

Az MNMKK OSZK Webarchívum 2026. júliusi és augusztusi hírei

Archiválás és nyilvántartás

Július 13. és augusztus 9. között lezajlott a nyári webtér aratás, melyet a korábbiakhoz hasonlóan most is három részletben, három önálló Heritrix job-ként futtatunk le, utána pedig még egy kiegészítő mentést is csináltunk a Browsertrix crawlerrel azoknak a webhelyeknek a kezdőlapjáról, amelyek esetében a Heritrix robotja valamilyen okból nem tudott elindulni. Az aratást hosszas előkészítő munka előzte meg, mely során egyrészt bővítettük a korábban használt seed-listákat a Common Crawl projektől kapott (legalább részben) magyar nyelvűként azonosított webcímekkel, valamint az elmúlt hónapokban archivált HTML fájlokban talált .hu végű linkekkel. Másrészt töröltünk minden olyan URL-t a kiinduló címek közül, amelyeket a webszerver által visszaadott státuszkód és/vagy a kezdőoldal fejlécében levő *title* adat alapján inaktívnak vagy gyűjtőkörön kívülieknek minősítettünk. Az előző években az előkészítés része volt a *robots.txt* fájl meglétének ellenőrzése is, mivel ennek hiányában más paraméterrel kell a Heritrixet futtatni, de most kiderült, hogy ezt már nem lehet minden esetben egyértelműen eldönteni, mert az AI feltanításához tartalmakat gyűjtő robotok miatt általánossá vált az ezek elleni védekezés. Szintén új tapasztalat, hogy az automatikus átirányítások gyakran nem működnek, ezért most *http* és *https* protokollal is felvettünk minden URL-t az aratáshoz használt seed-listákba.

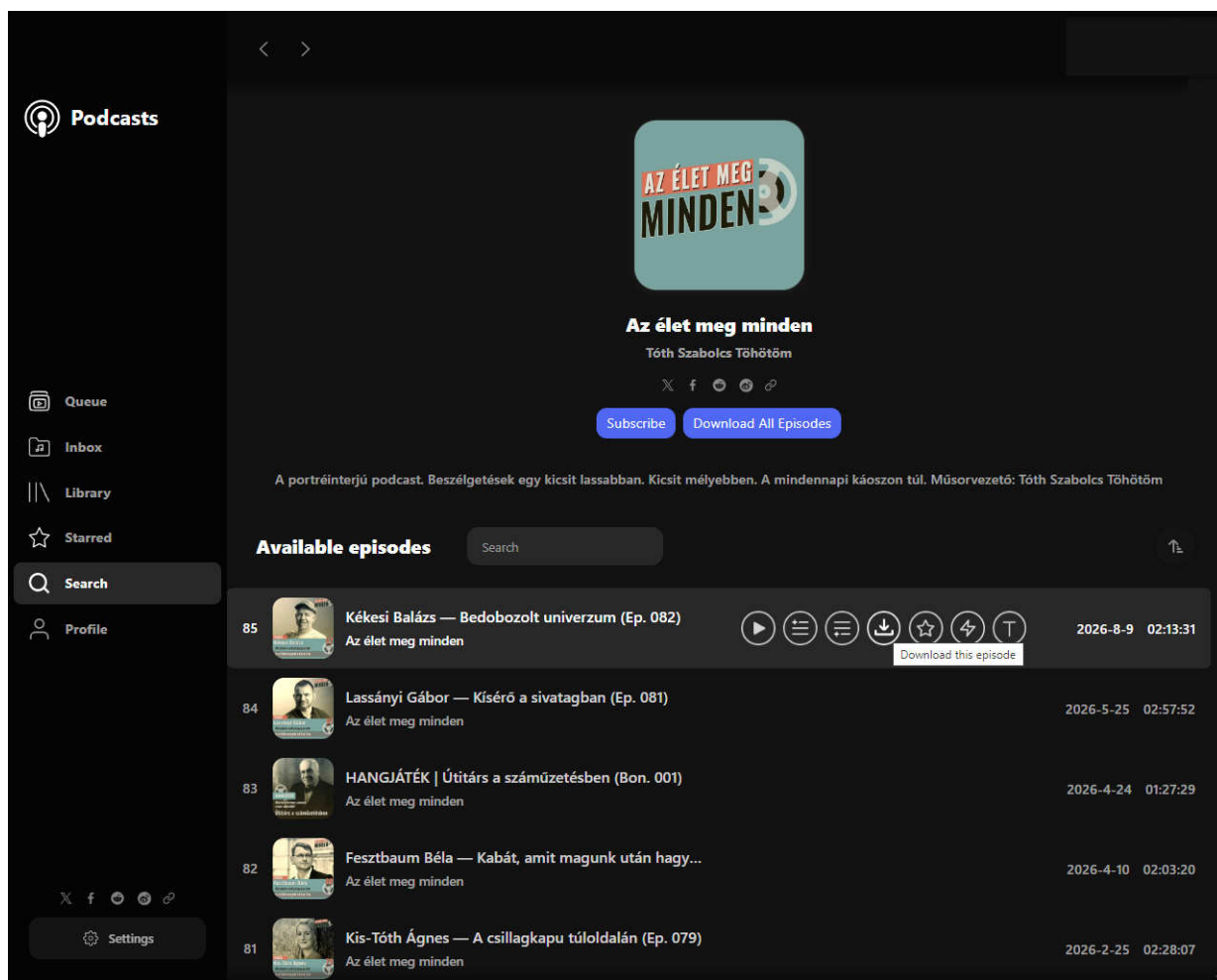
Az aktuális webtér nyilvántartás, melyben benne vannak a tömegesen generált aldomének és alternatív doménnevek is, 919.370 tételt tartalmaz, ami valamivel kevesebb, mint az egy évvel ezelőtt összeállított 942.323 URL-ből álló lista, de a 26 napnyi futásidő alatt letöltött fájlok száma (173,5 millió) és a tömörítés előtti új tartalom mérete (7,3 GB) kb. 37 százalékkal nagyobb lett a legutóbbi webtér aratáshoz képest. A részletes statisztika a honlapunkon érhető el: <https://webarchivum.oszk.hu/webter-szintu-aratasok/>

