404 Not Found – Ki őrzi meg az internetet? Konferencia és workshop 2025

2025/11/12
2025/10/29

Az OSZK Digitális Bölcsészeti Központja **november 12-én** rendezi meg digitális megőrzéssel foglalkozó éves rendezvényét, melynek kiemelt témája a webarchívum mint adat lesz.

[1]

A Magyar Nemzeti Múzeum Közgyűjteményi Központ Országos Széchényi Könyvtár (a továbbiakban: MNMKK OSZK) Digitális Bölcsészeti Központja 2025. november 12-én tartja meg az internetes tartalmak archiválásával, tágabban a digitális megőrzéssel foglalkozó éves rendezvényét, az immár kilencedik ***404 Not Found – Ki őrzi meg az internetet?*** **című konferenciát és workshopot.**

Időpont: **2025. november 12.**
Helyszín: MNMKK Országos Széchényi Könyvtár
(Budavári Palota F épület, 1014 Budapest, Szent György tér 4–5–6.) [2]
Fb-esemény [3]

A konferencia és workshop **jelenléti** formában valósul meg,
de az érdeklődők **online** is bekapcsolódhatnak az élő közvetítésbe
az MNMKK OSZK YouTube csatornáján. [4]

A rendezvény ingyenes, de regisztrációhoz kötött.
Regisztráció: https://oszk.hu/404-not-found-2025-regisztracio [5]

Minden érdeklődőt szeretettel várunk!

Az idei rendezvény kiemelt témája a webarchívum mint adat, a tervezett előadások majd mindegyike ehhez kapcsolódik.
A Skót Nemzeti Könyvtárral közösen elnyert CENL-pályázat keretében itt tartózkodó kollégák, valamint a könyvtár kutató ösztöndíjasa az általuk végzett tevékenységről és a skót webarchívum metaadatainak hasznosulásáról beszélnek. A Nemzeti Audiovizuális Archívum, valamint az MTA Könyvtár és Információs Központ a webtartalom archiválásának speciális változatait mutatják be. Az ELTE Nyelvtudományi Kutatóközpont az archivált webes tartalom magyar nyelvi modell fejlesztésében való felhasználásáról tart előadást.
Végül az MNMKK OSZK munkatársai a webarchiválási csoport éves tevékenységéről, valamint a hírportálok napi szintű mentésének megújításáról tájékoztatják a résztvevőket. Az utóbbi témához kapcsolódik a délutáni workshop, amelynek keretében a Browsertrix crawler használatába vezetik be az érdeklődőket.

Előadások
Workshop
Előadások címe és rövid leírása
Workshop címe és rövid leírása
Technikai segítség

ELŐADÁSOK

10.00 Köszöntők:
Gerencsér Judit (MNMKK OSZK), általános főigazgató-helyettes
Latorcai Csaba (KTM), parlamenti államtitkár,
közigazgatási és területfejlesztési miniszterhelyettes

10.20 MacGlone, Eilidh (NLS) – Marshall, Joseph (NLS):
National Library of Scotland: Our web archive (1999–)

10.50 Kocsis Andrea (UE):
Hozzáférés engedélyezve: Kreatív utakon a webarchívumokhoz

11.10 Bajnok László (NAVA):
A NAVA archiválási tevékenysége – Audiovizuális dokumentumok katalogizálása a NAVA-ban

11.30 Kávészünet

11.50 Holl András (MTA KIK):
Repozitóriumok – a digitális információk archiválásának eszközei

12.10 Ligeti-Nagy Noémi (ELTE NYTK):
Milyen magyar nyelvet tanul meg a mesterséges intelligencia? –
Az internet tartalma mint tanítóanyag

12.30 Drótos László (MNMKK OSZK):
Adatok az adatokról – Az MNMKK OSZK webarchívuma 2025-ben

12.50 Ebédszünet

13.40 Kalcsó Gyula (MNMKK OSZK):
Hírportálok napi mentése Browsertrix crawlerrel

WORKSHOP

14.00 Kalcsó Gyula (MNMKK OSZK):
Az archiválás új szintje – dinamikusan generált, interaktív webhelyek mentése böngészőalapú crawlerrel (Browsertrix)

