Stuck in the Middle:

Developing Research Workflows for a Multi-Scale Text Analysis

M. H. Beals, Loughborough University

There is a serendipity to historical research; the slow, often piecemeal gathering of ideas over a lifetime of reading can create strange and wonderful connections between seemingly disparate facts and, by extension, lead to new and unexpected insights. Laurel Thatcher Ulrich alluded to this process in the 1997 documentary *American Experience: A Midwife's Tale*. In a voiceover, she relates how one passage of Martha Ballad's diary seemed particularly "out of place" to her. [1] Later, she "began to do more research into Martha's life and into the incidence of epidemics in the eighteenth century and suddenly it just all fell together in one place"; the connection between the birth of Martha's daughter and the death of another child many years later was suddenly clear. This level of access to a historian's analytical process was unusual for me. It did not appear in the book—which provided the conclusion but not the journey to it; had I not seen the documentary I might never have understood fully the nuance of her interpretation.[2] I began to wonder: how many of my connections and assumptions do I make clear to my readers?

Although commonplace in modern historical documentaries, Ulrich's intertwining of the historical narrative with her narrative of discovery is not without criticism. Restricted by a processual framework, certain evidence can be unjustly foregrounded, distorting the analysis. Stylistically, it can be obtrusive or unnecessary. But is the invisible author any less problem-

[1] *American Experience: A Midwife's Tale* [DVD] Blueberry Hill Productions (1997). A transcription is available at <http://www.pbs.org/wgbh/amex/mwt/filmmore/pt.html> [accessed 1 August 2016].

[2] Laurel Thatcher Ulrich, *A Midwife's Tale: The Life of Martha Ballard, Based on Her Diary, 1785-1812* (New York: Random House, 1990), 43.

atic? However carefully we craft our methodologies, the explicit framework in which we place our evidence, our day-to-day experience of research, our workflow of discovery, is littered with innumerable, and often unconscious, choices on how to select, organise, contextualise and interrogate sources. Few if any of these choices are made explicit in the summative analysis. Is it therefore any wonder that two historians may come to radically different interpretations of the same evidence or that they may be at a loss as to understand the other's conclusion?

The use of digital methods and tools only complicates matters further. Most historians accept that any historical interpretation is shaped by the specific questions asked and is therefore only one of many possible conclusions that could be derived from a specific pool of evidence. In response, most have developed reading strategies to tease out these implicit aims or perspectives. Yet, these less effective when critiquing digital analyses; a re-examination of the sources or an analysis of the author cannot always shed sufficient light on the workflow that shaped that analysis. Moreover, historiographical conventions have encouraged authors to exclude seemingly mundane activities, such as data cleaning and mark up, when describing their research. Nonetheless, the choices made during these activities, and the assumptions that underpin them, often have dramatic consequences on the final analyses. Large-scale corpora cannot simply be re-read to reverse-engineer the researcher's point of view; if an interpretation is to be considered plausible, the author must leave behind not only a citation to the dataset used but detailed documentation of both their methodology and workflow, the specific choices made at each stage of gathering, processing, and analysing the data and the rationale behind them. For example, a note that quantitative evidence was regularised is meaningless without knowing the software with which it was regularised, the specific changes made to the data by that software, and the stage of research in which that regularisation took place. With this information, we can critique and refine the questions and assumptions made at every

stage of our research and improve our understanding of the past.

As an example of the minimum level of documentation required for a digitally augmented analysis, the following is a narrative of the workflow of the *Scissors and Paste Project*. [3] Before telegraphy, the dissemination of national and international news relied upon a global system of authorised and unauthorised copying, a process generally referred to as scissors-and-paste journalism. *Scissors and Paste* utilises the British Library Newspapers collection to explore the possibilities of mining large-scale corpora for these reprinted texts. As news content was time-sensitive, and more directly descriptive than miscellany or literary content, it was often reprinted in an uninterrupted chain of high-fidelity copies. Research into these dissemination pathways has led to the development a suite of tools and methodologies to identify reprint families and to suggest both directionality and branching within them. From these results, detailed analyses of additions, omissions and wholesale changes can be made, offering insights into the mechanics of reprinting that left behind few if any other traces in the historical record. Like many digital projects, the seemingly mundane choices made in obtaining, cleaning, processing and analysing the original dataset have affected my final conclusions. Moreover, because the analysis takes place at multiple scales, this project is particularly susceptible to hidden or non-intuitive workflow practices; by providing full documentation of my methodology on the project website, my interpretations can be better understood and refined as we develop a clearer understanding of this oft discussed by rarely documented journalistic practice.

[3] All derived datasets, methodologies and open source software described in this essay are available on the project website, "Scissors and Paste" (Open Science Framework, 2016) <osf.io/nm2rq> [accessed 1 August 2016].

