Az OSZK Webarchívum 2024. márciusi hírei

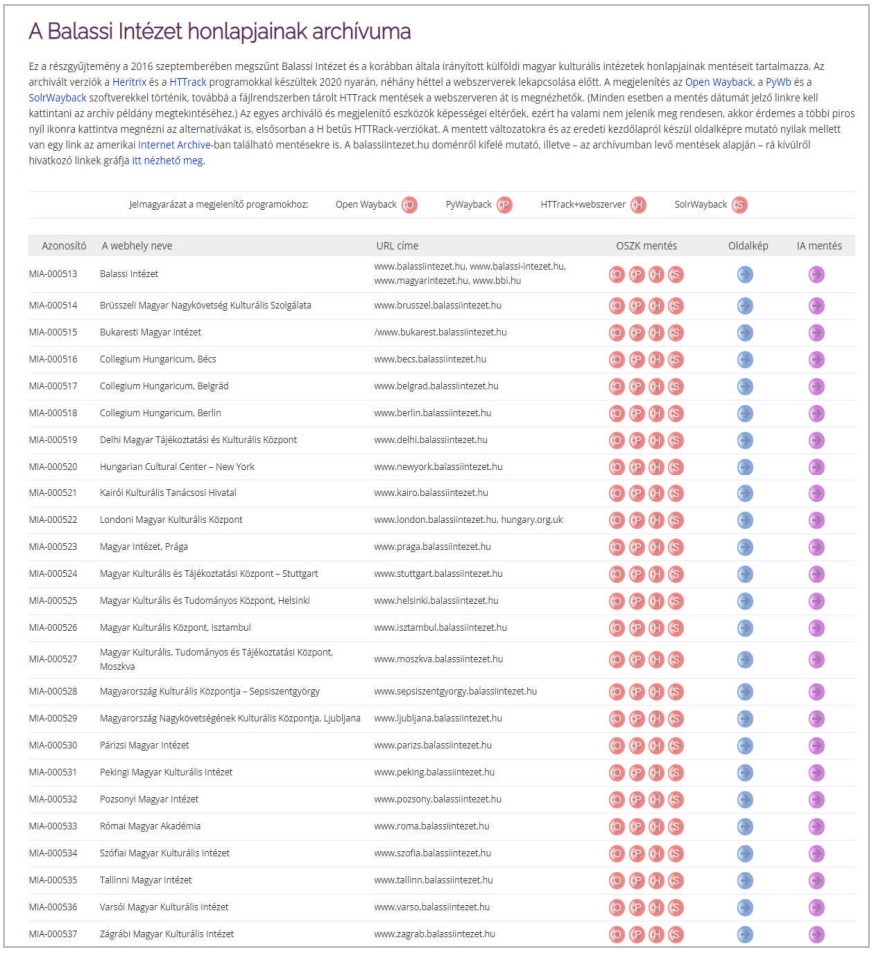
Archiválás és nyilvántartás

Március elején lefutott a mezőgazdasággal és az élelmiszeriparral kapcsolatos honlapokat és blogokat tartalmazó részgyűjtemény első aratása, ami 20 és fél óra után elérte az 500 gigabájtos mérethatárt. Ez alatt az idő alatt a kezdésnek megadott 2632 címről elindulva több mint 10 millió URL-t talált a Heritrix és ezekről a címekről 3,8 milliót fájlt tárolt el, a többi vagy duplumnak minősült, vagy a várakozási sorban maradt. Utóbbiak nagy része majd a negyedévente ismétlődő aratások során lesz letöltve. Elkészítettük a részgyűjtemény XML formátumú metaadat rekordját is, amit kitettünk a webarchívum honlapjára.

Építünk egy újabb tematikus válogatást SZOLGKER kódnéven, amibe a szolgáltatás és a kereskedelem mellett a szállítás és a közlekedés is beletartozik. Eddig 3750 webcímet gyűjtöttünk össze, egyelőre főleg a közlekedés témakörében. A gyűjtemény bővítése áprilisban is folytatódik még.

Elkezdtük a 2024. június 9-re kiírt önkormányzati és EP képviselői választással kapcsolatos hírek és weboldalak mentését. A már több mint 300 tételes lista első aratása (a közösségi oldalak nélkül) március 19-én indult és hetente ismételjük majd, várhatóan június végéig.

