Az OSZK Webarchívum 2022 áprilisi hírei

Archiválás

A nyilvános archívum 24 webhellyel bővült áprilisban, főként könyvtári és képzőművészeti témájú honlapokkal és blogokkal, egy civil szervezet oldalaival, valamint a KSH adatvizualizációival. Ezeket a mentéseket ellenőriztük is, és esetenként több napon keresztül próbáltuk optimalizálni az archiválási paramétereket a minél jobb minőség elérése érdekében. A zárt gyűjtemény tematikus seed listáiba is felvettünk néhány tucat új tételt, de ezeknél ebben a hónapban nem végeztünk tömeges címbővítést és ellenőrzést. Jelentősen nőtt viszont a kárpátaljai gyűjteményünk, egy hónap alatt több mint 500 címmel és most 1456 tételt tartalmaz. Ennek a fele Facebook oldal, melyek mindegyikéről egyedi mentéseket készítettünk az ArchiveWeb.page böngésző-alapú programmal. Lementettünk továbbá 11 Twitter fiókot és hashtag-alapú találati listát is. A hazai és a határon túli online sajtóban megjelenő ukrajnai híreket visszamenőlegesen is begyűjtöttük januárra és februárra vonatkozóan azokról a hírportálokról, amelyeknél volt erre lehetőség. A keletkezett WARC fájlokat továbbadtuk a Digitális Bölcsészeti Központ (DBK) munkatársainak, akik így ezeket is fel tudják használni a vizualizációhoz. (Az első szófelhőkből készített rövid videó az OSZK YouTube csatornáján már megnézhető: https://www.youtube.com/watch?v=tX3tTKBa6Oc.) Az áprilisi országgyűlési választás után szintén hashtag-ek alapján mentettünk az eseménnyel és a pártokkal kapcsolatos Facebook, Instagram és Twitter bejegyzéseket.

Elkezdtük előkészíteni a júniusra tervezett következő webtér-szintű aratás seed listáját. A december végi mintegy 450 ezer tételes listánkat kiegészítettük a korábban archivált weboldalakban levő, a .hu országdoménre mutató linkekből kigyűjtött fő- és aldoménekkel (1,4 millió cím), valamint az Internet Archive-tól megvásárolt és április közepén átadott nyilvántartással a .hu végű webhelyekről (közel 5 millió cím). Az egyesített és duplumoktól megtisztított lista csaknem 5 és fél millió tételes, de ebben még sok a szintaktikailag hibás és a tömegesen generált URL, ezért első lépésben ezeket javítjuk, illetve válogatjuk szét. A következő lépés annak letesztelése lesz, hogy melyek ezek közül az élő webhelyek, majd következik a *title* metaadat begyűjtése és az alapján még egy újabb adattisztítási fázis.

