Az OSZK Webarchívum 2023 januári hírei

Archiválás

2023. január 1-től megszűnt a koronavirus.gov.hu címen 2020 márciusa óta működő kormányzati tájékoztató oldal. A hivatalos járványügyi adatokat és híreket közlő honlapot rendszeresen archiváltuk a KORONAVIRUS2020 nevű, eseményalapú gyűjteményünk részeként, de a leállítás előtti napokban készítettünk róla néhány teljes mentést a zárt és a nyilvános webarchívumba is, így utóbbit bárki meg tudja nézni a webarchivum.oszk.hu/demo-kezdolap/#egeszseg oldalról elindulva. Mivel a Kormány 626/2020. (XII. 22.) számú rendeletének 6. § (3) pontja szerint a kormányzati és önkormányzati tartalomszolgáltatások esetében nem kell külön szerződést kötnie az OSZK-nak az archivált verzió nyilvános szolgáltatására, ezért tudtuk ilyen gyorsan újra elérhetővé tenni ezt a fontos információforrást.

Az eoldal.hu domén alatt levő honlapokon megjelent tájékoztatás szerint február 1-től megszűnik ez a népszerű magyar tárhelyszolgáltatás, ezért január 24-én elindítottunk egy aratást, ami több mint 16 ezer webhelyre terjedt ki. A seed-listát részben a webtér és a tematikus nyilvántartásainkban már szereplő, az eoldal.hu alá bejegyzett aldoménekből állítottuk össze (14214 db), részben pedig a DNS szervereket lekérdező *amass* programmal gyűjtöttük (13971 db), de felhasználtuk az eOldal saját honlapján található, a leglátogatottabb webhelyek címeit tartalmazó toplistákat is (2533 db). A duplumok kiszűrése után 16186 kiinduló címet adtunk meg a robotnak, ami kevesebb, mint másfél nap alatt közel egy millió URL-t mentett le, fél terabájt összméretben.

Ebben a hónapban is kaptunk néhány archiválási javaslatot, melyek egyedi megoldásokat igényeltek. Az egyik a OSZK és a Vrije Universiteit közös virtuális kiállítására vonatkozott, ami a 17. és 18. században Hollandiában peregrináló magyar diákok által kivitt könyvek és egyéb nyomtatványok digitalizált címlapjaiból készült és a holland egyetemi könyvtár szerverén található. Az interaktívan lapozható és nagyítható képnézegető miatt ezt a webes összeállítást csak a Conifer nevű szolgáltatással sikerült archiválni egy régebbi Chrome böngészőt futtatva a felhőben futó virtuális gépen. A másik tanulságos eset a SZTE Klebelsberg Könyvtár Képtár és Médiatéka szerverén található Karikó Katalin digitális gyűjtemény archiválásával kapcsolatos, melyet bár rendszeresen mentünk a természettudományi címlistánk részeként, de a minőségellenőrzés során kiderült, hogy a képek hiányoznak az archivált változatból, mivel a szerveren levő robots.txt fájlban ki vannak tiltva a robotok a képek többségét tartalmazó „/files/” alkönyvtárból. Ezért január 9-én a robots.txt megkerülésével indítottunk egy új mentést erről a virtuális kiállításról, amely már jobban sikerült, bár az interaktív képnagyító funkcióval itt is baj van: csak az Open Wayback megjelenítővel működik, a modernebb PyWb hibaüzenetet ad.

Az archiválási és metaadatolási munkálatokba januártól bekapcsolódott Marcin Mirski, aki már korábban is segített nekünk a címlisták és az oldalképek ellenőrzésében, az elmúlt hetekben pedig megtanulta a Web Curator Tool és a Kaptafa nevű rendszerekkel az aratások indítását, valamint az XML fájlok készítését.

