Az OSZK Webarchívum 2022 szeptemberi hírei

Archiválás

A több hétig tartó nagy méretű webtér aratás és a nyári szabadságok miatt kissé felborult menetrend után szeptemberben már helyreállt a tematikus aratások ütemezése, minden hét végén lefuttattuk egy-egy részgyűjtemény negyedéves újramentését a zárt archívumba. A nyilvános gyűjteményben viszont sajnos nem frissülnek a webhelyek a Web Curator Tool rendszer leállása miatt és az időközben megjelent új WCT verzióra sem tudtunk még áttérni. Ezért csak annak a két honlapnak (Pálos Rend és Ipari örökség) a mentései kerültek szolgáltatásba a beérkezett engedélyek után, amelyek a zárt gyűjteményt kiszolgáló szerveren készültek és onnan lettek áthozva.

Befejezéséhez közeledik a magyar podkasztok hosszú távú megőrzésére tett első kísérletünk. Szeptemberben 340 csatornáról több mint 20 ezer adást töltöttünk le, így a teljes állomány már ezernél is több podkaszt kb. 63 ezer epizódját tartalmazza. Általános tapasztalat, hogy ennél a műfajnál is nehezedik az archiválhatóság és nemcsak azért, mert a szolgáltató platformok webes felületei komplexitásukban egyre jobban hasonlítanak a podkaszt lejátszó mobil alkalmazásokhoz, vagy mert egyre többen csak a YouTube-ra töltik fel az adásaikat, akkor is, ha nincs hozzájuk videofelvétel, hanem azért is, mert egyes szolgáltatók elkezdték korlátozni a hangfájlok letölthetőségét: vagy egy olyan kódolt formátumban lehet menteni őket, melyet csak a platform saját alkalmazása tud lejátszani (pl. Spotify), vagy nincsen semmilyen letöltési lehetőség (pl. Mixcloud). A klasszikus értelemben vett podkasztokhoz kötelezően hozzátartozó RSS feed is sokszor hiányzik újabban, vagy ha van is, a benne levő linkek már nem működnek, így egyéb tárhelyekről (pl. az Internet Archive-ból vagy a YouTube-ról) kell összevadászni a régebbi adásokat, ha még egyáltalán elérhetőek valahol.

A hónap folyamán pozitív visszacsatolás érkezett a munkánkról egy felhasználótól, aki először archiválásra javasolta az Internetes Szinkron Adatbázist (http://iszdb.hu), majd később a német tulajdonú, de egy magyar zeneszerzőről szóló The World of György Szabados nevű honlapot is (http://györgy-szabados.com). A jelzésből érdekes levélváltás, majd személyes kapcsolatfelvétel kerekedett, melyek során kiderült, hogy régóta figyeli a munkánkat és használja is a szolgáltatásainkat – honlapunk mellett az egy ideje már az OSZK olvasóteréből elérhető zárt archívumot is. Mi nemcsak lementettük a két oldalt lehetőségeink szerint, de ezeken keresztül be is tudtuk mutatni az archiválási problémákat valakinek, aki közreműködője az egyik webhelynek.

