

# Machine Learning to Read Yesterday's News. How semantic enrichments enhance the study of digitised historical newspapers

**Bunout, Estelle**

estelle.bunout@uni.lu

Universität Luxemburg, Luxemburg

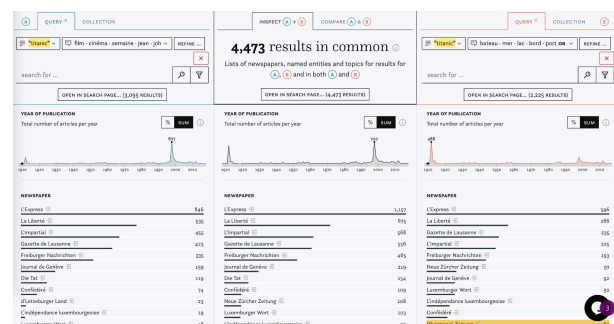
**Düring, Marten**

marten.during@uni.lu

Universität Luxemburg, Luxemburg

In this workshop we will use the *impresso* app to explore opportunities and challenges which accompany the semantic enrichment of historical newspapers. We will reflect on the added value of Natural Language Processing techniques such as topic modelling, text reuse detection and word embeddings for historians in conjunction with an introduction and critical assessment of design solutions for the scalable reading of such enriched sources.

We target researchers at all (digital) skill levels.



Historical newspapers shaped and reflect the political, moral, and economic environments in which they were produced. They hold dense, continuous, and multi-level information which can help us reconstruct how contemporaries represented and experienced their present. This makes them indispensable sources for research and their value is reflected in mass digitisation efforts over the past years.

As a consequence, researchers today face an abundance of materials which can no longer be managed with keyword search and basic content filtering alone even though only a fraction of the overall archival record has yet been processed. Digitisation also transformed analogue sources into

highly complex digital objects determined by multiple layers of technical processing. Subsumed under the notion of “digital hermeneutics”, scholars have pointed to the epistemological challenges inherent in such documents and called for a critical engagement with data provenance, its processing, and interfaces.

Our goal for this workshop is first for historians to familiarise themselves with the opportunities and pitfalls of semantically enriched sources. Second, we will introduce participants to the complexity of digitised newspaper collections and focus on key operations for their exploration and analysis. Third, we will encourage participants to discuss the capacities of the *impresso* app to support historical research and to offer transparency regarding data processing and interface functions. To this end we will focus on:

- Creation and comparison of user-generated content collections to reveal (dis)similarities (see e.g. screenshot above).
- Word embeddings to reveal synonyms, related terms, (historical) spelling variations and frequent OCR misspellings.
- Content filters based on topic models to include or exclude themes such as “sports”, “arts” or “foreign politics”.
- Content filters based on linked named entities to reveal the changing contexts in which entities such as persons, institutions and locations appear across time and newspapers.
- Article recommendations to identify potentially relevant content outside a researcher’s search scope.
- Dedicated exploration interfaces for text reuse clusters, n-grams and topics to reveal trends over time and newspapers and to assist query-building.
- Image similarity search reveals the distribution of similar images within the corpus.
- Visualisations of gaps, biases in the corpus and confidence scores for OCR and entities to better manage user expectations as to what can be found in the corpus and to judge the value of any finds.
- Educational materials to document data processing and interface functionalities.
- The interplay between these components allows researchers to address generic, yet complex historical questions such as: “How did the news about the Titanic catastrophe travel through the media sphere?” or “What constitutes a “crisis”? And why did it peak in 1932 in the press?”

We propose the following structure, for a group of about 20 participants:

1. Introduction: *impresso* project, newspaper collections and semantic enrichment
2. Small groups: Content retrieval with newspaper interfaces, e.g. Deutsches Zeitungsportal, ANNO, or eLuxemburgensia.

3. Demonstration and hands-on tutorial of the impresso interface with focus on: data criticism, content retrieval, comparison, result representativity, blind spots, user-generated collections.
4. Discussion: Experiences of working with two distinct newspaper interfaces and their search tools, compare and contrast. To which extent do digital tools empower new search and discovery workflows for newspapers? Which new skills do such tools require? How can we trust computationally generated information? What questions could not be answered that should have been (corpus stats, overview...)? What interaction was missing?
5. Outlook: follow-up project, impresso API + notebooks for data-driven research and impresso Powervis experimentation: navigate the graphs to get more general impressions of the corpus, linked to the pre-determined questions

Please note that participants can be at any level of digital literacy, the bigger barrier might rely in the language of the source material (French and German mainly).

Also, the newspapers collections will be handled only via the app, to enable the exploration of the existing collections. In the context of this workshop, the content extraction will consist in searching with keywords, using filters based on semantic enrichments, visualise, comparing the results of queries, etc.

No need to download or install anything to conduct the workshop, all is web-based. The needed material for participants is a laptop with internet connection.

## Convenors

Marten Düring, Assistant Professor in Digital History at the Luxembourg Centre for Contemporary and Digital History (C2DH), University of Luxembourg

Estelle Bunout, post-doctoral researcher at the C2DH.

Relevant publications:

- Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology, edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. Studies in Digital History and Hermeneutics. Berlin, Germany: De Gruyter, 2023.
- Düring, Marten, Roman Kalyakin and Daniele Guido. “Impresso Inspect and Compare. Visual Comparison of Semantically Enriched Historical Newspaper Articles.” Information 12, no. 9 (September 2021): 348.
- With Ehrmann, Maud, et al. Historical Newspaper User Interfaces: A Review. 2017. library.ifla.org, <http://library.ifla.org/2578/>.
- « The digitisation of newspapers: how to turn a page », From the archival to the digital turn · Ranke.2, 2019, <https://ranke2.uni.lu/lessons/>
- « Collections of Digitised Newspapers as Historical Sources – Parthenos training », 2019: <https://traiparthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/>

[ning.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/](https://traiparthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/)

## Bibliographie

**Alberto Romele, Marta Severo, and Paolo Furia,** “Digital Hermeneutics: From Interpreting with Machines to Interpretational Machines,” AI & SOCIETY 35, no. 1 (March 2020): 73–86, <https://doi.org/10.1007/s00146-018-0856-2>;

**Andreas Fickers and Juliane Tatarinov, eds.,** Digital History and Hermeneutics: Between Theory and Practice, Digital History and Hermeneutics, vol. 2 (De Gruyter Oldenbourg, 2022), <https://www.degruyter.com/document/isbn/9783110723991/html>;

**Chiel van den Akker et al.,** “Digital Hermeneutics: Agora and the Online Understanding of Cultural Heritage,” in Proceedings of the 3rd International Web Science Conference, WebSci ’11 (Koblenz, Germany: Association for Computing Machinery, 2011), 1–7, <https://doi.org/10.1145/2527031.2527039>.