Az OSZK Webarchívum 2021 szeptemberi hírei

Archiválási munkák

A negyedéves ütemezésű tematikus aratások közül szeptemberben hat futott le, összesen 17.560 URL címről kiindulva (a téma szerinti megoszlásukat lásd a beszámoló végén). A naplófájlokból készített statisztikák alapján folyamatosan próbáljuk finomhangolni az aratórobot működését, hogy lehetőség szerint releváns tartalmakat mentsünk, arányos mértékben. A hírek megritkulása miatt leállítottuk három esemény alapú részgyűjtemény heti vagy kétheti mentését: Nemzetközi Eucharisztikus Kongresszus, Nyári Olimpia és Paralimpia, Foci EB, viszont elkezdtünk összegyűjteni és heti rendszerességgel menteni a jövő évi parlamenti választásokkal, illetve az idei ellenzéki előválasztással kapcsolatos cikkeket és egyéb weboldalakat. (https://webarchivum.oszk.hu/bongeszes-orszaggyulesi-valasztas-2022/) A tömeges aratások mellett egyedi mentéseket is csinálunk, melyeknél módunk van alaposabb minőségellenőrzésre és a hibák okának feltárása után vagy változtatunk az archiválási technológián, vagy a webhely gazdájának segítségét kérjük. Három példa az elmúlt hetekből:

- A Berzsenyi Dániel Megyei Könyvtárral (BDMK) még augusztusban történt megbeszélésen szóba került a megújult honlapjuk archiválhatóságának kérdése. Mivel az archiválást végző szoftver lementi az akadálymentes változatot is és a robottal nem kompatibilis megoldás használata miatt ez a változat jelent meg visszanézéskor, a hibát jeleztük a kollégáknak, akik ennek javítását kérték a fejlesztőktől. Kétkörös módosítás során, amiket próbamentésekkel ellenőriztünk, több apró változtatást tettek, melyek eredményeképpen problémamentesen archiválhatóvá vált a megújult honlap. Ez az összedolgozás pozitív példával szolgál arra, hogy már a fejlesztés alatt érdemes odafigyelni az archiválhatóságra, valamint arra is, hogy kis változtatásokkal is jelentős javulás érhető el az eredményben. A www.bdmk.hu első sikeres mentése a nyilvános archívum oldalán is elérhető (http://webadmin.oszk.hu/pywb/*?url=https://www.bdmk.hu/).
- Még júliusban kereste meg osztályunkat a „Chess – Sakklista" levelezőcsoport gazdája, hogy a lista megszűnése előtt archiváljuk azt. Egyeztetések (nyilvánosság kérdése, technikai lehetőségek) és próbamentések után szeptemberben került sor a végleges archiválásra. Mivel a levélmellékleteket hozzáférés híján nem tudtuk lementeni, javasoltuk a jelszavas védelem feloldását vagy az egész levelezőlista kiexportálását és átadását megőrzésre. Az együttműködés további eredménye, hogy a fórum gazdájának egyéb munkái bekerülnek a MEK és az EPA gyűjteményeibe is.
- Szintén a tartalomtulajdonos kezdeményezte az OSZK-nál három honlapjának archiválását, amit a Digitálistartalom-fejlesztési és -szolgáltatási Osztály munkatársaival közösen végzünk el. A nyugalmazott erdőmérnök által gondozott webhelyek egyike lényegében egy digitális szakkönyv, így az a MEK-ben is megjelenik majd, a könyv PowerPoint mellékleteit pedig a DKA fogja feldolgozni. Itt is vannak technikai problémák (pl. nem relatív linkek a képgalériákban) és hozzáférési akadályok (jelszóval védett aloldalak), melyek szintén a tartalom gazdájával és a szolgáltatójával való együttműködést igénylik.

Címlisták bővítése és aktualizálása

2020 szeptembere óta a Kormányzati Informatikai Fejlesztési Ügynökség a sulinet.hu domén alá regisztrált web- és mail-szervereket fokozatosan átállította az edu.hu doménre. Bár még többnyire élnek a régi sulinetes címek is és általában át vannak irányítva az újra, de már mi is elkezdtük ezt a változást átvezetni a nyilvántartásunkban. A hónap elején elvégzett ellenőrzés szerint az OKTAT részgyűjteményünkben található 786 darab sulinet.hu végződésű URL közül 697 webhelynek van alternatív címe (637 az edu.hu-ra, a többi pedig más aldoménre költözött). Mindössze 89 olyan oktatási honlap maradt, amelynél nem sikerült más elérhetőséget találni és közülük 40-nél már a sulinetes cím sem működik.