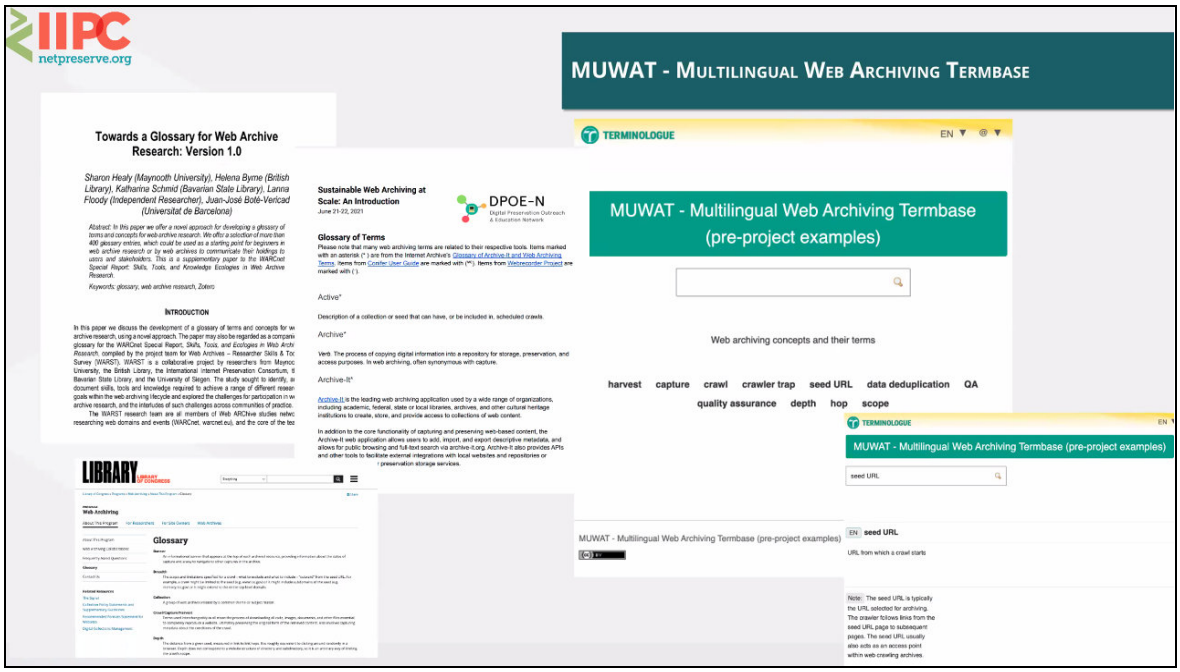
404-es workshop

Elkezdtük az idei, immár hatodik „404 Not Found – Ki őrzi meg az internetet?” konferencia és workshop előkészítését, amit szeretnénk újra jelenléti részvétellel megtartani, de az elmúlt két év tapasztalataiból tanulva biztosítani egyben az online kapcsolódás lehetőségét is. A programterv gyorsan összeállt, nehezebb diónak bizonyult viszont a megfelelő időpontot megtalálása. Figyelembe véve az előkészítésre szükséges időt és a már betervezett egyéb eseményeket, végül december 8-án kerül megtartásra a rendezvény. A konferencia a digitális megőrzés témáját, valamint az ehhez kapcsolódó hazai és nemzetközi tudományos projektek ismertetését helyezi középpontba: szó lesz az OSZK-ban zajló webarchiválási tevékenységről, a hazai könyvtárakkal formálódó együttműködésről, valamint az előállt digitális tartalom kutatási lehetőségeiről, különös tekintettel a digitális bölcsészet szakterületeivel való kapcsolódási pontokra.

Egyéb rendezvények

Az előző havi beszámolónkba már nem került bele az az augusztus 31-én tartott „Web Archiving the War in Ukraine” című webinárium, melyet az International Internet Preservation Consortium szervezett. Az első előadást az ukrajnai Lviv várostörténeti kutatóintézetének munkatársa tartotta arról, hogy hogyan mentik a háborúval foglalkozó vagy ahhoz valamilyen módon kapcsolódó témájú Telegram csoportok és hírcsatornák anyagát a rendszer saját export funkcióját kihasználva, és hogy milyen személyiségi jogi és biztonságpolitikai kérdéseket vet fel ennek az anyagnak a kutathatóvá tétele. Ez után Mark Graham, a Wayback Machine szolgáltatás vezetője számolt be arról, hogy az Internet Archive – számos szervezettel és önkéntessel együttműködve – hogyan menti egyrészt az orosz és ukrán médiát és webet, másrészt az ebben a régióban legfontosabb két közösségi platform, a VK.com és a Telegram anyagát. Előbbiről kb. 5 millió, utóbbiról 1,2 milliárd bejegyzést töltöttek le eddig, speciálisan erre a feladatra írt scriptek segítségével. Az utolsó előadás az IIPC saját projektjét ismertette, melyhez mi is küldtünk nyáron magyar webcímeket. Az Archive-It szolgáltatással három menetben, összesen max. 1 terabájt mennyiségben aratnak híroldalakat, civil és kormányzati webhelyeket, közösségi médiát és blogokat. A projekt egyik koordinátora, Kees Teszelszky a tervek szerint a mi 404-es workshopunkon is beszámol majd erről a munkáról.

Szeptemberben három újabb IIPC-s online eseményen vettünk részt. A most már rendszeresen meghirdetett „IIPC Updates: Call with Members” találkozón az amerikai Kongresszusi Könyvtár és a luxemburgi, illetve az új-zélandi nemzeti könyvtárak mellett mi is tartottunk egy pár perces beszámolót az elmúlt három hónap eredményeiről. A külföldi kollégák által használt rendszerek és munkamódszerek megismerése mellett itt értesülhettünk először a most formálódó MUWAT (Multilingual Web Archiving Termbase) projektről is, melynek célja egy többnyelvű értelmező szótár összeállítása a webarchiválással kapcsolatos szakkifejezésekből.