Metaadatolás

Decemberről áthúzódó feladatként még 860 újabb tétellel bővítettük a TARSTUD nevű részgyűjteményünk „Közgazdaság” kategóriáját, nagyrészt menedzsmenttel, marketinggel és számvitellel foglalkozó honlapok és blogok adataival. Szintén jelentős bővítés történt a PODCAST listában: egy 722 magyar csatornát tartalmazó oldalról a már ismertek és a már nem működők kiszűrése után 214 újabb podcast weblapját vettük nyilvántartásba. Január közepén elkezdtünk egy új részgyűjteményt kialakítani EGESZSEG néven, melybe eddig közel 1600 URL címet és webhely nevet válogattunk be. Elsőként az egészségügyi szervezetek, a kórházak és a rendelők honlapjait gyűjtjük össze, de ide

kerülnek majd a gyógyszerészeti, a pszichológiai/pszichiátriai, a természetgyógyászati és az állatorvostudományi oldalak, valamint a fogyatékkal élőket és a szociálisan rászorulókat segítő intézmények webhelyei is.

Elkészítettük a tavaly létrehozott három tematikus részgyűjtemény (TERMUSZ, TARSTUD, PODCAST) metaadat rekordjait és kitettük őket a honlapunkra. Folytattuk a nyilvános gyűjteményben levő, eddig még részletes adatokkal nem rendelkező webhelyek leírását is, ezek az XML fájlok folyamatosan kerülnek fel a webarchivum.oszk.hu/demo-kezdolap oldalra. A metaadatolás során egyben minőségellenőrzést is végzünk és a problémás eseteknél megpróbálunk más paraméterekkel vagy más szoftverrel jobb mentéseket készíteni, továbbá leállítjuk a már régóta nem frissülő webhelyek előre beütemezett aratásait.

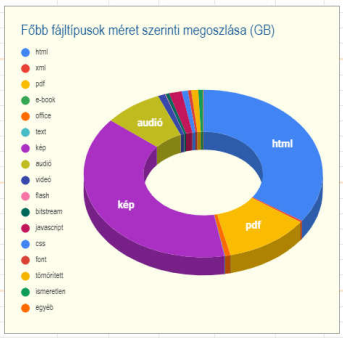
Statisztikák

A 2022. december 2. és 20. között három részletben lezajlott webtér szintű aratás összesített adatai felkerültek a honlapra. Mivel a fél évvel korábbihoz képest nem változtattunk a kiinduláskénti megadott címlistán és az aratási paramétereken, ezért a két archiválási ciklus eredménye jól összehasonlítható. A legfontosabb különbség az, hogy bár most valamivel kevesebb fájlt töltött le a robot, mint nyáron, de ebből darabszámra több volt az új vagy megváltozott tartalom, és összméretben is többet, a korábbi 6,1 helyett közel 6,7 terabájtot tárolt el a szerver.

Elkészült a tavalyi aratások összesített adatait tartalmazó táblázat és grafikon is, melybe a webtér, a tematikus részgyűjtemények, valamint a műfaji alapon válogatott e-periodikák és podkasztok weboldalai számítanak bele. (Nincsenek tehát benne a podkaszt csatornákról külön letöltött hangfájlok, a főbb híroldalak napi mentései, az események és földrajzi helyek alapján zajló aratások, valamint a nyilvános szerveren levő archív webhelyek.) 2022-ben szerettük volna legalább megduplázni az év folyamán begyűjtött tartalmat, ezért a tömeges aratásokat nagyobb mélységben és hosszabb ideig futtattuk, továbbá jelentősen bővítettük az emberi munkával válogatott és az automatikusan gyűjtött URL listákat. Bár utóbbi nagyon „szemetes” lett, ami meglátszik a sikertelenül bejárt seed címek magas számán, a kitűzött célt így is meghaladtuk: a 2021. évi 171,1 millió helyett tavaly 448,5 millió új vagy megváltozott címet mentett el a robot, 39,9 terabájt összméretben a korábbi év 16,2 terabájtjához képest.