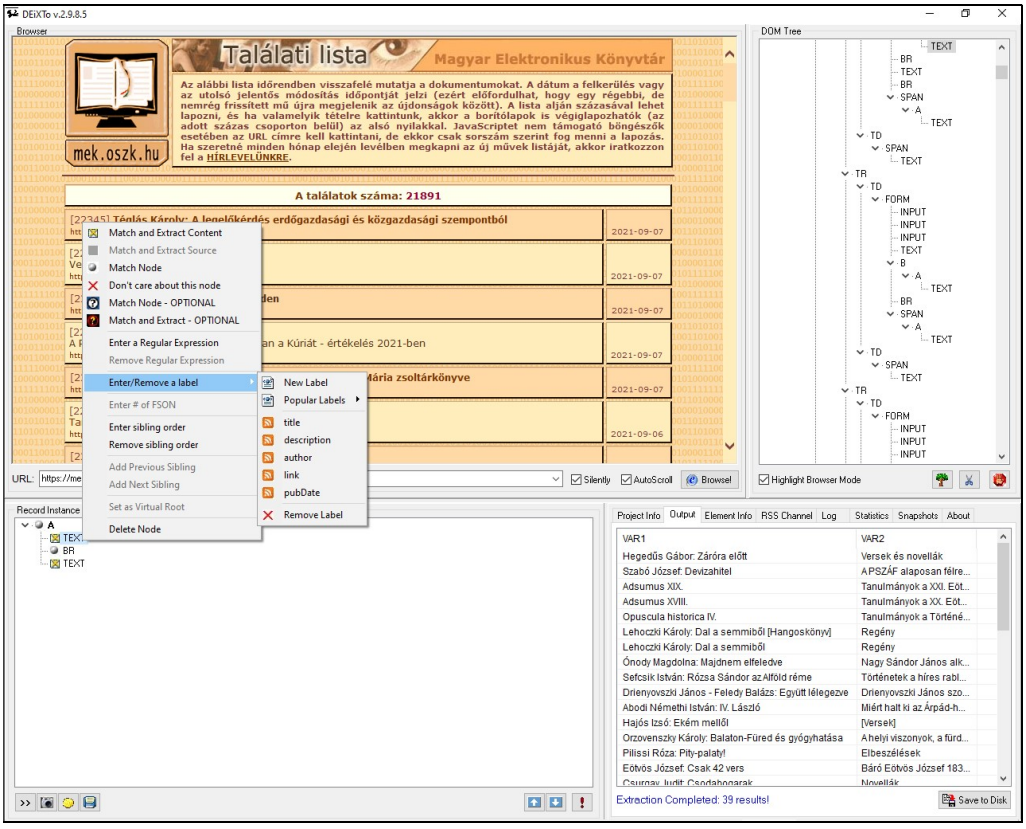
A Magyar Marketing Szövetség és az Internet Marketing Tagozat idén 20. alkalommal hirdeti meg az „Az Év Honlapja" pályázatát, melyre október 14-ig lehet nevezni a www.azevhonlapja.hu oldalon. A korábbi évek nyerteseinek listáját átnéztük és a 475 URL címből 66 olyan volt, amellyel bővíteni tudtuk valamelyik meglévő részgyűjteményünket. A többi vagy már ismert webhely volt számunkra, vagy olyan

webes applikáció, illetve technológiai megoldás, amit nem tudunk archiválni, vagy kívül esett a jelenlegi gyűjtőkörünkön, illetve már eltűnt az internetről.

Elkezdünk egy újabb tematikus részgyűjteményt építeni TERMUSZ néven, melybe természet- és műszaki tudományokkal kapcsolatos webhelyeket veszünk fel. A címlista még csak kb. 700 tételes, de folyamatosan bővítjük a következő hetekben és még az idén lefuttatjuk az első aratását. A címgyűjtés mellék-termékeként a már létező gyűjteményeink (elsősorban az oktatási és kutatási témájúak) több száz új webhellyel bővültek néhány hét alatt. Az elektronikus időszaki kiadványok nyilvántartása is főként ennek a munkának köszönhetően gyarapodott 138, eddig még ismeretlen tétellel szeptemberben.

