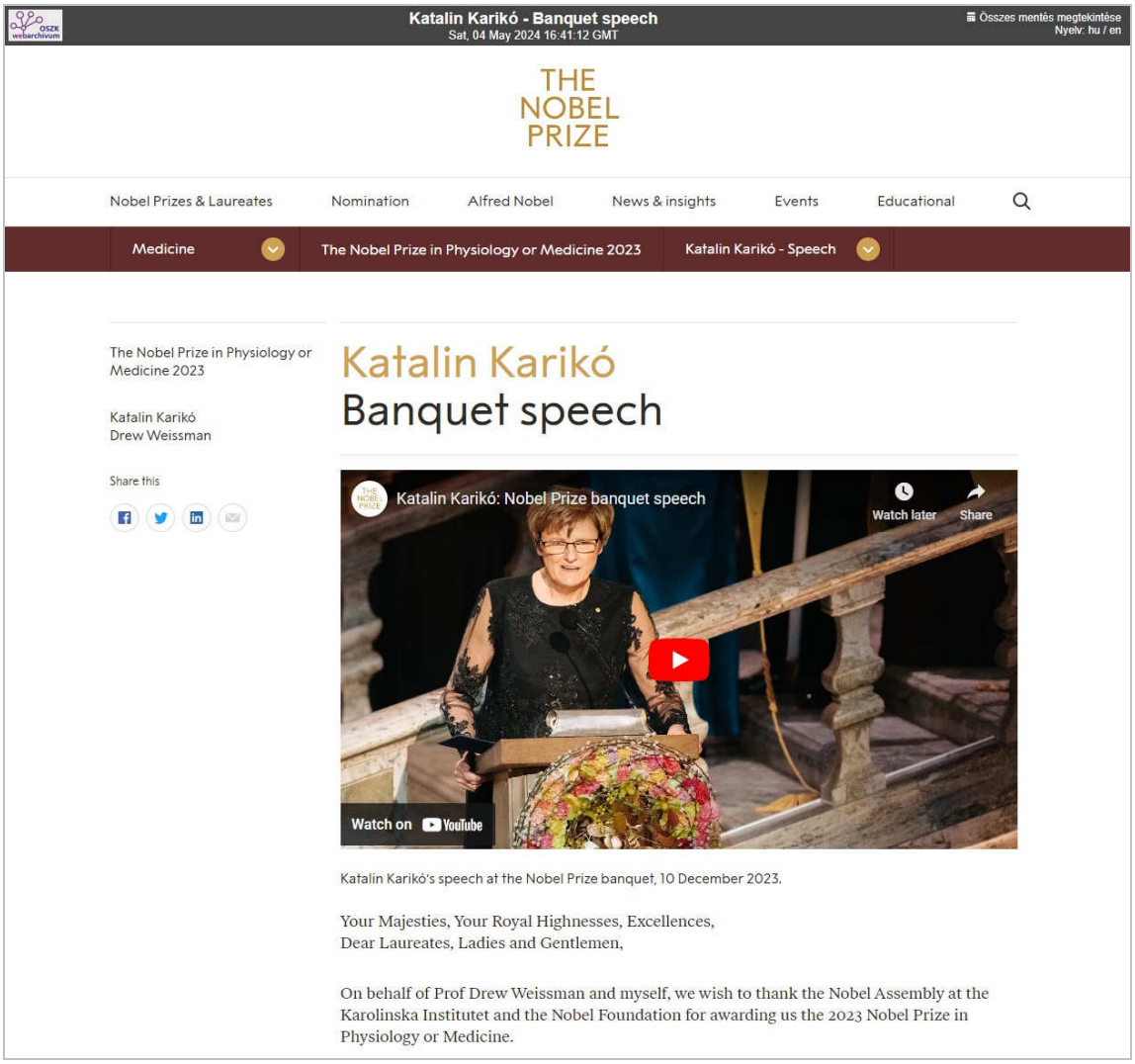
Az OSZK Webarchívum 2024. májusi hírei

Archiválás és nyilvántartás

A hónap elején jelentősen bővítettük a Karikó Katalinnal kapcsolatos cikkekből és egyéb weboldalakból álló különgyűjteményünket, melynek címlistáját közösen gondozzuk a Szegedi Egyetem Klebelsberg Könyvtárának munkatársaival. Most nagyrészt a Nobel-díj bejelentése, 2023 októbere óta megjelent hazai és külföldi híreket (kb. 3400 darabot), valamint videókat és közösségi média tartalmakat gyűjtöttük össze, majd megpróbáltuk lementeni őket a Browsertrix, illetve néhány esetben az ArchiveWeb.page programokkal. Mivel a Browsertrix Cloud felületről tömörített WACZ formátumban lehet letölteni a mentések eredményét, ami a weboldalak anyagát tároló WARC-ok mellett még további hasznos állományokat is tartalmaz (napló, index, képernyőfotó, nyers szöveg, JSON metaadat), ezért ezeket a WACZ fájlokat is elraktároztuk a jövőnek. A Karikó gyűjteménynek – a tavalyi válogatással egyesített és duplumszűrésen átesett – több mint 7800 tételes címlistája a honlapunkon elérhető és ugyancsak ott található az a felület, amivel a lementett weboldalak teljes szövegében lehet keresni. Jogi okokból csak a metaadatok jelennek meg a találatoknál, maguk az archivált oldalak az OSZK olvasótermében erre kijelölt gépeken nézhetők meg.



Karikó Katalin köszönő beszédének a Nobel-díj hivatalos honlapjáról archivált példánya a PyWb megjelenítőben

A nyilvánosan elérhető webarchívumba májusban 30 új tételt tettük ki, nagyrészt kormányzati honlapokat, melyeket a WCT-vel vezérelt Heritrix, néhány esetben pedig a HTTRack vagy a Browsertrix programokkal tudtunk elfogadható minőségben archiválni.

