

## Az OSZK Webarchívum 2021 májusi hírei

### Webtér aratás előkészítése

Elkezdtek bővíteni és tisztítani a június végére tervezett, a magyar webtér általunk eddig felderített részére kiterjedő – „pillanatfelvétel” jellegű – aratásához a kiinduló (seed) címeket tartalmazó listát. Első lépésben az idén és tavaly mentett WARC fájljokból kigyűjtöttük a .hu végű címekre mutató linkeket. Az így kapott, közel 717 ezer domén és aldomén nevet összevetettük a korábban különböző forrásokból már összegyűjtött 251 ezres címlistánkkal, majd az abban még nem szereplő tételeket részben automatikus szűrésekkel, részben emberi munkával tovább szűkítettük a relevánsnak tűnőkre (pl. töröltük a levelező szervereket, az egy webhelyhez tömegesen létrehozott aldoméneket). Az ezután megmaradt 356 ezer URL-nél a Chrome motorját használva lekérdeztük a webservertől a státusz kódokat és a kezdőoldal title adatát, ami (jó esetben) a webhely neve. E két adat alapján még további szűrésekre lesz szükség: töröljük a valamilyen hibakódot adó, illetve a title alapján érdektelen tételeket (pl. parkoló domén, megszűnt honlap, bejelentkezéshez kötött oldal, Facebook-ra átirányított cím). Várhatóan még így is nagyon „szemetes” lesz a lista, mert a valóban releváns, eddig még nem ismert webhelyek kiválogatásához egy-két százezernyi URL egyenkénti átnézésére lenne szükség, amire már nincs kapacitásunk.

### Twitter csatornák mentése

Az április végi tesztek után folytatódott a magyar Twitter csatornák gyűjtése és archiválása. Eddig 800 fiók bejegyzéseit mentettük le az ArchiveWeb.page (<https://archiveweb.page>) böngésző-kiegészítőjével, WARC formátumban, 8,5 gigabájt összméretben. Sajnos a Twitter kb. 750 tweetben korlátozza a visszagörgethető listát, így csak 468 esetben sikerült a fiók tulajdonosa által közzétett minden üzenetet archiválni, a többinél csupán az utolsó néhány hónap vagy év anyaga lett letöltve. A válogatás körét folyamatosan bővítjük, jelenleg főként közgyűjtemények, színházak, könyvkiadók és könyvesboltok, egyházak, egyetemek, tudományos és kulturális intézmények, kormányzati, illetve politikai szervezetek és személyek, médiaszolgáltatók, újságok és újságírók csatornái vannak ebben a részgyűjteményben: <https://webarchivum.oszk.hu/webarchivum/bongesztes-twitter/>

### Hírportálok közösségi oldalainak címgyűjtése

A nagyobb hírportálok közösségi felületeinek URL-címeit gyűjtöttük össze május utolsó hetében, kiterve a határon túli magyar tartalomszolgáltatásokra is. Mintegy negyven hírforrás közösségi média oldalát tártuk fel és fogjuk lementeni a jövőben.

### Új gyűjtemény: NEK2020

Egy újabb eseménnyel, a 2020-ról a járványhelyzet miatt 2021 szeptemberére halasztott 52. Nemzetközi Eucharisztikus Kongresszussal kapcsolatos híreket kezdtük el archiválni egyelőre havi gyakorisággal, augusztustól pedig majd hetente. Addig még a címlistát is tovább bővítjük releváns oldalakkal, cikkekkel. <https://webarchivum.oszk.hu/webarchivum/bongesztes-nemzetkozi-eucharisztikus-kongresszus-2020/>

### Covid archívum indexelése

A koronavírus járvánnyal kapcsolatos hír- és egyéb weboldalak heti gyakoriságú aratása 2020. február 3-án kezdődött (<https://webarchivum.oszk.hu/bongesztes-koronavirus-jarvany-2020/>) Az azóta összegyűjtött több mint ezer WARC fájl (527 gigabájt) eddig csak az URL cím alapján volt böngészhető a zárt archívumban, de májusban sikerült az archivált weboldalakat leindexelni a SolrWayback programmal, amellyel nemcsak a teljes szövegben előforduló szavakra, valamint a képekre lehet rákeresni, hanem különféle statisztikákat is lehet készíteni a szavak előfordulási gyakorisága, illetve a technikai metaadatok alapján.