A nyilvános archívumba 64 újabb tételt került ki, főként kormányzati és önkormányzati honlapok mentései. Megpróbáltuk őket a lehető legjobb minőségben archiválni, ezért volt olyan, amelynél 3-4 mentést is csináltunk különböző beállításokkal, illetve szükség esetén a Heritrix mellett a HTTrack robotját is bevetettük. Az OSZK által kiadott folyóiratok aktuális és archív weboldalairól is készítettünk másolatokat, melyek a honlapunk „OSZK-s webhelyek archívuma” menüpontja alatt a „Periodikák” kategóriában nézhetők meg.



A Balassi Intézet honlapjainak archívuma

Ez a részgyűjtemény a 2016 szeptemberében megszűnt Balassi Intézet és a korábban általa irányított külföldi magyar kulturális intézetek honlapjainak mentéseit tartalmazza. Az archivált verziók a Heritrix és a HTTrack programokkal készültek 2020 nyarán, néhány héttel a webszerverek lekapcsolása előtt. A megjelenítés az Open Wayback, a PyWb és a SolrWayback szoftverekkel történik, továbbá a fájlrendszerben tárolt HTTrack mentések a webszerveren át is megnézhetők. (Minden esetben a mentés dátumát jelző linkre kell kattintani az archív példány megtekintéséhez.) Az egyes archiváló és megjelenítő eszközök képességei eltérőek, ezért ha valami nem jelenik meg rendesen, akkor érdemes a többi piros nyíl ikonra kattintva megnézni az alternatívákat is, elsősorban a H betűs HTTRack-verziókat. A mentett változatokra és az eredeti kezdőlapról készült oldalképre mutató nyilak mellett van egy link az amerikai Internet Archive-ban található mentésekre is. A balassiintezet.hu doménről kifelé mutató, illetve – az archívumban levő mentések alapján – rá kívülről hivatkozó linkek gráfja itt nézhető meg.

Jelmagyarázat a megjelenítő programokhoz: Open Wayback (IO) PyWayback (IP) HTTrack+webszerver (IH) SolrWayback (IS)

Azonosító	A webhely neve	URL címe	OSZK mentés	Oldalkép	IA mentés
MIA-000513	Balassi Intézet	www.balassiintezet.hu, www.balassi-intezet.hu, www.magyarintezet.hu, www.bbi.hu	IO IP IH IS		
MIA-000514	Brüsszeli Magyar Nagykövetség Kulturális Szolgálata	www.brusszel.balassiintezet.hu	IO IP IH IS		
MIA-000515	Bukaresti Magyar Intézet	/www.bukarest.balassiintezet.hu	IO IP IH IS		
MIA-000516	Collegium Hungaricum, Bécs	www.becs.balassiintezet.hu	IO IP IH IS		
MIA-000517	Collegium Hungaricum, Belgrád	www.belgrad.balassiintezet.hu	IO IP IH IS		
MIA-000518	Collegium Hungaricum, Berlin	www.berlin.balassiintezet.hu	IO IP IH IS		
MIA-000519	Delhi Magyar Tájékoztatási és Kulturális Központ	www.delhi.balassiintezet.hu	IO IP IH IS		
MIA-000520	Hungarian Cultural Center – New York	www.newyork.balassiintezet.hu	IO IP IH IS		
MIA-000521	Kairói Kulturális Tanácsosi Hivatal	www.kairo.balassiintezet.hu	IO IP IH IS		
MIA-000522	Londoni Magyar Kulturális Központ	www.london.balassiintezet.hu, hungary.org.uk	IO IP IH IS		
MIA-000523	Magyar Intézet, Prága	www.praga.balassiintezet.hu	IO IP IH IS		
MIA-000524	Magyar Kulturális és Tájékoztatási Központ – Stuttgart	www.stuttgart.balassiintezet.hu	IO IP IH IS		
MIA-000525	Magyar Kulturális és Tudományos Központ, Helsinki	www.helsinki.balassiintezet.hu	IO IP IH IS		
MIA-000526	Magyar Kulturális Központ, Isztambul	www.isztambul.balassiintezet.hu	IO IP IH IS		
MIA-000527	Magyar Kulturális, Tudományos és Tájékoztatási Központ, Moszkva	www.moszkva.balassiintezet.hu	IO IP IH IS		
MIA-000528	Magyarország Kulturális Központja – Sepsiszentgyörgy	www.sepsiszentgyorgy.balassiintezet.hu	IO IP IH IS		
MIA-000529	Magyarország Nagykövetségének Kulturális Központja, Ljubljana	www.ljubljana.balassiintezet.hu	IO IP IH IS		
MIA-000530	Párizsi Magyar Intézet	www.paris.balassiintezet.hu	IO IP IH IS		
MIA-000531	Pekingi Magyar Kulturális Intézet	www.peking.balassiintezet.hu	IO IP IH IS		
MIA-000532	Pozsonyi Magyar Intézet	www.pozsony.balassiintezet.hu	IO IP IH IS		
MIA-000533	Római Magyar Akadémia	www.roma.balassiintezet.hu	IO IP IH IS		
MIA-000534	Szófiai Magyar Kulturális Intézet	www.szofia.balassiintezet.hu	IO IP IH IS		
MIA-000535	Tallinni Magyar Intézet	www.tallinn.balassiintezet.hu	IO IP IH IS		
MIA-000536	Varsói Magyar Kulturális Intézet	www.varso.balassiintezet.hu	IO IP IH IS		
MIA-000537	Zágrábi Magyar Kulturális Intézet	www.zagrab.balassiintezet.hu	IO IP IH IS		

