404 Not Found – Who Will Preserve the Internet?

2023/11/29
2023/11/10

On 29 November 2023, the National Széchényi Library's (OSZK) Digital Humanities Centre (DBK) will hold its seventh annual 404 Not Found – Who Will Preserve the Internet? conference and workshop on archiving content on the Internet.

[1]

For the seventh time, the NSZL's Digital Humanities Centre (DBK) is holding its annual event on archiving content on the Internet, the 404 Not Found – Who Preserves the Internet? conference and workshop. Everyone is welcome to attend!

Date: **29 November 2023, Wednesday 10 am**
Location: **NSZL, Northern Reading Room**
(Building F of the Buda Castle Palace, Szent György tér 4–5–6., 1014 Budapest)

Attendance is free, but is subject to prior registration.
Interested parties can also join the live stream online.

Registration is now closed, thank you for all the interest!

The event will focus on the renewal of the national library's web archiving activities, new tools, technologies and new institutional partnerships. The DBK will present the latest activities related to web archiving, targeted data retrieval using scraping techniques, the use of artificial intelligence, the Karikó collection. The conference will also be the closing event of an international call for proposals with the National Library of Luxembourg, which will include a presentation and a workshop on the technologies involved.

Presentations

10.00 Welcome speeches

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

11.20 László Drótos (OSZK): *Renewing the OSZK Web Archive*

11.40 Coffee break

12.00 Márta Éva Kiss – Anna Pálfy (University of Szeged Klebelsberg Library): *Dreams Come True – Progress Report on the Karikó Web Archiving in Szeged*

12.20 Gyula Kalcsó (OSZK): *The Use and Role of Scraping Technology in Web Archiving*

12.40 Eszter Simon (OSZK): *Automatic Processing of Texts Resulting from Web Harvesting*

13.00 Lunch break

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.

Title and abstract of the presentations

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

Since 2016, the National Library of Luxembourg [2] 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 [3], 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.

László Drótos (OSZK): *Renewing the OSZK Web Archive*

The web archiving activity of the OSZK was launched in 2017, and based on the experience of the six years since then, the situation is ripe for rethinking the organically developed system. The presentation will introduce the draft version 2.0 of the web archive, alongside the 2023 progress report, with the main aspects of automating workflows, improving the quality of archiving, a unified metadata inventory, and making the collection suitable for research and long-term preservation.

Márta Éva Kiss – Anna Pálfy (SZTE EK): *Dreams Come True – Progress Report on the Karikó Web Archiving in Szeged*

At last year's 404 conference, Dr. Károly Kokas spoke about the Karikó collection of the SZTE Klebelsberg Library and the main points of the planned web archive in his presentation entitled From virtual exhibition to archiving: in the footsteps of Katalin Karikó. Almost a year has passed since then, the collaboration has been established and archiving activities have commenced. In this year's presentation, you will hear about the further development of the virtual exhibition, the work done in the past year and upcoming tasks. We will also discuss the new and potential challenges that we will face in continuing to build the collection associated with Katalin Karikó, the recipient of the latest Nobel Prize.

Gyula Kalcsó (OSZK): *The Use and Role of Scraping Technology in Web Archiving*

The web archiving activities of the OSZK mainly consist of bulk web archiving with the Heritrix software, but we also save individual websites, parts of websites or even individual webpages and other files in smaller quantities, but with the aim of achieving the highest possible quality. At the same time, there are cases where it is not possible or necessary to archive and display the original interface, it is sufficient to collect only the relevant content and some metadata by a method called web scraping. Such a task could be, for example, to extend the podcast collection of the web archive, to collect articles for building text corpora or to enrich the digital photo archive (DKA) with freely usable digital photos. The presentation will demonstrate the first of these major scraping projects using the example of kozterkep.hu [4].

Eszter Simon (OSZK): *Automatic Processing of Texts Resulting from Web Harvesting*

In addition to the many other formats, web harvesting also results in a large amount of text. Using natural language processing (NLP) tools on this material, a huge corpus of text is created, from which a lot of valuable and interesting data can be extracted, and which can also serve as input and aid for further language processing steps. In my presentation, I will demonstrate the steps of processing texts resulting from web harvesting and outline the future directions of development that we plan to implement in the web archive. The latter include automated subject indexing and topic modelling, as well as the teaching of large language models (LLMs).

2023/11/25 - 15:30

Links:
[1] https://oszk.hu/sites/default/files/404-not-found_KONFERENCIA_231129_ENG_program.jpg
[2] https://bnl.public.lu/en.html
[3] https://www.webarchive.lu/
[4] http://kozterkep.hu
[5] https://oszk.hu/en/category/foszotar-es-pozicionalo/hirek
[6] https://oszk.hu/en/category/foszotar-es-pozicionalo/rendezvenyek