

Az MNMKK OSZK Webarchívum 2025. májusi és júniusi hírei

Archiválás és nyilvántartás

Az elmúlt hetek fő feladata a nyári webtér aratáshoz szükséges címlisták előállítása volt. Mivel ezeket elég régóta nem frissítettük, ezért most kétféle módszerrel is megpróbáltuk bővíteni az archiválásra alkalmas webhelyek URL címeit. Egyrészt összeszedtük az összes listát 2017, vagyis az OSZK-s webarchiválási projekt kezdete óta. Ezekben benne vannak a korábbi webtér aratásoknál használt címek, a tematikus, műfaji és esemény-alapú részgyűjtemények emberi munkával válogatott nyilvántartásai, valamint az internetes linkgyűjteményekből letöltött és a más forrásoktól átvett URL-ek – beleértve a korábban már egyszer megszűntnek vagy inaktívnak nyilvánított webhelyeket is, mert lehetséges, hogy azóta újraindult néhány közülük. Az adatok tisztítása után a végleges lista több mint 5,4 millió sorból állt. Másrészt egy python szkripttel kigyűjtöttük a webarchívumban található WARC fájlokból az összes olyan URL-t, ami .hu végű doménra vagy aldoménra mutat. A duplumok eltávolítása után ezek száma kb. 3 millió lett. A kétféle megközelítés eredményeként előállt listák egyesítése és ismételt duplumszűrése után 6,5 millió cím maradt. A következő lépés a webszerverek által visszaadott státuszkódok lekérdezése volt, mert csak azokat tekintjük „élőnek”, amelyek 200-as vagy – az átirányítások miatt – valamilyen 300-as kódot adnak. (Utóbbiaknál azokat az URL-eket is kigyűjtöttük, amelyekre az átirányítás mutat.) Ezt követően az így megmaradt kb. 3,2 millió webhely *title* metaadatát kérdeztük le, ami alapján ki lehet majd szűrni a gyűjtőkörbe nem tartozó és az üres (pl. parkoló) webhelyeket. További feladat lesz a tömegesen generált aldomének kiválogatása, melyeket kisebb mélységgel aratunk, mivel ezek általában csak aloldalakra mutatnak. A webtér-szintű aratás előkészítésének melléktermékeként egy olyan, több tízezer címet tartalmazó lista is előáll, amiben az eddig még nem nyilvántartott, de a *title* alapján potenciálisan fontos webhelyek vannak. A következő hónapok feladata lesz ezek emberi munkával való kiválogatása és besorolása a megfelelő részgyűjteményekbe. Az egyes részfeladatokhoz készült szkripteket és a teljes munkafolyamatot részletesen dokumentáltuk a Redmine feladatkezelő rendszer wikijében.

Június végén befejeztük a vallási és a közművelődési szervezetek Facebook fiókjainak mentését a fél-automatikusan működő ArchiveWeb.Page (AWP) programmal. Ez a munka még március közepén indult, de néhány hétig le kellett állni vele, mert a Meta cég módosította a Facebook és az Instagram webes felületének kódját és ezért az AWP használhatatlanná vált. A program fejlesztője a május 3-án kiadott verzióban ezt a problémát részben kijavította, de a legújabb vagy a kiemelt bejegyzésekben levő Facebook videóknál továbbra sem működött az AWP, ezért egy további szoftverfrissítésre volt szükség, ami sajnos még továbbra is hajlamos lefagyni a hosszabb hírfolyamok görgetése közben. Májusban és júniusban 483 fiókról készült mentés, melyek közül 322-nél sikerült a teljes hírfolyamot letölteni a legelső bejegyzésig visszamenőleg. Az archivált anyag több mint 400 gigabájt, ami a WACZ és WARC fájlok együttes mérete. A tematikus részgyűjteményekhez kapcsolódóan eddig nyilvántartásba vett 4153 Facebook fiók közül 3035-ről van már legalább részleges mentésünk. Ősszel a könyvkiadók és -kereskedők oldalainak archiválásával szeretnénk folytatni ezt a munkát.

Május végén befejeződött a „Könyvtári kihívás” program, ezért június 14-én mi is lezártuk az ezzel kapcsolatos hírek és közösségi média tartalmak heti rendszerességű mentését. 2024 decembere óta 256 egyedi weboldalt, 3140 Facebook és 1148 Instagram posztot, 41 YouTube és 86 Tiktok videót archiváltunk, szintén az AWP programmal. A webarchívumban használt WARC formátum mellett itt is megőriztük az AWP saját WACZ csomagfájljait is, így a teljes anyag közel 28 gigabájt lett.

Kezdetben heti gyakorisággal, a többször elhalasztott rajt óta pedig naponta aratjuk az Axiom Mission 4 küldetéssel és Kapu Tibor úrhajóssal, illetve általában a HUNOR Magyar Úrhajós Programmal foglalkozó híreket és egyéb webtartalmakat. A közösségi média platformokról (Facebook, Instagram, X, LinkedIn, YouTube, TikTok) eddig két AWP-s mentést csinálunk: június 9-én és 29-én, a harmadikra pedig a küldetés befejezése után kerül majd sor.

A tesztelési célokat szolgáló szerverünkre feltelepítettük a Browsertrix Crawler nevű szoftvert, amivel az AWP-hez hasonlóan egy böngészőn keresztül lehet weboldalt lementeni, de teljesen automatizált módon. Elsőként a fontosabb magyar hírportálok új tartalmainak napi szintű aratására fogjuk használni ezt az eszközt, de ehhez fel kell deríteni ezeknek az oldalaknak a szerkezetét, hogy hogyan lehet a legfrissebb híreket beazonosítani rajtuk. Ezt a feladatot egy diákmunkás bevonásával június utolsó hetében elkezdtük.

