404 Not Found – Ki őrzi meg az internetet?

2023/11/29
2023/11/10

Az OSZK Digitális Bölcsészeti Központja (DBK) **2023. november 29-én** tartja meg az internetes tartalmak archiválásával foglalkozó éves rendezvényét, a hetedik *404 Not Found – Ki őrzi meg az internetet?* konferenciát és workshopot.

[1]

Az OSZK Digitális Bölcsészeti Központja (DBK) hetedik alkalommal tartja meg az internetes tartalmak archiválásával foglalkozó éves rendezvényét, a *404 Not Found – Ki őrzi meg az internetet?* konferenciát és workshopot. Minden érdeklődőt szeretettel várunk!

Időpont: **2023. november 29., szerda 10 óra**
Helyszín: **OSZK, Északi olvasóterem**
(Budavári Palota F épület, 1014 Budapest, Szent György tér 4–5–6.)

Facebook-esemény [2]

A részvétel ingyenes, de regisztrációhoz van kötve. Az érdeklődők online is bekapcsolódhatnak az élő közvetítésbe.
A jelentkezés lezárult, köszönjük az érdeklődést!

Az esemény középpontjában a nemzeti könyvtár webarchiválási tevékenységének megújítása, újfajta eszközök, technológiák és az új intézményi kapcsolatok bemutatása áll. A DBK ismerteti a webarchívumhoz kapcsolódó legújabb tevékenységeket, a célzott adatmentést scrapingmódszerrel, a mesterséges intelligencia használatát, a Karikó-gyűjteményt. A konferencia egyúttal záróeseménye a Luxembourgi Nemzeti Könyvtárral közösen elnyert nemzetközi pályázatnak: előadáson mutatják be tevékenységüket, valamint workshopot tartanak a használt technológiákról.

A KONFERENCIA PROGRAMJA:

Előadások

10.00 Köszöntők

10.20 Luxemburgi webarchiválás
Ben Els (BNL): *Curatorial Aspects of The Luxembourg Web Archive*
László Tóth (BNL): *Technical Aspects of The Luxembourg Web Archive*

11.20 Drótos László (OSZK): *Az OSZK webarchívumának megújítása*

11.40 Kávészünet

12.00 Kiss Márta Éva – Pálfy Anna (SZTE EK): Megvalósult álmok – Helyzetjelentés a szegedi Karikó-webarchiválásról

12.20 Kalcsó Gyula (OSZK): *A scrapingtechnológia használata és helye a webarchiválásban*

12.40 Simon Eszter: *A webaratásból származó szövegek automatikus feldolgozása*

13.00 Ebédszünet

Workshop

14.00 László Tóth (BNL): *Browsertrix Cloud*

15.00 Ben Els (BNL): *From Luxemburgensia to Hungarica – Using AI, we follow the traces of Hungarian culture through the BnL's collections of digitised newspapers and web archives.*

Az előadások rövid leírása:

Ben Els (BNL): *Curatorial Aspects of The Luxembourg Web Archive*

Since 2016, the National Library of Luxembourg [3] preserves the Luxembourg web under national legal deposit. The Internet is changing at a rapid pace and there are a lot of obstacles in providing the best possible coverage for all websites in the Luxembourg web sphere. Archiving institutions have to balance between limited resources, in terms of budget, technical capacities and manpower, while also facing challenges in terms of tool development and accessibility for different user groups. This presentation will cover the operating modes and types of collections of the Luxembourg Web Archive [4], team setup and contracted services, our collection policy and plans for collaborative curation. We will have a look at the particularities of the Luxembourg legal deposit and its impact on the launch of a web archiving program at the National Library. We will cover different thematic and event collections and how they are presented on our information and participation platform webarchive.lu. Moreover, we will take a look at what examples of the Hungarian language, Hungarian websites and the Hungarian community in Luxembourg can be discovered in our current collections. We will dive into the different search options that help us to explore large web archive datasets.

László Tóth (BNL): *Technical Aspects of The Luxembourg Web Archive*

In this presentation, we will detail the Luxembourg web archive from a technical viewpoint. We will discuss our harvesting methods (seasonal, thematic, and behind-the-paywall harvests), our technical infrastructure (servers, configurations) as well as various statistics of our web archives. In particular, we will present our in-house Browsertrix-based crawler, that allows us to perform cross-crawl deduplication, i.e., harvesting only the resources that were not already harvested during previous campaigns. Our method allows us to harvest an entire website domain in less than an hour and, in some cases, as little as 5 to 10 minutes. This will be followed by a short description of our current, as well as new, hardware. Indeed, the BnL has recently upgraded its web archiving hardware, including 4 high-performance servers with 96 cores and 768 GB RAM each. These machines will host our new indexing and playback solutions, comprising of a hybrid system including PyWb, OutbackCDX and a Solr cluster used by SolrWayback. Our indexers are set to work non-stop for about three weeks to index our entire web archive of 300 TB WARC files into Solr, thus enabling full-text search and other advanced features. Finally, we will briefly speak about the migration of our entire web archives into a new storage based on IBM S3 object storage solution, and the subsequent adaptation of existing software such as PyWb and SolrWayback to be able to efficiently load WARC data directly from S3 buckets.

