

# Quo vadis digitised newspapers and radio? Next steps for the integration of western European collections via *impresso* II.

## Bunout, Estelle

estelle.bunout@uni.lu

Universität Luxemburg, Luxemburg, UL

## Düring, Marten

marten.during@uni.lu

Universität Luxemburg, Luxemburg, UL

## Clematide, Simon

simon.clematide@uzh.ch

Universität Zürich, UZH, Schweiz

## Ehrmann, Maud

maud.ehrmann@epfl.ch

Ecole Polytechnique Fédérale de Lausanne, EPFL, Schweiz

## Guido, Daniele

daniele.guido@uni.lu

Universität Luxemburg, Luxemburg, UL

## Ruppen Coutaz, Raphäelle

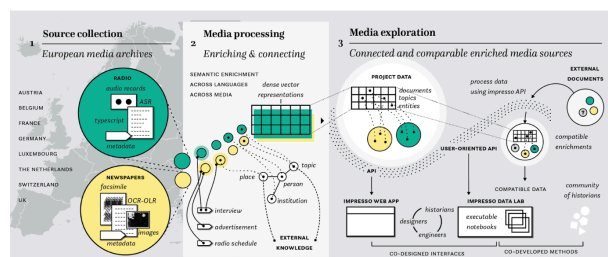
raphaelle.ruppencoutaz@unil.ch

Universität Lausanne, UNIL, Schweiz

## Beelen, Kaspar

kaspar.beelen@turing.ac.uk

The Alan Turing Institute, Großbritannien



Over the past few years, interdisciplinary research projects such as Oceanic Exchanges, Living with Machines, Numapresse or NewsEye (see links below) all worked

towards the semantic enrichment, integration and computational analysis of historical newspaper collections across institutional, national and linguistic boundaries. Outputs include shared tasks to advance the quality of semantic enrichment of historical text (Ehrmann, 2020), symposia to identify future research directions (Ehrmann, 2023), as well as a wide range of original historical research using computational methods together with reflections on their application (Keck et al., 2022; Oberbichler et al., 2021; Bunout et al., 2022). Most recently, the HAICu project announced the application of machine learning technologies to Dutch cultural heritage collections across different modalities on a national scale.

Against this background, we propose to present a new project: “*impresso* - Media Monitoring of the Past II. Beyond Borders: Connecting Historical Newspapers and Radio”, which builds on the first *impresso* project that compiled and semantically enriched a corpus of Swiss and Luxembourgish newspapers based on the collections of project partners such as the national libraries of Switzerland and Luxembourg, Neue Zürcher Zeitung, Le Temps, the Valais State Archives and the Swiss Economic Archive.

The first *impresso* application integrates new opportunities offered by semantic enrichments such as word embeddings, topic modelling, and the automated detection of text reuse and languages for content search and discovery as well as comparative and critical perspectives on the available data. Its design was driven by the principles of co-design, generosity (the provision of multiple entry points to the collection) and transparency (reflections on tools, document processing and data quality) (preprint Düring et al., 2023). The application is freely accessible.

This first project has demonstrated the added value of integrating sources from different languages into the same system in order to better facilitate their joint exploration and comparison. The corpus of the second *impresso* project will build on the first in several ways: First, it will broaden the corpus to include radio alongside newspaper sources. Second, it will expand to a Western European scale in partnership with national and state-level libraries as well as archives dedicated to the preservation of audiovisual materials together with new partners in Austria, Belgium, France, Germany, Luxembourg, the Netherlands and Switzerland. Many of our source material will be in German and be connected to contents in French, Dutch, English etc.

Third, these collections will be transformed from noisy and heterogeneous text in multiple languages into rich data, integrated and represented in a shared vector space. Fourth, it is our goal to develop an open and generic technological framework for the seamless exploration of semantically enriched and connected media archives. Fifth, the project will benefit from five (media) historical case studies which will exploit the newly available data under the shared research theme “influences” but also active engagement with research communities in digital humanities and history. Sixth, and finally, *impresso* will seek to actively support the research community by offering relevant data and interfaces for the exploration of its collections.

With the proposed poster we also seek to share a call for collaboration: Ingrained in the spirit of the *impresso* project is the belief in interdisciplinarity and collaborative design on the intersection between computational linguistics, computer science, history, digital humanities and design. We hope to reach out to researchers interested in contributing to the paradigm shift in the processing, representation and analysis of historical document collections.

## Project links:

impresso project : <https://impresso-project.ch/>  
Oceanic Exchanges : <https://oceanicexchanges.org/>  
Living with Machines : <https://livingwithmachines.ac.uk/>  
Numapresse : <https://numapresse.hypotheses.org/>  
NewsEye : <https://www.newseye.eu/>

## Bibliographie

**Bunout, Estelle, Maud Ehrmann, Frédéric Clavert, eds.** Digitised Newspapers – A New Eldorado for Historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization. Studies in Digital History and Hermeneutics. Berlin: De Gruyter, 2022.

**Ehrmann, Maud, Marten Düring, Clemens Neudecker, Antoine Doucet.** “Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292).” Dagstuhl Reports 12, no. 7 (2023): 112–79. <https://doi.org/10.4230/DagRep.12.7.112>.

**Ehrmann, Maud, Matteo Romanello, Stefan Bircher, Simon Clematide.** “Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers.” In Advances in Information Retrieval, 524–32. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020. [https://doi.org/10.1007/978-3-030-45442-5\\_68](https://doi.org/10.1007/978-3-030-45442-5_68).

**Keck, Jana, Mila Oiva, Paul Fyfe.** “Lajos Kossuth and the Transnational News: A Computational and Multilingual Approach to Digitized Newspaper Collections.” Media History (2022): 1–18. <https://doi.org/10.1080/13688804.2022.2146905>.

**Oberbichler, Sarah, Emanuela Boros, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, Mikko Tolonen.** “Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians.” Journal of the Association for Information Science and Technology, 2021, <https://doi.org/10.1002/asi.24565>.