A MUWAT projekt előkészítő anyagai

A másik két Zoom-meeting technikai jellegű volt. Az elsőn a Browsertrix nevű, böngésző-alapú archiváló eszköz újdonságait mutatták be a fejlesztők. Ezt a rendszert mi is teszteltük korábban és szeretnénk majd használni elsősorban a hírportálok és a közösségi média mentésére. A másodikon pedig a PyWb megjelenítő szoftver következő hónapokban várható új verziónak tervezett funkcióról kaptunk tájékoztatást. Ezek közül számunkra különösen érdekes, hogy a WARC mellett támogatni

fogja a WACZ formátumot is a ReplayWeb.page modul beépítésével, így remélhetően jobban visszanézhetők lesznek vele például a Facebook oldalak mentései. Fontos újdonság még, hogy szabályozható lesz a hozzáférés szintje, tehát nem kell külön rendszerben szolgáltatni a webarchívum nyilvános és csak helyben használható részeit.

2022. szeptember 14-én a KBR, vagyis a belga nemzeti könyvtár tartott egy hibrid rendezvényt „Wanted: social media data - Archiving practices and research use" címmel, melyen szintén online vettünk részt. Ez tulajdonképpen a közösségi média archiválására fókuszáló BESOCIAL projektjük záró eseménye volt. A projekt keretében egy munkacsoport felmérte az európai országok nemzeti könyvtárainak gyakorlatát (az OSZK részéről mi is kitöltöttük a kérdőívet) és megpróbáltak egy összképet felvázolni az archiválási lehetőségekről és problémákról, illetve az archivált anyag tudományos hasznosításának módjairól. A második és harmadik munkacsoportban kísérleti munkafolyamatok megalkotását tűzték ki célként, ideértve az aratást, a minőségellenőrzést és hosszú távú megőrzési tervet. A negyedik munkacsoport a jogi háttér felmérésével foglalkozott, az ötödikben pedig egy fenntartható archiválási gyakorlat kialakítására tettek kísérletet. A rendezvényen előadásokat tartottak a projekt irányítói, illetve néhány partner könyvtár és kutató intézmény képviselői is. A belga nemzeti könyvtárban újra megpróbálkoznak a webarchiválás informatikai hátterének megteremtésével és szeretnének a jövőben is foglalkozni a közösségi média archiválásának kérdéseivel.

A szeptember 20-21-én az OSZK-ban megrendezett, a bibliográfiai adatok és a szemantikus web témájával foglalkozó BIBFRAME Workshopon is részt vett Németh Márton. A rendezvény kitűnő lehetőség volt a webarchívumok metaadatolásáról szóló beszélgetésekre is. A hónap végén kollégánk a Kutatók Éjszakáján – szintén a nemzeti könyvtárban – tart egy előadást a webarchiválási munkafolyamatokról, a kutatási célú hasznosításról és a személyes archiválás fontosságáról.

Az őszi rendezvény dömping egyik bennünket is érintő eseménye november 10-én a pozsonyi Egyetemi Könyvtárban lesz. A hosszú távú digitális megőrzéssel foglalkozó CDA 2022 konferencián tartandó előadás publikálásra kerülő változatát a hónap elején elkészítettük és leadtuk „Web archiving research in the context of digital humanities" címmel. Felkérést kaptunk továbbá az ELTE november 23-25 között tartandó „DH_BUDAPEST_2022 & Dariah Days" című konferenciájára. Ez a rendezvény különböző digitális bölcsészeti témákkal foglalkozik, hazai és külföldi előadókkal és kerekasztal résztvevőkkel.

Együttműködések

Szeptemberben megkaptuk a győri Dr. Kovács Pál Megyei Könyvtár és Közösségi Tértől is a részletes választ, hogy mely területeken szeretnének részt venni a megyei hatókörű könyvtárakkal a webarchiválás területén kialakított együttműködésben. Mivel a „404-es" workshopig szeretnénk megkötni ezeket a szerződéseket, be kell gyűjtenünk a függőben lévő válaszokat is az ezzel még adós könyvtáraktól.

Haladás történt viszont a szegedi Klebelsberg Könyvtárral tervezett közös projektben, ami a Karikó Katalin gyűjtemény archivált webtartalmakkal való bővítését célozza. A könyvtár vezetőivel és a gyűjteményt gondozó munkatársával szeptember 7-én tartott megbeszélésen arról is szó esett, hogy a nemzeti könyvtár miben tudná segíteni ezt a munkát, majd ezt követően elkészült egy feljegyzés és egy szerződéstervezet is. Ez az első olyan projektünk, amit már a Jira rendszerben kezdtünk el dokumentálni, részeként a nyilvántartások egységesítésére irányuló törekvésnek a Digitális Bölcsészeti Központon belül.

Az elmúlt hetekben lefutott tematikus aratások

Sport, testkultúra (3490 db seed URL)
Vallások, hitrendszerek, egyházak (2741 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (913 db seed URL)
Kutatóintézetek, tudományos szervezetek (1157 db seed URL)
Egyetemek, főiskolák (3961 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