Journal of Open Humanities Data

Beals, M H 2017 Scissors and Paste: The Georgian Reprints, 1800–1837. *Journal of Open Humanities Data,* 3: 1, DOI: https://doi.org/10.5334/johd.8

DATA PAPER

Scissors and Paste: The Georgian Reprints, 1800–1837

M. H. Beals

Loughborough University, GB
m.h.beals@lboro.ac.uk

This dataset, part of the Scissors and Paste Project (https://osf.io/nm2rq), describes instances of reprinting and text reuse (scissors-and-paste journalism) in British newspapers between 1800–1837. It was derived from the 19th-Century British Library Newspapers, Part 1 digitised newspaper collection by using plagiarism detection software to identify instances of substantially similar text. It contains a series of manifests that describe a) instances of shared content b) the likely directionality of copying and c) which instances are evolutionary dead-ends and have no known reprints. It is comprised of 1,824 TSV files, divided into four directories, each representing one month between January 1800 and December 1837.

Keywords: Great Britain; news flow; newspapers; publishing industry

(1) Overview

Repository location

Beals MH. Scissors and Paste: The Georgian Reprints v1.0.0 [Dataset]. Loughborough; 2016 [cited 2016 Dec 12]. DOI: https://doi.org/10.5281/zenodo.200399

Context

Before 1837, electronic telegraphy was in its infancy and was not employed in the transmission of news content [1]. Instead, newspapers within and beyond Britain engaged in scissors-and-paste journalism, wherein one newspaper copied, in part or in whole, material from other publications. This created a highly decentralized, global news network [2]. Although the level of reprinting varied between titles, the practice was largely seen as mutually beneficial and in the wider public interest. To that end, both the British and United States governments subsidised postal exchanges between newspapers and larger papers employed exchange editors to curate incoming material for republication [3].

Scissors-and-paste practices can be seen in all types of periodical material, including news, correspondence, literature, poetry, jokes and advertisements [4]. The degree to which attribution was professionally expected is a matter of ongoing research, but even when attribution did take place, it was given unsystematically; sometimes newspapers listed the date and title of the original publication, rather than the one from which they had directly copied, while other times they offered only basic clues, such as 'a London paper' [5]. This has led to a sense of frustration, and several honest mistakes, by those using newspapers as indicators of local or regional public opinion; this lack of clear attribution, alongside anonymous or pseudonymous authorship, leaves the modern reader unsure as to the true origin of a given text. Matching texts within 19th-century corpora computationally allows us to work with reprinted and reworked materials with a greater confidence as to their provenance. News content, broadly defined as the time-sensitive recordings of events, was likely to be reprinted quickly and maintain a high fidelity regardless of the number of generations, making it particularly well suited for electronic discovery.

The Scissors and Paste Project (http://www.scissorsandpaste.net) tracks reprinting and reuse in the long-19th century (1783–1914) across the Anglophone world. The initial phase of the project involved the development of a suite of tools and methodologies to efficiently identify reprint families and then suggest both directionality and branching within these subsets. From these case-studies, detailed analyses of additions, omissions and wholesale changes can offer insights into the mechanics of reprinting that left behind few if any other traces in the historical record.

The Georgian Reprints represents the first discrete dataset to come out of this project, focusing on the years 1800–1837 within Great Britain. It is comprised of 1,824 monthly listings of reprinting within the 19th-Century British Library Newspaper collection. As the wider project progresses, it is expected that further datasets for additional years and wider corpora will be made available through the project website (https://osf.io/nm2rq). A fuller description of the methods used to create the dataset, and the rationale behind these, can be found in the related article within the *Journal of Victorian Culture* [6].

(2) Methods

Steps

The Source Data

The Georgian Reprints is derived from 226,507 page-level XML files from the 19th-Century British Library Newspapers, Part 1 collection [7]. The collection contains

transcriptions from 51 newspaper titles, regularised into 31 distinct publications, across 38 years. The original page-level XML transcriptions files were first transformed using XSL (included in the dataset) into non-encoded plaintext files; all metadata and XML tags were removed. The NormalisedDate, title, and pageSequence metadata tags were retained and used to name the plaintext files, which served as unique identifiers in the subsequent processing steps.