Obtaining Data

As the size and number of digitisation projects grow, researchers can theoretically access more information with the click of a mouse than their predecessors could in a lifetime of archival research. However, the inconsistent storage, dispersal and retrieval mechanisms for this information may negate many of the advantages of digitisation. Particularly egregious are digital periodical and newspaper collections. Despite being limited by the same curatorial factors that affect all digital materials, these archives often give the mistaken impression of completeness.[4] When hunting for specific, known documents, these problems are manageable but when making an initial sampling of a resource, or when attempting large-scale distant reading, the limitations of these seemingly full-text repositories becomes clear.

Despite these, *Scissors and Paste* relies upon digitised databases to map the dissemination of newspaper reprints on a previously unachievable scale. Although historians have long known about scissors-and-paste journalism, systematically recording instances of reprinting would have been unviable before digitisation.[5] With the advent of full-text searching, articles and their reprints can, and have, be found across multiple databases using a combination of specific phrases and chronological or geographical filters.[6] However, this method is both inefficient and unreproducible. With the notable exception of Trove (National Library of Australia), digitisers have produced the machine-readable (that is, searchable) versions of their newspaper collections optical character recognition (OCR) rather than manual transcription. This process produces many errors that would effectively prevent reliable, full-text

[4] James Mussell, *The Nineteenth-Century Press in the Digital Age* (Basingstoke: Palgrave, 2012), 41.

[5] B. Nicholson, 'Counting Culture; or, How to Read Victorian Newspapers from a Distance', *Journal of Victorian Culture*, vol. 17, 2 (2012).

[6] For examples of this case-study approach, see M. H. Beals, 'The Role of the Sydney Gazette in the Creation of Australia in the Scottish Public Sphere' in Catherine Feely and John Hinks (eds.), *Historical Networks in the Book Trade* (Oxford: Routledge, 2016) and Bob Nicholson, '"You Kick the Bucket; We Do the Rest!": Jokes and the Culture of Reprinting in the Transatlantic Press', *Journal of Victorian Culture*, 17.3 (2012), 273–86.

searching if a user's queries were applied directly. For this reason, many searches allow, or automatically apply, an error tolerance to the search terms, allowing it to return possible matches even if the transcriptions were flawed. The exact methods by which a provider applies these imprecise, or fuzzy, search parameters are usually unknown; consequently, those attempting to replicate or build upon digital research cannot be sure that they will obtain the same selection of results, nor can the original researcher be sure they are obtaining representative samples in different databases.

When working with collection indirectly through pre-defined search interfaces, this uncertainty can be partially mitigated by recording queries as well as results—how you search, not just what you find.[7] In this case, a single spreadsheet will often suffice—a table listing the database and the date of the search, the particular search parameters used, and the specific results obtained. This process of recording serves two important functions: for the researcher, it ensures that the specific keywords and limiting factors, such as date, geographical location, or title, are uniformly replicated in different sessions and in different databases; for those critiquing the work, it provides a greater understanding of the choices and assumptions that underpin the final analysis. Such tables do not guarantee that others can reproduce your search; they do, however, shed light on how the search was constructed and why it returned these specific results. If a conclusion appears unfounded, examining the search parameters may explain whether the researcher selected inappropriate parameters, the dataset was incomplete on the date of the search, or the search algorithms employed by the interface altered the query.

Although this method may suffice when undertaking a close reading of a well-defined

[7] For a fuller discussion of recording searches, see M. H. Beals, "Record How You Search, Not Just What You Find: Thoughtfully Constructed Search Terms Greatly Enhance the Reliability of Digital Research" *The LSE Impact Blog* (2013), <http://blogs.lse.ac.uk/impactofsocialsciences/2013/06/10/record-how-you-search-not-just-what-you-find/> [accessed on 1 August 2016].

case study, it is unlikely to fulfil either function on a larger scale. For distant reading, direct access to a static, self-contained and well-documented dataset may be required. With this, the results of any computer-aided analysis can be placed in a specific context and, more importantly, the results can be fully replicated. Whether the dataset is obtained from a third-party or created by the researcher, what is important is that the specific version of the dataset is recorded and that the data itself is securely and sustainably stored. Through the support of British Library Labs, *Scissors and Paste* was able to obtain and make use of this type of dataset, namely British Library Newspapers, Part I: 1800-1900.[8] As a static collection of public domain data, with strong claims to geographical and political representativeness, this collection facilitated the development of well-defined and reproducible methodologies and results. The remainder of this essay will discuss how to document the cleaning, processing, and analysis of fully accessible and self-contained datasets.

