

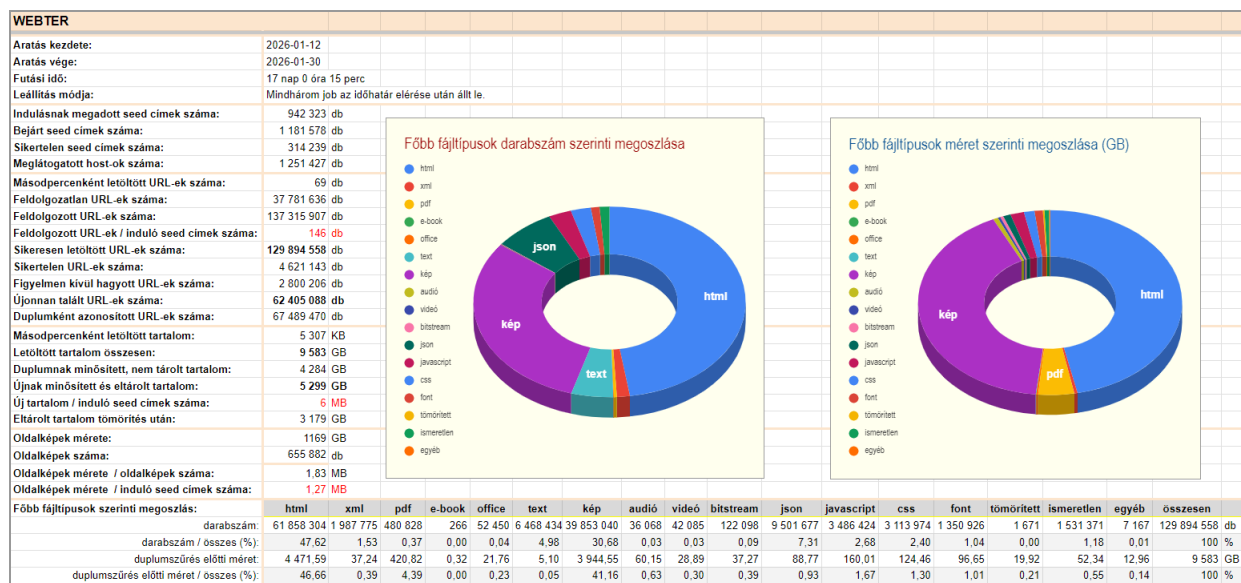
Az MNMKG OSZK Webarchívum 2026. januári és februári hírei

Archiválás és nyilvántartás

A webarchívum szervereinek és tárhelyeinek költöztetése miatt 2025 negyedik negyedévében nem tudtuk elvégezni a betervezett tömeges mentéseket, köztük a magyar webtér eddig felderített részéről készített „pillanatfelvételt” sem. Ezt most január 12. és 30. között pótoltuk oly módon, hogy a tavaly nyáron frissített és bővített címlistákról indítottuk el újra a Heritrix robotot. A három aratási job összesen 17 napig futott és a kiindulásnak megadott 942 ezer címről kiindulva csaknem 130 millió URL-t töltött le a szoftver, melyeknek közel fele volt új vagy megváltozott fájl, több mint 5 terabájt összméretben (tömörítés előtti érték). Sajnos egyre több az olyan szerver, ami akadályozza a robotokat, ezért február 3. és 18. között a böngészőt használó Browsertrix crawlerrel is csináltunk egy mentést arról a több mint 230 ezer webhelyről, amelyeknél a Heritrix *notcrawled* vagy *forbidden* státuszt írt a naplófájlba. Mivel a Browsertrix jóval lassabb és több tárhelyet fogyaszt, mert még nem tudja kiszűrni a duplumokat, ezért ezeknél csak a kezdőlapokat próbáltuk meg letölteni, ami 20 ezer cím kivételével sikerült is. A pót-aratás mérete 727 gigabájt lett.