15.30 A konferencia zárása: kérdések, hozzászólások, vita

ELŐADÁSOK CÍME ÉS RÖVID LEÍRÁSA

MacGlone, Eilidh (NLS) – Marshall, Joseph (NLS): *National Library of Scotland: Our web archive (1999–)*
A Skót Nemzeti Könyvtár egy kötelespéldány-könyvtár, amelynek székhelye Edinburgh-ban és Glasgow-ban található, az Egyesült Királyságban. 2013-ban a Kötelespéldány-könyvtárak (Nem nyomtatott művek) szabályzat felhatalmazott minket arra, hogy az engedélyeken alapuló gyűjteményünk mellett weboldalakat is kötelespéldányként archiváljunk. Ez lehetővé tette számunkra, hogy folytassuk munkánkat a skóciai online független nyomtatás és magánkiadás hagyományának megőrzése érdekében. Az azóta eltelt években kidolgoztuk saját megközelítésünket, gyakran kutatókkal és más könyvtárakkal együttműködve. Ez az előadás gyűjteményeinkből vett példákon keresztül mutatja be, hogyan jutottunk el a munkafolyamatainkhoz és módszereinkhez ebben a jogi keretrendszerben.

Kocsis Andrea (UE): *Hozzáférés engedélyezve: Kreatív utakon a webarchívumokhoz*
Az előadás három eltérő webarchívum-felhasználói csoport bevonásának lehetőségeit vizsgálja: az általános olvasókat, az adathasználókat és a „digitálisan kíváncsiakat” (digitally curious). A University of Edinburgh, a Cambridge University Libraries és a National Library of Scotland együttműködésében végzett kutatások példáira építve az előadás ajánlásokat fogalmaz meg a széles körű hozzáférést szem előtt tartó webarchívum-programokhoz.
Először az általános olvasók élményének gazdagítása érdekében játékosított (gamified) felületeket, valamint közösségi ismeretterjesztő eseményeket – például kiállítást – alkalmazunk, hogy növeljük a könyvtárhasználók tudatosságát és kíváncsiságát a rendelkezésre álló webarchívum-források iránt. Másodszor az adathasználói közösség igényeire reagálva a metadata datasets és a Datasheets for Data dokumentáció gondos kurálására helyezzük a hangsúlyt, ösztönözve a webarchívum-gyűjteményeken végzett kvantitatív kutatásokat. E munka része olyan közösségi kutatói események szervezése is, mint az adatvizualizáció vagy pilot study felhívások, amelyek eredményei később az általános olvasók számára elérhető forrásokba is beépülhetnek.
Végül a digitális kompetenciák hiányából fakadó akadályok leküzdése érdekében olyan forrásokat fejlesztettünk, amelyek kifejezetten a „digitálisan kíváncsi” közönségnek szólnak – azoknak, akik felismerik a webarchívumokban rejlő kvantitatív lehetőségeket, de nem rendelkeznek fejlett számítástechnikai ismeretekkel. Feltételezésünk szerint e csoport digitális készségeinek fejlesztése hozzájárulhat ahhoz, hogy érdeklődésük a webarchívumok felfedezése és használata felé nyíljon.
A bemutatott megközelítéseket két esettanulmányon keresztül szemléltetem: az Archive of Tomorrow és a Digital Ghost projektekkel.

Bajnok László (NAVA): *A NAVA archiválási tevékenysége –Audiovizuális dokumentumok katalogizálása a NAVA-ban*
Annak érdekében, hogy a jövő nemzedékei korhű képet kapjanak a jelen és a múlt eseményeiről, társadalmáról és kultúrájáról, a rádiós és televíziós műsorszámok gyűjtése, nyilvántartása, megőrzése és a nyilvánosság számára történő hozzáférhetővé tétele céljából az Országgyűlés

megalkotta a NAVA-törvényt. A Nemzeti Audiovizuális Archívum 2006. január elsején kezdte meg üzemszerű működését. Húsz éve gyűjti, digitálisan tárolja, katalogizálja a műsorszolgáltatók által sugárzott magyar gyártású vagy magyar vonatkozású műsorszámokat. A katalógusa nyilvános, a videók nagyobb része oktatási és kutatási célra dedikált hálózaton keresztül érhető el.