Matching with Copyfind64 v4.1.4

The plaintext files were first analysed, and instances of shared text found, using the plagiarism-detection software Copyfind64 v.4.1.4 by Lou Bloomfield [8]. The following settings were used:

- PhraseLength = 10
- WordThreshold = 200
- SkipLength = 20
- MismatchTolerance = 5
- MismatchPercentage = 50
- BriefReport = False
- IgnoreCase = True
- IgnoreNumbers = True
- IgnoreOuterPunctuation = True
- IgnorePunctuation = True
- SkipLongWords = False
- SkipNonwords = False

These settings were designed to be as forgiving as possible to OCR errors while not accruing an unmanageable number of false positives. The collection was divided into one-month sets across the 38-year period. Each of these monthly sets was compared against itself and the succeeding seven months. Manual testing suggested that any matches after 200 days were either false positives, annual notices, advertisements or miscellany content rather than news. The manifests outputted by Copyfind64—the Raw Matching Reports—were then further processed by a series of heuristic filters, described in the next section.

Accounting for False Positives

The raw matching reports were processed by two sets of heuristics, applied by the programmes Memetracker v1.0.0 (doi: 10.5281/zenodo.198542) and ReprintMapper v1.0.0 (doi: 10.5281/zenodo.198564), devised by the author.

Memetracker applied three, basic filtering heuristics to remove false positives from the raw matching reports. These heuristics are not set by the user, but are instead hard-coded into the software; these can, however, be modified by re-compiling the annotated source code available on Github. The first removed all self-matches—reprints in which the earlier and later instances were both from the same newspaper—through a simple deletion of these entries. Manual testing consistently indicated that these were advertisements, notices or other forms of boilerplate text. The second heuristic removed all matches that exceeded 200 days. This harmonized the data to precisely 200 days, inclusive, as the initial eight-month filter within Copyfind64 varied slightly throughout the year. The final heuristic further constricted the word count required for a match. The settings chosen for Copyfind required at least 200 matching words divided among phrases of no fewer than 10 words each. Memetracker, on the other hand, looked at the three quantitative similarity measures–the overall perfect match and the imperfect matching scores for each document—and filtered out those that had a perfect match of fewer than 160 words as well as an imperfect match of fewer than 90 words in both documents. These levels were chosen by testing the Raw Matching Report from the year 1815 to remove as many false positives as possible while retaining all true matches.

ReprintMapper, unlike Memetracker, describes specific ancestor-descendent relationships rather than all matching content. It applies identical heuristics as Memetracker before applying additional processing instructions. First, it removed all same-day matches. Although it was technically possible for one newspaper to reprint material from another on the same day, the lack of edition metadata and the paucity of newspapers printed in the same geographical location made such matches highly unlikely. Future iterations of this dataset may take geographical information into account more precisely.

Next, it compared all possible predecessors of each reprint on the number of matching words and the date difference. The match with the highest fidelity was determined to be the most likely ancestor of that reprint. Comparing computer and manually created stemma (trees) indicated that ordering on raw word matches resulted in identical or near-identical results to ordering based on close reading. Where two matches had identical fidelity, the earlier match was determined to be the ancestor as, in the absence of other information, it was logical to ascribe ancestry to the earliest possible source.

A manifest of these ancestor-descendent relationships was then outputted. A second manifest was also created of all pages that appeared to be evolutionary dead-ends; that is, where they did not appear to be the ancestor of any subsequent reprints.

Sampling strategy

The original XML dataset contained 42 corrupted files (out of a total corpus of 15700 files) that could not be transformed into plain text transcriptions; a full listing is available at https://github.com/mhbeals/BL19thC_Reprints/tree/master/Errors. All other files within the collection were analysed.

Quality Control

Over the 38-year period, several publications altered their title. During the initial comparison process, the title indicated by the XML "title" tag was used to prevent data loss. In the final derived dataset, these titles have been normalised to enable consistent analysis across all years. A full manifest of titles and their normalisations in the derived dataset has been included. Versions of this data without this normalisation can be found at the Scissors and Paste Project Website.

After running Copyfind64 on a five-year set of pages, I sampled those raw matches that were excluded by Memetracker and ReprintMapper. I found that fewer than 2% of matches were incorrectly removed from the dataset, occasionally by being over 200 days apart, but largely owing to being a very short articles. This represents a likely loss of 300 out of 15700 records across the 38-year period. There was no evidence that this percentage was higher or lower in particular years or titles. This was considered an acceptable false-negative rate as lowering the threshold would have significantly increased the rate of false positives.

(3) Dataset description

Object name
Scissors and Paste: The Georgian Reprints.

Format names and versions
TSV, XSL.

Creation dates
Start date 2016-07-07; end date 2016-12-01.

Dataset Creators
The dataset was devised and created by M. H. Beals, Loughborough University.

Language
The dataset contains 1,824 TSV files, divided into four directories. Each directory contains 456 files, each representing one month between January 1800 and December 1837. The headings for the TSV files in each directory are as follows:

RawMatchingReports
Files in this directory represent all matches as determined by Copyfind64, the month of the filename referring to that of the earlier of the two pages. See **Table 1**.