OGYVAL2026 kódnevvel létrehoztunk egy új esemény-alapú részgyűjteményt az áprilisi parlamenti választásokkal foglalkozó hírek, webhelyek, egyedi dokumentumok és webkettes tartalmak archiválása céljából. Előbbiek automatikus aratása hetente történik, a közösségi média fiókok és bejegyzések emberi közreműködéssel történő mentése pedig minden hónap végén. A címlistában február végén 123 hírportál és 73 egyéb weboldal volt, a seed URL címek száma – a különféle címkék és a lapozólinkek miatt – pedig meghaladta a 600-at. Más esemény-alapú részgyűjteményeinkhez képest ennél jóval magasabb a közösségi média aránya, mivel a választási kampány nagy része már ott zajlik. Eddig már több mint 250 Facebook, Instagram, TikTok és X fiókról, illetve YouTube lejátszási listáról csináltunk mentéseket az ArchiveWeb.page (AWP) nevű Chrome kiegészítővel, kb. 100 gigabájt összméretben. Ezeket a tartalmakat a WARC mellett WACZ csomagokban is eltároljuk abban a reményben, hogy a megjelenítő programok a jövőben ezt a formátumot is támogatni fogják és így jó minőségben visszanezhetők és elemezhetők lesznek ezek a dinamikus generált webkettes oldalak is.

Folytattuk az aktuális „Könyvtári kihívás” pályázattal kapcsolatos webtartalmak heti gyakoriságú mentéseit, szintén az AWP-vel. Januárban és februárban 3,6 gigabájt (72 hírt és egyéb weboldalt, illetve 1354 közösségi média bejegyzést) töltöttünk le.



A 2026 januárjában lefutott webtér aratás eredményének statisztikája

Több ezer tétellel bővültek a témájuk szerint részgyűjteményekbe sorolt webhelyek címlistái az év első két hónapjában. Ezúttal elsősorban az egészségüggyel, a társadalomtudományokkal és a médiával foglalkozó honlapok/blogok közül válogattunk be korábban még nem nyilvántartottakat. Átnéztük és besoroltuk a WARC fájlokból kigyűjtött .hu végű URL-ekhez tartozó *title* metaadat alapján beazonosított társaságok, alapítványok, egyesületek és egyéb civil, egyházi, politikai, sport és tudományos szervezetek webhelyeit is.

A bővítés mellett most volt kapacitás az időközben eltűnt, megváltozott, vagy elköltözött szolgáltatások kiválogatására is. Elsőként a 2023 első félévében létrehozott, elég vegyes tartalmú, ELETMOD kódnevű részgyűjtemény címlistáját ellenőriztük és a mintegy 10 ezer webhelynek közel az egy tizedét kellett megszüntetnünk nyilvánítani. Még rosszabb volt az arány a kulturális intézményeket és szervezeteket 2019 májusa óta nyilvántartó MUVHAZ gyűjtemény esetében, ahol a kb. 1000 címből 185 bizonyult eltűntnek.

Február elején frissítettük az UKRAJNA2022 nevű részgyűjteményünk címlistáját, ami most kb. 160 hazai és határon túli információforrást tartalmaz, valamint az ezekhez tartozó több mint 900 címkés- vagy keresőlinket. Az orosz-ukrán konfliktussal, majd háborúval kapcsolatos híreket 2022 eleje óta heti gyakorisággal aratjuk és időről-időre felülvizsgáljuk a témával kapcsolatos kulcsszavakat, amelyeket a magyar hírportálok használnak. Most már a *béketerv*, *béketárgyalások*, *békefolyamat* stb. címkéket tartalmazó URL címek is szerepelnek a listában.

Szintén február első hetében kerültek ki a honlapunkra az utolsó XML metaadat fájlok, így most már minden archivált webhelynél van részletes leírás a nyilvános [demó gyűjtemény](#) weboldalon, ami – HTML-re konvertálva – a barna színű nyíl gombra kattintva nézhető meg. Az XML fájlokban keresni is lehet a [Webarchívum/Keresés](#) menüpont alatt elérhető egyszerű űrlap segítségével.

Egyéb ügyek

Elkészítettük a 2025. évi tömeges aratások összesített adatait (beleszámítva a 2026 januárjára átcsúszott második féléves webtér aratást is) és kitettük a honlapunk [„Alapinformációk és -adatok”](#) nevű oldalára. Az MNMKK OSZK éves jelentéséhez összeállítottuk a webarchiváló csoport tavalyi tevékenységéről szóló részletes beszámolót és kitöltöttük a Könyvtári Intézet kérdőívét az innovációkról és a jó gyakorlatokról. Megírtuk továbbá a 2026-os munkatervet, amiben a két legfontosabb feladat a metaadat nyilvántartó adatbázis létrehozása és a jelenlegi informatikai rendszer megújítása, melyekkel kapcsolatban heti rendszerességgel egyeztetünk az informatikus kollégákkal.

