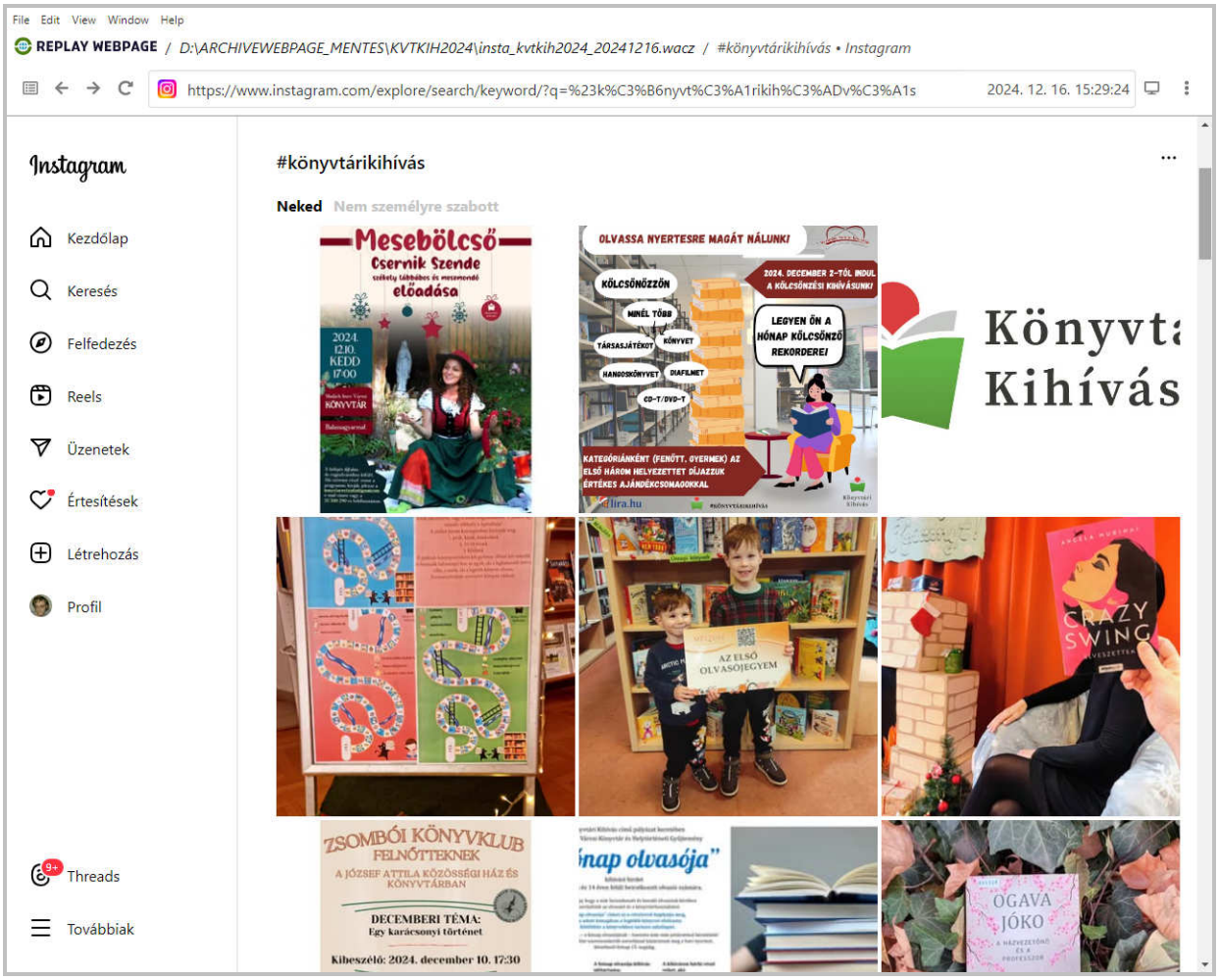
Az MNMKK OSZK Webarchívum 2024. decemberi hírei

Archiválás és nyilvántartás

December első hetében még folytatódott a podcast csatornákról a 2023 ősze óta közzétett új adások lementése, de most ez a munka az év végi webtér aratással kapcsolatos feladatok miatt egyelőre szünetel. Ez alatt a néhány nap alatt 46 csatornáról 1718 hangfájlt sikerült letölteni 99 gigabájt összméretben.

Létrehoztunk egy új esemény-alapú részgyűjteményt a Kulturális és Innovációs Minisztérium által októberben meghirdetett „Könyvtári Kihívás” című pályázattal kapcsolatos hírek és posztok összegyűjtésére. A honlapunkon is megtekinthető listában weboldalak (48 hírportál és 16 további honlap) URL címei, valamint webkettes platformokra (Facebook, Instagram, X, YouTube, TikTok) mutató keresőlinkek találhatók. Első alkalommal december 15. és 19. között készítettünk róluk mentést, az ArchiveWeb.page böngészőkiegészítőt használva. A közösségi média bejegyzésekből és videókból ekkor kb. 760-at találtunk. A teljes archívum tömörített mérete jelenleg 1,8 gigabájt.