A Balassi Intézet által gondozott honlapok mentései

Készítettünk egy különgyűjteményt a 2016 szeptemberében megszűnt Balassi Intézet és a korábban általa irányított külföldi magyar kulturális intézetek honlapjainak mentéseiből, amelyek a Heritrix és a HTTrack programokkal készültek 2020 nyarán, néhány héttel a webszerverek végleges lekapcsolása előtt. A magyar és angol nyelvű összeállítás nyilvánosan is megtekinthető a honlapunkon.

Ismeretterjesztés, konferencia és tanulmány

„Az internet archiválása mint közgyűjteményi feladat" című tanfolyamunk megújításával kapcsolatos kérdőívünket, ami a Katalist levelezőcsoportban lett meghirdetve, két hét alatt 30-an töltötték ki. A lehetséges résztémák mellett rákérdeztünk a tanfolyam preferált módjára is. Ezzel kapcsolatban az egymás utáni 4 nap helyett többen a kétszer 2 napos változatra szavaztak, és a jelenléti mellett a hibrid formára is jöttek voksok.

Az Országos Széchényi Könyvtár „csevej" című podkaszt sorozatának 20. adásában Kalcsó Gyulával, a webarchiválási csoport vezetőjével hallható egy félórás beszélgetés, főként a webscraping technológiáról és az archivált tartalmak kutathatóvá tételéről.

Szintén ő tart előadást április 4-én az idei Networkshop konferencián az Eszterházy Károly Katolikus Egyetemen a scrapingről, és ismerteti a Köztérkép honlapról ezzel a technológiával történt képarchiválás projektjét; a hónap végén pedig részt vesz az IIPC szervezet éves közgyűlésén és konferenciáján Párizsban.

Tervezünk egy cikket a közép- és kelet-európai webarchívumokról, ezért kiküldtünk egy körlevelet az érintett nemzeti és egyetemi könyvtárakban ezzel foglalkozó kollégáknak, melyben egyebek mellett az általuk használt technológiáról, a minőségellenőrzésről, a hosszú távú megőrzésről és a gyűjteményhez való hozzáférésről érdeklődtünk. Rákérdeztünk arra is, hogy van-e magyar vonatkozású tartalom a webarchívumukban.

Az elmúlt hetekben lefutott tematikus aratások

Vallások, hitrendszerek, egyházak (2946 db seed URL)
Sport, testkultúra (3837 db seed URL)
Egészségügy, szociális szféra (8668 db seed URL)
Idegenforgalom, vendéglátás (7585 db seed URL)
Kutatóintézetek, tudományos szervezetek (1274 db seed URL)
Egyetemek, főiskolák (4288 db seed URL)
Mezőgazdaság és élelmiszeripar (2632 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