



How are we working with digitized newspapers?

Workshop - 31.05.2022

Estelle Bunout, Marten Düring

Newspapers' collections are massive, it is their recurrent characteristic alongside diversity of shapes and contents. The attractiveness of this source relies on its frequency, diversity of contents to enable longitudinal, local, and national queries. Newspapers are used by many researchers, as a central source, as complementary source, as a substitution for missing archives on particular issues, or for particular periods. Their massive nature makes it impossible for individual researchers to carry the necessary digitisation, OCR and even less so, index the digitised content to make it searchable, perform NLP or other text mining tools that would make them explorable in their digitised form.

Added to that discrepancy between what is technically possible and practically available, the debates surrounding the digitisation of newspapers brought to light issues connected with the representativity of the digitised corpus and the impact of the search tools in the building of research corpora. Andreas Fickers argued for a practice of the digital source criticism embedded in the customary training of historians, enhanced with better understanding of how these sources are produced¹ and heralded the challenges of dealing with their abundance². All the while, Ian Milligan demonstrated the silent shift towards digitised historical newspapers, the digitisation distorting the collections used in PhDs in Canada³. Indeed, the digitisation shuffles the search practice: putting small articles and highlighted articles on the same plane, based rather on their algorithmic relevance than on their physical visibility or historical relevance. The questions raised a decade ago are slowly finding their ways into the methodological discussion of researchers relying on these sources.

Another crucial element needs to be underlined: the space of potential search and the granularity of interaction with the actual existing source material does not correspond necessarily to the options offered to the users of search and viewing interfaces. The words typed in the search bar are looked for in the indexed part of the digitised material. The step of indexation of OCR output can imply lemmatisation, which means that all forms of a given word might be search then. For instance, one might search for "foi", which means faith in French, but a stemmed indexation will also return hits for "foie", which means liver. The documentation is not always available, so, this is one element that the users have to test out or be aware of, when assessing their results.

¹ Andreas Fickers, 'Veins Filled with the Dilluted Sap of Rationality. A Critical Reply to Rens Bod', *BMGN-Low Countries Historical Review* 128, no. 4 (2013): 155–63.

² Andreas Fickers, 'Towards A New Digital Historicism? Doing History In The Age Of Abundance.', *VIEW Journal of European Television History and Culture* 1, no. 1 (2012): 19–26, <https://doi.org/10.18146/2213-0969.2012.jethc004>.

³ Ian Milligan, 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010', *The Canadian Historical Review* 94, no. 4 (November 2013): 540–69.

The question is not only how to find relevant elements: experiences with the digitisation pipelines and interface design have produced many answers, even if they are not implemented everywhere. About ten years ago, Robert B. Allen and his co-authors asked, “what to do with a million pages of digitized historical newspapers?”⁴ in the context of an already massive effort of digitization of these materials. The paper laid out tools ranging from text extraction to image clustering, to deal with the advertisements published in the historical newspapers, highlighting this rich but arduously exploitable source via the dominant text-based search tools, namely the keyword search. Many of these proposed tools have been proved efficient by several initiatives but hardly implemented in the existing collections, in their publicly available interfaces.

What is also rather undefined is what to do with the collected material and how to discuss the findings, rooted in digital source criticism, including the technical biases and the collection biases.

Based on workshop gathering experienced researchers, we propose to discuss a series of challenges posed by the availability of these rich collections: what are the workflows that are technically possible and how to integrate the epistemological issues into the discussion of the findings?

This workshop led to the design of a checklist of questions on each step of the workflow (hereafter as appendix):

- Representativity of the digital collection compared to the analogue
- What is behind the interface? How do the search engine settings, indexation and pre-processing of the collection impact the query?
- What can be done in the interface?
- Contextualisation: what are the actionable content-based metadata (e.g. via NLP outputs) that can help contextualise the result list.

⁴ Robert B. Allen, Weizhong Zhu, and Robert Sieczkiewicz, ‘What to Do With a Million Pages of Digitized Historical Newspapers?’, February 2010, <https://www.ideals.illinois.edu/handle/2142/14932>.

Appendix: Questionnaire to what we need (to know) when working with digitised newspapers

- **Collection description**
 - What titles and what time span are covered?
 - What is the context of the creation of these digitised (sub)collections?
 - What regional/national diversity is represented in the collection?
 - How many languages are present in the titles?
 - Does the collection specialise in newspapers or contain other documents, archives?

- **Digitisation context and output**
 - What is the institutional frame of the digitisation (public/private institution, GLAM/project)?
 - Is the digitisation ongoing? Are the titles digitised entirely or only for limited portions?
 - Is the content searchable with keywords (OCR)?
 - Is there another portal that has published this collection, and is it processed differently?
 - Are there corpus statistics available (numbers of tokens, pages, issues)?

- **Search settings and user interactions**
 - Can I store the resulting hits? Can I add labels to selected items?
 - Can I export the resulting hits? In bulk/individually? Which format?
 - What are the effects of lemmatisation and indexation of the digitised content on the search results? Do the queries "apples" and "apple" have the same results? Can I search for two consecutive words (e.g. Spanish flue) within one article, or will they appear on one page/issue?
 - Can I test the "vulnerability" of the chosen keywords? Is it prone to OCR misidentification - and what options are there in the search page to mitigate this?
 - Can I limit my search to a date range, a type of item (articles, advertisement, obituaries, tables...)?
 - Can I search for images (adverts, maps, photographs, cartoons)?

- **Access:** what type of access is made possible with the digitised newspaper collection?
 - Can I access it via an interface and/or via an API?
 - Are there ready-made datasets to download, and what do they contain (images, text, METS/ALTO files)?
 - Are there other portals of the same institution that offer different interactions than the general interface (via "labs" for data visualisation, use of NLP outputs such as named entities)?