

Az MNMKK OSZK Webarchívum 2026. májusi és júniusi hírei

Archiválás és nyilvántartás

Május 12-én, az új kormány megalakulása után lezártuk a 2026. évi országgyűlési választásokkal kapcsolatos hírek, weboldalak és közösségi média tartalmak január 19. óta folyó mentését. Ennél a részgyűjteménynél első alkalommal töltöttünk le Reddit fórumokat, valamint nagy számban TikTok és 20 percnél rövidebb YouTube videókat is. Utolsó lépésként pedig csináltunk egy aratást a kormány.hu honlapról és néhány további megszűnő vagy átalakuló intézményi és média oldalról.

Befejeződött a „Könyvtári Kihívás” pályázattal foglalkozó webtartalmak heti rendszerességgel archiválása is, amit még 2025 decemberében kezdtünk el a böngészőbe beépülő ArchiveWeb.page (AWP) programmal. Június 14-ig összesen 112 sajtóhírt, 143 könyvtári és 25 kormányzati vagy önkormányzati, valamint 13 egyéb típusú weboldalt töltöttünk le. Ezen felül 3645 Facebook, 1009 Instagram és 23 X posztot, valamint 91 YouTube és 98 TikTok videót is archiváltunk.

Az Egyesült Államok július 4-én ünnepelte függetlenné válásának 250. évfordulóját és mivel az ott élő magyarság számára ez egy fontos esemény, ezért május 22-én elkezdtük összegyűjteni a hazai és külföldi online médiában, valamint a különböző szervezetek és intézmények internetes felületein megjelenő híreket, továbbá a rendezvénysorozattal foglalkozó YouTube videókat. Az AWP-vel WACZ formátumban archivált tartalom mennyisége június végére meghaladta az 5,5 gigabájtot, a kis felbontású MP4 fájlok összmérete pedig a 2,5 gigát. A gyűjteményhez egy önálló weboldalt készítettünk a honlapunkon belül (webarchivum.oszk.hu/america250), ahol az egyes WACZ csomagok a ReplayWeb.page megjelenítővel nézhetők meg (jogi okokból csak az OSZK-ban levő dedikált gépekről), de vannak rajta képernyőfotók is, továbbá linkek az Internet Archive mentéseire, illetve az eredeti oldalakra. A szöveges dokumentumokhoz egy SolrWayback-alapú teljes szövegű keresőt is beállítottunk. A gyűjtést és az archiválást még néhány hétig tovább folytatjuk.

Az America250 ünnepségsorozat magyar vonatkozásai



Ez a különgyűjtemény az Egyesült Államok 250 éves fennállása alkalmából rendezett ünnepségsorozattal kapcsolatos magyar nyelvű vagy magyar vonatkozású híreket, dokumentumokat, blogbejegyzéseket, videókat és közösségi média tartalmakat foglalja magában. A válogatás a hazai webhelyek mellett kiterjedt az amerikai magyar szervezetek oldalaira is. Az archiválás 2026 nyarán történt az ArchiveWeb.page programmal, az eredeti URL címek és a mentések pontos dátumai [ebben a táblázatban láthatók](#). A lementett fájlok a ReplayWeb.page megjelenítővel nézhetők vissza az OSZK olvasótermében erre kijelölt gépekről. A közösségi platformok esetében több bejegyzés is van egy csomagban és ha megnyitjuk valamelyiket, akkor a ReplayWeb.page bal felső ikonjával (Browse Contents) megjeleníthető oldalsávbán tudjuk böngészni a többi. Az archivált weboldalak szövegében keresni is lehet az alábbi képre kattintva, de maguk a találatok – a szerzői jogok tiszteletben tartása érdekében – szintén csak a dedikált munkaállomásokról nyithatók meg.

Keresés

SOLRWAYBACK

☐ GROUPED SEARCH (1) ☐ IMAGE SEARCH (1) ☐ URL SEARCH (1)

Keresés a weboldalak teljes szövegében

Böngészés

[Híroldalak, tévé- és rádiócsatornák](#) [Egyéb weblapok, blogok és dokumentumok](#) [Közösségi média tartalmak](#) [YouTube videók](#)

Híroldalak, tévé- és rádiócsatornák

Azonosító	Cím	Kiadvány / Webhely	OSZK mentés	Képernyőkép	IA mentés	Eredeti
USA250-100001	250 éves az USA, grandiózus fényjátékkal nyitottak	Cenweb Magazin				
USA250-100002	250 éves fennállását ünnepli az Amerikai Egyesült Államok	Vajdasági Rádió és Televízió				
USA250-100003	Amerika 250 éves, a Hennessy 1627 Nm-rel gratulál	Totalcar				
USA250-100004	Amerika 250 éves. A jubileumi év első ünnepi eseménye: CES 2026	Prim hírek				

