Az MNMKK OSZK Webarchívum 2026. márciusi és áprilisi hírei

Archiválás és nyilvántartás

A tömegesen aratott tematikus és műfaji részgyűjteményekben nagy különbségek vannak a nyilvántartott seed URL-ek számában (a legkisebb 850, a legnagyobb 13.000 címből áll), amit eddig a futásidő hosszával (3, 4 vagy 5 nap) próbáltunk kompenzálni. De a 2022 elején kialakított szisztéma óta jelentősen megnőtt néhány gyűjtemény mérete, valamint azt is tapasztaltunk, hogy a nagyobbak akár egy nap alatt meghaladják a minden aratásra egységesen vonatkozó 500 GB-os mérethatárt, míg mások ugyanezt csak több nap alatt érik el, vagy pedig meg sem közelítik. Ráadásul a nagyoknál jelentős számú URL maradt várakozó állapotban, amiket nem töltött le a robot. Persze a kiinduló címek száma nem feltétlenül jellemzi, hogy az adott gyűjtemény webhelyein mekkora mennyiségű tartalom van, de azért logikus, hogy eszerint állapítsuk meg a mérethatárokat. Ezért áprilistól ökölszabályként 1000 URL-enként 100 GB méretet határoztunk meg, azzal a megkötéssel, hogy az eddigi eredmények és a gyűjtemény jellege szerint egyenként állapítjuk meg a tényleges mérethatárt, igazodva a paraméterezésre használt Kaptafa programban már meglévő értékekhez. Az egyéb aratási beállítások (mélység, futásidő stb.) nem változnak.

Az április 12-i választás utáni napokban sűrítettük az OGYVAL2026 kódnevű részgyűjteményben levő hírportálok és egyéb webhelyek aratási gyakoriságát. A közösségi média emberi munkával való archiválása is intenzívebb volt az elmúlt két hónapban, közel 200 gigabájtot mentettünk le 187 db WACZ fájlban az ArchiveWeb.Page böngészőkiegészítővel. Érdekesség, hogy a Facebook, az Instagram, az X és a TikTok mellett első alkalommal töltöttünk le Reddit csoportokat és egyedi posztokat, valamint YouTube lejátszási listákat, illetve 20 percnél rövidebb videókat. Utóbbiak száma 543 és vegyesen vannak benne magyar és idegen nyelvű tudósítások, reakciók, elemző beszélgetések stb. Az országgyűlési választásokkal és az új kormány megalakulásával kapcsolatos hírek és webkettes tartalmak archiválását május 11-ig folytatjuk. A lementett fájlokból és metaadataikból kutatható adathalmazok létrehozását is tervezzük.

Az idei „Könyvtári kihívás” pályázathoz kapcsolódó weboldalak és posztok heti gyakoriságú archiválása eredményeként márciusban és áprilisban 126 db WACZ csomag keletkezett, 5,5 gigabájt összméretben.

