

Az MNMKK OSZK Webarchívum 2025. szeptemberi és októberi hírei

Archiválás és nyilvántartás

Szeptember folyamán és október első két hetében gyorsított ütemezéssel sikerült befejezni a harmadik negyedéves tematikus aratásokat, éppen még a szerverek és a tárhelyek költöztetésének elindulása előtt. A KIFÜ megszűnése miatt ugyanis a régi C4E (Cloud for Education) nevű felhőből mindent át kell telepíteni az új infrastruktúrára. Az öt szerverünk átvitele október végére megtörtént, de a legnagyobb, 300 terabájtos tárterület tartalmának átmásolása akár januárig is eltarthat és mivel a szervereken futó programok frissítése, majd beállítása is több hetet vesz igénybe, ezért a tömeges aratásokat egyelőre szüneteltetnünk kell és fennakadások várhatók a nyilvános, valamint a csak helyben használható szolgáltatásainkban is.

A teszt szerveren futó Browsertrix crawlerrel egyelőre még tudunk napi gyakorisággal híreket menteni néhány fontosabb portálról, és megpróbáljuk majd az ukrajnai háborúval és béketárgyalási kísérletekkel kapcsolatos hírek heti szintű mentéseit is megoldani ezen a gépen, hogy ne maradjon ki egy hosszabb időszak ennél a fontos részgyűjteménynél.

Szintén a Browsertrix segítségével (két menetben) sikerült mintegy 800 gigabájtnyi tartalmat letölteni az elmúlt hetekben azokról a webhelyekről, amelyeknél a Heritrix valamilyen okból nem tudott elindulni a nyári webtér-szintű aratás alkalmával. Bár ez a hiánypótlás idő- és tárhelyhiány miatt most csak a kezdőoldalakra terjedt ki, de így is nagyon hasznos volt olyan szempontból is, hogy a böngészőn keresztül való archiválás visszamenőleg megjavított egyes régebbi mentéseket, amelyeknél a Heritrix ugyan letöltötte a HTML oldalakat, de például a külalakot meghatározó CSS vagy a navigációhoz szükséges Javascript fájlokat nem.

A webtér aratás seed listájának bővítése során a WARC fájlokból kigyűjtött, eddig még önállóan nem nyilvántartott címek közül az elmúlt két hónapban kb. 7 ezret sikerült átnézni és közel 5 ezret be is soroltunk valamelyik tematikus válogatásba. A legnagyobb bővítés az EGESZSEG, a TARSTUD és a TURIZMUS kódjelű részgyűjtemények esetében történt. Ez a munka a következő hónapokban is folytatódik még.

A nyilvános webarchívum oldalára 90 darab részletes, XML formátumú metaadat fájl került ki ebben az időszakban a Digitális Képtárház katalógizáló könyvtárosainak jóvoltából.

Rendezvények

Az MNMKK OSZK Digitális Bölcsészeti Központja (DBK) 2025. november 12-én tartja meg az internetes tartalmak archiválásával, tágabban a digitális megőrzéssel foglalkozó éves rendezvényét, az immár kilencedik „404 Not Found - Ki őrzi meg az internetet?” című konferenciát és workshopot. Az idei rendezvény kiemelt témája a webarchívum mint adat, a tervezett előadások majd mindegyike ehhez kapcsolódik. A Skót Nemzeti Könyvtárral közösen elnyert CENL-pályázat keretében itt tartózkodó kollégák, valamint a könyvtár kutató ösztöndíjasa az általuk végzett tevékenységről és a skót webarchívum metaadatainak hasznosulásáról beszélnek. A Nemzeti Audiovizuális Archívum, valamint az MTA Könyvtár és Információs Központ a webtartalom archiválásának speciális változatait mutatják be. Az ELTE Nyelvtudományi Kutatóközpont az archivált webes tartalom magyar nyelvi modell fejlesztésében való felhasználásáról tart előadást. Végül a DBK munkatársai a webarchiválási csoport éves tevékenységéről, valamint a hírportálok napi szintű mentésének megújításáról tájékoztatják a résztvevőket. Az utóbbi témához kapcsolódik a délutáni workshop, amelynek keretében a Browsertrix crawler használatába vezetik be az érdeklődőket. A részletes program az OSZK honlapján olvasható: <https://oszk.hu/rendezvenyek/404-not-found-ki-orzi-meg-az-internetet-konferencia-es-workshop-251112>



