Az OSZK Webarchívum 2024. februári hírei

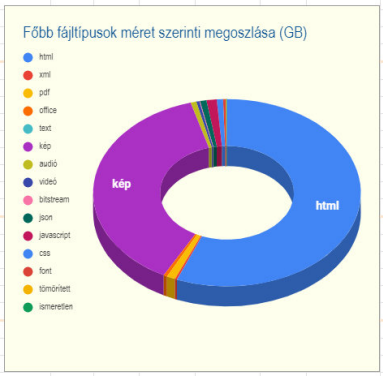
Archiválás és nyilvántartás

A januárban csaknem 24 napig futó webtér szintű jobok mellett nem akartuk elindítani az eredetileg az év első hónapjára beütemezett tematikus aratásokat is, hogy ne terheljük túl a szervert, ezért ezeket a februáriakkal együtt, a korábbi gyakorlathoz képest sűrűbben indítottuk el. A hónap végére sikerült behozni a tavalyi technikai problémák miatt kialakult lemaradást, így márciustól újra az eredeti ütemben folynak majd a tömeges mentések. A februári aratások felsorolása a beszámoló végén olvasható, részletes statisztikáik pedig a szintén ott található linken nézhetők meg.

Február 5. és 8. között első alkalommal mentettük önálló részgyűjtemény formájában a CSALAD kódnevű címlistában szereplő 2 ezer webhelyet. Ez a válogatásunk a következő főbb témakörökből áll: párkapcsolat, esküvő, házasság, szexualitás, erotika, születés, anyaság, család, gyereknevelés, idősek, halál. A három napnyi futásidő alatt 5 millió URL-t mentett le a robot 338 gigabájt összméretben, amiből 303 volt az új tartalom. Elkészítettük a CSALAD és a tavaly összeállított ELETMOD részgyűjtemények XML formátumú metaadat rekordjait is, melyeket feltöltöttünk a honlapunkra.

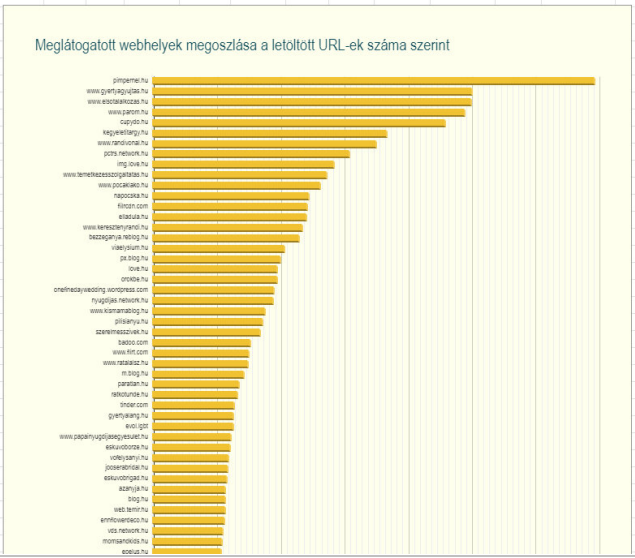
CSALAD

Aratás kezdete:	2024-02-05	
Aratás vége:	2024-02-08	
Futási idő:	3 nap 0 óra 0 perc	
Leállítás módja:	Az aratás az időhatár elérése után magától leállt.	
Indulásnak megadott seed címek száma:	2 069	db
Bejárt seed címek száma:	2 462	db
Sikertelen seed címek száma:	127	db
Meglátogatott host-ok száma:	10 055	db
Másodpercenként letöltött URL-ek száma:	19	db
Feldolgozatlan URL-ek száma:	925 797	db
Feldolgozott URL-ek száma:	5 176 160	db
Feldolgozott URL-ek / induló seed címek száma:	2 502	db
Sikeresen letöltött URL-ek száma:	5 046 358	db
Sikertelen URL-ek száma:	25 555	db
Figyelmen kívül hagyott URL-ek száma:	104 247	db
Újonnan talált URL-ek száma:	3 946 469	db
Duplumként azonosított URL-ek száma:	1 099 889	db
Másodpercenként letöltött tartalom:	1 368	KB
Letöltött tartalom összesen:	338	GB
Duplumnak minősített, nem tárolt tartalom:	35	GB
Újnak minősített és eltárolt tartalom:	303	GB
Új tartalom / induló seed címek száma:	150	MB
Eltárolt tartalom tömörítés után:	182	GB
Oldalképek mérete:	5,6	GB
Oldalképek száma:	1973	db
Oldalképek mérete / oldalképek száma:	2,91	MB
Oldalképek mérete / induló seed címek száma:	2,77	MB