Holl András (MTA KIK): ***Repozitóriumok – a digitális információk archiválásának eszközei***
A 2019-es 404 Not Found – Ki őrzi meg az internetet? konferencián tartott előadásom (noWebarchive – Amikor nem a webarchiválás a megfelelő módszer a web archiválására) alapvetését követve, a 2025-ös Networkshopon elhangzott, az Informatikatörténeti Fórum webes adatbázisának archiválásával foglalkozó előadás nyomán számolok be komplex webes tartalmak repozitóriumi archiválásáról. Gyakorta a web dokumentumok hozzáférhetőségének és logikai rendszerbe szervezésének eszköze, ezen információk megőrzésének egyik eszköze a webarchiválás. Ugyanezen dokumentumokat és a kontextusukat megteremtő, ám további információkat is hordozó HTML-oldalakat tervezetten repozitóriumba lehet helyezni, alternatívát kínálva a webarchiválásnak. Az előadás egyúttal megemlékezés Dömölki Bálintról, a NWS2025-ös előadás első szerzőjéről.

Ligeti-Nagy Noémi (ELTE NYTK): ***Milyen magyar nyelvet tanul meg a mesterséges intelligencia? – Az internet tartalma mint tanítóanyag***
Az ELTE Nyelvtudományi Kutatóközpontjának Nyelvtechnológiai Kutatócsoportja több éve foglalkozik nyelvmodellek fejlesztésével. A PULI nyelvmodellcsalád tagjai között találunk egynyelvű modellt és háromnyelvű modellt, generatív alapmodellt és chatmodellt, GPT-alapú és Llama-alapú modellt; közös azonban ezekben a különös gonddal előállított magyar nyelvű tanítóanyag. Előadásomban bemutatom, honnan és hogyan gyűjtjük a mára már 70 milliárd szavasra duzzadt tanítókorpuszunkhoz az adatokat; milyen kihívásokkal szembesül az, aki nekiáll az internetes szöveggyűjtésnek; és milyen adattisztítási lépésekre van szükség ezeknél az adatoknál. A fentiek átgondolt tervezése és kivitelezése tette azt lehetővé, hogy a legújabb modellünk, a ChatPULI, a GPT-modelleknél is fluensebben, hibátlanul tudjon magyarul „beszélni”.

Drótos László (MNMKK OSZK): ***Adatok az adatokról – Az MNMKK OSZK webarchívuma 2025-ben***
A nemzeti könyvtár Digitális Filológiai és Webarchiválási Osztályán folyó archiválási tevékenység eredményeképpen 2025-ben is több terabájttal nőtt a letöltött webtartalom. Az előadás számokkal és grafikonokkal illusztrálva mutatja be az újonnan létrehozott, illetve a jelentősen bővített részgyűjteményeket, és külön kitér az idei első webtérszintű aratás előkészítő munkálataira, valamint annak eredményére. A Heritrix robottal való archiválás mellett sok gigabájtnyi mentés készült a Browsertrix Crawler nevű programmal és nagy mennyiségű a félautomatikus megoldásokkal letöltött közösségimédia-tartalom is. A nyilvános gyűjtemény esetében több száz részletes metaadat-leírás született a Digitális Képarchívum munkatársainak köszönhetően. Az informatikai infrastruktúra és a munkafolyamatok megújításának részeként pedig elkészült annak az adatbázisnak a terve, amely az archiválandó és az archivált tartalmak adatainak nyilvántartása mellett egyes részfeladatok automatizált vezérlésére is használható lesz.

Kalcsó Gyula (MNMKK OSZK): ***Hírportálok napi mentése Browsertrix crawlerrel***
Az előadás bemutatja az MNMKK OSZK DBK Digitális Filológiai és Webarchiválási Osztály webarchiválási csoportjának új eljárását válogatott magyar hírportálok napi szintű mentésére. A webarchiválásban használt aratórobotunk, a Heritrix bizonyos típusú tartalmakat gyengébb minőségben tud menteni, ugyanakkor sokkal kisebb az erőforrásigénye, ezért a tömeges aratásokban célszerűbb használni. A hírportálok dinamikus vagy felhasználói interakciót igénylő oldalait böngészőalapú crawlerrel pontosabban lehet archiválni. Erre a célra a Webrecorder-projekt Browsertrix nevű crawlerét használjuk, amelyet az egyes portálok esetén célzottan konfigurálunk annak érdekében, hogy megfelelő minőségű mentésekhez jussunk. Az előadás bemutat néhány példát arra, hogy milyen beállításokat kell alkalmazni a cél elérése érdekében, és hogyan történik a hírportálok címlapjának képernyőképpel történő mentése, valamint az aktuális napi hírek megtalálása és archiválása. Bemutatja azokat a felhasználói interakciókat szimuláló szkripteket is, amelyek segítségével egyes portálokon a görgetést és lapozást igénylő hírfolyamokat is menteni tudjuk.