A „webarchívum mint adat” tematikájú idei rendezvényünk meghívója

Az IIPC szervezet 2026. április 21-22-én Brüsszelben rendezi meg éves konferenciáját, melyre „Storing URLs, targets, and other time-varying entities in a database as a path to sustainable recordkeeping” címmel adtunk be egy 5 perces villámelőadás-javaslatot arról, hogy hogyan tervezzük nyilvántartani az időben változó metaadatokat, valamint egy-egy URL cím különböző verziói közötti kapcsolatokat.

Az IIPC négy webináriumot is szervezett az első két őszi hónapban. Szeptember 18-án a Browsertrix robotjának viselkedését szabályozó „behaviors” funkcióról beszéltek a szoftver fejlesztői, továbbá ennek egy konkrét alkalmazását ismerhettük meg a Stanford University Libraries webarchiváló szakemberének előadásából. Egy héttel később a szokásos „Call with Members” találkozón jelentették be, hogy a részben nyilvános, részben pedig csak a konzorcium tagjai számára elérhető, a University of North Texas digitális könyvtárban nyilvántartott videók, szoftverek, oktatási anyagok és az internetes tartalmak megőrzésével kapcsolatos egyéb dokumentumok száma elérte az 1000 tételt. Bemutakozott továbbá az IIPC-hez nemrég csatlakozott National Library of Medicine webarchívuma; a UK Government Web Archive munkatársai pedig a kutatói igények felméréséről és a szolgáltatófelület újratervezéséről számoltak be. Október 9-én a közösségi média archiválásáról zajlott egy beszélgetés holland, luxemburgi és brit kollégák részvételével. Az október 28-i webináriumon a Browsertrix-nek, mint a legújabb és legfejlettebb archiváló eszköznek a meglévő munkafolyamatokba való beépítéséről hangzott el két előadás. A luxemburgi nemzeti könyvtárban az előfizetéshez kötött hírportálok, a kormányzati oldalak és speciális webhelyek mentéséhez használják a három szerveren kialakított, Kubernetes-alapú „Browsertrix farm”-ot. A letöltött weblapok ellenőrzéséhez saját megoldást dolgoztak ki, amivel egy diákmunkás napi szinten tudja felügyelni és minőségi kategóriákba sorolni az egyes mentéseket. A dániai webarchívumnál a híroldalakon megjelenő új cikkek felderítésével és a Browsertrix crawlerrel való letöltésével kísérleteznek, továbbá az általuk használt, a Heritrix robotot futtató NetarchiveSuite keretrendszerrel való integrálásra és a WACZ csomagokból kitömörített WARC fájlok automatikus indexelésére és raktározására dolgoztak ki helyi megoldásokat.

Az elmúlt hetekben lefutott tematikus aratások

Vallások, hitrendszerek, egyházak (3500 db seed URL)
Mezőgazdaság és élelmiszeripar (2903 db seed URL)
Kutatóintézetek, tudományos szervezetek (1370 db seed URL)
Egyetemek, főiskolák (4575 db seed URL)
Sport, testkultúra (3918 db seed URL)
Képző-, előadó-, zene- és filmművészet (9581 db seed URL)
Természet- és műszaki tudományok, szakterületek (2949 db seed URL)
Bölcsészet- és társadalomtudományok, szakterületek (6542 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1586 db seed URL)
Irodalom, irodalomtudomány és -történet (1639 db seed URL)
Egészségügy, szociális szféra (11190 db seed URL)
Életmód, szabadidő, hobbi (9782 db seed URL)
Könyvtárak, levéltárak, múzeumok és galériák (2294 db seed URL)
Párkapcsolat, család (2159 db seed URL)
Történelem, hely- és családtörténet (1582 db seed URL)
Közoktatás és egyéb képzések (7579 db seed URL)
Idegenforgalom, vendéglátás (8309 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