

Az OSZK Webarchívum 2022 januári hírei

Archiválási munkák

2021. december 26. és 2022. január 3. között lefutott a félévente végzett webtér szintű aratás, mely közel 450 ezer domén és aldómén címre terjedt ki. A robot 74 millió URL-ről töltötte le a fájlkat, ebből 30 millió volt új vagy a korábbi mentéshez képest megváltozott. Az eltárolt tartalom kb. 2,4 terabájt lett. Az aratás után oldalképek is készültek a webhelyek kezdőlapjáról, közel 400 ezer képfájl 700 gigabájt összméretben. Szintén az év elejére is áthúzódó feladat volt az online időszaki kiadványok címlistájának bővítése különböző linkgyűjtemények, valamint a nemzetközi és az OSZK-ban működő nemzeti ISSN központ adatai alapján. Január folyamán 329 címet vettünk fel ebbe a részgyűjteménybe, melynek következő aratására a hónap végén kerül sor. Két új esemény alapú részgyűjteményt alakítottunk ki és kezdtünk el heti rendszerességgel menteni (férfi kézilabda EB és pekingi téli olimpia), és jelentősen kibővítettük a 2022. évi országgyűlési választásokkal kapcsolatos internetes források még tavaly összeállított címlistáját. Utóbbi január végétől májusig szintén hetente aratjuk majd.

Statisztikák, éves beszámolók és munkatervek

A heti tematikus aratások és a webtér szintű aratás részletes statisztikai táblázatai és grafikonjai mellett éves összesítéseket is készítettünk a webarchívum tavalyi bővüléséről és aktuális méreteiről (a főbb adatokat lásd a <https://webarchivum.oszk.hu/alapinformaciok-es-adatok/> oldalon), valamint a Webarchiválási Osztály egyéb számszerűsíthető tevékenységeiről. Összeállítottuk továbbá a 2021. évi munkabeszámolóinkat és az idei terveket.

Felhasználási szerződések megújítása

Múlt év decemberében kerestük meg azokat a partnereinket a nyilvános szolgáltatáshoz szükséges felhasználási szerződések határozatlan idejű meghosszabbításával, akiktől a webarchiválás projekt pilot szakaszában még csak határozott időre kértünk engedélyt. Az év vége ellenére folyamatosan érkeztek be az engedélyek és január végére a megkeresett partnereink körülbelül felétől kaptunk pozitív választ, ráadásul többször további webtartalmakra is megadva a hozzájárulásukat. Sajnos egy-egy elutasítás is történt, de ezek inkább egyéb okokra vezethetők vissza, semmint a szándék hiányára, és óhatatlan az is, hogy az elmaradó válasz miatt kelljen majd kivenni archivált anyagot a nyilvános szolgáltatásból. Mivel az év végi leállás és kollégánk januári betegsége is lassította az ügymenetet, ezért csak február elején tudjuk kiküldeni az újabb emlékeztető leveleket a szerződés meghosszabbításáról.

WCT tesztelés

Informatikai szempontból a hónap fő feladata a Web Curator Tool keretrendszer új verziójával kapcsolatos problémák megoldása volt, mivel ezt használjuk a nyilvános gyűjteményrésznél az aratások paraméterezésére és ütemezésére. Hosszas hibakeresés, tesztelgetés és a holland fejlesztővel való levelezés után végül sikerült kijavítani a hibákat, így februártól már újra üzemszerűen folyhat az archiválás a webadmin szerveren és bővíteni is tudjuk a gyűjteményt azokkal a webhelyekkel, melyekre a szerződések megújításán túl szintén engedélyt kaptunk a tulajdonosaiktól.

Helyben használat

A webarchívumok hasznosulása szempontjából fontos, hogy az archivált tartalmak a felhasználók számára is elérhetőek legyenek, azonban az online közzététel az esetek nagy részében engedély-

köteles. Az egyedi engedélyeztetés nyilván nem járható út a teljes magyar webtér esetében, ezért felértékelődik, hogy legalább helyben használatos szolgáltatásként elérhető legyen az érdeklődők számára az archivált tartalom. Azonban ebben az esetben is biztosítani kell az állomány megfelelő védelmét és a jogszerű használatot, hogy például ne lehessen letölteni vagy másolatot készíteni a védett tartalmakról. Ezért egy terminál szerver lett kialakítva erre a célra az OSZK informatikusai által, melyhez készítettünk egy egyszerű felhasználói felületet is. Ezt jelenleg a tájékoztató könyvtáros kollégák tesztelik, mivel elsősorban ők segítik majd az olvasókat és csak ezután lehet élesíteni az új szolgáltatást.

