Boutique Big Data: reintegrating close and distant reading of 19th-Century newspapers

This item was submitted to Loughborough University's Institutional Repository by the/an author.

Citation: BEALS, M.H., 2016. Boutique Big Data: reintegrating close and distant reading of 19th-Century newspapers. [Poster delivered at:] Digital Humanities 2016, Jagiellonian University & Pedagogical University, Krakow, 11th-16th July 2016

Additional Information:

- This poster was presented at Digital Humanities 2016 on 13th July 2016.

Metadata Record: https://dspace.lboro.ac.uk/2134/23544

Version: Published

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the published version.
In the 19th century, most English-language newspapers were run by owner-operators who could not afford to employ foreign correspondents, even as readers demanded a wide selection of local and international news. In response, editors took advantage of postal subsidies to exchange copies of their newspaper with other editors, snipping the most interesting bits and inserting them into their own issues. This process, known as *scissors-and-paste journalism*, often created *viral texts* that appeared worldwide.

**DETECTING PLAGIARISM**

Scissors-and-paste networks transferred information in several distinct ways: *direct reprinting*, *paraphrasing*, and *compilation*. Each practice left different markers and requires different text-mining processes to identify. The first and last can be recognized by *plagiarism detection software* such as the open-source programme, *Copyfind*. As illustrated below, the software compares the texts on both a word-and-character basis, allowing a user-dictated number of mistakes, indicated in blue italics. This compensates for a large percentage of OCR errors. It then outputs a list of likely matches, alongside the number of matching words and characters.

Although they cannot fully represent the true historical network, visualisations can suggest the degree to which different communities shared a common perspective. Below are Force Atlas 2 projections of the British Library 19th-Century Newspapers Collection, showing how shared content changed over time; the earlier network *became more integrated and new outliers developing*. This, however, only illustrates connections, not directionality.

**IMAGINED COMMUNITIES**

Although they cannot fully represent the true historical network, visualisations can suggest the degree to which different communities shared a common perspective. Below are Force Atlas 2 projections of the British Library’s 19th-Century Newspapers Collection, showing how shared content changed over time; the earlier network became more integrated and new outliers developing. This, however, only illustrates connections, not directionality.

**MAPPING DIRECTIONALITY**

After removing advertisements and false positives, the next step was to determine directionality: Chronology offers some evidence of *who copied whom* but reprinting was not a linear process; multiple versions were replicated at the same time. Manual transcriptions were taken of select case studies and examined for additions, omissions, and restyling. This, alongside chronology, suggested that a reprint’s direct ancestor was the version with the *smallest number of individual characters changed*. In the case of identical scores, the ancestor was assumed to be the earliest version. Using these results, a filtering algorithm was designed to determine likely *ancestor-descendant relationships* within the corpus. A directed network for 1800-1820 appears below.

**UNDERSTANDING EVOLUTION**

Although *Copyfind* provides a broad analysis of directionality, understanding *the processes taking place during reprinting* requires a smaller-scale analysis. The coding of case studies suggests which changes were typically made and which of these were evolutionary successful or most likely to be reprinted themselves. Below, variants of a story were coded on a word-by-word basis. This table of binary values was inputted into *Mesquite*, which modelled several possible dendrograms, or *evolutionary trees*.

**REFERENCES**


Blomfield, L. “Copyfind64 version 4.1.4.” https://osf.io/nm2rq