Az MNMKK OSZK Webarchívum 2024. szeptemberi hírei

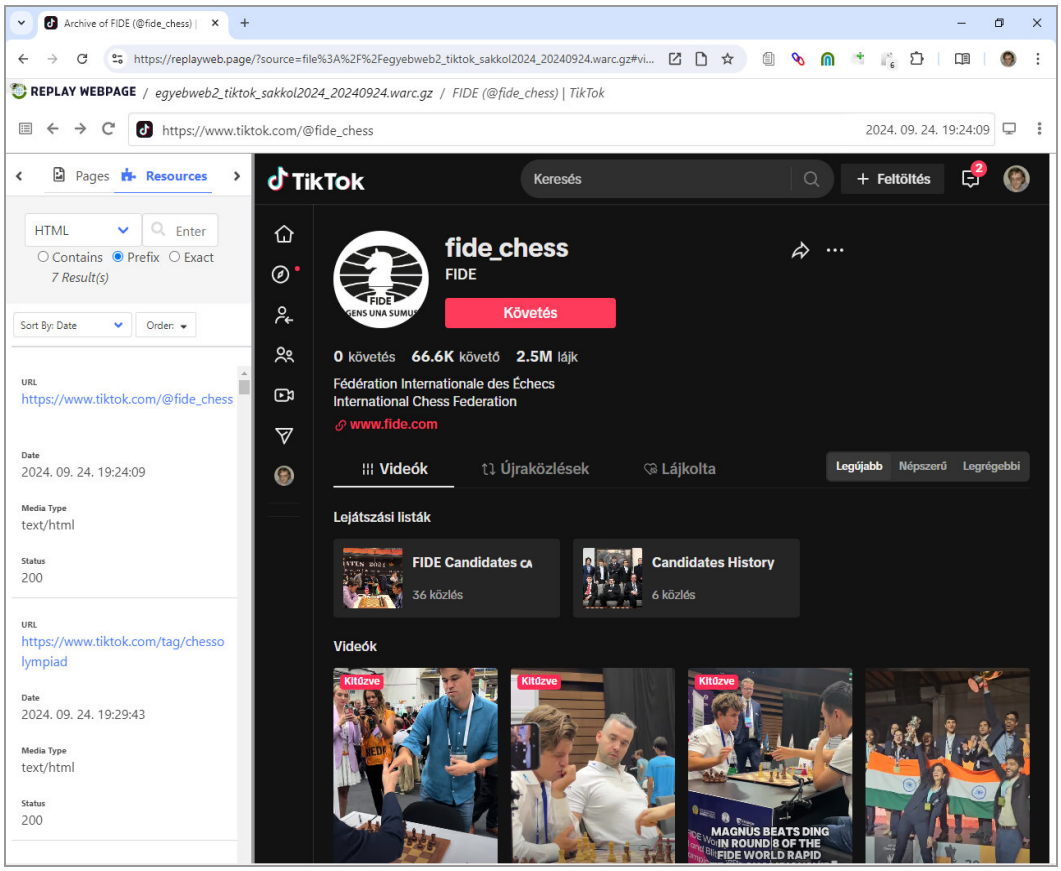
Archiválás és nyilvántartás

Az elmúlt napokban a webharvest2 szerverünkön futó Heritrix szoftver memóriahiány és egyéb hibák miatt többször is leállt, így a szeptember utolsó hetére ütemezett aratások nem, vagy csak töredékesen futottak le, ezért rövidebb a beszámoló végén levő lista a szokásosnál.

A Kormányzati Informatikai Fejlesztési Ügynökség év végi megszűnése miatt szeptember 12-én egy új aratást indítottunk arról a címlistáról, amiben a KIFÜ mellett az elődjének számító NIIF webhelyei is benne vannak, továbbá a nyilvános archívumunkban már korábban is elérhető kifu.gov.hu honlapról megpróbáltunk egy minél teljesebb mentést készíteni a WCT rendszerben.

A párizsi olimpiát követő paralimpiáról szóló híreket is learattuk a NYAROL2024 kódjelű gyűjteményünk részeként. A rendezvény zárása után, szeptember 10-én pedig a paralimpiáról megjelent magyar nyelvű közösségi média tartalmakból is megpróbáltunk minél többet letölteni az ArchiveWeb.page nevű Chrome kiegészítővel. Az Instagram, a Facebook, a Threads, az X és a YouTube szervereiről tömörítve több mint 1,8 gigabájtnyi szöveget, képet és videót sikerült összegyűjteni ezzel a módszerrel.