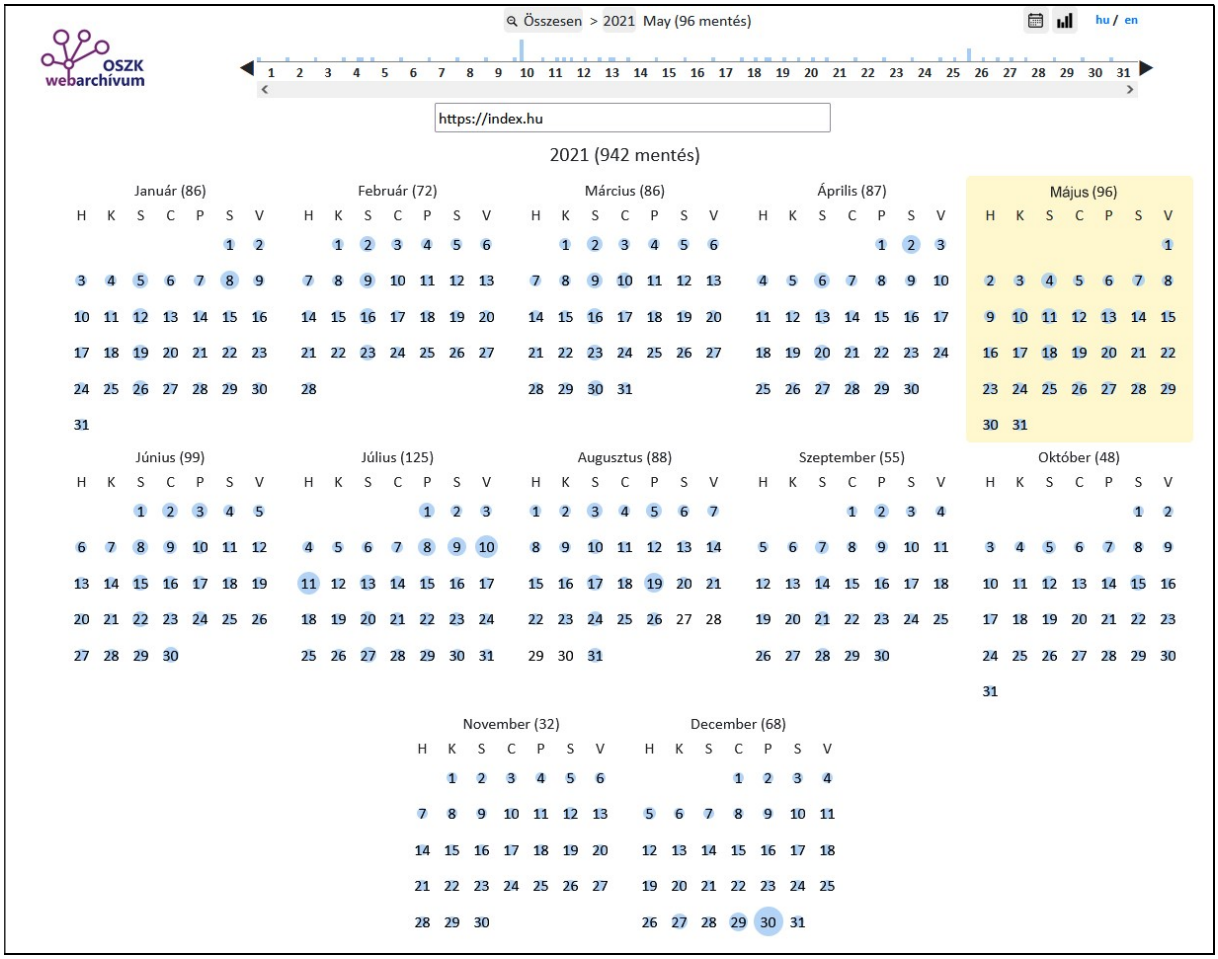
Technológia, szolgáltatás

A különféle Linux-alapú szoftverek tesztelésére és a tanfolyamon való bemutatására használt, VirtualBox alatt Windows-os PC-n is futtatható virtuális gépünk kereteit mostanra már kinőttük, ezért egy új rendszert állítottunk össze, melyben több a memória és dinamikusan bővíthető a tárhely, majd feltettünk rá különféle archiváló és megjelenítő szoftvereket. Utóbbiak közül a PyWb 2.7-es béta verzióját tesztelés és magyarítás után a rendszergazdánk a zárt, majd a nyilvános archívumot üzemeltető szerverekre is feltelepítette. Ebben a 2.7-es változatban számunkra a legfontosabb újdonság a korábbinál áttekinthetőbb és több funkciót biztosító naptárnézet és fejléc. Ezenkívül lehetőség van a felület teljes áttervezésére és a hozzáférés szabályozására is.

Az Internet Archive magyar keresőjének felületét is tovább teszteltük az elmúlt hetekben, ellenőriztük az elkészült javításokat és jeleztünk még néhány további – főként fordítási – hibát, valamint megírtuk a szolgáltatás súgóját magyar és angol nyelven, melyek a honlapunkról vannak belinkelve az amerikai szerveren futó portálba.

Az archív anyagot tartalmazó WARC fájlok mellett a jövőben szeretnénk majd kutatási célokra jobban megfelelő WET és WAT állományokat is generálni. Előbbi csak a weboldalak formázatlan szövegét tartalmazza, utóbbi pedig a WARC konténerekben levő egyedi fájlok metaadatait. Jelenleg az UKRAJNA és a KARPATALJA nevű részgyűjteményekhez készülnek ilyen WET és WAT verziók, amiket továbbítunk a feldolgozással foglalkozó digitális bölcsész kollégáknak.