Memes
List those pairs of pages that share a significant amount of content. Individual matches are only listed once; that is, B–A and not also A–B. See **Table 2**.

AncestorDescendent
Files in this directory list every page, linking it to the one match that is most likely its direct ancestor (though not necessarily its direct predecessor). If there are no later variants of the page, the page is excluded from the list. The column headers are identical to those in Memes.

Deadends
Files in this directory describe pages that do not have any descendent pages, as determined by ReprintMapper. See **Table 3**.

License
CC-BY 4.0.

Repository name
Zenodo.

Publication date
2016-12-13.

(4) Reuse potential

The dataset was created to explore trends and correlations in text reuse within 19th-century British newspapers. By understanding the extent to which identical, or near-identical, texts spread in rapid succession, it becomes clearer the degree to which Britain shared a common knowledge of domestic and global events. By understanding the general directionality of this news flow, we are also able to better understand the power relationship between metropolitan and provincial newspapers, as well those in port, industrial and agricultural communities. The dataset was also created to supplement existing knowledge about the political and commercial alignments of individual newspapers by allowing for high-resolution, longitudinal studies of shared content. Thus, there is particular potential for reuse of the dataset in periodical studies. It provides a quantitative context for any discussions of the influence of a particular newspaper, especially if it is further filtered to articles known to have originated in that title.

Other potential uses are as a reference text and as a basis for further research into specific memes or reprint families. As a reference text, *The Georgian Reprints* is currently the largest index of reprints within British periodicals. Any individual working with the British Library newspaper collection, in whatever context and from whichever disciplinary background, can look up the individual pages they are working with and see if that content is a reprint or was reprinted elsewhere. As explicit attribution was rare in this period, evidence indicating the possible origins of a text can help inform users as to its usefulness or fundamentally change the arguments based upon it.

Those researching particular events or texts can also further develop the dataset by filtering for texts on a particular topic (manually or through topic modelling of the original collection) and then adding specific descriptions to the pair listings. These augmented datasets could then be used to qualify the trends and correlations seen across the wider dataset. For example, news of a certain genre or regarding a particular topic may have a different pattern or rate of dissemination than the corpus as a whole. Likewise, although Memetracker filtered out the majority of advertisements by removing same-title matches, a large number of national advertisements for books, patent medicines and the lottery are also listed. Filtering for these entries using full-text searching within the original collection could offer new insights into Georgian advertising. Likewise, filtering for only same-title matches in the Raw Matching Reports is likely to return a corpus largely composed of local advertising.

Limitations and Provisos
Although a complete representation of the original digitised newspaper corpus, there are some key limitations to the data within the Raw Matching Reports. First, the 19th-Century British Library Newspapers, Part 1 collection contains only 31 titles and does not represent a complete corpus of the British press for this period. Careful examination of which titles are included is recommended. Second, the Raw Matching Reports

Field	Description
PM	The number of perfectly matching words in phrases of at least 10 words.
OL	The number of perfectly and imperfectly matching words in phrases of at least 10 words in the later (Reprint) document.
OR	The number of perfectly and imperfectly matching words in phrases of at least 10 words in the earlier (Original) document.
RYEAR	This column indicates the year in which the later of two matching pages was printed.
RMONTH	This column indicates the month in which the later of two matching pages was printed.
RDAY	This column indicates the day in which the later of two matching pages was printed.
RTITLE	This column indicates the title of the later of two matching pages.
RFILENAME	This column indicates the filename of the later of the two matching pages. Filenames are formatted as (YYYY.MM.DD _Title_page).
OYEAR	This column indicates the year in which the earlier of two matching pages was printed.
OMONTH	This column indicates the month in which the earlier of two matching pages was printed.
ODAY	This column indicates the day in which the earlier of two matching pages was printed.
OTITLE	This column indicates the title of the earlier of two matching pages.
OFILENAME	This column indicates the filename of the earlier of the two matching pages. Filenames are formatted as (YYYY.MM.DD_Title_page).

Table 1: Fields used in RawMatchingReports.

Field	Description
RYEAR	This column indicates the year in which the later of two matching pages was printed.
RMONTH	This column indicates the month in which the later of two matching pages was printed.
RDAY	This column indicates the day in which the later of two matching pages was printed.
RTITLE	This column indicates the title of the later of two matching pages.
RPAGE	This column indicates the page number of the later of two matching pages. Page numbers were given an S prefix w when two editions of the same date-title combination were discovered within in the original XML collection. In the original collection, these may have been designated with either an S or a V in the XML filename.
OYEAR	This column indicates the year in which the earlier of two matching pages was printed.
OMONTH	This column indicates the month in which the earlier of two matching pages was printed.
ODAY	This column indicates the day in which the earlier of two matching pages was printed.
OTITLE	This column indicates the title of the earlier of two matching pages.
OPAGE	This column indicates the page number of the earlier of two matching pages. Page numbers were given an S prefix w when two editions of the same date-title combination were discovered within in the original XML collection. In the original collection, these may have been designated with either an S or a V in the XML filename.