Az „America250” évforduló alkalmából készült különgyűjtemény weboldalának részlete

Elkezdtek a tavalyi mentés óta megjelent új podcast részek egyedi hangfájlok formájában való letöltését és raktározását. Június 18. és 30. között 162 csatornáról 8122 adást archiváltunk 594 gigabájt összméretben. A 162 podcast közül 13-at első alkalommal mentettünk le. Ez a munka a következő hónapokban is folytatódni fog.

Májusban két honlappal bővült a nyilvános webarchívum: Tanítsunk Magyarorszáért (tanitsunk.hu) és Jász-Nagykun-Szolnok Megyei Könyvtárportál (zounuk.hu).

Az adatbázisba való betöltés előkészítése miatt átnéztük és egységesítettük a BALASSI, 56INT, KIFU és MIADEMO nevű részgyűjtemények táblázatait.

Szintén az adatkonszolidáció keretében újabb részgyűjtemények címlistáinak ellenőrzése történt meg. Az elmúlt hetekben a TARSTUD, KUTINT, IRODALOM, SPORT, KONYVKIAD, EGESZSEG és KORMONKOR kódjelű táblázatokban tárolt sok tízezer URL közül több ezret kellett megszüntetnek nyilvánítani, miután a böngészőben megnyitva őket vagy nem jelent meg az eredetileg nyilvántartásba vett webhely, vagy pedig jelentősen megváltozott a tartalma. A további aratásokból törölt tételek pótlása érdekében folyamatosan veszünk fel címeket a különböző tematikus részgyűjteményekbe. Májusban és júniusban közel 3 ezer webhelyet soroltunk be valamelyik nyilvántartásba.

Elkezdődött a nyári webtér-szintű aratás előkészítése és az egyes munkafázisok részletes dokumentálása a Redmine rendszerben. Ezúttal is kikerestük az elmúlt időszakban keletkezett WARC fájlokban levő, eddig még általunk nem ismert .hu végű doméneket (kb. 64 ezer), továbbá kaptunk egy 2,8 milliós listát a San Francisco-ban alapított Common Crawl nevű non-profit szervezettől, amely 2011 óta aratja a webet. Ez a lista azokat a – nemcsak .hu végű – doméneket tartalmazza, amelyeknél a Common Crawl rendszere magyar nyelvű tartalmat (is) azonosított. Az URL és az aratás éve mellett egy százalékos értéket is megadtak, ami a magyar szavak arányát mutatja a webhelyről lementett szöveges tartalommal belül. A többszöri előfordulások és az általunk már nyilvántartott címek kiszűrése után közel 48 ezer ismeretlennek tűnő webhely maradt, melyeknél először lekérdeztük a *title* adatot, majd ez alapján a gyanúsakat emberi munkával ellenőrizve besoroltuk négy csoportba: gyűjtőkörön kívüli (7762), tömegesen generált aldomén (901), már nem működő vagy bejelentkezést igénylő szerver (22297), hungarikumnak számító és nagyobb mélységben aratandó webhely (16796). A nyelvi elemzésnek köszönhetően sok olyan külföldi szerverről szolgáltatott, részben vagy teljes egészében magyar nyelvű webhellyel tudtuk bővíteni a webtér nyilvántartást, amelyekre más módon nagyon kis eséllyel találtunk volna rá. Hasonlóképpen megcsináltuk a *title* lekérdezést és a négy csoportba sorolást a WARC-okban levő weboldalakban talált .hu végű doménekre és a 2025-ös webtér aratásoknál használt címlista esetében is. Végül is szinte pontosan ugyanannyi tavaly óta megszűnt címet kellett kivenni a listából, mint amennyivel bővíteni tudtuk azt, úgyhogy júliusban 723 ezer seed URL-ről tudjuk majd elindítani a robotot. (Ebben nincsenek benne a tömegesen generált aldomének, amelyek ellenőrzése még folyik.) Az idej előkészítő munka fontos tanulsága, hogy az AI tanításához használt, webes tartalmakat begyűjtő robotok elszaporodása miatt a szolgáltatók egyre több és egyre kifinomultabb megoldásokat használnak a scraper és a crawler szoftverek távoltartására, ami nemcsak a webarchiválást nehezíti meg, hanem az olyan egyszerűnek tűnő részfeladatokat is, mint amilyen a *title* metaadat lekérdezése vagy a *robots.txt* fájl meglétének ellenőrzése.