2022. ÉVI ÖSSZESÍTETT ADATOK

Aratások száma:	**68**	**db**
Indulásnak megadott seed címek száma:	2 981 058	db
Bejárt seed címek száma:	3 609 167	db
Sikertelen seed címek száma:	683 528	db
Meglátogatott host-ok száma:	4 662 930	db
Másodpercenként letöltött URL-ek száma:	62	db
Megtalált URL-ek száma:	916 299 829	db
Megtalált URL-ek / induló seed címek száma:	254	db
Sikeresen letöltött URL-ek száma:	**880 540 487**	**db**
Sikertelen URL-ek száma:	9 521 729	db
Figyelmen kívül hagyott URL-ek száma:	26 237 613	db
Újonnan talált URL-ek száma:	**448 535 952**	**db**
Duplumként azonosított URL-ek száma:	431 830 877	db
Másodpercenként letöltött tartalom:	7 309	KB
Letöltött tartalom összesen:	**78 818**	GB
Duplumnak minősített, nem tárolt tartalom:	38 920	GB
Újnak minősített és eltárolt tartalom:	**39 898**	**GB**
Új tartalom / induló seed címek száma:	11	MB
Eltárolt tartalom tömörítés után:	23 939	GB
Tömörített tartalom / induló seed címek száma:	8	MB
Oldalképek mérete:	3369,7	GB
Oldalképek száma:	2 653 664	db
Oldalképek mérete / oldalképek száma:	1,21	MB
Oldalképek mérete / induló seed címek száma:	0,79	MB



Főbb fájltípusok szerinti megoszlás:	html	xml	pdf	e-book	office	text	kép	audió	videó	flash	bitstream	javascript	css	font	tömörített	ismeretlen	egyéb	összesen	
darabszám:	404 992 207	20 357 259	6 226 792	23 822	1 035 520	21 075 291	297 664 211	654 073	49 910	236 902	690 893	100 832 268	15 248 065	5 357 066	125 919	5 861 643	108 646	880 540 487	db
darabszám / összes (%):	45,99	2,31	0,71	0,00	0,12	2,39	33,80	0,07	0,01	0,03	0,08	11,45	1,73	0,61	0,01	0,67	0,01	100	%
duplumszűrés előtti méret:	27 215,07	236,78	9 035,93	6,22	611,33	37,26	30 568,58	6 246,48	769,57	55,39	409,83	1 335,91	656,17	377,27	742,26	464,36	52,93	78 821	GB
duplumszűrés előtti méret / összes (%):	34,53	0,30	11,46	0,01	0,78	0,05	38,78	7,92	0,98	0,07	0,52	1,69	0,83	0,48	0,94	0,59	0,07	100	%

A tavalyi 68 darab tömeges aratás adatainak összesítése

Az idei aratásokról készülő statisztikákban történt néhány kisebb változtatás. Például ezentúl nyilvántartjuk a robot által felderített, de nem letöltött URL-ek darabszámát is a „Feldolgozatlan URL-ek száma” nevű sorban. Ezek olyan címek, amelyeket az archiváló szoftver kigyűjtött a lementett weboldalakban talált linkekből és betette őket a várakozási sorba, de végül nem kérte le őket a

webszerverekről, általában azért, mert időközben az aratás elérte az előre beállított méret- vagy időhatárt. Módosult kicsit a MIME típusok szerinti megoszlást mutató táblázat és grafikon is, mert külön vettük a JSON (JavaScript Object Notation) fájlokat a Javascriptektől, mivel a JSON jelenleg már egy nyelvfüggetlen adatcsere formátum.

Informatikai ügyek

Ebben a hónapban is több kisebb-nagyobb technikai problémát jeleztünk az OSZK-s informatikusoknak, és karbantartás miatt volt egy leállás is január 21-22-én, ami a tömeges archiválást végző virtuális gépet érintette. A tárhelyet is bővíteni kellett a zárt archívumot és a honlapunkat szolgáltató szervereken, előbbi plusz 100, utóbbi pedig 5 terabájtot kapott.

A webarchívum szerverein futó szoftverekkel foglalkozó informatikus kollégánk sikeres teszteket végzett a PyWb megjelenítőbe tavaly beépített „access control" funkcióval, amivel szabályozni lehet, hogy mely domének és aldomének nézhetők meg nyilvánosan és melyek csak helyben, az erre a célra dedikált gépekről. Terveink szerint a jövőben így szolgáltatnánk a webarchívum nyilvános részét a jelenlegi külön szerveres megoldás helyett.

A webarchívum honlapját működtető WordPress rendszerben létrehoztunk egy új webhelyet, ami a leendő Széchényi Ferenc digitális gyűjteményt fogja szolgáltatni. A Rákóczi-emlékév alkalmából 2019-ben készült összeállításunkhoz hasonlóan ez is egy hibrid gyűjtemény lesz, melyben digitálisan született és digitalizált dokumentumok egyaránt megtalálhatók. A Széchényi Ferenc életét és hagyatékát bemutató képek, szövegek, weboldalak válogatása és metaadatolása január közepére befejeződött, így már csak a szolgáltatófelület kialakítása van hátra.