A Heritrix által készített seeds-report.txt fájlban a „notcrawled” státuszú URL-eknél egy scripttel beazonosítottuk az eredeti seed URL-eket (átirányítások és tokenek nélkül 141 ezer cím), majd ezeket átadtuk a Browsertrix böngésző-alapú robotjának, ami 135 ezer esetben tudott valamit letölteni, 572 GB összméretben. Az újonnan archivált tartalmak indexelése és az oldalképek készítése jelenleg még folyamatban van, akárcsak a WEBTER nevű gyűjtemény keresőjének a frissítése, mert ehhez előbb manuálisan ellenőrizzük és pótoljuk a hiányzó, vagy téves, vagy csak hibaüzenetet tartalmazó *title* metaadatokat.

Július 28-án lezártuk az USA250 kódnevű különgyűjtemény (webarchivum.oszk.hu/america250) további frissítését, amelyben az Egyesült Államok függetlenné válásának 250. évfordulóját ünneplő eseménysorozat magyar nyelvű vagy magyar vonatkozású híreit, közösségi média tartalmait és videóit gyűjtöttük össze. A május 22-én indult munka során 327 forrást archiváltunk, a WACZ állományok összmérete 7,7 GB lett, az MP4 fájlok pedig 3,2 GB-nyi tárhelyet foglalnak el. A WACZ formátumú csomagok a ReplayWeb.page programmal nézhetők vissza, a teljes szövegű keresést pedig a SolrWayback rendszer biztosítja.

Befejeződött a magyar podcast csatornákon az elmúlt 1-2 évben megjelent új adások letöltése a Podcasts nevű Chrome bővítmény segítségével. Mivel a podcast platformok robottal nem, vagy csak nagyon töredékesen arathatók, ezért ezt a munkát emberi közreműködéssel, böngészőn keresztül végezzük – általában a nyári „uborkaszegzonban”. A humán intelligenciára azért is szükség van, mert egyes csatornákon több műsorsorozat is fut, melyeket szét kell válogatni, illetve más forrásokból (pl. a podcast saját honlapjáról vagy a YouTube-ról) lehet pótolni az RSS feedből esetenként hiányzó részeket. Július elejétől szeptember első hetéig 433 csatornáról 21.261 hangfájl mentettünk le és raktároztunk el 1174 GB összméretben.

Folytatódott az adatkonszolidáció, vagyis a jelenleg táblázatokban tárolt metaadatok előkészítése az adatbázisba való betöltéshez. Az elmúlt időszakban az OSZK saját webhelyeinek nyilvántartása lett egységesítve.



Egy podcast csatorna legfrissebb adásainak letöltése a Podcasts kiegészítővel

Egyéb ügyek

Frissítettük az OSZK webarchívumára vonatkozó adatokat az IIPC „Tools Survey” nevű táblázatában, melyben a nemzetközi szervezet azt tartja nyilván, hogy az egyes tagintézményekben milyen szoftvereket használnak az archiváláshoz, a lementett weboldalak visszanevezéséhez és egyéb munkafolyamatokhoz.

Ugyancsak frissült az a szerződésminta, ami a <https://forms.office.com/e/fUqMLzb2a> címen elérhető űrlap oldaláról tölthető le. A webhelyek gazdái itt adhatják meg az adatokat az archivált példányok nyilvános szolgáltatásához szükséges felhasználási szerződéshez.

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