Az OSZK Webarchívum 2023 júniusi hírei

Archiválás, metaadat nyilvántartás

A tömeges aratások ebben a hónapban is szüneteltek, mert az állomány új tárhelyre való költöztetése még nem fejeződött be. Sikerült viszont megoldani a Web Curator Tool keretrendszerben beragadt job-ok problémáját, így legalább a nyilvános gyűjtemény mentései újra menetrendszerűen lefutnak. Néhány újabb – főként OSZK-s – webhelyet is felvettünk a WCT-be, ezek mentett változatai már visszanézhetők a honlapon. A régebbieknél pedig folytattuk a hiányzó XML leírások pótlását.

Az Európa Kulturális Fővárosa eseménysorozattal kapcsolatos információforrásokból több mint 70 URL címet válogattunk össze két nap alatt, főként hírportálok és közösségi média oldalak tematikus címkéit. Ugyancsak aratásra kész az ELETMOD kódnevű, sokféle szabadidős tevékenységgel és életvezetéssel kapcsolatos területet magába foglaló részgyűjteményünk, amely most 7809 webhely adatait tartalmazza. (A hónap folyamán kb. 1500 címmel bővítettük a listát, elsősorban kreatív alkotás, lakberendezés, szépségápolás és divat témákban.) Mindkét új gyűjtemény első mentésével meg kell várnunk a tárterület cserét, akárcsak a webtér szintű aratással, melyet eredetileg június végére terveztünk. Az elektronikus periodikák nyilvántartásába 66 kiadványt vettünk fel a hónap folyamán, többségükben az ELETMOD címlista gyűjtése során talált magazinokat. Az ELPERI mellett a TORTENELEM is jelentősen bővült, több mint 180 tételt soroltunk be különböző alkategóriákba.

Újra elkezdtünk podcast adásokat menteni. A tavaly óta talált kb. 230 újabb podcast csatorna közül az elmúlt napokban ötvennek a hanganyagát mentettük le egy böngészőkiegészítő segítségével (2045 db fájl 132 gigabájt összméretben), köztük az OSZK saját podcastját, aminek 2022-ben még csak néhány kísérleti adása volt. Ez a munka a következő hetekben még folytatódik.