Új mentéseket csináltunk az OSZK saját online szolgáltatásairól, amelyek szintén megnézhetők a honlapunk Webarchívum/Részgyűjtemények menüpontja alatt. Ezúttal a Browsertrix robotjának adtunk meg 150 seed URL-t, ami csak egy ugrásnyi mélységig követte a kezdőoldalakon levő linkeket, de így is mintegy 5 ezer aloldalt töltött le 19,4 GB összméretben. A mentéseket egyenként ellenőriztük és a nagyon hibásak esetében megpróbáltunk az ArchiveWeb.page nevű böngészőkiegészítővel készíteni egy jobbat. Ugyancsak ezt a programot használtuk a nemzeti könyvtár közösségi oldalainak archiválására, de ezek visszanézése a jelenlegi megjelenítőinkkel nem lehetséges a felhasználói interaktivitásra épülő platformokon alkalmazott megoldások miatt.

Folyamatos feladat a meglévő tematikus részgyűjtemények bővítése, amit jelentősen meggyorsít, ha rátalálunk valahol egy-egy nagyobb linkgyűjteményre. Pár hete az egyik legnagyobb hazai termékkereső és -összehasonlító oldal, az arukereso.hu több mint 4 ezer tételes partnerlistáját kezdtük el átnézni és felvenni a saját nyilvántartásunkba azokat, amelyeket korábban még nem, vagy legfeljebb csak a webtér szintű aratások keretében mentettünk. Eddig a lista felét ellenőriztük és így is több mint ezer olyan webshopot találtunk, amelyek még nem voltak benne a SZOLGKER nevű gyűjteményben, amihez a szolgáltatás és közlekedés mellett a kereskedelem témája is tartozik. Az EGESZSEG és az ELETMOD is bővült egykét száz tétellel, előbbi főleg gyógyszerek és táplálékkiegészítők forgalmazóival, utóbbi pedig kertészeti és állateledel boltokkal.