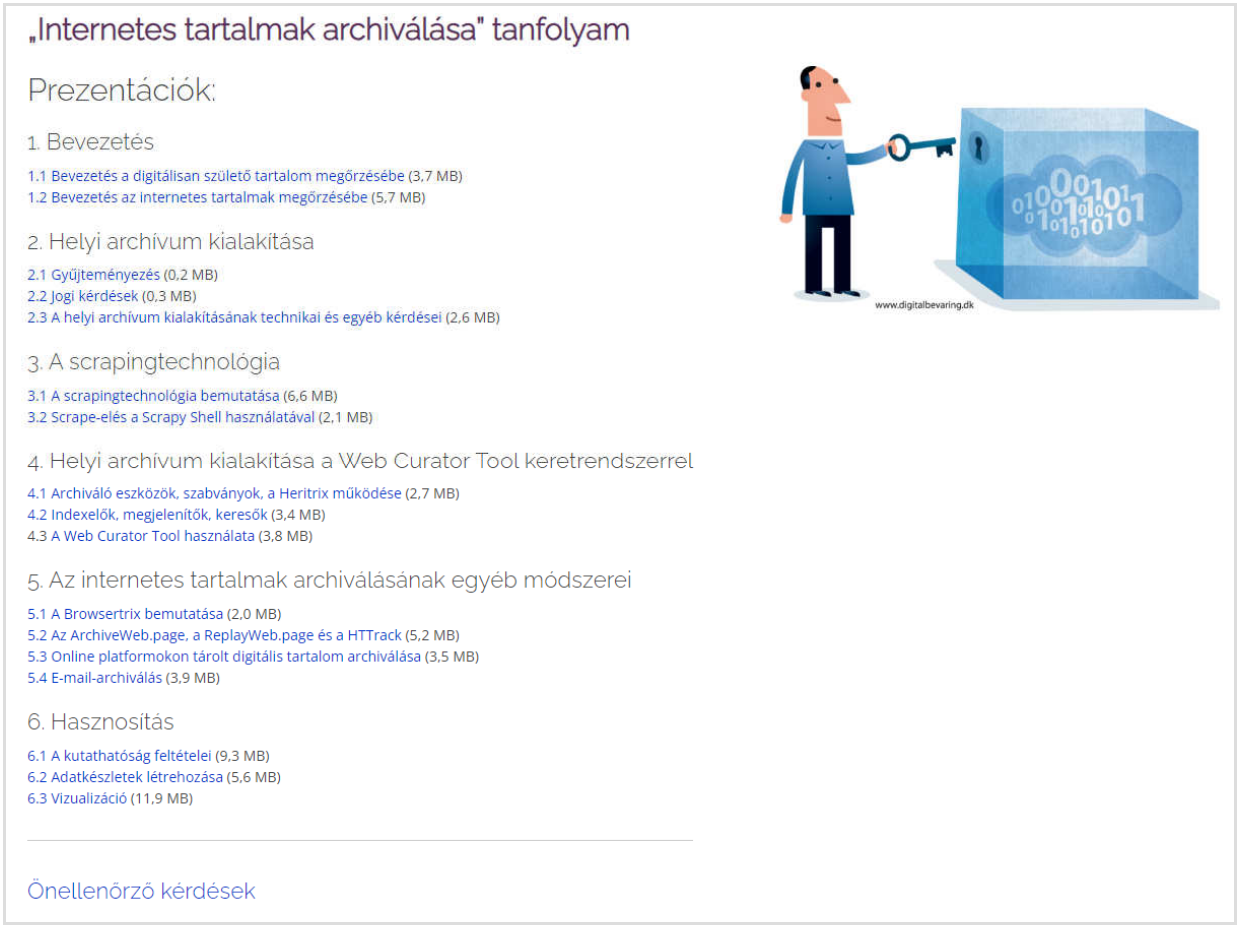
Folytatódott a tematikus részgyűjtemények bővítése a korábbi mentésekben levő linkekből kigyűjtött .hu végű URL címekkel. Egyebek mellett több mint ezer civil, vallási, sport és politikai szervezet honlapja és blogja, és mintegy 1300 iskolai és egyéb oktatási témájú webhely került besorolásra két hónap alatt.

Folyik az adatkonszolidáció is, hogy az új adatbázisba lehetőség szerint egységes és aktuális információk kerüljenek. Egyrészt átnéztük a VALLAS, MEDIA, MEZGAZ és TURIZMUS jelű részgyűjtemények címlistáit és megszűnt státuszra állítottuk az elérhetetlen, üres, vagy tartalmában jelentősen megváltozott URL-eket, továbbá megpróbáltuk kideríteni, hogy nem költöztek-e el ezek a webhelyek valamilyen más címre. Másrészt elkezdtük átmásolni egy átmeneti táblázatba a már betölthető metaadatokat az adatbázis mezőinek megfelelő formában. Eddig a KOZGYUJT és a TORTENELEM készült el és folyamatban van a nyilvános DEMO gyűjtemény adatainak előkészítése is. További feladat a duplumok kezelése, vagyis egységes névvel való ellátása azoknak a webhelyeknek, amelyeket több részgyűjteménybe is besoroltunk, illetve annak eldöntése, hogy melyik legyen az az elsődleges részgyűjtemény, amelynek keretében aratjuk őket. Ez a munka a tematikus címlisták esetében április végére befejeződött, de még hátra van a periodikák és a podkasztok átnézése, mivel ezek speciális megoldást igényelnek.