Szoftverek és módszerek

A járvány okozta utazási korlátozások miatt egyre gyakoribb és egyre jobban megszervezett IIPC webináriumok közül kettőn is részt vettünk az elmúlt hetekben. Az első augusztus 31-én volt, ezért a múlt havi beszámolónkban még nem szerepelt. Ezen a PyWb nevű megjelentő eszköz új, 2.6-os verzióját mutatta be az amerikai programozója, valamint annak működését az izlandi nemzeti és egyetemi könyvtár munkatársa. A fő újdonságok: áttekinthetőbb naptárnézet, nyelvi lokalizálási lehetőség, hozzáférés szabályozás. A PyWb-t mi is használjuk, ezért elkezdtük tesztelni az új változatot, egyelőre a zárt archívum szerverén, és írtunk egy levelet a fejlesztőnek az általunk tapasztalt problémákról és a javasolt további funkciókról. A másik webinárium témája a szelektív aratásokhoz használt címlisták gondozása és az archivált anyag minőségbiztosítása volt. Portugál, francia, brit, amerikai, holland és új-zélandi kollégák mutatták be, hogy milyen, részben „házilag” fejlesztett eszközöket és módszereket használnak ezekhez a munkafolyamatokhoz. Számunkra a legérdekesebb a Web Curator Tool rövidesen megjelenő 3.1-es, illetve a néhány hónap múlva várható 3.2-es verziója volt. A WCT új funkciói sokat segítenek majd a lementett tartalom ellenőrzésében, a hibák vagy hiányok megtalálásában és ezek korrigálásában, illetve pótlásában. Szintén az IIPC által támogatott projekt a SolrWayback kereső- és megjelenítő-rendszer, melyből szeptember 8-án ismét kaptunk egy javított, 4.2.1-es változatot a dán fejlesztőtől, amit a nyilvános szerverünkön már be is üzemeltünk (http://webadmin.oszk.hu/solrwayback/). A kisebb, speciális feladatokra szolgáló programok tesztelése is tovább folyt, a hónap első felében ezeket a szoftvereket próbáltuk ki Ubuntu Linux vagy Windows alatt: Browsertrix Crawler, EIS Archiver, DEiXTo, ItSucks, IRobotSoft.



Könyvek szerzőjének és címeinek kigyűjtése a MEK találati listájából a DEiXTo web scraperrel.

Ismeretterjesztés és oktatás

A szoftvertesztek a MIA Wiki-hez is hasznosak, melybe a hónap folyamán újabb 29 szócikket írtunk (https://webarchivum.oszk.hu/mediawiki/). Elkészült továbbá egy 11 oldalas összefoglaló a webtartalmak archiválásáról, ami a közgyűjtemények számára módszertani útmutatóként szolgáló Fehér Könyv új kiadásában fog megjelenni. Szintén megjelenés előtt áll „A webarchiválás oktatásának nemzetközi keretei és hazai tapasztalatai" című cikkünk, ez a Könyvtári Figyelő 2021/3 számában lesz olvasható. A szlovák „ITlib – Informačné technológie a knižnice" folyóiratban angol nyelvű tanulmányunk jelent meg a Rákóczi emlékév alkalmából digitalizált vagy digitálisan született, valamint webarchiválással lementett dokumentumokból összeállított gyűjteményünkről „Rákóczi thematic digital archive at the National Széchényi Library" címmel (https://doi.org/10.52036/1335793X.2021.1-2.42-45). A Könyvtári Intézet által szervezett „Az internet archiválása mint közgyűjteményi feladat" tanfolyam végül online formában kerül megtartásra szeptember 28. és október 1. között, mert a jelenléti oktatásra nem volt elég jelentkező. Elkezdtük szervezni a szokásos éves „404 Not Found – Ki őrzi meg az internetet?" rendezvényünket, ami várhatóan november 23-25. között lesz egy vagy két napos formában, szintén Teams környezetben.

Külföldi kapcsolatok

Felvettük a kapcsolatot a pozsonyi Egyetemi Könyvtár, illetve a cseh, horvát, osztrák, szlovén nemzeti könyvtárak webarchívumaival, hogy jobban megismerjük egymást, illetve megvitassuk a szorosabb együttműködés lehetőségeit. Az idén a tavalyihoz hasonlóan online megrendezendő „404-es workshopon" egy teljes napot szánunk erre a programra. Mindegyik intézmény örömmel fogadta a kezdeményezést és szívesen részt vesznek a rendezvényen bemutatkozó előadás, illetve kerekasztal beszélgetés formájában.