Table 2: Fields used in Memes.

Field	Description
YEAR	This column indicates the year in which the page was printed.
MONTH	This column indicates the month in which the page was printed.
DAY	This column indicates the day in which the article page printed.
TITLE	This column indicates the title in which the article page printed.
PAGE	This column indicates the page number. Page numbers were given an S prefix w when two editions of the same date-title combination were discovered within in the original XML collection. In the original collection, these may have been designated with either an S or a V in the XML filename.

Table 3: Fields used in Deadends.

only list those document pairs for which 200 words of shared content can be computationally identified. As the machine-readable transcriptions of these pages were obtained through optical character recognition (OCR) from digitised images, the accuracy rate can vary significantly; some pages are largely illegible. While these errors are unlikely to result in false positives, they may have caused a large number of false negatives. Therefore, the number of true matches is certainly higher than those recognised by the comparison process; these manifests should, therefore, be considered a minimum rather than an average or maximum reprinting rate for any given title or period. Likewise, ReprintMapper can only find the best match within the corpus. If two descendants of a single ancestor are present, but their common ancestor is not, ReprintMapper will link the later to the earlier version, even if these actually represent two different branches. This false positive must be excluded manually by using contextual knowledge. Finally, documentation as to the editions or individual copies digitised by the original British Library project were not indicated in the page-level metadata and could not be accounted for in the text comparison process.

An important final proviso is that the transcriptions used in the text comparison process were at page rather than article resolution; that is, each file representing a whole page of text rather than a smaller subdivision of it. This decision was taken owing to (a) the imprecision of computational subdivision for newspapers from this period and (b) the improvement in the matching of ancestors and descendants when there was evidence of multiple reprints from a single source. However, if a page has reprints from two separate ancestors within the corpus, ReprintMapper will only link it to source with the larger match; the other connection will be lost. Iterations of the process at article level would produce additional pairs but lose the added certainty obtained from matching multiple-article reprints.

Acknowledgements

The derived dataset was created using the XML transcriptions of the 19th-Century British Library Newspapers, Part 1 collection. I am grateful to James Baker and Mahendra Mahey of British Library Labs for their support of the project and their work in securing the necessary permissions for use of the original XML collection.

Competing Interests

The author has no competing interests to declare.

References

1. **Barton, RN.** New Media: The birth of telegraphic news in Britain 1847–68. *Media History*. 2010; 16(4): 379–406. DOI: https://doi.org/10.1080/13688804.2010.507475
2. **Beals, MH.** The Role of the Sydney Gazette in the Creation of Australia in the Scottish Public Sphere. Historical Networks in the Book Trade. In: Feely, C, Hinks, J, (eds.). Basingstoke: Routledge; 2016. p. 148–70.
3. **Silberstein-Loeb, J.** The International Distribution of News: The Associated Press, Press Association, and Reuters, 1848–1947. Cambridge: Cambridge University Press; 2014.
4. **Brake, L** and **Demoor, M.** Dictionary of Nineteenth-century Journalism in Great Britain and Ireland. Gent: Academia Press; 2009.
5. **Georgian Pingbacks.** [Internet]. Loughborough (UK): Scissors and Paste; [cited 2016 Dec 13]. Available from: http://www.scissorsandpaste.net/georgian-pingbacks.
6. **Beals, MH.** Stuck in the Middle: Developing Research Workflows for a Multi-Scale Text Analysis. *Journal of Victorian Culture*. 2017; 22. DOI: https://doi.org/10.1080/13555502.2017.1301178
7. **British Library Newspapers.** Part I: 1800–1900. [Internet]. Andover (UK): Gale Cengage Learning; [cited 2016 Dec 13]. Available from: http://gale.cengage.co.uk/british-library-newspapers/19th-century-british-library-newspapers-part-i.aspx.
8. **Bloomfield, L.** Copyfind64 v.4.1.4. Charlottesville, VA; 2016; [cited 2016 Dec 13]. Available from: http://plagiarism.bloomfieldmedia.com/wordpress/software/copyfind/.

How to cite this article: Beals, M H 2017 Scissors and Paste: The Georgian Reprints, 1800–1837. *Journal of Open Humanities Data*, 3: 1, DOI: https://doi.org/10.5334/johd.8

Published: 05 April 2017

]u[*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS