Az OSZK Webarchívum 2024 januári hírei

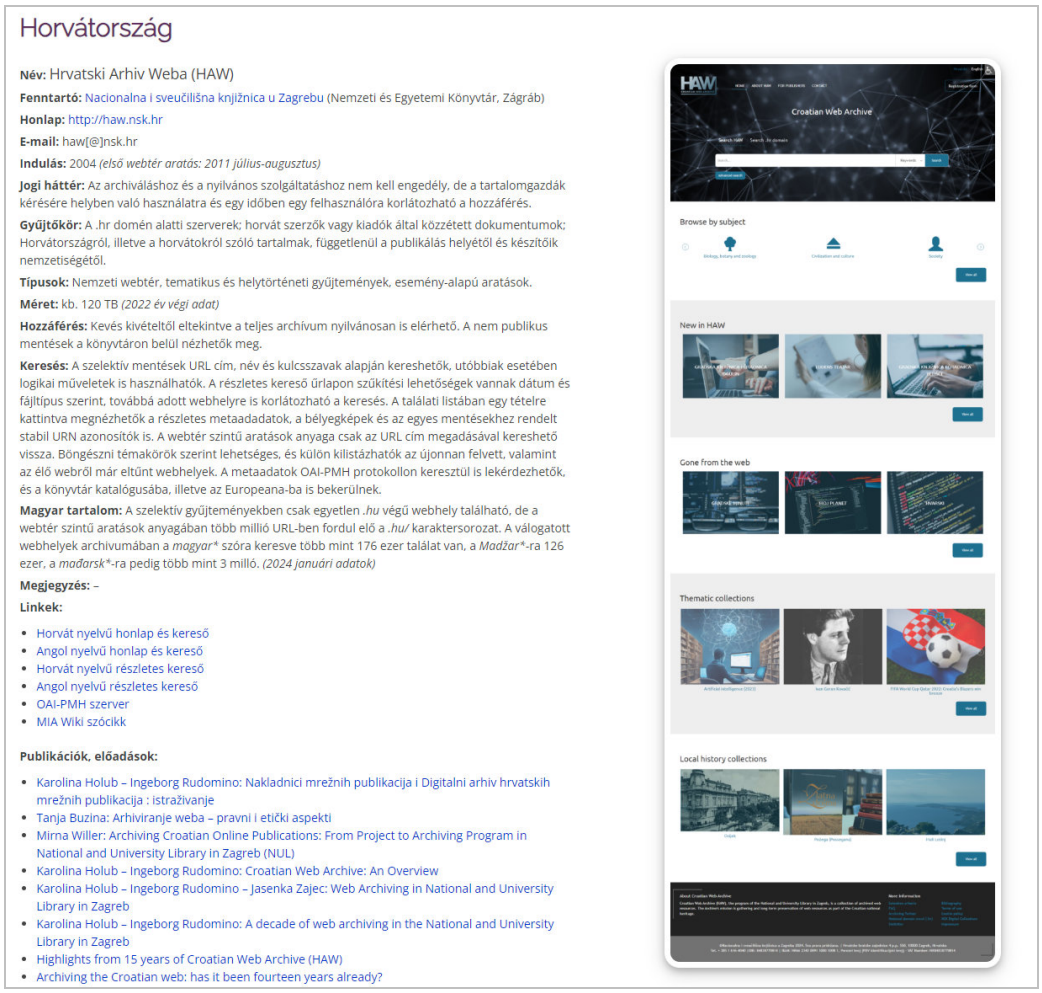
Archiválás és nyilvántartás

A tavalyi technikai nehézségek okozta csúszás miatt január elején indítottuk el az eredetileg decemberre tervezett webtér szintű aratást, amely három részletben zajlik. Az első job 10 napig futott és ez alatt 65 millió URL címet töltött le a Heritrix robot, ebből 30 millió volt az új vagy megváltozott fájl, 1,8 terabájt összméretben. A második részaratás még fut, ez 8 nap alatt 3,5 terabájtnyi tartalmat gyűjtött be.

Az elmúlt hetekben öt új honlappal gyarapítottuk a nyilvános gyűjteményt, melyek szolgáltatására engedélyt kaptunk és ugyanennyi tételt vettünk fel az OSZK saját webhelyeinek (többségében korlátozott mélységű) mentéseit tartalmazó, nagyrészt szintén publikus részgyűjteménybe, köztük a Magyar Elektronikus Könyvtár nemrég megújult oldalát.

Lementettük a decemberben talált 8 podcast csatorna hanganyagát, közel 3 ezer fájlt 129 gigabájt méretben.

Az újonnan kialakított, „Párkapcsolat, család” elnevezésű gyűjtemény mérete elérte a 2 ezer címet, a januári bővítés során mintegy 500 új tételt vettünk fel a nyilvántartásba. A címlista első aratására februárban kerül majd sor. A CSALAD mellett az ELPERI és a TORTENELEM gyűjtemények gyarapodtak jelentősebb mértékben ebben a hónapban. Az elektronikus periodikák közé 182 db eddig nem ismert kiadvány weboldalát vettük fel, a történelmi, család- és helytörténeti témájú webhelyek száma pedig 120 tétellel nőtt.