**SOLRWAYBACK**

astrazeneca vérrög

GROUPED SEARCH [?] IMAGE SEARCH [?] URL SEARCH [?]

SEARCH WITH UPLOADED FILE GPS IMAGE SEARCH TOOLBOX

**FACETS**

**domain**

- hazipatika.com (1,411)
- mandiner.hu (969)
- 444.hu (254)
- hvg.hu (183)
- origo.hu (145)
- pestisracok.hu (135)
- nepszava.hu (116)
- egeszsegkalauz.hu (109)
- webbeteg.hu (78)
- 24.hu (63)
- promotions.hu (49)
- infostart.hu (47)
- koronavirus-kisokos.eu (47)
- blikk.hu (46)
- hirstart.hu (31)
- ripost.hu (31)
- vg.hu (29)
- portfolio.hu (28)
- nlc.hu (25)
- euronews.com (22)
- telex.hu (22)
- azonnali.hu (18)
- mfor.hu (17)
- informed.hu (15)
- wikipedia.org (14)
- magyarhirlap.hu (8)
- qubit.hu (8)
- napi.hu (7)
- vasarhely24.com (7)
- magyar nemzet.hu (6)
- parameter.sk (6)
- ujszo.com (6)
- fevidek.ma (5)
- noizz.hu (5)
- intima.hu (4)
- life.hu (4)
- merce.hu (4)

**RESULTS** [1]

Showing 0 - 20 of 2,278 unique entries matching query (total hits: 3,997).

Previous 20 Next 20

**#1** **ASTRAZENECA VÉRRÖG | 24.HU** score: 75.11163

Type: html, web page @ 24.hu  
Date: 13/04-2021  
Url: https://24.hu/tag/astrazeneca-verrog/

**Highlighted content:**

"vérrög Ausztrál AstraZeneca-méreg: 700 ezer oltott, két vérrögösödés Ausztráliában sem tűnik kevésbé"

View data fields

**#2** **AZ ASTRAZENECA-VAKCINA ÉS A VÉRRÖGKÉPZŐDÉS: VILÁGOS A KAPCSOLAT :: EGÉSZSÉGPOLITIKA - INFORMED ORVOSI ÉS ÉLE TMÖD PORTÁL :: KORONAVÍRUS JÁRVÁNY, ASTRAZENECA, VÉRRÖG, THROMBÓZIS** score: 59.545773

Type: html, web page @ informed.hu  
Date: 13/04-2021  
Url: https://www.informed.hu/egpol\_new/az-astrazeneca-vakcina-es-a-verrogekpezodes-vilagos-a-kapcsolat-242464.html

**Highlighted content:**

"| koronavírus járvány, AstraZeneca, vérrög, thrombózis 2021-04-07 13:54:17 Kapcsolódó cikkek Az AstraZeneca"

View data fields

**#3** **ASTRAZENECA ÚJABB KÖZLEMÉNYE AZ OLTÓANYAGRÓL - NLC.HU** score: 58.80651

Type: html, web page @ nlc.hu  
Date: 16/03-2021  
Url: https://nlc.hu/ezvan/20210315/gyarto-ujabb-kozlemeny-astrazeneca-oltas-verrogekpezodes/

**VISUALIZATION OF SEARCH RESULT BY DOMAIN**

## Networkshop 2021 előadások

Visszanézhető a Videotoriumban az idei Networkshop konferencián tartott előadásaink:

- A közösségi média archiválásának nehézségei: <https://kifu.videotorium.hu/hu/recordings/42378>
- A COVID járvánnyal kapcsolatos webarchiválási tevékenységek – nemzetközi és hazai körkép: <https://kifu.videotorium.hu/hu/recordings/42429>

## Előadás az ELTE Könyvtár- és Információtudományi Intézetében

Az Országos Széchényi Könyvtárral való együttműködési megállapodás keretében az ELTE Könyvtár- és Információtudományi Intézetében az alapképzésben zajló "Tájékoztatás, Információforrások 2." kurzus vendégeként Németh Márton, az OSZK Webarchiválási Osztályának munkatársa tartott előadást "Bevezetés a webarchiválásba – A webarchívumok működése, szolgáltatásai, az archivált webes anyagok szolgáltatásának alapelemei" címmel. A prezentáció innen tölthető le:

[https://webarchivum.oszk.hu/wp-content/uploads/2021/05/Bevezetes-a-Webarchivalasba-ELTE\\_KIT.pptx](https://webarchivum.oszk.hu/wp-content/uploads/2021/05/Bevezetes-a-Webarchivalasba-ELTE_KIT.pptx)

Az előadás keretében érintett témakörök:

- A webarchiválás alapjai
- Tájékoztató munka során hogyan tudják a hallgatók alkalmazni a webarchiválás eredményeit?
- Tájékoztató könyvtárosok milyen módon tudják támogatni a webarchiválási munkákat?
- Az intézet hallgatói hogyan tudnak részt venni az OSZK webarchiválási tevékenységében?