Cleaning Data

Digital analyses require a significant amount of data cleaning before any actual analysis can take place, with surveys suggesting that it can account for up to 80% of the total time expended on research, regardless of the source or nature of the data.[9] Whether stored in tabular, graph, relational, hierarchical or long form, most databases will need to be at least partially transformed and this will affect any subsequent analyses; by focusing on the most relevant data or organising it in the most appropriate format for the specific questions being asked, these transformations inevitably bias the shape, and possibly the content, of the final results. Moreover, no dataset is unimpeachable—experimental and survey data is subject to variances

[8] For details on this dataset, see <http://gale.cengage.co.uk/british-library-newspapers/19th-century-british-library-newspapers-part-i.aspx> [accessed on 1 August 2016].

[9] CrowdFlower, *2016 Data Science Report* (2016), 6, <http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf> [accessed on 1 August 2016].

in its collection and all data is vulnerable to input errors, whether at the collection stage or a later point of data consolidation or compartmentalisation. Documentation of the provenance of these transformed datasets place them in a clear context for later analysis.

Like most researchers, my workflow began with the transformation my dataset into a format that was compatible with my analytical framework. Owing to the size of the collection, its manner of storage at the British Library, and the nature of the analysis I intended to undertake—comparing full pages of text against the entire corpus—the dataset could not be accessed or efficiently queried through an online Application Programming Interface (API). Instead, it needed to be transferred and stored on an external hard drive. Compressed in standard ZIP files, the dataset was roughly 250 GB. Once decompressed, it increased in size to 920 GB. A copy of the compressed and uncompressed data was stored on a single 3TB external hard drive (allowing its use on several different workstations) while the original compressed data remained on a separate 500GB hard-drive. The latter allowed me to return to the original data if the larger drive failed or the data became otherwise corrupted—a situation that occurred on two separate occasions owing to mechanical and software errors.

The uncompressed data consisted of a collection of eXtensible Markup Language (XML) files, each containing the metadata and text for a single page within the digital newspaper collection. These files were contained in separate directories, derived from individual digitisation batches rather than bibliographical details, though the individual filenames did indicate their contents. Because I was comparing documents based on date, this processual division was a hindrance. I merged the files into a single directory through a simple command line function—**FOR /R x:\original %G IN (*.xml) DO move %G x:\new**—wherein /R searches through the entire tree of subdirectories and %G stands in for the various filenames. Working on the command line was important considering the sheer size of the database; copying or moving these through the graphical user interface (Windows 10) often re-

sulted in the system becoming permanently unresponsive (hanging) during the enumerating process, which counted the files before they were moved. Once I had reorganised the data into a single directory, I could reformat them *en masse*.

The XML files, as created through the optical character recognition (OCR) processes, contained a significant amount of metadata that was unnecessary for my analysis. Moreover, the text was formatted in a way that it could not be effectively parsed by standard plagiarism detection software. For example, instead of displaying "will take place in the PublicRooms", the XML displayed the first few words of the page as:

```
<pageWord coord="792,626,865,663">will</pageWord>
<pageWord coord="867,626,939,665">take</pageWord>
<pageWord coord="166,656,248,692">place</pageWord>
<pageWord coord="254,656,300,692">in</pageWord>
<pageWord coord="291,655,346,693">the</pageWord>
<pageWord coord="334,655,554,693">PublicRooms</pageWord>
```

Excusing the OCR error, which failed to include the space between "Public" and "Rooms", the real difficulty was the inclusion of coordinates for each word. These are used to highlight the text on the corresponding image of the page, but hindered my analysis by breaking up the text I was attempting to compare. Likewise, the metadata preceding each text was important for establishing the provenance of the file but irrelevant to my textual analysis. Therefore, using eXentsible Systle Language (XSL) scripts, I transformed the data from complex, metadata-rich XML into smaller plain-text files.[10] These files were given human-readable names—**Title_YY_MM_DD_Page.txt**—and contained only the raw text. This transformation was done at the command line using the Saxon XSL converter—**java net.sf.saxon.Transform -s:WO1_ANJO_1830_01_06-0001.xml -XSL:"x:\stylesheet\BL_TEXT.xsl" -o:"x:\plaintext\"**.

[10] For a further description of this process, see M. H. Beals, "Transforming Data for Reuse and Re-publication with XML and XSL," *Programming Historian* (07 July 2016), <http://programminghistorian.org/lessons/transforming-xml-with-xsl> [accessed on 1 August 2016].

At this stage, the value of systematic data cleaning comes to the fore. It quickly became apparent that a significant, if seemingly random, number of XML files were corrupted, halting the transformation sequence. There were three key errors: incomplete files, files with corrupted characters and files that resulted in duplicate human-readable filenames. The last of these was the simplest to correct. These were the result of multiple editions of certain pages. They had been given the prefix S and could be quickly moved into a subdirectory for separate processing. The first two faults, however, were not easily traceable as the errors were not consistent. Each time the process failed, the corrupted file was removed and the process re-started. Fewer than 500 of the two million XML files generated this type of error, but their random dispersal slowed the automation process. Because the errors were idiosyncratic, a list of excluded files has been created, but these files cannot, as yet, be reintegrated into the wider analysis. Nonetheless, these exclusions are recorded in detail, qualifying the current results and allowing the data to be reintegrated into the set at a later date.

