Az OSZK Webarchívum 2023 februári hírei

Archiválás, metaadatolás

Az archiváló szerverrel kapcsolatos technikai problémák miatt az ütemezett tömeges aratások egy részét el kellett halasztanunk a tervezett időponthoz képest, de február végére sikerült behozni a lemaradást.

Az orvostudományi, gyógyszerészeti, egészségügyi és szociális témájú webhelyek nyilvántartását jelentős mértékben tudtuk bővíteni a hónap folyamán, elsősorban a webtér szintű aratáshoz használt lista alapján. Egy hónapja kb. 1600 webhely adatait tartalmazta ez az új válogatás, jelenleg pedig már több mint 7500-at. Az EGESZSEG nevű részgyűjtemény első aratására márciusban kerül sor.

A februári intenzív gyűjtőmunka „melléktermékeként" több más tematikus címlistánk is bővült közel 200 új webhellyel. Például az elektronikus periodikák közé 30 oldalt vettünk fel ebben a hónapban. Ezek részben az egészségügyi intézmények és szervezetek honlapjain talált szakmai lapok, részben pedig az OSZK-ban működő ISSN Irodától és az Elektronikus Periodika Archívumtól érkező értesítésekben szereplő, a webarchívumban eddig még nem nyilvántartott folyóiratok és egyéb időszaki kiadványok.

Folytattuk a nyilvános archívumban már régóta szolgáltatott, de még nem metaadatolt webhelyek leírását, továbbá felvettünk tucatnyi újabb „közpénzes" honlapot a WCT adatbázisába, melyekről elkészültek az első próbamentések. Ezek ellenőrzés után szintén a nyilvános felületen fognak megjelenni. Mivel idén szeretnénk megszüntetni a párhuzamosságot a publikus és a csak helyben használható archívumrészek között, ezért elkezdtük tesztelni a PyWb megjelenítő hozzáférést szabályozó funkcióját. Erre a célra kiválogattunk néhány olyan állami vagy önkormányzati fenntartású webhelyet, amelyek viszonylag jól menthetők és a webarchiválást szabályozó kormányrendelet értelmében nem kell egyedi szerződést kötni a lementett tartalom nyilvános szolgáltatásához.