Az ELTE KIT blogján megjelent hír:

<http://elte-lis.blogspot.com/2021/05/nemeth-marton-vendegeloadasa.html>

## **BeSocial report**

Május elején megjelent a közösségi média archiválásával foglalkozó belga BeSocial projekt első összefoglaló jelentése „Towards a sustainable social media archiving strategy for Belgium: WP1 report : an international review of Social Media Archiving initiatives” címmel. A 2020-ban végzett kutatás több nemzeti könyvtárra (köztük a magyarra) terjedt ki és azt igyekezett felmérni, hogy hol milyen módon és milyen jogi környezetben válogatják, mentik, őrzik és szolgáltatják Twitter, Facebook, Instagram, Youtube és egyéb szervereken közzétett nyilvános tartalmakat. A jelentés tartalmazó szöveges PDF dokumentum mellett egy Google táblában közzétették az erre a célra használható speciális szoftverek jellemzőit is: <https://researchportal.unamur.be/en/publications/towards-a-sustainable-social-media-archiving-strategy-for-belgium>

<https://docs.google.com/spreadsheets/u/1/d/1nGuTC9Ww5yWZQ0wSUPPnIITMJBF1JyEaDOCO0Ve-09U/>

## **CHANSE projekt támogatás kérés**

Niels Brügger, az aarhusi egyetem professzora néhány munkatársával együtt azzal a kérdéssel fordult hozzánk, hogy járuljunk hozzá a webarchívumok tudományos célú hasznosítását vizsgáló nemzetközi projekt „Best friends forever: a European history of online friendship (from the 1980s to the 2020s)” című pályázatához egy szándéknyilatkozattal, hogy kutatásaik során használhatják az OSZK Webarchívum állományát is. A könyvtár vezetése kinyilvánította a segítő szándékát, a részletekről a pályázat remélhetőleg pozitív elbírálása után fogunk tárgyalni.

## **Ajánlatkérés az Internet Archive-től**

2018-ban már egyszer már váltottunk levelet az Internet Archive-val, melynek eredményeként konkrét üzleti ajánlatot tettek nekünk a náluk őrzött magyar anyag kereshetővé tételére, illetve a seed-listák átadására. Az OSZK részéről most jutottunk el odáig, hogy a szerződés megköthető lenne, ezért az ajánlat frissítését kértük az IA-tól. Egy udvarias hangú levélben máris tájékoztattak, hogy dolgoznak rajta és hogy nagyságrendileg nem fog látványosan nőni a három éve a szolgáltatások ellenértékéért felvázolt összeg.

## **IIPC Training Working Group online ötletbörze**

2021. május 18-án az IIPC Training Working Group aktuális összejevetelén ötletbörzét tartottunk a tervezett középhaladó tananyag témaköreinek felvázolására. Webkurátorok és informatikus szakemberek együtt vitatták meg a különféle igényeket, melyek a Google Jamboard nevű szolgáltató felületén digitális post-it címkék formájában is rögzítésre kerültek. A csoport vezetése a júniusi IIPC közgyűlésre előterjeszt egy munkatervet a további teendőkről. A Training Working Group weboldala:

<https://netpreserve.org/about-us/working-groups/training-working-group/>

## **#EWA4DH 2021 konferencia felhívás**

Az archivált webtartalmak adatelemzésére fókuszáló „Engaging with Web Archives 4 Digital Humanities” konferencia augusztus folyamán kerül megrendezésre az online térben. További részletek ezen az oldalon találhatóak: <https://ewaconference.com/ewa4dh-call-for-papers/ewa-postgrads-ecr/>