Mivel idén szeretnénk megújítani a metaadat-nyilvántartásunkat, ezért elkezdtünk ismerkedni az XML fájlok kezelését segítő Oxygen, illetve a megosztott munkát és verziókövetést lehetővé tevő GitLab rendszerekkel.

Tudományos munka, rendezvények

A hónap végén egy újabb megbeszélést tartunk a Könyvtári Szabványosítási Iroda munkatársával a webhelyek RDA-alapú leírásáról, ami szintén a metadatolás megújításához kapcsolódó feladat. A KSZI vezetőjén keresztül január közepén jutott el hozzánk az a munkaanyag, amit az RDA fejlesztését irányító bizottság (RDA RSC) a MARC 21 bibliográfiai adattárolási és adatcsere szabvány fejlesztőitől kapott véleményezésre. A szabványban tervezett változtatások közt van egy új, 857-es számú mező, melyben az archivált webtartalmak vagy digitális dokumentumok metaadatai tárolhatók. A 857-es almezőit és indikátorait összevetettük a mi metaadat struktúránkkal és megpróbáltuk megfeleltetni őket. A kérdéses pontokat jeleztük és javasoltunk egy további almezőt, amibe az archivált webhely minőségére, az esetleges hibákra és hiányokra vonatkozó információk írhatók.

Január 26-án Németh Márton és Kalcsó Gyula „Digital Humanities Research in Context of Web Archiving in Library Environment" címmel online előadást tartott az Oslo Metropolitan University által szervezett idei BOBCATSSS konferencián. (A prezentáció letölthető a honlapunkról.) Ugyanezen a napon került sor Szegeden a XIX. Magyar Számítógépes Nyelvészeti Konferencián Kalcsó Gyula, Mihály Eszter és Szűcs Kata közös előadására és bemutatójára „Korpuszépítés és -feldolgozás learatott webes tartalomból" címmel.

A hónap folyamán az IIPC konzorcium két webináriumot is szervezett, ezeket mi is figyelemmel követtük. Az elsőt a Harvard University Library Innovation Lab részlegének szoftverfejlesztője tartotta arról, hogy a Twitter körüli bonyodalmakat és tömeges leiratkozást/letiltást látva kifejlesztettek egy „thread-keeper" nevű alkalmazást. A social.perma.cc oldalon ingyenes online szolgáltatásként elérhető, de saját szerverre is telepíthető rendszerrel Twitter üzenetváltások archiválhatók időbélyeggel és hitelesítő tanúsítvánnyal ellátott PDF fájlok formájában, melyekből a reklámokat és egyéb

fölösleges elemeket előzetesen eltávolítja a program. A másik online előadás a közösségi média archiválásával foglalkozó projekteket felmérő kutatás főbb tanulságait foglalta össze, melyet a University of London School of Advanced Study egyik PhD-aspiránsa készített. Tavaly mi is kitöltöttük a kérdőívét és részt vettünk egy online interjún, melyen beszámoltunk a saját kísérleteinkről és tapasztalatainkról ezen a területen. Összesen 33 projekttől érkeztek válaszok a felmérés öt hónapja alatt. A legtöbb helyen Twittert archiválnak, ezt követi a Facebook, majd pedig az Instagram, végül jóval kisebb arányban a többi webkettes platform. (A kérdőíven még nem szerepelt, de a prezentációt követő beszélgetésből kiderült, hogy egy-két helyen már a Twitter egyik lehetséges utódjaként emlegetett Mastodon archiválását is elkezdték.) A kutatás egyik tanulságos megállapítása, hogy bár egyre szaporodnak a közösségi média archiválásával foglalkozó projektek, ezek gyakran függetlenek a közgyűjteményektől.

Az elmúlt hetekben lefutott tematikus és webtér aratások

Elektronikus periodikák (9681 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (6686 db seed URL)
Podkasztok (3788 db seed URL)
Történelem, hely- és családtörténet (1212 db seed URL)
Média, sajtó, műsorszórás (894 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1561 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