A *könyvtárikihívás* hashtag alapján archivált Instagram posztok a ReplayWeb.page megjelenítőben

A hónap fő feladata a webtér-szintű aratás előkészítése és lebonyolítása volt, amire elég kevés időnk maradt, mivel a Kormányzati Informatikai Fejlesztési Ügynökség megszűnése miatt bizonytalanná vált, hogy meddig működnek még a KIFÜ által üzemeltetett szervereink és ezért nem akartuk, hogy az aratási jobok befejezése átcsússzon januárra. Az első feladat az archiváló robotnak kiindulásként megadott seed-

listák aktualizálása és bővítése volt. Utóbbit ezúttal úgy oldottuk meg, hogy a 2023 előtt keletkezett WARC fájlokból kigyűjtöttük a .hu végű domén és aldomén neveket, majd összevetettük őket az idei nyári webtér aratásnál használt címlistával és az abban nem szereplő webhelyeknél lekértük a státuszkódot előbb protokoll nélkül, majd http és https protokollal is. A valamilyen 2xx vagy 3xx kódot visszaadó szervereknél ezután a *title* metaadatot is lekérdeztük, amely alapján tovább próbáltuk szűkíteni a valóban „élő" és gyűjtőkörbe tartozó, eddig még nem nyilvántartott webhelyek halmazát.

Első körben 53 421 olyan működőnek tűnő szervert találtunk, amelyek nem szerepeltek a nyáron aratott címek között, de mint később kiderült, ezek között elég sok volt a valójában már ismert domén/aldomén, csak vagy másként kódolt ékezetekkel tartottuk őket nyilván, vagy .hu helyett .com végződéssel szerepeltek a Google Blogger platformja által generált *blogspot*-os aldomének közt, vagy korábban már nem gyűjtőkörbe tartozónak minősítettük őket. A valóban új és archiválásra érdemes címek kiválogatását az is nehezítette, hogy a *title* lekérdezéséhez használt *curl* parancs több ezer esetben vagy egyáltalán nem tudta ezt a metaadatot begyűjteni (leginkább az átirányítások miatt), vagy a szerveren levő tanúsítványhoz tartozó másik aldomén *title* adatát adta vissza. Ezeket az eseteket részben a *headless browser*-t futtató Screaming Frog nevű programmal, részben pedig – Chrome-ban megnyitva – emberi munkával tudtuk szétválogatni három csoportba: aratandó címek, inaktív vagy bejelentkezést igénylő webhelyek, nem gyűjtőkörbe tartozó címek. Utóbbiak közé került egyébként sok blogspot.hu végű aldomén is, melyeket külföldi bloggerek használnak (a 3438-ból 1576 darab), sajnos ezek kiszűrését sem tudjuk még automatizálni. A „természetes intelligenciával" való ellenőrzésnek viszont megvolt az az előnye, hogy a böngészőben megnyitott közel 17 ezer webhely között több mint 3 ezer olyat találtunk, amelyet a webtér aratás seed-listája mellett érdemes valamelyik tematikus vagy műfaji részgyűjteményünkbe is felvenni. (Ezeknek a téma- és altémakörökbe való besorolása elkezdődött az elmúlt napokban és várhatóan január közepére be is fejeződik.) Az eredetileg talált 53 421 új címből végül 11 677 maradt, melyekkel bővíteni tudtuk a nyáron használt fő seed-listát, és további 17 560 – webhopokhoz, cégadatbázisokhoz stb. tartozó – tömegesen generált aldomén is volt köztük, amiket szintén aratunk, csak kisebb mélységgel.

A következő lépés a nyári seed-listában szereplő címek státuszának ellenőrzése volt, hogy valóban működnek-e még. Sajnos több mint két és félszer annyi szerver szűnt meg fél év alatt, mint amennyi új címet ki tudtunk bányászni a régi WARC fájlokból: egészen pontosan 39 923 „normál", illetve 36 830 „tömeges" szerver adott vissza 0-ás vagy valamilyen 4xx/5xx számú hibakódot a december 6-án végzett lekérdezéskor. Végül azt is ellenőriztük, hogy a működő webszervereknél van-e robots.txt fájl, mert ahol nincs, ott nem kell ezt figyelembe vennie a Heritrix robotnak.

Maga az archiválás három menetben zajlott. Az első job december 19-én indult és a tömegesen generált aldoménekre terjedt ki, két szintig követve a linkeket. Ez kevesebb mint 2 nap alatt lefutott és ez alatt 91 gigabájtot mentett le a Heritrix, amelyből 30 gigabájt volt az új tartalom. A második aratás december 22-én kezdődött és a robots.txt nélküli webhelyeket célozta meg, három szint mélységig, 10 napos futásidővel. Ennél a letöltött fájlok összmérete meghaladta a 2,4 terabájtot, melyből a korábbi mentésekhez képest a duplumokat eltávolítva 1,1 terabájt maradt. A robots.txt fájllal rendelkező webszervereket arató harmadik jobot csak többszöri újrapróbálkozás után sikerült elindítani és nem is állt le magától 10 nap után. Ez volt a legnagyobb címlista és így a letöltött tartalom mennyisége is több mint 5,7 terabájt lett, amiből 3 terabájt került eltárolásra a duplumszűrés után, miközben még 45 millió már felderített, de nem letöltött URL maradt a várakozási sorban. Jelenleg az oldalképek készítése folyik, ami várhatóan még hetekig eltart, amennyiben addig nem kezdődik meg a szervereink átköltöztetése.