A szeptember 10. és 23. között Budapesten megtartott sakkolimpiával kapcsolatos hírekből is összeállítottunk egy részgyűjteményt, melyeket részben robottal (102 GB), részben emberi munkával (10 GB) mentettünk. Utóbbinál szintén az AWP böngészőkiegészítőt használtuk és a fent említett webkettes platformokon kívül a TikTokról, valamint a Nemzetközi Sakkszövetség, a FIDE honlapjáról is archiváltunk vele egyedi dokumentumokat.



A FIDE TikTok csatornájának mentése a ReplayWeb.page megjelenítőben

Az Instagramról nagyobb mennyiségű tartalmat utoljára 3-4 évvel ezelőtt töltöttünk le, akkor még a Webrecorder programot használva. Ez a munka azért maradt abba, mert az Instagram szervere egyre sűrűbben tiltotta le az archiváláshoz használt gép IP címét. Most az ArchiveWeb.page és egy véletlenszerű működésre állított billentyűmakró segítségével próbálunk minél több, elsősorban intézmények vagy szerkesztőségek által közzétett Insta-posztot archiválni. Szeptember utolsó hetében 171 darab, felsőoktatási intézményekhez köthető fiókról töltöttünk le ilyen tartalmakat, melyek közül 125-öt első alkalommal mentettünk. Mintegy 70 ezer digitális képpel és rövid videóval sikerült gyarapítani a gyűjteményünket ezzel a félautomatikus módszerrel. A válogatás és az archiválás a következő hetekben tovább folytatódik, az egyetemek és a főiskolák után a közgyűjtemények és a hírportálok Instagram oldalai következnek. Sajnos az így létrejött WARC fájlok tartalma a hagyományos megjelenítőkkel nem nézhető vissza, ehhez a ReplayWeb.page programot kell majd feltelepítenünk a szerverünkre és alternatívaként beépíteni a szolgáltató felületbe.

Szeptember 12. és 17. között a podcast gyűjteményünk bővítésének egy újabb üteme zajlott le. Ezúttal 104, korábban még nem ismert csatornát vettünk fel a nyilvántartásba és egy Podcasts nevű Chrome bővítménnyel visszamenőleg az adásaikat is lementettük (6520 db hangfájl, 271 GB összméret). Letöltöttük továbbá az elmúlt hónapokban kiadott új részeket az OSZK saját podcast csatornájáról.

Átnéztünk egy 588 tételes címlistát, ami a nyári webtér aratáshoz összegyűjtött *title* metaadatok alapján lett leválogatva a „kézműves” szóra keresve és a korábban még nem ismert webhelyeket besoroltuk a megfelelő tematikus részgyűjteménybe.

Ismeretterjesztés

Az „Internetes tartalmak archiválása” című tanfolyamhoz elkészült három újabb modul prezentációja, melyek a webarchiválás jogi kérdéseivel, a gyűjteményépítéssel, valamint a felhőszolgáltatásokban tárolt saját tartalmak exportálásával foglalkoznak. Kialakítottuk továbbá a honlapunkon a tananyag (egyelőre még nem nyilvános) weboldalát. A tanfolyamhoz is használt MIA wikit augusztus végén és szeptember elején 27 új szócikkel bővítettük és frissítettünk körülbelül egy tucat régebbit.

Elkezdtük szervezni az idei „404 Not Found - Ki őrzi meg az internetet” rendezvényünket, ami a tervek szerint november 27-én lesz megtartva. A délelőtti konferencia és a délutáni kerekasztal beszélgetés tematikája már nagyjából összeállt. Utóbbira a Magyar Nemzeti Múzeum Közgyűjteményi Központ tagintézményeinek képviselőit kértük fel, és ehhez kapcsolódóan elkezdtük összegyűjteni az MNMKK-hoz köthető webhelyek és közösségi média oldalak címeit. Ezek mentéseiből egy különgyűjteményt szeretnénk majd kialakítani.