## **WARCNET projekt újdonságok**

Niels Brügger, a projekt főkoordinátora az irányító testülettel egyeztetve arról tájékoztatott minket, hogy a projektből hátralévő másfél éves időtartamra két újabb munkacsoport létrehozását kezdeményezik. Az egyik a webarchívumokban tárolt adatok kezelésével kapcsolatos eljárások, szabályzatok, szabványok összegyűjtésével foglalkozik majd. A másik pedig az EU Horizon programjához kapcsolódó új pályázati lehetőségek feltárására és pályázatok előkészítésére fog összpontosítani, hogy a webarchívumok tudo-

mányos célú hasznosításával kapcsolatos munkálatok folytatódhassanak a WARCNET projekt kifizése után is: <https://cc.au.dk/en/warcnet/>

### **Segítségkérés az IIPC levelezőlistáján a válogatási gyakorlatról**

Május 25-én az IIPC levelezőlistáján kéréssel fordultunk nemzetközi partnereink felé, azt tudakolva, hogy ők milyen elveket követnek az archiválandó webhelyek manuálisan válogatása során. Még ugyanaznap részletes áttekintést kaptunk a British Library-ből, a finn és a svájci nemzeti könyvtáraktól, valamint a pozsonyi Egyetemi Könyvtártól, később pedig a washingtoni Kongresszusi Könyvtártól is, a válogatást meghatározó jogszabályokról és az intézményükben folytatott gyakorlatról. Ezeket az információkat a saját gyűjtőkori szabályzatunk összeállítása során szeretnénk felhasználni.

### **Együtműködés kialakítása könyvtárakkal**

A webarchiválás komplex jellege és szerteágazó feladatai miatt célszerű más memória intézményeket is bevonni a tevékenységbe úgy, hogy ennek eredménye mindkét fél számára hasznos legyen (<http://ojs.elte.hu/kf/article/view/2297>). Májusban érkezett végre az a pillanat, amikor már úgy tudunk megbeszéléseket folytatni érdeklődő partnerekkel, hogy nem csak általánosságban volt szó a különféle lehetőségekről, hanem a konkrét terveket, feladatokat tudtunk egyeztetni, és ezeket alapján előkészíteni az intézmények közötti megállapodás tervezetét. Első körben azokat a megyei hatókörű könyvtárakat kerestük meg, akik a tavalyi KDS lezárultakor jelezték jövőbeni együttműködési szándékukat, de szeretnénk megszólítani a többi megyei könyvtárat is a közeljövőben, távolilag pedig szakkönyvtárakat és más közgyűjteményeket is. Elsősorban öt archiválási részfeladatban számítunk partnereink közreműködésére: 1. gyűjtőkörükhöz illeszkedő gyűjtemény kialakítása és gondozása; 2. minőségellenőrzés; 3. metaadatolás; 4. közreműködés a nyilvános szolgáltatáshoz szükséges engedélyeztetési folyamatban és végül 5. kisebb, nem hardverigényes, inkább élőmunkát igénylő archiválási feladatok végzése (pl. web2-es tartalmak mentése). Megyei könyvtárak esetében kézenfekvő területi elv szerint helyi gyűjteményt kialakítani, míg egy szakkönyvtárnál annak specializációja mentén. Ezek az OSZK Webarchívumban virtuális gyűjteményként szerepelnének, vagyis az összegyűjtött webhelyek a tematikus aratásokban kerülnek archiválásra, önálló gyűjteményként történő mentésre csak a partnerek számára kerülne sor, saját archívum kialakítása céljából. Az összegyűjtött állomány mindkét fél katalógusában szerepelne adatszinten, és hosszú távon célszerű lenne a nyilvánosan nem szolgáltatható anyag helyi hozzáférést is kialakítani. Az első három egyeztetés a kecskeméti Katona József Könyvtár, az egri Bródy Sándor Megyei és Városi Könyvtár és a Fővárosi Szabó Ervin Könyvtár kollégáival történt a héten. Utóbbi két okból is érdekes, egyrészt mert az FSZEK nem csak regionális vonatkozásban, hanem mint szociológiai szakkönyvtár is tud gyűjteményt létrehozni és gondozni, másrészt külön öröm, hogy ők keresték meg a Webarchívumot az együttműködés szándékával.

### **Az elmúlt hetekben lefutott aratások**

Képző-, előadó-, zene- és filmművészet (7552 db seed URL)

Irodalom, irodalomtudomány és -történet (1259 db seed URL)

Könyvtárak, levéltárak, múzeumok és galériák (1544 db seed URL)

Történelem, hely- és családtörténet (541 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: [mia@mek.oszk.hu](mailto:mia@mek.oszk.hu)