A decemberi webtér aratás munkafázisait és az ezekhez készült scripteket részletesen dokumentáltuk a Redmine wikijében, az újonnan nyilvántartásba vett .hu végű doménekkel és aldoménekkel pedig bővítettük a honlapunkon elérhető seed-keresőt, mellyel az URL címekben és a *title* metaadatokban lehet keresni.

Informatikai ügyek

A KIFÜ megszűnésével a webarchívum szervereinek üzemeltetését a NISZ (Nemzeti Infokommunikációs Szolgáltató) veszi majd át 2025-től. Ennek az átállásnak az előkészítéseként összeállítottunk egy igénylistát arról, hogy hogyan lehetne kiépíteni egy új, a mostaninál jobban menedzselhető és skálázható hardver infrastruktúrát és milyen szoftvereket kellene rá feltelepíteni. Az új rendszer kiépítése közben természetesen a jelenlegi munkafolyamatok és szolgáltatások közül legalább a legfontosabbakat (tömeges aratások, honlap, nyilvános archívum) továbbra is működtetni kell.

Az üzemeltetési feladatok átvételével és a metaadatok adatbázisba való áttöltésével kapcsolatban két megbeszélést is tartottunk az OSZK informatikusaival december első felében.

Ismeretterjesztés, nemzetközi kapcsolatok

A Háromká (Könyv, Könyvtár, Könyvtáros) lap 2024. évi utolsó, dupla számában megjelent Kalcsó Gyula írása a helyi vonatkozású webtartalmak archiválásáról és az együttműködés fontosságáról ezen a téren.

Az IIPC három webináriumot is tartott ebben a hónapban. Az elsőre december 5-én került sor a nemzetközi könyvtári szervezet, az IFLA hírmédiával foglalkozó szekciójával közösen. A videóbeszélgetés első felében az IIPC tevékenységét mutatták be az IFLA tagok számára, majd pedig az online hírforrások archiválásának gyakorlatáról tartottak rövid előadásokat a francia és az izlandi nemzeti könyvtárak, valamint az amerikai Kongresszusi Könyvtár munkatársai. December 9-én a szokásos „member updates" beszámolókra került sor. Elsőként az IIPC-hez nemrég csatlakozott Common Crawl Foundation mutatkozott be, majd pedig a luxemburgi nemzeti könyvtár, az oxfordi Bodleian Libraries, a Columbia Egyetemen működő Ivy Plus Libraries Confederation, valamint az osztrák nemzeti könyvtár webarchívumainak újdonságairól hallhattunk érdekes információkat. Erre az alkalomra mi is szerettünk volna bejelentkezni, hogy beszámoljunk a webtér aratás előkészítéséről, de addigra már betelt az előadói létszámkeret. Az utolsó 2024-es IIPC webinárum december 18-án zajlott le. Ezen a norvég, az ausztrál és a brit könyvtárak szakemberei ismertették a jelenlegi legkorszerűbb webarchiváló eszközzel, a Browsertrix rendszerrel szerzett eddigi tapasztalataikat és a köré épített saját fejlesztéseiket.

A belga Ghent University Media, Innovation and Contemporary Technologies kutatócsoportjának munkatársai „The evolving landscape of web and social media archiving: a comprehensive review of current practices" címmel összeállítottak egy kutatási jelentést, melyben – kérdőíves felmérések és interjúk alapján – 12 ország (köztük hazánk) aktuális gyakorlatát mutatják be a web és a közösségi média archiválásának területén. December 19-én ellenőrzés céljából megkaptuk ennek az anyagnak a kéziratát és összeírtuk a bennünket érintő részekben talált hibákat, félreértéseket.

December 15-től egy hónapig az OSZK Digitális Filológiai és Webarchiválási Osztályán és Humántudományi Bibliográfiai Koordinációs Osztályán végzi kutatómunkáját egy görög diák az Erasmus program keretében. A mi részünkről bemutatjuk neki a webarchívum munkafolyamatait és szolgáltatásait, a kutatási feladata pedig a „born-digital" tartalmak megőrzésével foglalkozó görögországi projekt összefoglalása lesz.

Az elmúlt hetekben lefutott aratások

Mezőgazdaság és élelmiszeripar (2764 db seed URL)
Sport, testkultúra (3667 db seed URL)
Egészségügy, szociális szféra (9086 db seed URL)
Idegenforgalom, vendéglátás (7883 db seed URL)
Vallások, hitrendszerek, egyházak (2986 db seed URL)
Egyetemek, főiskolák (4405 db seed URL)
Kutatóintézetek, tudományos szervezetek (1337 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