Az ukrajnai híreket tartalmazó részgyűjteményhez elkészült egy publikus teljes szövegű kereső a SolrWayback programmal, amit annak dán fejlesztője módosított számunkra, úgy, hogy maguk az archivált hírek a nyilvános oldalon ne legyenek megnézhetők a szerzői jogi korlátok miatt. Tesztelés után még néhány további kérést küldtünk neki és a rendszergazdánknak, ezek megoldását követően kerülhet ki ez az új kereső a honlapunkra.



A PyWb megjelenítő új naptárnézete

Ismeretterjesztés

Az idén Debrecenben megrendezett Networkshop konferencia második napján Németh Márton az OSZK webarchívumának nemzetközi kapcsolatairól beszélt, Visky Ákos László pedig a Bács-Kiskun megyei városi és községi könyvtárosok szakmai továbbképzésén „A webarchiválás kihívásai” címmel tartott előadást április 25-én Kecskeméten. Mindkét prezentáció letölthető a honlapunkról: https://webarchivum.oszk.hu/eloadasok-prezentaciok-publikaciok/#2

Az IIPC májusi konferenciájának szervezőbizottsága elfogadta Németh Márton és Kalcsó Gyula (DBK) közös előadásjavaslatát, melynek címe: „Data extraction and visualization of harvested WARC files of thematic collection on Ukrainian War at the National Széchényi Library”. A szervezők a web-archívumok kutatásával foglalkozó szekcióba és *lightning talk* típusba, vagyis a rövid, áttekintő jellegű beszámolók közé sorolták be az előadást.

A dániai Aarhus egyetemén működő School of Communication and Culture munkatársai által koordinált nemzetközi WARCnet projekt egy tanulmánykötet kiadását tervezi. A projektben az OSZK is részt vesz annak indulásától kezdve és mivel a WARCnet a webarchívumok kutatási célú hasznosításával foglalkozik, ezért a kötet egyik témájaként javasoltuk az ukrajnai háborúval kapcsolatos online hírek elemzését.

Együttműködések, megbeszélések

Rendszeres értekezleteket tartunk a DBK munkatársaival, melyeken az ukrajnai és kárpátaljai részgyűjtemények feldolgozása során felmerülő aktuális feladatokon túl egyéb témák is szóba kerülnek a webarchívum kutathatóvá tételével kapcsolatban. Megismertetjük egymást a két osztályon használt szoftverekkel és munkafolyamatokkal, és igyekszünk majd átvenni a jó gyakorlatokat (pl. a DBK által az XML fájloknál alkalmazott verziókövetést, ami a webarchívum metaadat rekordjainál is nagyon hasznos lenne).

Újabb Teams megbeszélés tartottunk Takács Dániellel, az ELTE Állam- és Jogtudományi Kar könyvtárának vezetőjével, aki a jogi témájú webhelyeken túl a hazai podkasztok weboldalainak összegyűjtésében is felajánlotta a segítségét és már mintegy félszáz új címmel bővítette is a MEDIA gyűjteményben levő nyilvántartásunkat.

E hónap végén a Szegedi Tudományegyetem Klebelsberg Kuno Könyvtárának informatikai főigazgatóhelyettesével, Kokas Károllyal egyeztetünk arról, hogy a náluk tervezett Karikó Katalin különgyűjteményt hogyan lehetne archivált webtartalmakkal kiegészíteni.

Április 25-én Beatrice Canelli-vel, a University of London olasz PhD hallgatójával beszéltünk Zoom-on, aki a közösségi média archiválásáról érdeklődött, mivel erről írja a dolgozatát. Megosztottuk vele az eddigi tapasztalatainkat és a témával kapcsolatos írott anyagainkat.

Az elmúlt hetekben lefutott tematikus aratások

Történelem, hely- és családtörténet (1125 db seed URL)
Könyv- és egyéb kiadók, kereskedők (1513 db seed URL)
Média, sajtó, műsorszórás (1022 db seed URL)
Kormányzat, önkormányzatok, politikai és civil szervezetek (6278 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