Az OSZK Webarchívum 2024. júniusi hírei

Archiválás és nyilvántartás

A nyári webtér-szintű aratás miatt előrehoztuk az eredetileg júliusra beütemezett tematikus aratásokat, ezért ilyen hosszú a beszámoló végén levő felsorolás. Ezek mellett a 2024-es Labdarúgó Európa Bajnoksággal kapcsolatos hírforrásokból egy esemény-alapú részgyűjteményt is kialakítottunk és elkezdtük menteni június 5-től heti rendszerességgel.

A több mint 1,37 millió URL-ből álló WEBTER címlista utoljára 2022-ben volt frissítve és ellenőrizve, ezért most elkezdtük ezt aktualizálni részben saját scriptekkel, részben ingyenes segédprogramokkal. Első lépésként kigyűjtöttük az idei mentésekben található linkeket, melyek *.hu* végződésű doménekre vagy aldoménekre mutatnak és megnéztük, hogy melyek azok, amik nincsenek benne a WEBTER listában. Összesen 14.568 jelenleg még működő címet találtunk, de ezek között volt 1888 olyan tétel is, amiket az elmúlt két év során már más módon is megtaláltunk és felvettünk valamelyik tematikus vagy műfaji részgyűjteménybe. A webszerverek által küldött *title* metaadatot lekérdezve tovább szűkítettük a WARC fájlokban talált linkek körét: szétválogattuk a tömegesen generált aldoméneket, a parkoló doméneket és a bejelentkezést igénylő oldalakat, valamint a gyűjtőkörön kívüli webhelyeket, így végül 6250 „új" cím maradt. Ezek böngészőben való egyenkénti ellenőrzését elkezdtük, mivel elég sok olyan van köztük, amelyeket érdemes valamelyik tematikus gyűjteménybe is beválogatni.

A 2022-es listában szereplő URL-eknél is lekérdeztük az egyes webszerverek státuszát, majd pedig a *title* adatát is. Ezek alapján a tömeges aldomének 30 százaléka szűnt meg vagy lett inaktív két év alatt. A többi domén és aldomén esetében is jelentős a csökkenés, a mintegy 1 millió cím 11 százaléka adott vissza valamilyen 400-as vagy 500-as hibakódot, vagyis vált véglegesen vagy ideiglenesen elérhetetlenné. A még működő címek emberi munkával való ellenőrzése folyamatban van, ezeknél várhatóan 20 százalék körül lesz a veszteség. A jellemzően webshopok, cégadatbázisok, szálláshely nyilvántartások stb. aloldalaira mutató, tömegesen generált aldomének kisebb mélységű aratása július 24. és 26. között lefutott, a robot 102 gigabájtnyi fájlt mentett le, ebből 47 GB volt az új vagy megváltozott tartalom. A nem-tömeges címek aratására júliusban kerül sor.

A hónap folyamán több mint 3 ezer tétellel bővítettük a „kézzel válogatott" és témák szerint besorolt webhelyek nyilvántartását, részben a fent említett módon, vagyis az idén archivált weboldalakból kigyűjtött linkek átnézésével, részben pedig az arukereso.hu portálon regisztrált webshopok listája alapján. Ennek a munkának köszönhetően június végén a 21 tematikus és a szintén robottal aratott 2 műfaji részgyűjtemény (periodikák és podkasztok) összmérete elérte a 100 ezer tételt. (A százezredik nyilvántartásba vett oldal a *lurdymozi.hu* volt.) Valójában ennél több címet válogattunk már össze 2017 óta, amennyiben beleszámoljuk az időközben megszűnt és ezért törölt státuszt kapott webhelyeket is, de persze a jelenleg aktívként nyilvántartott szerverek között is lehet több ezer inaktív, melyeket csak a következő címellenőrzés során fedezünk majd fel és teszünk át a töröltek közé.