Horvátország

Név: Hrvatski Arhiv Weba (HAW)

Fenntartó: Nacionalna i sveučilišna knjižnica u Zagrebu (Nemzeti és Egyetemi Könyvtár, Zágráb)

Honlap: http://haw.nsk.hr

E-mail: haw[@]nsk.hr

Indulás: 2004 *(első webtér aratás: 2011 július-augusztus)*

Jogi háttér: Az archiváláshoz és a nyilvános szolgáltatáshoz nem kell engedély, de a tartalomgazdák kérésére helyben való használatra és egy időben egy felhasználóra korlátozható a hozzáférés.

Gyűjtőkör: A .hr domén alatti szerverek; horvát szerzők vagy kiadók által közzétett dokumentumok; Horvátországról, illetve a horvátokról szóló tartalmak, függetlenül a publikálás helyétől és készítőik nemzetiségétől.

Típusok: Nemzeti webtér, tematikus és helytörténeti gyűjtemények, esemény-alapú aratások.

Méret: kb. 120 TB *(2022 év végi adat)*

Hozzáférés: Kevés kivételtől eltekintve a teljes archívum nyilvánosan is elérhető. A nem publikus mentések a könyvtáron belül nézhetők meg.

Keresés: A szelektív mentések URL cím, név és kulcsszavak alapján kereshetők, utóbbiak esetében logikai műveletek is használhatók. A részletes kereső űrlapon szűkítési lehetőségek vannak dátum és fájltípus szerint, továbbá adott webhelyre is korlátozható a keresés. A találati listában egy tételre kattintva megnézhetők a részletes metaadatadatok, a bélyegképek és az egyes mentésekhez rendelt stabil URN azonosítók is. A webtér szintű aratások anyaga csak az URL cím megadásával kereshető vissza. Böngészni témakörök szerint lehetséges, és külön kilistázhatók az újonnan felvett, valamint az élő webről már eltűnt webhelyek. A metaadatok OAI-PMH protokollon keresztül is lekérdezhetők, és a könyvtár katalógusába, illetve az Europeana-ba is bekerülnek.

Magyar tartalom: A szelektív gyűjteményekben csak egyetlen *.hu* végű webhely található, de a webtér szintű aratások anyagában több millió URL-ben fordul elő a *.hu/* karaktersorozat. A válogatott webhelyek archívumában a *magyar** szóra keresve több mint 176 ezer találat van, a *Madžar**-ra 126 ezer, a *maďarsk**-ra pedig több mint 3 milló. *(2024 januári adatok)*

Megjegyzés: –

Linkek:

- Horvát nyelvű honlap és kereső
- Angol nyelvű honlap és kereső
- Horvát nyelvű részletes kereső
- Angol nyelvű részletes kereső
- OAI-PMH szerver
- MIA Wiki szócikk

Publikációk, előadások:

- Karolina Holub – Ingeborg Rudomino: Nakladnici mrežnih publikacija i Digitalni arhiv hrvatskih mrežnih publikacija : istraživanje
- Tanja Buzina: Arhiviranje weba – pravni i etički aspekti
- Mirna Willer: Archiving Croatian Online Publications: From Project to Archiving Program in National and University Library in Zagreb (NUL)
- Karolina Holub – Ingeborg Rudomino: Croatian Web Archive: An Overview
- Karolina Holub – Ingeborg Rudomino – Jasenka Zajec: Web Archiving in National and University Library in Zagreb
- Karolina Holub – Ingeborg Rudomino: A decade of web archiving in the National and University Library in Zagreb
- Highlights from 15 years of Croatian Web Archive (HAW)
- Archiving the Croatian web: has it been fourteen years already?

A közép-európai webarchívumokról készült egyik adatlap

Technikai ügyek

Az eddig a Scrapy futtatására használt tesztszerverre feltelepítettük a Web Curator Tool keretrendszert is, mert szeretnénk kipróbálni, hogy hogyan lehetne vele egyedi mentéseket vezérelni arról a több tízezer webhelyről, amelyekről eddig csak tömeges aratások készültek.

Több hetes futási idő után elkészültek azoknak a CDX index és cheksum fájloknak a tömörített állományai, amelyeket a Nyelvtudományi Kutatóközpontnak adunk majd át az OSZK-val kötött együttműködési megállapodás részeként.

Kapcsolatok

A honlapunk „Szakembereknek” menüjében megjelent egy új alpont a közép-európai webarchívumokról (https://webarchivum.oszk.hu/szakembereknek/cewa/) Itt a 2021-ben tartott „404-es” rendezvényünkön bemutatkozott öt, környező országbeli nemzeti archívum adatlapja érhető el. Az egyes projektek ismertetése mellett megpróbáltuk azt is felderíteni, hogy mennyi magyar nyelvű vagy magyar vonatkozású tartalom lehet ezekben a gyűjteményekben.

Formálódik az együttműködés a Vörösmarty Mihály Könyvtárral, ahol egy Fejér vármegyei regionális webarchívum létrehozását tervezik. Az elmúlt napokban beszéltünk a székesfehérvári kollégákkal és elkészítettünk egy feljegyzést a VMK és az OSZK által vállalt feladatokról.

A projekt hírei a https://webarchivum.oszk.hu/a-projektrol/hirek-esemenyek/ oldalon kísérhetők figyelemmel. Kapcsolati cím: webarchivum@oszk.hu