



Improving the Web Archiving Infrastructure of the National Széchényi Library and the National Library of Luxembourg

**CENL Erland Kolding Nielsen Grant Project Report
by Dávid Rózsa, General Director of NSZL**

Annual General Meeting, Warsaw, 16–18 June 2024

BACKGROUND

- The National Széchényi Library started its **web archiving activities in 2017**
- Digital Humanities Centre Department of **Digital Philology and Web Archiving**
- **Subcollections:**
 - Thematic subcollections (approx. 100,000 sites)
 - National web (approx. 1.3 million domains)
 - Event-based and other special subcollections (1400 Transcarpathian sites)
- The overall size is more than **100 terabytes**
- Stored in standard **WARC** format

APPLICATION

- **10th of March 2023:** joint application by the National Széchényi Library and the National Library of Luxembourg (Bibliothèque Nationale du Luxembourg, BNL)
- **Aim:** improving the web archiving infrastructure of the NSZL and the BNL
- **The collaboration intended to advance four of the objectives** supported by the Erland Kolding Nielsen Grant:
 - Infrastructure development
 - Development of digital collections
 - Improving research
 - Promoting interlibrary staff communications

CONTRIBUTIONS PROPOSED

- Visiting each other to **share knowledge**
- Organising a **conference and workshops** to present our activity to each other
- **Communicating** the outcomes of the collaboration to the international community
- **Sharing code** with each other and with the international community

NSZL STAFF VISIT IN LUXEMBOURG



BNL STAFF VISIT IN BUDAPEST



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
11:40 **László Drótos (NSZL):** Renewing the NSZL Web Archive
11:40 Coffee break
12:00 **Márta Éva Kiss - Anna Pálffy (USZ KLI):** Dreams Come True - Progress Report on the Karió Web Archiving in Sarged
12:20 **Gyula Kalcsó (NSZL):** The Use and Role of Scraping Technology in Web Archiving
12:40 **Eszter Simen:** 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 Luxembourg to Hungary - Using AI, we follow the traces of Hungarian culture through the BNL's collections of digitised newspapers and web archives

REGISTRATION



LATEST CONFERENCES



OUTCOMES

- Our main objective, to improve the infrastructure of the two web archives, has been fully achieved: BNL has installed a new instance of SolrWayback, NSZL started using Browsertrix.
- We have shared a lot of knowledge and good practices that we can incorporate into both institution's web archiving activities (e.g. configuring SolrWayback, harvesting behind-paywall news sites).
- During the visits, we have built up working partnerships that promise to go well beyond the scope of the grant cooperation (a new cooperation between the University of Luxembourg and NSZL).
- We have also delivered on our commitment to code-sharing by exchanging and open-sourcing important code that will greatly contribute to the significant advancement of the web archiving activities of both national libraries (links to shared code on the last slide).

Thank You for Your Attention!

Blog post: <https://tinyurl.com/yda9ey3j>

Conference and Workshop Web Site:
<https://tinyurl.com/53kuhxdc>

Shared Code:

<https://tinyurl.com/y3ae9xz5>

<https://tinyurl.com/tndwhasf>

<https://tinyurl.com/47y2jx9f>