Főbb fájltípusok szerinti megoszlás:	html	xml	pdf	e-book	office	text	kép	audió	videó	bitstream	json	javascript	css	font	tömörített	ismeretlen	egyéb	összesen	
darabszám:	2 664 758	123 037	2 907	0	807	20 604	1 476 249	328	2 779	34 154	559 639	60 305	45 919	17 384	19	37 439	30	5 046 358	db
darabszám / összes (%):	52,81	2,44	0,06	0,00	0,02	0,41	29,25	0,01	0,06	0,68	11,09	1,20	0,91	0,34	0,00	0,74	0,00	100	%
duplumszűrés előtti méret:	189,75	0,92	3,44	0,00	1,50	0,04	127,74	2,42	1,21	0,31	2,58	4,20	2,60	1,03	0,12	0,41	0,00	338	GB
duplumszűrés előtti méret / összes (%):	56,09	0,27	1,02	0,00	0,44	0,01	37,76	0,72	0,36	0,09	0,76	1,24	0,77	0,30	0,04	0,12	0,00	100	%

Webhelyek legalább 1400 letöltött URL-rel:	webhely	URL-ek	tartalom (MB)
	pimpernel.hu	138647	7751,29
	www.gyertyagyujtas.hu	100158	3060,07
	www.elsotalalkozas.hu	100002	2495,86
	www.parom.hu	97766	4858,47
	cupydo.hu	91680	2060,76
	kegyeletitargy.hu	73413	2178,79
	www.randivonal.hu	69953	5374,76
	pctrs.network.hu	61554	1743,75
	img.love.hu	56731	604,12
	www.temetkezesszolgaltatas.hu	54404	2557,49
	www.pocaklako.hu	52407	1137,90
	napocska.hu	48835	2596,20
	fiilrcdn.com	48455	675,85
	elladula.hu	48130	1324,66
	www.keresztenyrandi.hu	46770	1789,59
	bezzeganya.reblog.hu	45804	1709,95
	viaelysium.hu	41176	1637,05
	px.blog.hu	40044	18,18
	love.hu	38903	641,10
	orokbe.hu	38878	10240,19
	onefinedaywedding.wordpress.c	37993	3633,10
	nyugdijas.network.hu	37617	1607,11
	www.kismamablog.hu	35019	1204,00
	pillisanyu.hu	34285	940,15
	szerelmesszivek.hu	33601	2059,57
	badoo.com	30579	3022,84
	www.flirt.com	29904	972,13
	www.ratalalsz.hu	29744	1000,56
	m.blog.hu	28357	6471,52
	paratlan.hu	27013	1497,26
	ratkotunde.hu	26324	1219,44
	tinder.com	25437	10240,21
	gyertyalang.hu	25115	1139,49



A „Párkapcsolat, család” nevű részgyűjtemény első aratásának statisztikája

Létrehoztunk egy újabb válogatást MEZGAZ néven, amiben a mezőgazdasággal és az élelmiszeriparral kapcsolatos honlapokat és blogokat gyűjtjük. A címlista már több mint 2,5 ezer tételt tartalmaz, első aratására március elején kerül majd sor. Minden ilyen gyűjtési időszakban, így most is, több száz tétellel gyarapodott a nyilvántartásunk olyan címekkel, amelyeket eddig nem ismertünk és valamilyen más témakörbe vagy műfajba tartoznak. Az elektronikus időszaki kiadványok táblázatába például 114 URL került be február folyamán.

Az elmúlt hetekben 44 olyan webhely mentéseivel bővült a webarchívum nyilvános része, amelyek szolgáltatáshoz nem kell egyedi engedélyeket kérni. Főként települési és vármegyei önkormányzatok, könyvtárak és művelődési házak, illetve nemzeti parkok honlapjainak archivált verziói kerültek ki a publikus felületre, ahol három-négy megjelenítővel is visszanézhetők ezek. Az első tesztmentés után minden esetben megpróbáltuk optimálisan paraméterezni a Web Curator Tool keretrendszerben a Hertirix robotot és néha a PC-n futtatható HTTrack programmal is készítettünk egy alternatív mentést, aminek az eredményét WARC formátumra konvertáltuk. Ennek ellenére az archív változatokban vannak hibák és hiányok, amelyek részben az archiváló és/vagy a megjelenítő szoftverek képességeivel magyarázhatók, részben pedig az egyes fájlok és az egyes aratások méreteire, a futási időre és a linkek követési mélységére beállított korlátokkal. Tanulságos megfigyelni, hogy ugyanaz a mentés mennyire másként jelenik meg a különböző visszanéző szoftverekben, illetve hogy mennyire függ az archiválás sikeressége az eredeti webhelyen használt technikai megoldásoktól.

