Historical Newspaper Content Mining: Revisiting the *impresso* Project's Challenges in Text and Image Processing, Design and Historical Scholarship[1]

Maud Ehrmann[2]
Estelle Bunout[1]
Simon Clematide[3]
Marten Düring[1]
Andreas Fickers[1]
Daniele Guido[1]
Roman Kalyakin[1]
Frédéric Kaplan[2]
Matteo Romanello[2]
Paul Schroeder[1]
Philipp Ströbel[3]
Thijs van Beek[1]
Martin Volk[3]
Lars Wieneke[1]

[1] *Luxembourg Centre for Contemporary and Digital History (C²DH)*
[2] *École polytechnique fédérale de Lausanne (EPFL)*
[3] *Institut für Computerlinguistik der Universität Zürich*

Abstract

impresso. Media Monitoring of the Past[2] is an interdisciplinary research project in which a team of computational linguists, designers and historians collaborate on the datafication of a multilingual corpus of digitized historical newspapers. The primary goals of the project are to improve text mining tools for historical text, to enrich historical newspapers with (semi-) automatically generated data and to integrate such data into historical research workflows by means of a newly developed user interface[3].

impresso addresses the challenges posed by large-scale collections of digitized and datafied newspapers. These can be grouped into five categories:

[1] Long abstract presented at the Digital Humanities (DH) Conference, Ottawa, 2020. See the book of abstracts: https://dh2020.adho.org/abstracts/ (PDF not available as of March 2021), and the online presentation.
[2] https://impresso-project.ch/
[3] https://impresso-project.ch/app/#/

1. Newspaper silos: Due to legal restrictions and digitisation policy constraints, portals for digitized newspapers are bound to provide incomplete, non-representative collections which have been subjected to varying automated processing of varying quality.
2. Big, messy data: Newspaper digital items are characterised by incompleteness, inconsistencies and duplicates.

Noisy, historical text: imperfect OCR, faulty article segmentation and lack of appropriate linguistic resources greatly affect image and text mining algorithms' robustness.

3. Generosity [1]: Search and discovery of relevant content within such large and heterogeneous corpora.
4. Transparency: Critical assessment of inherent biases in exploratory tools, digitized sources and data extracted from them.

In this paper we discuss our efforts to overcome these challenges and to integrate text mining and data visualisation applications in general historical research practices which are characterised by search operations as well as the need to create topical collections [2]. This work is part of a larger academic and institutional endeavour carried out by several large consortia projects that seek to apply computational methods to digitized historical newspapers [3]. We will begin with an overview of the document and text processing steps. For each of them, the presentation will feature the following facets: why it is useful, which results were obtained, how difficulties were overcome, and the road ahead.

- **Data acquisition and pre-processing**: Digitized newspapers are scattered across many institutions with different access policies and facilities, in an opaque landscape of collection silos. On the administrative and legal sides, we will report on our strategies to inventory and acquire digital newspaper collections, as well as to agree on rights statements with data providers. On the technical side, we will present the workflows that were developed to cope with heterogeneous image and OCR (Optical Character Recognition) formats, and our efforts towards standardization[4].
- **Digital document processing**: *impresso's* starting point is - ideally - text and text blocks as outputted by OCR and OLR (Optical Layout Recognition). However, because of the aforementioned heterogeneity, text and layout elements are usually noisy and require to be carefully assessed, corrected, and sometimes re-OCRized. During the project we tested and implemented multilingual OCR quality assessment, Black Letter Neural HTR systems [4], and newspaper semantic segmentation using both textual and visual features [5].
- **Lexical processing:** Once the text has been acquired from images, the first steps consist in applying a series of linguistic pre-processing, including language identification, tokenization, historical spelling normalisation and lemmatization. The historical and multilingual nature of the *impresso* corpus naturally complicates some of these tasks. As complementary resources they are useful for subsequent text processing tasks. For direct usage within the user interface, distributed representations of words (word embeddings) are computed for each language in the corpus. In N-gram visualisations, users can explore the changes in word usage in our corpus over time; corpus-specific word embeddings allow the user to efficiently retrieve and compare related words, even across languages.

[4] https://github.com/impresso/impresso-schemas

- **Named entity processing**: Referential units such as names of persons, locations and organizations underlie the semantics of texts and guide their interpretation. Their automatic recognition and disambiguation greatly support information retrieval and exploration of large-scale textual collections.
- **Topic modelling**: We determine which 'topics' occur in newspaper collections to allow users to derive thematics. To this end, several topic models (across the whole corpus, per newspaper, per time period) are computed and topics are assigned to newspaper articles.
- **Text reuse** detects and aligns similar text passages and yields clusters of reused passages embedded within longer documents in large collections. *impresso* uses the *passim* [6][5] implementation.
- **Image similarity**: A visual search engine based on [7] complements text search and renders visual elements, e.g. advertisements, photographs, drawings and maps searchable.
- **System architecture**: Text and image processing components need to be integrated in a modular system architecture, which also includes an API, a middle layer and a frontend. We published a technical cookbook[6] which includes all recipes to apply *impresso* processings and to deploy the *impresso* interface.

