régebbi URL listánkat és az archívumban levő WARC fájlokból kigyűjtött .hu végű domének/aldomének címeit, a duplumok és a webszerverek által visszaadott státuszkódok alapján nem működőnek tekinthető webhelyek kiszűrése után is egy hatalmas, mintegy 3,2 millió tételes halmaz maradt. Ezt kellett az URL-ek és a kezdőlapok fejlécében talált *title* metaadatok alapján különböző kategóriákba szortírozni. Először megpróbálkoztunk az automatizált szétválogatással, de az erre a célra kísérletképpen összerakott scriptek kimenete olyan sok fals pozitív és fals negatív esetet tartalmazott, hogy egyelőre maradunk az emberi intelligenciánál.

Idő- és munkaspórolás céljából a 2024-ben már valamelyik kategóriába beosztott címeket újra ugyanabba a csoportba tettük, de azért nagyjából ezeket is átnéztük és a korábban hibásan kategorizált vagy időközben megváltozott webhelyeket átsoroltuk. Ezt követően a tavalyi nyilvántartásban nem szereplő több mint 1,5 millió URL-t kellett szétválogatni négy részhalmazba: tömegesen generált aldomének, gyűjtőkörön kívüli (pl. NSFW) oldalak, bejelentkezést igénylő szerverek vagy inaktív (pl. parkoló) domének, működőnek tűnő „valódi” webhelyek. A válogatást több dolog is nehezítette. Egyrészt a *title* metaadat elég sok esetben hiányzott vagy nem hordozott érdemi információt, esetleg nem jól lett lekérdezve és egy másik webhelyhez tartozott. Tovább bonyolította a munkát a többféle ékezetkódolás, amiket nem minden esetben sikerült automatikusan egységesíteni, valamint az a tény, hogy ezúttal a *title* lekérdezésekor megpróbáltuk az átirányításokat is követni, így sok olyan külföldi céges webhely is bekerült a címlistába, amelyeknek nincs magyar vonatkozása, csak a tulajdonosaik megvették a .hu végű doménnevet is. Hasonló a helyzet a Google által 2003-ban felvásárolt Blogger platformmal, amelynél a listánkban található kb. 155 ezer blogspot.hu végződésű URL automatikusan a blogspot.com címre irányítódik át, de ezeknek a blogoknak csak egy kis része magyar. További feladat volt az alkönyvtárakat is tartalmazó URL-ek közül törölni azokat, amelyek nem egy önálló webhelyre mutatnak, hanem egy honlap valamelyik aloldalára. A manuális ellenőrzés során kiderült az is, hogy egyre több az olyan webszerver, amely – a robotok és a túlterheléses támadások elleni védekezés céljából – valamilyen hibaüzenetet ad vissza vagy egy captcha-t dob fel, ha úgy érzékeli, hogy nem egy grafikus böngészőből érkezik a kérés, és gyakori a rosszul konfigurált vagy lejárt tanúsítvánnyal rendelkező szerver is.

Az előkészítő munkafolyamat utolsó fázisa a robotok viselkedését szabályozó *robots.txt* fájl meglétének ellenőrzése volt, amit ezúttal szintén új módszerrel, a *title* lekérdezéséhez hasonlóan négyféle előtaggal (<http://> <https://> <http://www.> és <https://www.>), valamint az esetleges átirányításokat követve és a kapott választ kielemezve próbáltunk megoldani. Mint kiderült, ez az új megoldás ugyan lassította a lekérdezést, de valószínűleg pontosabb eredményt kaptunk, mert a korábbiakhoz képest most sokkal kisebb lett a *robots.txt* nélküli szerverek aránya. Ennek megfelelően módosítottuk az egyes aratások futásidejét is: a *norobots* job 5 napig ment, a *robots.txt*-vel rendelkező webhelyek *normal* nevű aratása 15 napig, a tömegesen generált aldoménekre kiterjedő *tomeges* jelzésű job pedig 2 nap után állt le. Ez utóbbit indítottuk el először (július 28-án), ezt követte a *norobots* (július 31-én) és végül a *normal* job (augusztus 1-én). Sajnos az archiváló szerverünk erőforrásai kevésnek bizonyultak a két utóbbi aratás egyidejű futtatásához, úgyhogy a *normal* jobot le kellett állítani, majd a *norobots* befejeződése és a Heritrix újraindítása után előlről elkezdeni (augusztus 6-án). Így végül augusztus 21-én fejeződtek be az aratási munkálatok, de az oldalképek elkészítése még tovább tartott, a WARC fájlok indexelése pedig jelenleg is folyik. Az első féléves webtér aratás eredményéről részletes statisztikát készítettünk, a részfeladatokat és az azokhoz használt scripteket pedig a Redmine wikijében dokumentáltuk, továbbá frissítettük a honlapunkon elérhető seed-kereső indexállományát. Összefoglalásként elmondhatjuk, hogy megérte az előkészítésbe fektetett rengeteg munka, mert a decemberihez képest csaknem 125 ezerrel több seed URL-t tudtunk megadni a Heritrix robotnak, és bár a teljes letöltött tartalom most valamivel kevesebb volt, de ebből az újnak vagy megváltozottknak minősített 694 gigabájtal több.