WORKSHOP CÍME ÉS RÖVID LEÍRÁSA

Kalcsó Gyula (MNMKK OSZK): ***Az archiválás új szintje – dinamikusan generált, interaktív webhelyek mentése böngészőalapú crawlerrel (Browsertrix)***
A workshop célja, hogy bemutassa a résztvevőknek a Browsertrix crawler használatának alapjait. A Browsertrix böngészőalapú eszközként képes összetettebb, modern felépítésű webhelyek jó minőségű archiválására. Azok a webhelyek, amelyek felhasználói interaktivitás függvényében építik fel a böngészőben megjelenő struktúrájukat, hagyományosabb webarchiváló eszközökkel nem, vagy csak gyenge minőségben menthetők. A workshop résztvevői saját eszközeiken kipróbálhatják a Browsertrix telepítésének és konfigurálásának a lépéseit, és konkrét webhelyek próbamentésén és visszajátszásán keresztül tapasztalatot szerezhetnek a crawler használatában, ezáltal képessé válnak önálló mentések készítésére is.

TECHNIKAI SEGÍTSÉG

Amennyiben interaktívan szeretne részt venni a workshopon, javasoljuk, hogy eszközére telepítse fel a Docker Desktopot az alábbiaknak megfelelően:

Windows 10/11:

Töltse le a Docker Desktop szoftvert innen: https://www.docker.com/products/docker-desktop/ [6] (a legtöbb gépen a Download for Windows AMD64 letöltési linken található verzió lesz a megfelelő).

Futtassa a letöltött telepítőfájlt (rendszergazdai jogosultság szükséges). Telepítés után indítsa el, és ellenőrizze, hogy működik-e: nyisson egy Powershellt vagy parancssort (Win + X, Parancssor vagy Parancssor (rendszergazdai)), és írja be, majd nyomjon Entert:

docker version

Ha a telepítés sikeres volt, akkor a képernyőn a Docker Desktop szoftverre vonatkozó információkat fog látni.

MacOS Monterey (12)/Ventura (13)/Sonoma (14):

Töltse le a Docker Desktop szoftvert innen: https://www.docker.com/products/docker-desktop/ [6] (válassza a processzorának megfelelő verziót: Mac with Apple silicon/Mac with Intel chip).

Nyissa meg a letöltött .dmg fájlt. Húzza a Docker ikont az Applications mappába, indítsa el a Docker.appot az Applicationsből. Az első indításkor kérhet rendszergazdai jelszót, kérheti, hogy engedélyezze a „privileged access”-t (ez kell a konténerek futtatásához), illetőleg előfordulhat, hogy újra kell indítani a Docker Desktopot.

Az Applications (Alkalmazások) mappából indítsa el a Terminalt. Írja be, majd üssön Entert:

docker version

Ha a telepítés sikeres volt, akkor a képernyőn a Docker Desktop szoftverre vonatkozó információkat fog látni.
2025/11/09 - 13:12

Forrás webcím: https://oszk.hu/rendezvenyek/404-not-found-ki-orzi-meg-az-internetet-konferencia-es-workshop-251112

Hivatkozások:
[1] https://oszk.hu/sites/default/files/404_2025_FB_jav.jpg
[2] https://www.google.hu/maps/place/Orsz%C3%A1gos+Sz%C3%A9ch%C3%A9nyi+K%C3%B6nyvt

%C3%A1r/@47.4954231,19.0370263,17z/data=!3m1!4b1!4m6!3m5!1s0x4741dc3d2cbe10bd:0x95a7b3e3f4b9f737!8m2!3d47.4954231!4d19.0396012!16zL20vMDZycnls?hl=hu&entry=ttu&g_ep=EgoyMDI0MTEyNC4xIKXMDSoASAFQAw%3D%3D
[3] https://www.facebook.com/events/822274537166413
[4] https://www.youtube.com/@nemzetikonyvtar/streams
[5] https://oszk.hu/404-not-found-2025-regisztracio
[6] https://www.docker.com/products/docker-desktop/
[7] https://oszk.hu/category/foszotar-es-pozicionalo/hirek
[8] https://oszk.hu/category/foszotar-es-pozicionalo/rendezvenyek