Living with Machines: Exploring bias in the British Newspaper Archive

This paper will investigate the British Newspaper Archive as a corpus for historical research. We analyse the extent to which the digital collection is representative of the wider newspaper landscape. We achieve this by providing a rigorous analysis of selection bias (related to the composition of the corpus, i.e. the over- and underrepresentation of certain voices) and technological bias (imbalances caused by the technologies used to transform the printed newspaper into a machine-readable version). While the digital turn has made search and exploration of vast historical resources a much less painful (or at least less time-consuming) process, from a scholarly perspective, such databases remain problematic objects of scholarly inquiry given the manifold transformations and mediations that accompany the process of digitization (Fyfe, 2016; Mussell, 2012).

Following Fyfe (2016) we treat the electronic archive as a digital artefact that incorporates biases that arose from the opaque interplay between human (scholarly, institutional and corporate) choices and technological affordances (technologies used for storage and conversion of the printed newspapers).

To assess the impact of these choices on research outcomes, we firstly determine the extent to which the digital collection is representative of the wider newspaper landscape, focusing on the contrasting English counties of Lancashire and Dorset. Even though the number of digitized pages in the entire British Newspaper Archive is impressive (31,621,132 pages and counting) it constitutes only 6.6% of the total. More problematically, at present we lack a good estimate of how well the characteristics of the digital sample resemble those of population. To address this issue, we compare the digital newspaper collection to those publications listed in the Newspaper Press Directories (NPDs).

The NPDs were originally invented to provide 'a more dignified and permanent record' of British Newspapers--in 1861 they were recognized by the Post Office as an authoritative list of newspapers. The directories systematically collected contextual information such as price, geographic reach, political leaning, religious affiliation, ownership and other variables ((see Gliserman, 1969) and (Tom O'Malley, 2015) for a comprehensive overview of the NPDs' content and historical evolution). In this project, we consider the NPDs as our 'master list' to which we compare the metadata of the digital archive.[1] By comparing the distribution of specific variables (political leaning, price, region) we can estimate where and (how much) the digital sample deviates from the wider landscape, and subsequently simulate how these biases may affect research results.

Secondly, we investigate the content for technological bias, i.e. the presence of OCR-errors. The original XML documents come with a reported OCR-quality score at the word level. However, these scores cannot be taken at face-value (because they are non-transparent and produced by different processing pipelines or software (versions)). To get a grip on this measure of quality, we compare it to other indicators

[1] The NPDs are comprehensive but not definitive. Strictly speaking, there is no ultimate 'landscape' or 'universe' we can compare with, only a set of moving targets, which themselves are prone to bias.

(such as trigram and word counts) and subsequently study their distribution over the corpus. This enables us to gauge if some parts of corpus are effectively "silenced" by the OCR software (i.e. became unusable or unreadable), probe the explanations for these quality differences (e.g. was the newspapers scanned from microfilm; what the original price, etc).

Digital archives provide a very particular snapshot of the past: particular in terms of its technological makeup, its vantage point and perspective. The overarching aim of our endeavour is to pinpoint these particularities by putting the spotlight on the present biases. In practice, this boils down to characterising the voices captured by the collection, as well as determining those which are absent or hardly audible (because of the screeching OCR). This will allow historians and digital humanists to properly contextualize the results derived from the digital newspapers. We demonstrate how the focus on bias informs historical research within the Living with Machines project, elaborating on a case study that scrutinizes the changing concept of technology.

References

Fyfe, Paul. "An archaeology of Victorian newspapers." Victorian Periodicals Review 49, no. 4 (2016): 546-577.

Gitelman, Lisa, and Virginia Jackson. "Introduction to “Raw Data” Is an Oxymoron, edited by Lisa Gitelman, 1–14." (2013).

Gliserman, Susan. "Mitchell's" Newspaper Press Directory": 1846-1907." Victorian Periodicals Newsletter 4 (1969): 10-29.

Mussell, James. The Nineteenth-Century Press in the Digital Age. Springer, 2012.

O'Malley, Tom. "Mitchell's Newspaper Press Directory and the Late Victorian and Early Twentieth-Century Press." Victorian Periodicals Review 48, no. 4 (2015): 591-606.