Az eddigi webtér aratások legfontosabb adatainak összehasonlítása

A webtér-szintű aratáshoz szükséges címlista jelentős bővítése azzal a járulékos nyereséggel is járt, hogy az újonnan talált honlapok és blogok közül a fontosabbakkal gyarapítani tudjuk a nagyobb gyakorisággal és mélységgel aratott tematikus és műfaji részgyűjteményeinket is. Ez a munka augusztusban megkezdődött és még több hónapig is eltarthat. Első lépésben azokat az eseteket néztük át, amelyek egy .hu végű doménről egy nem magyar doménre vannak átirányítva, de ott magyar nyelvű vagy vonatkozású tartalom található. Ezt követően az EGESZSEG kódjelű részgyűjteményhez kezdtünk el válogatni eddig még nem nyilvántartott orvosi, állatorvosi, természetgyógyászati, gyógyszerészeti és szociális tematikájú webhelyeket. A hónap folyamán több mint 3 ezer tétellel sikerült így bővíteni az egyedileg ellenőrzött és altémakörökbe sorolt címeket tartalmazó nyilvántartásainkat.

A szerverter leterhelő webtér aratás lezárulta után újra elindítottuk a tematikus és műfaji részgyűjtemények aratásait, melyeket gyorsított ütemben futtatunk le augusztusban és szeptemberben. Csináltunk egy új mentést a nyilvános gyűjteményünk még „élőnek” tekinthető webhelyeiről is a WCT keretrendszerrel, mivel az automatikus ütemezést a várható infrastruktúra-váltás miatt korábban leállítottuk.

Július 20-án az ArchiveWeb.page Chrome-kiegészítőt használva frissítettük a HUNOR nevű magyar úrhajós programmal és Kapu Tibor úrutazásával foglalkozó közösségi média tartalmak archívumát, augusztus 26-án pedig lezártuk az ezekkel kapcsolatos hírek Heritrix robotval való mentését is.

Sikeres tesztekert csináltunk a Browsertrix crawlerrel, amely lényegében az ArchiveWeb.page automatizált változata: egy Chrome-alapú böngészőn keresztül tölti be a weboldalakokat, végiggörgeti azokat, elindítja a videókat és rákattint a linkekre, közben pedig eltárolja a webszervertől kapott fájlokat. Először a 24.hu hírportálhoz hoztunk létre egy olyan konfigurációt, ami lementi a címlapot (és egy oldalképet is készít róla), valamint a hírfolyamból az adott napon közzétett híreket. Ősztől már napi rendszerességgel szeretnénk ezzel a módszerrel menteni 10-20 fontosabb hazai és határon túli híroldal anyagát.

Az elmúlt két hónapban 39 db részletes, XML formátumú metaadat leírás készült el és vált visszakereshetővé a nyilvános webarchívum felületén.

Egyéb ügyek

A Könyvtári Intézet honlapján október elsejei kezdéssel ismét meghirdetésre került az „Internetes tartalmak archiválása” című tanfolyamunk, amire szeptember 10-ig lehet jelentkezni.

Elkezdjük a 2025-ös „404 Not Found – Ki őrzi meg az internetet?” konferencia és workshop előadásainak tervezgetését. Az idei rendezvényre várhatóan november első felében kerül sor.

Kitöltöttük az IIPC konzorcium kérdőívét arról, hogy milyen eszközöket használunk a különböző web-tartalmak begyűjtésére, szolgáltatására, kereshetővé és kutathatóvá tételére, illetve hogy melyek a problémás és melyek a fejleszteni kívánt területek.

Július 2-án az IIPC tagok szokásos, egy órás időtartamú „Member Update” videóbeszélgetésén néhány percben mi is beszámoltunk a nálunk zajló aktuális munkálatokról, a webtér-szintű aratás előkészítéséről.

Az elmúlt hetekben lefutott aratások

Magyar webtér (942323 db seed URL)

Kormányzat, önkormányzatok, politikai és civil szervezetek (7557 db seed URL)

Szolgáltatás, kereskedelem, szállítás, közlekedés (12259 db seed URL)

Elektronikus periodikák (11119 db seed URL)

Podkasztkok (4382 db seed URL)

Média, sajtó, műsorszórás (1028 db seed URL)

Kulturális intézmények, művelődési házak, rendezvényhelyszínek (1053 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