A Lengyel Állami Levéltár részéről megkeresés érkezett az OSZK felé, hogy ők is webarchiválási tevékenységet terveznek és a jó gyakorlatok megvitatására november közepén Varsóban konferenciát szerveznek (a járványhelyzet kedvező alakulása esetén személyes részvétellel). Őket is meghívtuk a „404-es workshop" résztvevői közé, ahol további értékes tapasztalatokkal gazdagodhatnak.

A WARCnet projekt tevékenysége háttérmunkákkal folytatódott e hónapban tovább. November elejére terveznek a dániai Aarhusba (a lakosság rendkívül magas szintű átoltottságát kihasználva) személyes részvételen alapuló találkozót, ahol a további teendőket lehet egyeztetni.

Az IIPC-nek a tagintézmények számára rendezett online találkozóján lehetőségünk nyílt beszámolni az utóbbi hónapok főbb tevékenységeiről, illetve meghallgatni a partnerintézmények hasonló beszámolóit. A közösségi média archiválása mindenütt nagyon előtérbe került, erről külön online tapasztalatcserék szervezése várható a későbbiekben. Hasonló egyeztetésre a tagintézmények között pedig negyedévente fog sor kerülni a jövőben. Megújul a konzorcium végrehajtó bizottsága is, a jelölési folyamat lezárult, a szavazatainkat október közepéig kell leadni.

Kitöltöttünk egy kutatók, illetve könyvtárosok számára összeállított nemzetközi kérdőívet „Web Archives: Researcher Skills & Tools" címmel, melynek középpontjában a webarchívumok kutatási célú felhasználása meglévő gyakorlatainak felmérése, a illetve potenciális lehetőségek meghatározása állt.

Az Internet Archive elküldte hivatalos formában az árajánlatát, mely egy olyan portál létrehozásáról és üzemeltetéséről szól, amin keresztül OSZK-s arculati elemekkel ellátott felületen lenne elérhető a magyar nyelvű anyaguk teljesszövegű indexeléssel, részletes keresési opciókkal, API funkciókkal. Emellett megkapnánk a .hu doménre vonatkozó, általuk összeállított címlistát is. Az ajánlatot továbbítottuk az OSZK illetékes vezetői felé a megfelelő engedélyek beszerzése és a megállapodás előkészítése céljából.

Könyvtári együttműködés

Ebben a hónapban is folytatódtak az egyeztetések a megyei könyvtárakkal a partnerkapcsolatok kialakításáról, a webarchiválásban történő együttműködésről. Öt megyei könyvtár munkatársaival került sor az első megbeszélésre szeptemberben: Somogyi Károly Városi és Megyei Könyvtár (SKVMK), Balassi Bálint Megyei Könyvtár (BBMK), II. Rákóczi Ferenc Megyei és Városi Könyvtár (RFMVK) és Vörösmarty

Mihály Könyvtár (VMK); a Bródy Sándor Megyei és Városi Könyvtár (BSMVK) munkatársaival pedig folytattuk a májusban megkezdett egyeztetést. Mindannyian egyetértettek a kezdeményezés, a webes tartalmak megőrzésének érdekében kialakítandó együttműködés fontosságával és az OSZK által javasolt munkamegosztással. Első lépésként kialakítottuk a megyei munkatáblázatokat, leválogattuk és feltöltöttük a vonatkozó címeket a nyilvántartásunkból. A megbeszélések állása szerint a hónap végén 5 db véglegesített szerződést továbbíthatunk jogi fogalmazásra, négy esetben pedig már csak a visszajelzésre, jóváhagyásra várunk. Öt pozitívan reagáló könyvtárral technikai okok miatt még nem tudtuk elkezdeni a párbeszédet, reméljük hamarosan ezekre is sor kerül. A nyári megkeresésre nem reagáló könyvtárakkal a pedig újra megpróbáljuk felvenni a kapcsolatot.

Az elmúlt hetekben lefutott tematikus aratások

Kutatóintézetek, tudományos szervezetek (1.079 db seed URL)
Egyetemek, főiskolák (3.788 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (889 db seed URL)
Vallások, hitrendszerek, egyházak (2.704 db seed URL)
Sport, testkultúra (3.453 db seed URL)
Idegenforgalom, vendéglátás (5.647 db seed URL)

A tematikus 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: mia@mek.oszk.hu