A hónap utolsó napjaiban elkezdtünk a nyári webtér aratás előkészítésén dolgozni. Első lépésben kigyűjtjük az idén készült WARC fájlokban található URL címekből a *.hu* végű doméneket és aldoméneket, majd összevetjük őket a korábban használt seed listáinkkal, hogy vannak-e közöttük számunkra eddig még ismeretlenek. A következő feladat a webhelyek működésének és a *robots.txt* fájl meglétének ellenőrzése lesz, majd pedig a *title* metaadat begyűjtése következik, amire a nem releváns oldalak kiszűréséhez és a honlapunkon elérhető seed-kereső frissítéséhez van szükség.

Oktatás és ismeretterjesztés

Újraterveztük a korábban „Az internet archiválása mint közgyűjteményi feladat” címen futó 30 órás továbbképző tanfolyamunk tematikáját, mivel annak akkreditációja lejárt. A Könyvtári Intézet várhatóan szeptemberben hirdeti majd meg az új tanfolyamot, amiben a korábbinál nagyobb hangsúlyt kap a helyi webarchívumok kialakítása, a scraping és a böngészőn keresztül való archiválás, valamint a saját tartalmak kimentése a felhőalapú platformokról. A résztémakörök és az arányok kialakítása során figyelembe vettük a tavaszi kérdőíves felmérésünkre érkezett 31 választ is.

A tanfolyam egyik fontos háttérforrása a MIA Wiki, amiben jelenleg 874 szócikk található. Ezek közül 34 az elmúlt hetekben született, főként az IIPC áprilisi konferenciáján bemutatott új projektekről és szoftverekről. Tucatnyi régebbi szócikket is frissítettünk, leginkább újonnan talált forrásokat linkeltünk hozzájuk.

Kapcsolatok

Május 2-én a belga Universiteit Gent médiával, innovációval és kommunikációs technológiával foglalkozó kutatócsoportjának egyik tagjától kaptunk egy részletes kérdőívet, amivel a weboldalak és a közösségi média archiválásának és szolgáltatásának aktuális helyzetét mérik fel. Ez egy hosszú távú kutatás része, amiben már a korábbi években is részt vettünk. Ezúttal is elküldtük az OSZK jelenlegi webarchiválási gyakorlatát és a magyar jogi környezet ismertető válaszainkat, melyek kiértékelését követően várhatóan egy távinterjúra is sor kerül majd.

Folytatódtak a személyes és az online megbeszélések a székesfehérvári Vörösmarty Mihály Könyvtár munkatársaival a Fejér vármegyei webarchívumról. Ebben a hónapban főleg gyűjteményépítési és technikai kérdésekről (WCT, OutbackCDX, PyWb) volt szó.

Május 21-én Nagy Mihály, a Digitális Örökség Nemzeti Laboratórium munkatársa, az ELTE doktorandusza tett látogatást a webarchívumban, abból a célból, hogy a doktori témájával kapcsolatos kutatási tevékenységét megtervezzük. A webarchívum az ukrajnai háborús hírek részgyűjteményből kinyert szövegek és/vagy WARC-ok kutathatóvá tételével tud hozzájárulni a feladataihoz. Ehhez azonban szükséges a kutatási célú felhasználás jogi hátterének a kidolgozása.

Az elmúlt hetekben lefutott tematikus aratások

Bölcsészet- és társadalomtudományok, szakterületek (6196 db seed URL)
Könyvtárak, levéltárak, múzeumok és galériák (2155 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (987 db seed URL)
Irodalom, irodalomtudomány és -történet (1433 db seed URL)
Képző-, előadó-, zene- és filmművészet (8792 db seed URL)
Közoktatás és egyéb képzések (7255 db seed URL)
Természet- és műszaki tudományok, szakterületek (2379 db seed URL)
Párkapcsolat, család (2074 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