Egyéb ügyek

A régi szerver közeljövőben várható leállítása miatt átmásoltuk az OSZK webarchiválási projektjének 2017 elején indult első honlapját a mekosztaly.oszk.hu/mia/ címről a webarchivum.oszk.hu/mia-regi-honlap/ címre és átírtuk a belső linkeket. A honlap korábbi verzióinak mentései a webarchivum.oszk.hu/a-projekt-regi-honlapja/ oldalról érhetők el.

Március 24. és 27. között tartottuk meg az „Internetes tartalmak archiválása” című tanfolyamunkat, melyen 12 közgyűjteményi szakember vett részt és vizsgázott le sikeresen. A megújított tananyag mellett azért is volt érdekes ez a kurzus, mert első alkalommal tartottuk hibrid formában: 2 jelenléti és 2 online napból állt. Az egyes napok prezentációi – köztük egy részletes ismertető a WCT keretrendszerről – és az elsajátítás szintjét felmérő önellenőrző kérdések elérhetőek a honlapunkon.



Az új tananyag prezentációi és a felmérő űrlap linkje a webarchívum honlapján

A honlapunk „Előadások, prezentációk, publikációk” oldalára felkerült két PowerPoint prezentáció a fejlesztés alatt levő metaadat nyilvántartó rendszerről, melyeket Kalcsó Gyula mutatott be áprilisban a debreceni Networkshop, illetve a brüsszeli IIPC konferencián. Az első címe: „A magyar webarchívum új nyilvántartó adatbázisa”, a másodiké pedig „Storing URLs, targets, and other time-varying entities in a database as a path to sustainable recordkeeping”. Az IIPC szervezet idei rendezvénye különösen hasznosnak bizonyult számunkra és nemcsak a sok érdekes újdonságot tartalmazó előadások és poszterek miatt, hanem a szünetekben zajló kapcsolatépítésnek köszönhetően is. Ígéretet kaptunk például, hogy megkapjuk a világméretű Common Crawl projekt keretében aratott – legalább részben – magyar nyelvű webhelyek URL címeit, melyekkel kiegészíthetjük a saját nyilvántartásunkat.

Március 12-én közös webináriumot tartott az AI4LAM, az IIIF és az IIPC, vagyis az Artificial Intelligence for Libraries, Archives & Museums, az International Image Interoperability Framework és az International Internet Preservation Consortium. A másfél órás eseménynek több mint száz résztvevője volt a világ minden részéről, akik betekintést nyerhetnek a három szervezet tevékenységébe, megismerkedhettek a különböző projektekkel és beszélgethettek az együttműködési lehetőségekről.

Az elmúlt hetekben lefutott webtér és tematikus aratások

Podkasztok (4402 db seed URL)
Elektronikus periodikák (11273 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1597 db seed URL)
Szolgáltatás, kereskedelem, szállítás, közlekedés (12616 db seed URL)
Életmód, szabadidő, hobbi (9337 db seed URL)
Média, sajtó, műsorszórás (970 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (8324 db seed URL)
Kutatóintézetek, tudományos szervezetek (1387 db seed URL)
Egyetemek, főiskolák (4660 db seed URL)
Vallások, hitrendszerek, egyházak (3017 db seed URL)
Mezőgazdaság és élelmiszeripar (3074 db seed URL)
Sport, testkultúra (4451 db seed URL)
Egészségügy, szociális szféra (13964 db seed URL)
Idegenforgalom, vendéglátás (13020 db seed URL)
Képző-, előadó-, zene- és filmművészet (9778 db seed URL)
Bölcsészet- és társadalomtudományok, szakterületek (7530 db seed URL)
Párkapcsolat, család (2214 db seed URL)
Közoktatás és egyéb képzések (7809 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (945 db seed URL)
Irodalom, irodalomtudomány és -történet (1667 db seed URL)
Természet- és műszaki tudományok, szakterületek (3095 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