Szerzői jogi tanfolyam

Januárban lehetőségünk nyílt részt venni a Könyvtári Intézet „A szerzői jog a könyvtárakban” című négy napos online képzésén, melyet osztályunkról Visky Ákos László végzett el. A tanfolyamon a téma szakértői ismertették ennek a területnek a szabályozását, és példákon keresztül bemutatták a könyvtárak számára fontos kérdéseket, illetve vitás eseteket. A tanultak az általános jogi ismeretbővülés mellett a nyilvános szolgáltatáshoz szükséges engedélyeztetésben jelentenek segítséget.

Oktatás és ismeretterjesztés

A Könyvtári Intézet január 26-án meghirdette „Az internet archiválása mint közgyűjteményi feladat” elnevezésű tanfolyamunkat, az első időpont március 8-11. Elkezdődött az idei konferenciákon való részvételek tervezése is. Osztályunkról Németh Márton eddig három előadástervet adott be a következő rendezvényekre: LIBER – az Európai Tudományos Könyvtárak Egyesülete 2022. évi konferenciája a dániai Odensében (az előadás címe: „The opportunities of web archiving in research libraries”); BOBCATSSS – könyvtár- és információtudományi felsőoktatási konferencia Debrecenben (az előadás címe: „Web Archiving in Higher Education”); és Networkshop 2022, szintén Debrecenben (az előadás címe: „Az OSZK Webarchívum nemzetközi kapcsolatai”). A tavaly novemberi „404 Not Found – Ki őrzi meg az internetet?” rendezvényünkön tartott előadásaink átdolgozott változatai a Könyvtári Figyelőben, illetve a Könyv, Könyvtár, Könyvtáros folyóiratokban jelennek meg rövidesen. Utóbbi kiadvány január közepén publikált 2021/9. számában olvasható Drótos László cikke „Crawlerek és scraperek” címmel azokról a programokról, melyeket az elmúlt év folyamán teszteltünk a Webarchiválási Osztályon. Az MR1 Kossuth Rádió „Napközben” című műsorában 2022. január 18-án beszélgetés hangzott el korábbi osztályvezetőnkkel, Moldován Istvánnal az OSZK-ban folyó webarchiválásról.

IIPC
netpreserve.org

INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

ARCHIVING SOCIAL MEDIA WORKSHOP: SESSION 2
SCOPING POLICIES AND SELECTION CRITERIA
19 JANUARY 2022, 15:30-17:00 CET

SUMMARY OF SESSION 1: TOOLS AND WORKFLOWS

1. MAIN CHALLENGES

- NEED FOR MORE RESOURCES TO DEVELOP AND IMPROVE TOOLS, ADAPTING TO FREQUENT CHANGES IN PLATFORMS
- APPROACHING SOCIAL MEDIA PLATFORMS AS NATIONAL LIBRARY
- CHALLENGES IN SCOPING, STORAGE, TIME-CONSUMING BESPOKE CRAWLS
- DESIRES: TAILOR WEBRECORDER FOR LARGE-SCALE HARVESTING, BE ABLE TO USE USE HEADLESS BROWSER WITH HERITRIX3

2. CHOICE OF THE BEST TOOL

- TWITTER API, SOCIAL FEED MANAGER, HERITRIX, WEBRECORDER, BRANDWATCH
- CRITERIA: AVAILABILITY, TECHNICAL SKILLS NEEDED, COST, EASE OF USE, RESOURCE-INTENSIVENESS
- EXPERIMENTATION, RECOMMENDATIONS FROM PEERS

3. NON-OBVIOUS SOCIAL MEDIA PLATFORMS

- EXPERIMENTS WITH SNAPCHAT, TUMBLR, REDDIT
- NEED FURTHER INVESTIGATION: TWITCH, TIKTOK, TELEGRAM, BITCHUTE

BIT.LY/2022-SOCIAL-MEDIA-WORKSHOP-PROGRAMME
@NETPRESERVE
@NSK_ZAGREB
@KBRBE

THIS WORKSHOP IS ORGANISED BY THE IIPC IN PARTNERSHIP WITH THE NATIONAL AND UNIVERSITY LIBRARY IN ZAGREB AND KBR | ROYAL LIBRARY OF BELGIUM