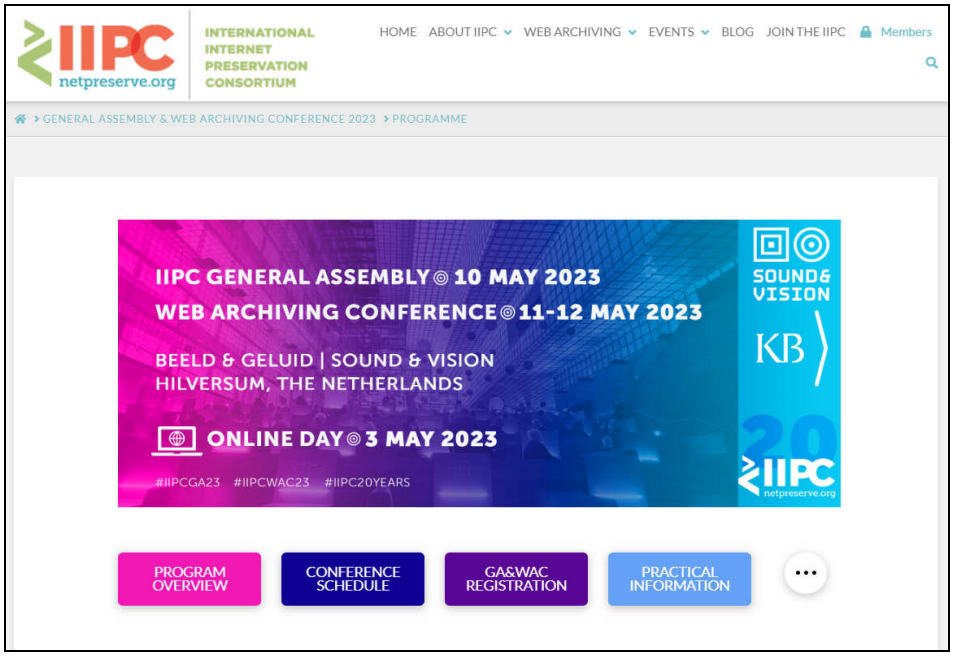
Tervek

Február első hetében összeállítottuk a webarchiválási csoport 2022. évi tevékenységéről szóló jelentést, valamint az idei munkatervet. Az utóbbiban szereplő feladatokkal kapcsolatban több megbeszélést is folytattunk, például a metaadat-nyilvántartás átalakításáról, a Digitális Képarchívummal való kapcsolatról, a közép-európai webarchívumok közötti együttműködési kezdeményezésről. Az egyik legsürgősebb teendő a Széchényi Ferencről szóló digitális gyűjtemény publikálása, amiben archivált weboldalak is vannak. Ennek a szolgáltató felületét elkészítettük a hónap folyamán, de még hátra van a kereső és böngésző funkciók fejlesztése, a weboldalak nyilvános szolgáltathatóságának engedélyeztetése, valami az angol verzióhoz szükséges fordítás.

Informatikai ügyek

Mivel a KIFÜ által üzemeltetett C4E felhőben futó archiváló szerverünk tárterülete nagyon lecsökkent és egyébként is másik zónába kell költöztetni a tárolókat és a szervereket, ezért a rendszergazdánk elkezdte átmásolni az archívum tartalmát, ami várhatóan több hetes munka lesz. A tárhely telítettsége miatt több leállás is volt ebben a hónapban, melyek után újra kellett indítani egyes automatizált folyamatokat. A hibák ellenőrzése során derült ki, hogy a fontosabb magyar hírportálok kezdőlapjának napi szintű mentésére használt Brozzler robot sem futott április óta, így ezekről az oldalakról is csak negyedévente készültek mentések a Heritrix programmal, az ELPERI részgyűjtemény aratása során.

A Brozzlert már tavaly szerettük volna leváltani a korszerűbb és gyorsabb Browsertrix crawlerre, amihez egy felhasználóbarát adminisztrációs felület is van, de ezt akkor nem sikerült megoldani. Néhány napja került fel a GitHubra a Browsertrix legfrissebb verziója, melyhez egy tesztszerver is tartozik, így egyelőre ezen próbálgatjuk az új funkciókat. A saját rendszerünkbe való beépítéshez a luxemburgi nemzeti könyvtár szakembereitől kapunk segítséget, akik legutóbb február 17-én, az IIPC által szervezett webináriumon számoltak be arról, hogy hogyan alakítottak ki egy automatizált keretrendszert a hírportálok hatékony, jó minőségű archiválására és a lementett oldalak szolgáltatására.



Rendezvények, tanfolyam

Az IIPC konzorcium, melynek az OSZK is tagja, májusban tartja éves közgyűlését és konferenciáját. A részletes program és az előadások absztraktjai a https://netpreserve.org/ga2023/programme/ oldalon már megnézhetők. A magyar Networkshop konferencia április 11-14. között kerül megrendezésre Veszprémben (https://nws.comp-rend.hu). Ezen munkacsoportunkat Kalcsó Gyula képviseli, aki a könyvtári témájú Katalist fórum anyagának archiválásáról tart előadást.

A Könyvtári Intézet március 6. és 9. közé hirdette meg az „Az internet archiválása mint közgyűjteményi feladat” című tanfolyamunkat, melynél gyorsan betelt a létszámkeret. Február utolsó napjaiban elkezdtük rá a felkészülést és a tananyag aktualizálását.

Az elmúlt hetekben lefutott tematikus és webtér aratások

Könyvtárak, levéltárak, múzeumok és galériák (2046 db seed URL)
Természet- és műszaki tudományok, szakterületek (2008 db seed URL)
Irodalom, irodalomtudomány és -történet (1455 db seed URL)
Bölcsészet- és társadalomtudományok, szakterületek (5702 db seed URL)
Közoktatás és egyéb képzések (6839 db seed URL)
Képző-, előadó-, zene- és filmművészet (8275 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