The ‘Environmental Scan’ at work: radical contextualisation of newspaper collections for new historical research

Beelen, Kaspar
kbeelen@turing.ac.uk
The Alan Turing Institute, United Kingdom

Lawrence, Jon
j.lawrence3@exeter.ac.uk
University of Exeter

McDonough, Katherine
kmcdonough@turing.ac.uk
The Alan Turing Institute, United Kingdom

Westerling, Kalle
kalle.westerling@bl.uk
The Alan Turing Institute, United Kingdom; British Library

Wilson, Daniel C.S.
dwilson@turing.ac.uk
The Alan Turing Institute, United Kingdom

This paper will demonstrate the practical potential of the 'Environmental Scan' – a new approach to working with large newspaper collections. The Environmental Scan uses newly digitised nineteenth-century reference works to bring powerful contextual information into connection with historical research using newspapers for the first time. Our method is based on British collections, but could be replicated for other nations and languages. The approach we used was only possible due to an ongoing collaboration between librarians, historians and data scientists, and their respective institutions. Large full-text newspaper collections – such as those held by the British Library – contain enormous amounts of data, but also highly complex metadata, which traditionally exists in catalogue records at different levels of detail and granularity. Nonetheless, a relatively limited amount of newspaper metadata has hitherto been captured in library catalogues. Among the many challenges of information management in relation to newspaper collections, is the complex and shifting metadata that results from the serial nature of newspaper publication. The limitations of record management systems and standards have, for example, made it difficult to capture changing titles as publications merge, separate, or change their geographical circulation. Other kinds of title-specific metadata such as price-points, ownership, target audience and political leaning would not normally be described in library records, and yet are fundamental attributes of these items.

The Environmental Scan undertakes the simple, yet Herculean task of locating all of the above information (and more) using contemporary Press Directories which we have digitised and processed. From the resulting parsed data, we have painstakingly identified and linked the enhanced metadata to the correct newspaper titles, and matched titles as they changed over time. The result is a set of highly valuable, enriched metadata which can be used by researchers to understand the context of newspapers they may be querying digitally, or in hardcopy. Does a physical or digitised collection exclude the penny press? Is it biased towards newspapers from particular regions? In the case of digitised collections, as the text-mining of the newspapers becomes ubiquitous, the perennial question of audience looms even larger: who was reading which newspapers, and for any given sample, how representative is it of the larger newspaper readership, and of the public as a whole? The 'Environmental Scan' can help researchers address these questions by allowing the creation of representative subsamples of newspaper collections according to price-point, geography and – crucially – political leaning. In this way, using enriched metadata helps avoid the 'bag of words' problem faced by so much research based on the indiscriminate text-mining of large text corpora. By re-connecting the content of newspapers (text) with the rich information about their social and economic context (metadata) we can move towards more meaningful results to a variety of domain-specific questions.

This paper develops and extends the approach set out in Beelen et al. (2023) and showcases the ways in which the principle of the 'scan' can be put to work for new forms of historical research. We focus on two of the many variables created by the Environmental Scan: a) political leaning and b) circulation geographies. In each case we demonstrate the power of this approach by visualising the evolution of the press through the prism of variables derived from 'Victorian' metadata gleaned from the press directories. We inspect the changing practices of political categorisation of newspapers, calculating the distribution of labels such as 'liberal' or 'independent' over time, but also scrutinise the stability of these labels. How often did directories register a change in a newspaper's professed political orientation, and can we observe a historical pattern? Secondly, we turn to circulation: using a new toponym identification and linking tool specifically tailored to working with historical newspaper data (in English), we attempt to refine our spatial understanding of readership by inspecting the localities, both big and small, in which newspapers claimed to have an audience, which we are now able to visualise in innovative ways.

Linking contemporary context about newspapers with pre-existing catalogue records and the full-text data enriches our view of the world of news, as it still exists in libraries. Political alignment and the spatial footprint of a newspaper are basic categories of information that, when added to existing metadata about publication dates or titles can dramatically improve meaningful discovery and analysis of the content within these collections.

Bibliography

Beelen, Kaspar et al (2023): ‘Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan’, *Digital Scholarship in the Humanities*. DOI: https://doi.org/10.1093/llc/fqac037

Brake, Laurel (2015): “Nineteenth-Century Newspaper Press Directories: The National Gallery of the British Press.” *Victorian Periodicals Review* 48.4: 569–90. DOI: https://doi.org/10.1353/vpr.2015.0055

Coll Ardanuy, Mariona et al (2019); “Resolving Places, Past and Present: - Toponym Resolution in Historical British Newspapers Using Multiple Resources.” In *Proceedings of the 13th Work-*

shop on Geographic Information Retrieval: 1-6. DOI: https://doi.org/10.1145/3371140.3371143

Fyfe, Paul (2018): 'Access, Computational Analysis, and Fair Use in the Digitized Nineteenth-Century Press', *Victorian Periodicals Review* 51.4: 716–37. DOI: https://doi.org/10.1353/vpr.2018.0051

Fyfe, Paul (2016); 'An Archaeology of Victorian Newspapers', *Victorian Periodicals Review* 49.4: 546–77. DOI: https://doi.org/10.1353/vpr.2016.0039

Hobbs, Andrew (2018): *A Fleet Street in Every Town: The Provincial Press in England, 1855-1900*. Cambridge: Open Book Publishers.

Lansdall-Welfare, Thomas, et al. (2017): "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences* 114.4: E457–65. DOI: https://doi.org/10.1073/pnas.1606380114

Mussell, James (2014): "Elemental Forms: The Newspaper as Popular Genre in the Nineteenth Century." *Media History* 20.1: 4–20. DOI: https://doi.org/10.1080/13688804.2014.880264

Neudecker, Clemens and Apostolos Antonacopoulos (2016): "Making Europe's Historical Newspapers Searchable." In *12th IAPR Workshop on Document Analysis Systems*: 405–10. DOI: https://doi.org/10.1109/DAS.2016.83

Verheul, Jaap, et al. (2022): "Using Word Vector Models to Trace Conceptual Change over Time and Space in Historical Newspapers, 1840–1914." *Digital Humanities Quarterly* 16.2. DOI: http://www.digitalhumanities.org/dhq/vol/16/2/000550/000550.html