A közösségi média archiválásáról szóló IIPC webinárium első napjának témái

Nemzetközi ügyek

Az IIPC szervezet, melyben az OSZK idén újabb 5 évvel hosszabbította meg a tagságát, január 19-én kétrészes webináriumot tartott a közösségi média archiválásáról. Ezen amerikai, kanadai, brit, belga, francia, dán, lengyel és új-zélandi kollégák beszéltek a Twitter, a Facebook, az Instagram, a Tumblr, a YouTube és egyéb nehezen megőrizhető platformokkal kapcsolatos tapasztalataikról, a technológiai, jogi és munkaszervezési kihívásokról. Bár mi még csak tesztekét csináltunk az elmúlt években ezen a területen, de a prezentációkat követő beszélgetéseken azért tudtunk érdekes információkat megosztani a webinárium résztvevőivel, továbbá kitöltöttük a rendezvényhez kapcsolódó kérdőívet is a saját gyakorlatunkról. Szintén az IIPC felhívására magyar online forrásokkal bővítettük a téli olimpiával foglalkozó weboldalak nemzetközi archívumát. A „404 Not Found” workshop kerekasztal beszélgetésén javasolt közép-európai együttműködés elindításához összeállítottunk egy feladattervet, aminek legfontosabb eleme egy IIPC pályázat előkészítése a közös portálról és keresőfelületről. Az Internet Archive és az OSZK közötti szerződés december végi aláírása után januárban elkezdődött az abban foglaltak megvalósítása. Ezzel kapcsolatban is írtunk egy összefoglaló anyagot, valamint megkaptuk az amerikai fél által készített ütemtervet is. E szerint több fázisban, március végére készül el az a portál, ahol az Internet Archive gyűjteményében levő, a .hu doménről 1996 óta 1,3 milliárd URL címről begyűjtött több mint 67,5 terabájnyi webtartalommal lehet majd keresni és a találatokat többféle szempont szerint szűrni, illetve statisztikákat készíteni belőlük. Az OSZK részéről a design elemeket és az intézményi környezetre szabható funkciókat kell meghatározni, majd magyarra fordítani és tesztelni a felhasználói felületet, végül pedig beépíteni a nemzeti könyvtár szolgáltatási portfóliójába.

Hazai együttműködések

A hónap végén kaptuk meg a jóváhagyást a Szegedi Tudományegyetem Klebelsberg Kuno Könyvtárától az együttműködési megállapodás véglegesítéséhez. A néhány ponton módosított szöveg reflektál arra a speciális helyzetre, hogy nemcsak az OSZK és az SZTE KK együttműködéséről van szó a regionális webtartalom feltárásában és archiválásában, hanem a térség két meghatározó könyvtára, az egyetemi és a megyei közötti kooperációról és profil szerinti feladatmegosztásról is. Folytatódtak a megbeszélések a székesfehérvári Vörösmarty Mihály Könyvtár munkatársaival is a tervezett Aranybulla Emlékév tematikus archívuma kapcsán, főleg könyvtárosi és technikai kérdések vonatkozásában. Az OSZK MEK-EPA-DKA gyűjteményeiből kiválogattuk számukra a releváns digitális és digitalizált dokumentumokat, valamint megosztottuk a Rákóczi Emlékév alkalmából készített hasonló „hibrid” archívum fejlesztése során szerzett tapasztalatainkat. Ebben a hónapban az egyéb sürgős feladatok és a megbetegedések miatt nem volt módunk újabb együttműködési megbeszélések kezdeményezésére, de amint lehetőség nyílik rá, folytatni szeretnénk a kapcsolatfelvételt az eddigi körökből kimaradt megyei könyvtárakkal.

Az elmúlt hetekben lefutott tematikus aratások

Elektronikus periodikák (9.221 seed URL)

Történelem, hely- és családtörténet (1.114 db seed URL)

Média, sajtó, műsorszórás (1.011 db seed URL)

Könyv- és egyéb kiadók, kereskedők (1.508 db seed URL)

Kormányzat, önkormányzatok, politikai és civil szervezetek (6.217 db seed URL)

A tematikus aratások részletes statisztikai adatai a <https://webarchivum.oszk.hu/szelektiv-aratasok/> weblapon nézhető meg. A projekt hírei a <https://webarchivum.oszk.hu/a-projektrol/hirek-esemenyek/> oldalon kísérhetők figyelemmel. Kapcsolati cím: mia@mek.oszk.hu