A második világháború befejezésének 80. évfordulója alkalmából az International Internet Preservation Consortium (IIPC) közzétett egy felhívást, amiben arra kérte a szervezet tagjait, hogy javasoljanak olyan weboldalt és webhelyeket, amelyeket az Archive-It szolgáltatással lementhetnek júniusban, augusztusban és szeptemberben a „World War Two 80th Anniversary Commemoration Web Archive” elnevezésű gyűjteménybe. Mi 80 tematikus és hírdalt válogattunk be ebbe a közös címlistába, amelyek között hangsúlyosan szerepelnek a nemzetközileg is ismert események – mint a magyarországi holokauszt – mellett az esetleg kevesebb figyelmet kapó, de nem kevésbé fontos mozzanatok, mint Budapest ostroma, a szovjet megszállás vagy a megtorlások.

Adatbekérő űrlap a nyilvános szolgáltatáshoz szükséges felhasználási szerződéshez

A Magyar Nemzeti Múzeum Közgyűjteményi Központ Országos Széchényi Könyvtár (a továbbiakban: MNMCK OSZK) Felhasználási szerződést köt azokkal a Jogtulajdonosokkal, akik/amelyek hozzájárulnak weboldalaik archivált változatainak az MNMCK OSZK online felületein történő publikálásához, nyilvános szolgáltatásához. A Felhasználási szerződés részletesen szabályozza a Felek jogait és kötelezettségeit a hatályos jogszabályokkal összhangban. Az űrlapon keresztül a Jogtulajdonos jelezni tudja szerződéskötési szándékát, és meg tudja adni az ahhoz szükséges adatokat is. A kitöltött űrlap alapján elkészült Felhasználási szerződést két példányban fogjuk elküldeni leendő partnerünk számára aláírásra, amelyből egy aláírt példányt kell visszaküldeniük az MNMCK OSZK számára.

Levelezési cím: Magyar Nemzeti Múzeum Közgyűjteményi Központ Országos Széchényi Könyvtár Digitális Filológiai és Webarchiválási Osztály, 1014 Budapest, Szent György tér 4-5-6.

A Felhasználási szerződés mintája ezen a linken érhető el: https://webarchivum.oszk.hu/wp-content/uploads/2025/06/webarchivalas_felhasznalasi_szerzodes_minta.pdf

Amikor elküldi ezt az űrlapot, az nem gyűjti automatikusan az Ön adatait, például a nevét és az e-mail-címét, hacsak Ön nem adja meg azt.

* Kötelező

1. Kérjük, adja meg, hogy magánszemélyként vagy jogi személyként kíván szerződést kötni! *

☐ Magánszemély

☐ Jogi személy

Következő

1 / 4 oldal

Az új felhasználási szerződéshez szükséges adatok megadására szolgáló űrlap első oldala

Egyéb ügyek

Megújul a jogvédett webhelyek archivált példányainak nyilvános szolgáltatásához szükséges felhasználási szerződés szövege és megkötésének módja. Az új rendszer szerint a jogtulajdonosnak csak néhány adatot kell megadnia egy webes űrlapon keresztül, magát szerződést az MNMCK OSZK készíti el és küldi ki aláírásra. Az új űrlap és a szerződés mintája rövidesen elérhető lesz a honlapunkon.

A május 27-i kezdéssel meghirdetett „Internetes tartalmak archiválása” című tanfolyamunkra nem gyűlt össze az indításhoz szükséges minimális létszám, így az most elmaradt. Legközelebb októberben hirdeti meg újra ezt a képzést a Könyvtári Intézet.

A tanfolyam anyagának egyik fontos modulja a Web Curator Tool keretrendszer oktatása, amihez egy részletes, magyar nyelvű prezentáció készült el, közel száz képernyőfotóval.

Május 14-én a győri Networkshop 2025. konferencián Kalcsó Gyula „A magyar webtér aratásával kapcsolatos kurátori feladatok” címmel tartott előadást. A prezentáció letölthető a honlapunkról.

Az IIPC konzorcium és az IFLA News Media Section munkacsoportja május 15-én az audiovizuális hírtartalmak archiválásáról szervezett egy online workshopot, melyen két előadás hangzott el. Az elsőt a kolumbiai Pontificia Universidad Javeriana munkatársa tartotta a Caracol Televisión archívumáról, amely az analóg hordozókon tárolt régi audiovizuális dokumentumok őrzése és digitalizálása mellett tévé- és rádióműsorok, valamint online videók, podcastok és webtartalmak gyűjtésével és szolgáltatásával is foglalkozik. A második esettanulmány az amerikai Library of Congress gyakorlatát mutatta be. A National Audio-Visual Conservation Center az általa kezelt több mint 10 millió hagyományos mozgóképfelvétel, illetve a hozzájuk kapcsolódó képi és szöveges dokumentumok mellett podcast- és online rádiócsatornákat is archivál, továbbá befogad csomagként beadott digitális videó- és hanganyagokat is. Június 17-én ennek a workshopnak a folytatásaként mexikói, kolumbiai és spanyolországi szakemberek beszéltek a spanyol nyelvű webes média archiválásáról és az archivált tartalmak kutatási célú hasznosításáról.

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