Moving from the processing to the interface, we developed novel search and discovery strategies. Inspired by user feedback collected during multiple workshops and motivated by the overarching goal to seamlessly shift between close and distant reading perspectives, the interface was designed to allow manifold combinations of the following features:

- **Creation and comparison of user-generated collections** to reveal (dis)similarities between them.
- **Keyword suggestions** based on word embeddings point to synonyms and related terms, reveal (historical) spelling variations and frequent OCR misspellings.
- **Content filters based on topic models** allow the inclusion or exclusion of specific groups of articles such as "sports", "arts" or "legal" and reveal multiple facets within search results and collections.
- **Content filters based on linked named entities** reveal the changing contexts in which entities such as persons, institutions and locations appear across time and newspapers.
- **Article recommendations** identify potentially relevant content outside a user's search scope and are based on topic modeling, metadata, named entity recognition and text reuse.
- **Exploratory interfaces for text reuse clusters, n-grams and topics** reveal patterns across time and newspapers and assist query-building.
- **Image similarity search** reveals the distribution of similar images within the corpus but also allows to check if a given image is present in the corpus. A third usage is keyword search in automatically detected image captions.
- **Visualisations of gaps, biases in the corpus and confidence scores for OCR and entities** help to better manage user expectations as to what can be found in the corpus and to judge the value of any finds.

[5] https://github.com/dasmiq/passim
[6] Private at the moment of submission, will be open to the public by the time of the conference.

Finally, *impresso* seeks to integrate text mining in historical research workflows. We will reflect on our efforts to foster skills development in text mining and data literacy for untrained users, against the backdrop of a user evaluation.

Bibliography

[1] M. Whitelaw, "Generous Interfaces for Digital Cultural Collections," *DHQ*, vol. 9, no. 1, 2015, http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html.

[2] R. Allen and R. Sieczkiewicz, "How Historians use Historical Newspapers," *Proceedings of the American Society for Information Science and Technology*, vol. 47, 2010.

[3] M. Ridge, G. Colavizza, L. Brake, M. Ehrmann, J.-P. Moreux, and A. Prescott, Eds., "The Past, Present and Future of Digital Scholarship with Newspaper Collections," in *DH 2019 Book of Abstracts*, 2019, https://infoscience.epfl.ch/record/271329.

[4] P. Ströbel, S. Clematide, "Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images," presented at the Digital Humanities 2019 : Complexities, Utrecht, 07-Dec-2019.

[5] R. Barman, "Historical newspaper semantic segmentation using visual and textual features," 2019, https://infoscience.epfl.ch/record/271306.

[6] D. A. Smith, R. Cordell, and A. Mullen, "Computational methods for uncovering reprinted texts in Antebellum newspapers," *American Literary History*, vol. 27, no. 3, pp. E1–E15, 2015.

[7] B. L. A. Seguin, "Making large art historical photo archives searchable," EPFL, Lausanne, 2018, http://infoscience.epfl.ch/record/261212.

Acknowledgements

We warmly thank our archive and library partners[7] for having embarked on impresso's adventure and for having shared their digitized newspaper collections. We also want to thank our historian partners (UNIL and infoclio) for their guidance and support during workshops, as well as the many associated researchers who agreed to give early feedback on *impresso* interface prototypes. Our thanks also go to David Smith for his help and guidance with - among others - text reuse, and Benoit Seguin for his invaluable contributions on the visual part of the search engine. The authors gratefully acknowledge the financial support of the Swiss National Science Foundation (SNSF) for the project "*impresso* – Media Monitoring of the Past" under grant number CR-SII5_173719.

To go further

- Watch the online presentation of this paper: https://www.youtube.com/watch?v=mfiSBcl2EA8&list=PLB45F159nVx-lEm_U8zTeqq95Q92oj08r

- Visit the *impresso* app and get a user account: https://impresso-project.ch/app

[7] SNSF Associated partners: Swiss National Library, National Library of Luxembourg, State Archives of Valais, Swiss Economic Archives, Ringier *Le Temps*, Neue Zürcher Zeitung; and further contributors.

- Watch the *impresso* clip: https://www.youtube.com/watch?v=2njluhEd3pg

- Watch a series of introductory videos: https://www.youtube.com/watch?v=y6Dfj49XWu8&list=PLB45F159nVx9CwVvXx1vYEBN--BWHurnn

- Visit the impresso's Zenodo and GitHub organisation accounts