Processing Data

Once the all XML data had been transformed into plain-text file containing only the newspaper text, these were inputted into Copyfind, an open-source plagiarism detection software programme.[11] Careful documentation of this latter process is required as finding “matches" is a subjective process. What qualifies as a match? A certain number of identical words or characters? A general composition or narrative structure? In addition to making its source code publicly available, Copyfind also allows the user to specify variables such as phrase length, overall minimum word count, mismatch tolerance and the decision to ignore or enforce matching case, punctuation, and non-words. Subtle changes to these can produce very different results. For *Scissors and Paste*, the variables were set to a relatively high matching re-

[11] Lou Bloomfield, Copyfind (version 4.1.4). Windows. Charlottesville, Va., 2016.

quirements (at least 20 instances of ten-word phrases) but with a very high tolerance for small character differences. This configuration, chosen through trial-and-error, lowered the possibility of false positives, those with common short phrases or boilerplate prefaces, while simultaneously allowing for errors in the OCR transcriptions. As the size of the dataset required matching to be done in one-month increments, careful record keeping of these iterations, and the version of this still-evolving software, helped ensure consistent analysis. Likewise, such records assist in contextualising and qualifying the computer-generated results, making clear the assumptions I made when defining 'a match'

Copyfind provides two sets of results. These two outputs have two different uses and are therefore stored separately. The first output is a collection of HTML files, colour-coded and hyperlinked to allow for visual comparison of the texts of any supposed match, navigable through a hyperlinked list of each match within the corpus. The second output is a tab-separated (TSV) version of this manifest. The HTML files were used to verify matches manually as well as to determine heuristics, or generalised rules, for efficiently removing false positives. The number of matches necessitated a team of data checkers, namely undergraduate research assistants.[12] These researchers accessed the comparison texts via a private, institutionally hosted website and recorded their results in a shared Google Drive spreadsheet; this proved an efficient means of checking a random sample of the derived data.

The second set of results was the basis for the definitive list of reprints. This single file, a plain-test list of matches, was first transformed with a series of regular expressions (ReGex), parsing the list of filenames into individual columns of metadata. This expanded table was then processed with OpenRefine, an open-source tool for cleaning data, to regular-

[12] The author would like to acknowledge the efforts of the undergraduate researchers at Loughborough University: Will Dickinson, Alice Gilbert, Ollie Luhrs, Alex Mackinder, Pooja Makwana, Matthew McCulloch, Jonny Ord, Emily Stanyard and Rebecca Thompson.

ise the newspaper titles, which changed over time and in different processing batches. This cleaned table was then filtered through ReprintMapper, an open-source tool I created to consistently apply the edit rules developed above. This final list was then stored on Github, which provides an open, versioned record of the computer-generated results.

Analysing Data

It is from this list, rather than samples and soundings, that I now derive my case studies. Once a sufficiently reprinted item is found (one with at least three versions), highly accurate manual transcriptions of these articles are taken. These are compared against the OCR database using Copyfind, occasionally returning new matches, and keyword searches are undertaken within other digitised newspaper collections. Once these digital search options are exhausted, a set of transcriptions are made and digitally compared; the individual discrepancies—conscious alterations, typographical errors and the implementation of house style—are recorded as a data matrix using an open-source python script. Each word or punctuation mark is listed and given a binary value—zero or one—denoting whether it appears in that version of the text. This matrix can then be used to create statistical models of the evolution of these texts or to provide detailed evidence for qualitative discourse analysis.

At this point, my research flows into more traditional dissemination channels and the consistent documentation of my workflow provides a solid basis for defending my final conclusions; nagging questions about the robustness of my literature review or the comprehensiveness of my data can be precisely answered. Although the answers are not always to my liking, I can compensate for known deficiencies in a way that I could not for implicit connections between data and my wider knowledge. Moreover, because I have made this documentation available online, others can build upon or critique my work, confident in their understanding of how I reached my conclusions; flaws in my methodology or dataset can be identi-

fied, and corrected, rather than my work being summarily dismissed as contrary to other interpretations.

In the end, however an individual research workflow is developed, a twenty-first-century researcher should be mindful of the choices they make when obtaining, ordering and analysing their sources. Whether information was discovered through an online search or by exploring unmarked boxes in a local archive, keeping careful records of these discoveries and choices allows for confidence in not only in your conclusions but in the development of new research questions to pursue in the future. Therefore, methodological critique may provide digital historians the most fertile ground for contributing to wider historiographical debates as well as the most straightforward way for all historians to engage with the digital humanities.