Technikai ügyek

A tesztszerveren a Web Curator Tool beüzemelésével kapcsolatos problémák okainak felderítése folyt. Az ékezetkódolási hibát végül sikerült elhárítani, de a WCT és a Heritrix közötti kommunikáció még nem stabil. A következő feladat a WCT tesztadatokkal való feltöltése lesz.

Újabb megbeszélések voltak az érintett informatikusokkal arról, hogy a kozterkep.hu oldalról begyűjtött mintegy 265 ezer, Creative Commons licenc alatt közzétett fénykép metaadatait hogyan kellene betölteni a Digitális Képarchívum adatbázisába és hogy milyen további részfeladatok kellenek még ahhoz, hogy ezek a köztéri műalkotásokat ábrázoló fotók megjelenhessenek a DKA-ban. Megállapodtunk abban is, hogy a konvertálást és az adatok betöltését először az új MEK fejlesztésére szolgáló, de a DKA rendszerének másolatát is tartalmazó tesztszerveren végezzük el, hogy ne akadályozzuk a hagyományos katalogizálást és az esetlegesen felmerülő hibákkal ne veszélyeztessük az éles szolgáltatást.

A Nyelvtudományi Kutatóközponttal való együttműködésünk keretében elkészültek és az átadás idejére egy ideiglenes depozit tárhelyre kerültek azok az index- és cheksum-fájlok, melyek a 2018 óta végzett webtér aratások technikai metaadatait tartalmazzák. Az OutbackCDX programmal készült indexek összmérete 321 gigabájt lett, a duplumszűrésre használt ellenőrző összegeket tartalmazó JDB adatbázisok pedig 544 gigabájtot foglalnak el egybecsomagolás és tömörítés után.

Ismeretterjesztés

Február 12-én online „gyorstalpaló” tanfolyamot tartottunk a Vörösmarty Mihály Könyvtár informatikusának a WCT használatáról, mivel a székesfehérvári kollégák is ezt a keretrendszert fogják használni a Fejér vármegyei regionális webarchívumhoz.

A honlapunk „A projektről” menüpontja alatt található „Előadások, prezentációk, publikációk” oldalon egy külön szekciót alakítottunk ki az OSZK webarchiválási projektjéről (is) szóló blogbejegyzésekről. A lista jelenleg 16 posztot tartalmaz, közülük kettő angol nyelvű, melyek az International Internet Preservation Consortium, illetve a Society of American Archivists blogjaiban jelentek meg.

Összeállítottunk egy kérdőív tervezetet, amivel azt szeretnénk felmérni, hogy a megújítás alatt álló, „Az internet archiválása mint közgyűjteményi feladat” elnevezésű tanfolyamunkkal kapcsolatban milyen igényei vannak a leendő résztvevőknek, melyek azok a résztémák, amelyek iránt leginkább érdeklődnek.

Az OSZK podcast csatornáján megjelent egy 40 perces interjú a webarchívum munkatársával, Drótos Lászlóval „Az internet úttörője” címmel.

Az elmúlt hetekben lefutott tematikus aratások

Könyvtárak, levéltárak, múzeumok és galériák (2131 db seed URL)
Irodalom, irodalomtudomány és -történet (1431 db seed URL)
Kulturális intézmények, művelődési házak, rendezvényhelyszínek (966 db seed URL)
Természet- és műszaki tudományok, szakterületek (2253 db seed URL)
Képző-, előadó-, zene- és filmművészet (8733 db seed URL)
Közoktatás és egyéb képzések (7239 db seed URL)
Bölcsészet- és társadalomtudományok, szakterületek (6131 db seed URL)
Életmód, szabadidő, hobbi (8597 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (7020 db seed URL)
Történelem, hely- és családtörténet (1465 db seed URL)
Podkasztok (4721 db seed URL)
Elektronikus periodikák (8737 db seed URL)
Média, sajtó, műsorszórás (937 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1477 db seed URL)
Párkapcsolat, család (2069 db seed URL)

Az egyes 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: webarchivum@oszk.hu