Kaptunk két archiválási javaslatot is (media.harmattan.hu, illetve chess.hu), melyeket egyelőre csak részlegesen sikerült lementeni az oldalakon használt technikai megoldások miatt. A nemrég megújult „Jeles napok" honlapról (jelesnapok.oszk.hu) viszont elég jó mentést tudtunk csinálni a Heritrix robottal, ez nyilvánosan is megnézhető az OSZK-s webhelyeket tartalmazó archívumunkban.

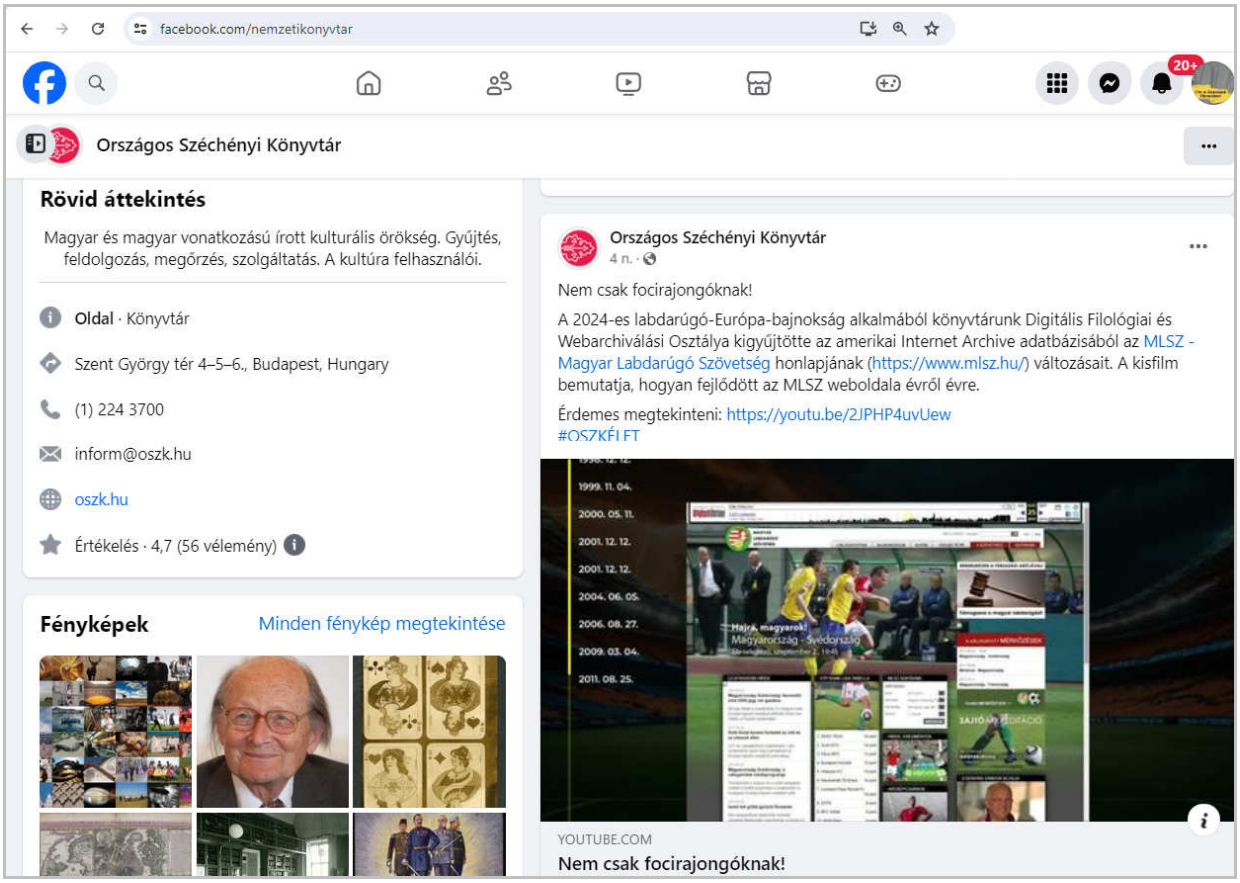
Oktatás és ismeretterjesztés

Összeállítottuk az „Internetes tartalmak archiválása" című tanfolyam tematikáját és kitöltöttük az akkreditáláshoz szükséges űrlapot. A közgyűjteményi dolgozók számára ajánlott 30 órás tanfolyam az alapozó ismeretek és a legújabb technológiák bemutatása mellett a helyi archívumok kialakításával és azok hasznosításával, valamint a saját webkettes tartalmak kimentésének lehetőségeivel is foglalkozik majd.

Beadtunk egy előadásjavaslatot a 2025 áprilisában Londonban megrendezésre kerülő Born-Digital Collections, Archives and Memory konferenciára „Methods of Archiving Various Web Content in the Hungarian National Library” címmel, mely jelenleg még elbírálás alatt van. Az előadás az OSZK gyakorlata alapján mutatja be a különféle internetes tartalmak megőrzésének módszereit.

A Foci EB alkalmából készítettünk egy PowerPoint animációt arról, hogy hogyan változott a Magyar Labdarúgó Szövetség honlapja (mlsz.hu) 1998 óta napjainkig. Az oldalképeket az Internet Archive mentései alapján készítettük. A prezentáció elérhető a honlapunkról, a videóváltozat pedig az OSZK Facebook oldalán és YouTube csatornáján nézhető meg.

Az Országos Széchényi Könyvtár július 1-től a Magyar Nemzeti Múzeum Közgyűjteményi Központ tagintézményeként működik tovább. A holdinghoz csatlakozó múzeumi kollégák számára készítettünk egy infografikát a *born digital* tartalmak típusairól és az OSZK jó gyakorlatairól ezek megőrzésében.