Elindult az adatkonszolidáció, vagyis a jelenlegi nyilvántartások egységesítése. Elsőként a részgyűjtemények metaadatait másoltuk át egy olyan munkalapra, amiből már áttölthetők az új adatbázisba. Kigyűjtöttük továbbá azokat az eseteket, amikor ugyanazt az URL címet több helyen is nyilvántartjuk és elkezdtük a webhelyek neveinek egységes formára való alakítását, valamint annak eldöntését, hogy az adott címet melyik részgyűjtemény keretében aratjuk a jövőben. A KOZGYUJT nevű nyilvántartásnál ez a munkafázis már el is készült és ha megtörténik a tematikus kategóriák és alkategóriák megfeleltetése a tezauruszból választott deskriptoroknak, akkor ez a táblázat is betölthető majd az adatbázisba. A következő lépés az adatbeviteli felület kialakítása lesz, amihez több ingyenes eszközt is szeretnénk kipróbálni.

Változás történt a szolgáltatási szerződéseknél, mert a korábbi gyakorlattól eltérően ezentúl nem egy Word vagy PDF dokumentumot kell szerkeszteniük azoknak a természetes vagy jogi személyeknek, akik szeretnék, hogy a webhelyükről készült mentések elérhetőek legyenek a nyilvános webarchívumban is, hanem elég csak egy online adatlapot kitölteni és ezután az MNMKK OSZK elkészíti, majd két példányban postázza a felhasználási szerződést, melyek közül – aláírás után – az egyiket vissza kell juttatni. Az űrlap és egy szerződésminta a [„Tartalomgazdának”](#) menüpont alatt érhető el a honlapunkon.

A 2025. november 12-én megrendezett „404 Not Found” konferencia és workshop videofelvételei felkerültek a [rendezvény weboldalára](#) az előadások rövid kivonatai és prezentációi mellé.

A lengyelországi Maria Curie-Skłodowska Egyetem egyik oktatója a webarchiváló projektekről folytat kutatást és az ezzel kapcsolatban feltett kérdéseire mi is válaszoltunk, melyek az OSZK webarchívum méretére, az aratások gyakoriságára és mélységére, a használt technológiára, a metadatokra stb. vonatkoztak.

Mindkét hónapban egy-egy érdekes IIPC webinariumon vettünk részt hallgatóként. Január 14-én a Browsertrix szoftver *Behaviors* nevű funkcióját mutatta be annak fejlesztője. A Browsertrix eleve tartalmaz beépített (pl. görgetés, videók elindítása) és webhely-specifikus (pl. Facebook, TikTok) viselkedéseket, melyeket a program használója tovább bővíthet. Ezek olyan Javascriptek, amikkel utánozni lehet az emberi interakciókat, például a *További hírek betöltése* gombra való kattintgatást. A másik webinariumot közösen szervezte az International Internet Preservation Consortium és a Digital Preservation Coalition február 19-án, melyen a Library of Congress 25 éve kezdődött webarchiválási projektjét mutatta be a könyvtár három munkatársa. A LoC webarchívuma jelenleg 204 részgyűjteményből áll és a teljes mérete 5,8 petabájt.

Az elmúlt hetekben lefutott webtér és tematikus aratások

Webtér szintű aratás (942323 db seed URL)

Történelem, hely- és családtörténet (1613 db seed URL)

Könyvtárak, levéltárak, múzeumok és galériák (2310 db seed URL)

Könyv- és egyéb kiadók, kereskedők (1598 db seed URL)

Média, sajtó, műsorszórás (1202 db seed URL)

Életmód, szabadidő, hobbi (9944 db seed URL)

Szolgáltatás, kereskedelem, szállítás, közlekedés (12593 db seed URL)

Kormányzat, önkormányzatok, politikai és civil szervezetek (7794 db seed URL)

Podcastok (4402 db seed URL)

Elektronikus periodikák (11250 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