Egyéb ügyek

Május végén megválaszoltunk egy kérdőívet, amely a törökországi Marmara University egyik PhD hallgatójától érkezett, aki a webarchiválásról írja a doktori disszertációját. A kérdések a technikai, szolgáltatási, jogi és adminisztratív területek mellett kiterjedtek az adatok integritásának biztosítására és a hosszú távú megőrzésre is.

Bekerültek az OSZK Digitális Képtárházába (DKA) a kozterkep.hu képmegosztó oldalról a 2024-ben scraping technológiával begyűjtött, köztéri alkotásokat ábrázoló fényképek és a hozzájuk tartozó – a DKA-ban használt formátumra átkonvertált és ellenőrzött – adatok. A metaadatok részletes keresőfelülete a https://dka.oszk.hu/reszkatalogus/prog/index_din.phtml?gyujtemeny=KOZTERKEP címen található és ugyancsak ezen az oldalon van a 41.156 tételt kilistázó gomb is.

Az április elején életbe lépett új SZMSZ szerint a Digitális Megőrzési és Webarchiválási Osztály részeként folyik tovább a webarchiválás, az osztály vezetője Kalcsó Gyula, a felettes szervezeti egység pedig a Digitális Tartalom- és Szolgáltatásfejlesztési Főosztály, mely a Digitális Tartalomfejlesztési Igazgatóság része. További változás, hogy a szervezeti átalakítás következményeként a DKA már nem a mi osztályunkhoz tartozik.

Május 21-én a Conifer nevű – korábban általunk is használt – felhőalapú webarchiváló rendszer fejlesztői „Conifer Twilight” címmel Zoom-on tartottak tájékoztatást arról, hogy a rövidesen megszűnő szolgáltatásból hogyan lehet lementeni és lokálisan elérhetővé tenni vagy a Browsertrix Cloudba migrálni az archivált webtartalmakat, és hogy ennek milyen előnyei és hátrányai várhatók.

A nemzetközi IIPC szervezet az áprilisi hagyományos után május 31-én egy online közgyűlést is tartott elsősorban azok számára, akik Brüsszelbe nem tudtak elmenni. Az elnökség, majd pedig az egyes tiszviselők és munkacsoport-vezetők beszámolóit után arra is jutott idő, hogy néhány tagintézmény munkatársai bemutassák az aktuális fejlesztéseiket, melyek többsége a webarchívumok jobb kutathatóságát célozza. A június 18-án tartott IIPC webinárium témája pedig egy konkrét kutatási projekt volt, ami a dán Aarhus University munkatársainak koordinálásával indult „Web Child” néven. A projektben részt vevő szakemberek azt kutatják, hogy a World Wide Web megjelenése és elterjedése hogyan befolyásolta az akkori gyermekek életét Dél-Koreában, Dániában és az Egyesült Államokban. Ehhez a korabeli sajtó és a hagyományos média áttekintése mellett az 1995 és 2005 között archivált weboldalakat is elemzik a SolrWayback rendszerrel és egyéb szoftverekkel.

Az elmúlt hetekben lefutott webtér és tematikus aratások

Szolgáltatás, kereskedelem, szállítás, közlekedés (12861 db seed URL)
Média, sajtó, műsorszórás (979 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1447 db seed URL)
Történelem, hely- és családtörténet (1618 db seed URL)
Életmód, szabadidő, hobbi (9576 db seed URL)
Podkasztkok (4407 db seed URL)
Elektronikus periodikák (11397 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (8358 db seed URL)
Kutatóintézetek, tudományos szervezetek (1135 db seed URL)
Egyetemek, főiskolák (4749 db seed URL)
Vallások, hitrendszerek, egyházak (3060 db seed URL)
Egészségügy, szociális szféra (13337 db seed URL)
Mezőgazdaság és élelmiszeripar (2866 db seed URL)
Idegenforgalom, vendéglátás (11888 db seed URL)
Sport, testkultúra (4338 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (947 db seed URL)
Képző-, előadó-, zene- és filmművészet (9942 db seed URL)
Párkapcsolat, család (2214 db seed URL)
Közoktatás és egyéb képzések (9617 db seed URL)
Bölcsészet- és társadalomtudományok, szakterületek (7561 db seed URL)
Természet- és műszaki tudományok, szakterületek (3114 db seed URL)
Irodalom, irodalomtudomány és -történet (1667 db seed URL)
Könyvtárak, levéltárak, múzeumok és galériák (2305 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