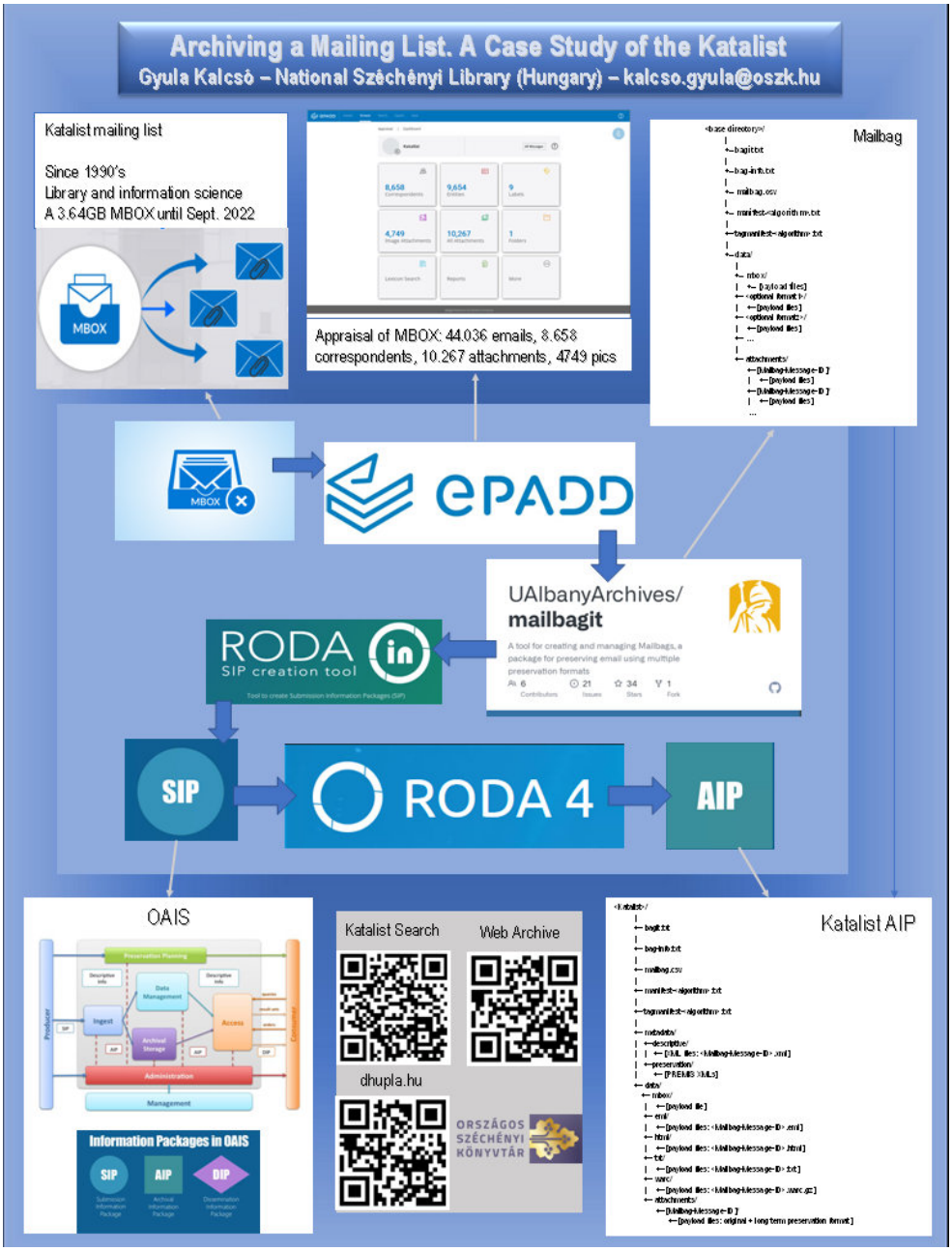
Technikai ügyek, újratervezés

A WCT rendbetétele mellett a nyilvános gyűjtemény metaadatkeresőjét is sikerült megjavítania az informatikusnak, mert – mint utólag kiderült – a Széchényi archívumhoz szükséges módosítások miatt az hibás lett. A webarchívum korábbi honlapja is elérhetetlen volt néhány hétig egy szerverbeállítás miatt, de most már újra működik a http://mekosztaly.oszk.hu/mia/ oldal, amin még vannak a projekt története szempontjából érdekes régi hírek és dokumentumok.

A webarchívum hat év után megérett az újratervezésre, időszerű az informatikai rendszer és egyes részfeladatok átalakítása. Ezért készítettünk egy listát a jelenlegi és a tervezett új munkafolyamatokról, továbbá kaptunk egy szervert, amin tesztelni lehet az elképzeléseinket, valamint eddig még nem használt eszközöket. Utóbbiak közül egyet, a Scrapy nevű webscrapert, valamint az ennek vezérlésére képes SpiderKeepert már fel is telepítette az informatikusunk. A scraper programok nem teljes webhelyek vagy weboldalak archiválására valók, hanem csak bizonyos fájlok és szövegek/adatok „összegereblyézésére” az internetről. Mi például a podcast adások hangfájljainak és metaadatainak automatikus begyűjtésére tudnánk használni a Scrapy-t, továbbá képmegosztó oldalakról digitális fotók és leírásuk lementésére, de hasznos lehet a közösségi média archiválására is, amennyiben az amúgy is nehezen megőrizhető külalak helyett csak a tartalmat szeretnénk letölteni.

Ismeretterjesztés, együttműködések

A június 6. és 9. között Budapesten megrendezett, digitális bölcsészeti témájú DARIAH konferencia poszter szekciójában Kalcsó Gyula „Archiving a Mailing List. A Case Study of the Katalist” címmel mutatta be egy levelezőcsoport anyagának archiválási folyamatát.



Kalcsó Gyula prezentációja a Katalist archiválásáról

A Tudományos és Műszaki Tájékoztatás folyóirat idei 2. számában megjelent Drótos László „Webarchiválás a nemzeti könyvtárban – Helyzetkép hat év után” című írása, ami a székesfehérvári Vörösmarty Mihály Könyvtár „Távcső a történelemre” című konferenciáján 2023. április 24-én elhangzott előadás szerkesztett változata.

A Könyvtári Figyelő 2022-es évfolyama most került fel a kiadvány honlapjára és az EPA-ba. Az 1. és 2. számban a 2021 novemberi „404 Not Found” rendezvényünkön elhangzott előadások egy részének szöveges verziói olvashatók, a 3. számban pedig Drótos László és Visky Ákos László „Webhelyek archiválási problémái” című cikke.

A Csorba Győző Könyvtár kérésére újra leválogattuk a Baranya megyei településnevek alapján a nyilvántartásunkban szereplő webhelyek adatait, mert tavaly óta elég jelentősen bővült a lista. A pécsi kollégák kiindulásként tudják majd használni ezt a táblázatot egy regionális gyűjtemény kialakításához.

A Conference of European National Librarians szervezet által meghirdetett pályázaton az OSZK és a Luxemburgi Nemzeti Könyvtár tervezete is támogatást nyert, melynek témája a két intézmény webarchiválási infrastruktúrájának és munkafolyamatainak fejlesztése, valamint a tudásmegosztás. Júniusban két online megbeszélést is tartottunk a luxemburgi szakemberekkel az együttműködés részleteiről. Az első személyes találkozó várhatóan szeptemberben lesz, amikor is megnézzük az ő rendszerüket, különös tekintettel a paywall mögötti híroldalak mentésére kidolgozott megoldásukra. A látogatás viszonzására az idei „404 Not Found" konferenciánk alatt kerül majd sor, melyen előadást is tartanak a vendégeink.