Az mlsz.hu honlap változásait bemutató videó az OSZK Facebook oldalán

Külföldi kapcsolatok

Rózsa Dávid, az OSZK főigazgatója június 18-án Varsóban a Conference of European National Librarians (CENL) éves közgyűlésén „Improving the Web Archiving Infrastructure of the National Széchényi Library and the Bibliothèque Nationale du Luxembourg” címmel tartott előadást a magyar és a luxemburgi nemzeti könyvtár által elnyert pályázat keretében megvalósult együttműködésről.

Június 11-én online interjú formájában válaszoltunk a belga Universiteit Gent kutatójának, valamint egyik gyakornokának kérdéseire az OSZK webarchiválási tevékenységét felmérő, májusban általunk kitöltött kérdőívvel kapcsolatosan.

Június 13-án volt az IIPC konzorcium által is támogatott Browsertrix Cloud archiváló keretrendszer minőségbiztosítási funkcióinak bemutatója egy több mint egy órás Zoom közvetítés formájában. A legfontosabb újdonság, hogy a rendszer össze tudja hasonlítani az egy archiválási menet (*crawl*) alatt elmentett weboldalakat azok eredeti, „élő" példányaival a képernyőfotók és a nyers szöveg alapján, majd jelzi az esetleges eltéréseket, hiányokat.

Az elmúlt hetekben lefutott tematikus aratások

Történelem, hely- és családtörténet (1478 db seed URL)
Média, sajtó, műsorszórás (972 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1519 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (7251 db seed URL)
Elektronikus periodikák (1079 db seed URL)
Podkasztok (4742 db seed URL)
Vallások, hitrendszerek, egyházak (2350 db seed URL)
Mezőgazdaság és élelmiszeripar (2695 db seed URL)
Egészségügy, szociális szféra (8872 db seed URL)
Sport, testkultúra (3900 db seed URL)
Kutatóintézetek, tudományos szervezetek (1280 db seed URL)
Egyetemek, főiskolák (4296 db seed URL)
Idegenforgalom, vendéglátás (7704 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