Drótos László (OSZK): *Az OSZK webarchívumának megújítása*

Az OSZK webarchiválási tevékenysége 2017-ben indult, az azóta eltelt hat év tapasztalatai alapján megérett a helyzet az organikusan kialakult rendszer újragondolására. Az előadás a 2023. évi helyzetjelentés mellett bemutatja a webarchívum 2.0-s verziójának tervezetét, melynél a főbb szempontok a következők: a munkafolyamatok automatizálása, az archiválás minőségének javítása, a metaadatok egységes nyilvántartása, a gyűjtemény kutatásra és hosszú távú megőrzésre alkalmassá tétele.

Kiss Márta Éva – Pálfy Anna (SZTE EK): *Megvalósult álmok – Helyzetjelentés a szegedi Karikó-webarchiválásról*

A tavalyi 404-konferencián dr. Kokas Károly *A virtuális kiállítástól az archiválásig: Karikó Katalin nyomában* című előadásában beszélt az SZTE Klebelsberg Könyvtár [5] Karikó-gyűjteményéről és a tervezett webarchívum kialakításának főbb pontjairól. Azóta eltelt majd egy év, az együttműködés

létrejött, és elindult az archiválási tevékenység is. Idei előadásunkban a virtuális kiállítás továbbfejlesztéséről, az eltelt évben végzett munkáról és a jövőbeli feladatokról hallanak majd. Szó lesz az új és lehetséges kihívásokról is, melyekkel találkozhatunk a frissen Nobel-díjat nyerő Karikó Katalin személyéhez kapcsolódó gyűjtemény továbbépítése során.

Kalcsó Gyula (OSZK): *A scrapingtechnológia használata és helye a webarchiválásban*

Az OSZK webarchiválási tevékenysége elsősorban tömeges webaratást jelent a Heritrix szoftverrel, ezenkívül azonban kisebb mennyiségben, de minél jobb minőségre törekedve mentünk egyedi webhelyeket, webhelyrészeket vagy akár egyedi weboldalakat és egyéb fájlokat is. Ugyanakkor vannak olyan esetek, amikor nem lehet vagy nem szükséges az eredeti felületet archiválni és megjeleníteni, hanem elegendő csak a releváns tartalmat és bizonyos metaadatokat begyűjteni úgynevezett web scraping módszerrel. Ilyen feladat lehet például a webarchívum podcastállományának bővítése, cikkek begyűjtése szövegkorpuszok építéséhez vagy a DKA gyarapítása szabadon felhasználható digitális fotókkal. Az előadás az első ilyen nagyobb scrapingprojektünket mutatja be a kozterkep.hu [6] példáján keresztül.

Simon Eszter: *A webaratásból származó szövegek automatikus feldolgozása*

A webaratásból a sok egyéb formátum mellett nagy mennyiségű szöveges anyag is keletkezik. Ezen az anyagon a természetesnyelv-feldolgozás (natural language processing, NLP) eszközeit alkalmazva egy hatalmas méretű szövegkorpusz jön létre, amiből egyrészt sok hasznos és érdekes adatot lehet kinyerni, másrészt bemenete és segédeszköze lehet további nyelvfeldolgozó lépéseknek. Előadásomban a webaratásból származó szövegek feldolgozásának lépéseit mutatom be, valamint felvázolom azokat a jövőbeli fejlesztési irányokat, amelyeket tervezünk a webarchívumon csinálni. Utóbbiak közé tartozik az automatikus tárgyszavazás és topikmodellezés, valamint nagy nyelvi modellek (large language models, LLMs) tanítása is.

Kapcsolódó tartalmak:

OSZK Webarchívum [7]

Az OSZK és a Szegedi Tudományegyetem kutatási és könyvtári együttműködése a humanizmus korától a mesterséges intelligenciáig [8]

2023/11/20 - 15:30

Forrás webcím: https://oszk.hu/rendezvenyek/404-not-found-ki-orzi-meg-az-internetet-23

Hivatkozások:
[1] https://oszk.hu/sites/default/files/404-not-found_KONFERENCIA_231129_fb.jpg
[2] https://www.facebook.com/events/151524918022393/?ref=newsfeed
[3] https://bnl.public.lu/en.html
[4] https://www.webarchive.lu/
[5] http://www.ek.szte.hu/
[6] http://kozterkep.hu
[7] https://webarchivum.oszk.hu/
[8] https://www.oszk.hu/hirek/az-orszagos-szechenyi-konyvtar-es-szegedi-tudomanyegyetem
[9] https://oszk.hu/category/foszotar-es-pozicionalo/hirek

[10] https://oszk.hu/category/foszotar-es-pozicionalo/rendezvenyek