A jövő áprilisi londoni Born-Digital Collections, Archives and Memory konferenciára „Methods of Archiving Various Web Content in the Hungarian National Library” címmel beadott előadásjavaslatunkat a szervezők sajnos nem fogadták el.

Az IIPC szintén áprilisi, Osloban tartandó konferenciájára egy poszterrel („Web Scraping in the Hungarian Web Archive”) és egy 15 perces prezentációval („Curatorial Tasks in the Top Domain Level Harvesting of a National Web Archive”) jelentkeztünk szeptember 17-én. Ezek elbírálása még nem történt meg.

A European Language Data Space és a Nyelvtudományi Kutatóközpont közös szervezésében október elsejére meghirdetett LDS Country Workshopon Kalcsó Gyula „A Magyar Nemzeti Könyvtár webarchiválási tevékenysége” címmel tart előadást.

Leadtunk egy rövid cikket a megújult Háromká folyóiratba („A helyi vonatkozású webtartalmak archiválása”), ami a regionális webarchívumok kialakításának fontosságáról és az OSZK ezzel kapcsolatos koordináló tevékenységéről szól.

Szeptember 4-én volt a legutóbbi „IIPC Updates: Call with Members” online megbeszélés, melyen ezúttal főként a konzorcium jövőjéről, a következő időszak terveiről volt szó. Az egyik részfeladat a netpreserve.org honlapon található információk aktualizálása. Ezzel kapcsolatban mi is küldtünk néhány módosítási javaslatot.

Belső tájékoztatásként összeállítottunk egy ismertetőt az OSZK webarchívumáról és egy 25 tételes listát a folyó, illetve az idén befejezett vagy a közeljövőben tervezett részprojektjeinkről.

Együttműködések

Két online megbeszélést tartottunk a HUN-REN Nyelvtudományi Kutatóközponttal, amelyek során szó esett a webarchiválással kapcsolatos együttműködésről is. Annak érdekében, hogy fel tudjuk mérni, mely webhelyekről vannak mentéseink, abban állapodtunk meg, hogy megosztjuk egymással a seed-listáinkat. Az archivált tartalom feldolgozása, valamint a nyelvimodell-építéshez történő felhasználása érdekében a kutatóközpont segítséget fog nyújtani. Ez elsősorban a szöveg kinyerését és nyelvi elemzését jelenti. Felmerült a topikmodellezés, valamint a tárgyszavazás kérdése is. A speech2text technológia fejlesztésének érdekében átadjuk a mentett podcastállományunkat is.

Egyre több kutatói igénnyel keresik meg a webarchívumot. A hónap folyamán Péter Róbert, a Szegedi Egyetem kutatója jelezte, hogy az Egészségbiztonsági Nemzeti Laboratórium szeretne webarchivált tartalmon elemzéseket végezni a vakcinákra vonatkozó attitűd témájában, amihez a korábbi hírportálos mentéseinket tudnánk a rendelkezésükre bocsátani. Az együttműködés jogi kereteit kell tisztázni (az SZTE-vel van aláírt együttműködési megállapodásunk).

Ugyancsak megkeresett minket Nagy Mihály, az ELTE doktorandusza, a Digitális Örökség Nemzeti Laboratórium munkatársa, aki már korábban jelezte, hogy az ukrajnai háborúval kapcsolatos mentéseinket szeretné használni topikmodellezésre. Az ilyen jellegű kutatói kérések jogi keretei is tisztázatlanok még.

A Digitális Örökség Nemzeti Laboratórium felhagy a webaratási tevékenységével, ezért megkerestek minket, hogy átadnák az anyagaikat. Tartottunk egy megbeszélést a részletek tisztázása érdekében, és abban maradtunk, hogy a WARC-fájljaikat és esetleg a TEI-XML-fájljaikat adják át nekünk.

Érkezett egy megkeresés az MTA által támogatott webhelyek archiválásáról is, ezzel kapcsolatban megkezdtük az egyeztetést Bilicsi Erikával, az MTA KIK Szakinformatikai Osztályának vezetőjével.

Az elmúlt hetekben lefutott aratások

Egészségügy, szociális szféra (9084 db seed URL)
Idegenforgalom, vendéglátás (7865 db seed URL)
Kutatóintézetek, tudományos szervezetek (1337 db seed URL)
Egyetemek, főiskolák (4402 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