Text Visualisation Tool for Exploring Digitised Historical Documents

Olivia Vane
Department of Innovation Design Engineering
Royal College of Art
Kensington Gore, London SW7 2EU
olivia.fletcher-vane@network.rca.ac.uk

DIS'18 Companion, *June 9–13, 2018, Hong Kong*
© 2018 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-5631-2/18/06.
https://doi.org/10.1145/3197391.3205428

Abstract

This paper describes a prototype timeline tool designed for humanities researchers exploring digitised historical documents. The tool visualises keyword instances in context mapped by date, and can be used to explore commentary around themes through time. Through designing the tool and evaluating it with humanities scholars, the role of the designer in the digital humanities is explored. Interview evaluation with historians provides evidence for the tool's capacity to support historical research, but also raises design issues by pointing to the value of simple, minimal design in this domain for interpretability.

Author Keywords

Design; HCI; Time; Timeline; Visualisation; Digitised collections; Digital Humanities.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

Digitised historical documents and archives are increasingly available to humanities researchers. With these new resources, there is a need for interface designs serving the requirements of scholarly research. There is also potential for novel interface designs to

Medical Officer of Health (MOH) Reports

The Medical Officer of Health reports 1848-1972 collection is roughly 5,500 documents. Reports were produced annually by the MOH for each London borough and consist of both narrative text concerning public health and tables of statistics. The reports vary in length (over 100 pages to fewer than 20) and in structure and content.

The Wellcome Library digitised the collection by photographing the reports cover-to-cover, applying Optical Character Recognition (OCR) and post-OCR manual text proofing; the resulting text data has a high degree of accuracy [20]. The date assigned to each document is its publication year.

I chose to work with this collection because it included high quality, complete transcribed text data for each report, and was sufficiently large to address my line of inquiry but not so large as to be unwieldy to work with.

explore new ways of accessing and exploring these collections.

The work presented here builds on a series of projects exploring temporal visualisation of historic data [1, 3, 15, 16]. This previous work aims to offer interactive visualisations that are sophisticated enough to use in investigating research questions, and to demonstrate what can be achieved when designers work with collection experts (curators, scholars etc.) to build tools. These projects also seek to better understand the role of the designer in digital humanities research [2]: an area of activity that increasingly involves developing interfaces and visualisations, and yet generally is absent of designers [5].

The prototype timeline tool presented here is designed for historians working with digital text collections. The purpose of the tool is to facilitate tracing commentary around themes through time across texts: an important activity in historical research, particularly when surveying primary sources at the beginning of a research project. A preliminary evaluation with historians brings up design considerations for this domain concerning interpretability and perceived trustworthiness of the interface.

Related Work

In visualisation research, keyword visualisations have been used for exploring document collections. Word clouds [13] reveal the important keywords in a corpus and Phrase Net [21] visualises the keywords joined by a connecting term such as “and” or “at”. Neither, however, shows the keywords’ usage context, limiting the degree to which a user can engage with themes or commentary across documents.

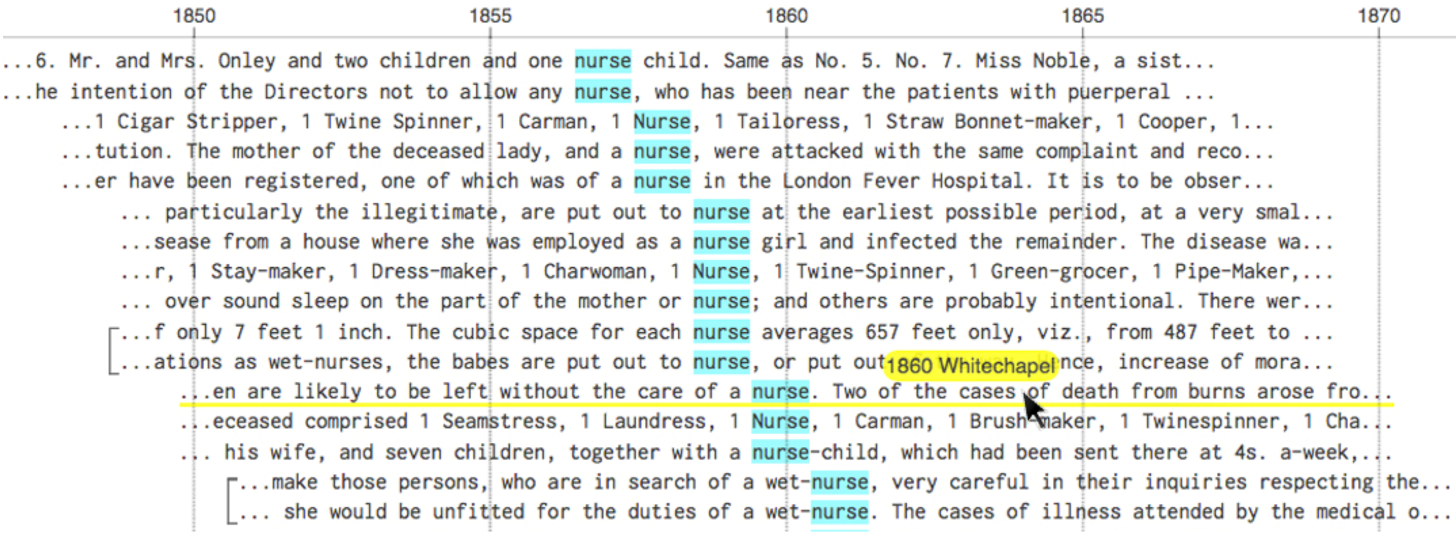
Word Tree [22] offers more context by visualising a central keyword with the terms that follow it across documents. Sentences where a common prefix is shared are collapsed to form a branching structure. While this tool allows users to examine the ways that a particular word or phrase is used in a text, the layout disconnects the text chunks that precede and follow the keyword. When surveying documents from a historical perspective, the full and original context for the keyword can be critical for sense making. For humanities scholars, it is very important for the context of visualised data to be available to aid fluid exploration and to help avoid misinterpretations [14]. Word Tree does not visualise the temporal dimension.

Topic modeling has been used in text visualisation to explore themes across texts. The temporal variation of ‘topics’ (statistically-identified recurring patterns of co-occurring words [4]) across documents has been visualised as columns of keywords [6], and as a river [7, 8, 9, 12, 17, 18]. There are calls for caution, however, concerning the application of topic modeling in the digital humanities. As Schmidt argues, ‘topics’ are ambiguous and can be misleading. The lack of transparency in the outputs of these complex algorithms obscures the subjective specifics of the data and closes off scrutiny from fellow scholars [19].

Design

The prototype tool presented here is designed to offer scholars a way to overview commentary across documents through time, while staying close to the original texts. The prototype tool is demonstrated with a collection of historical public health reports: the Medical Officer of Health (MOH) reports [23]. Here I walk through the design:

Figure 2: Zoomed out visualisation of MOH reports data: 'putrid', 1900-1972.



...6. Mr. and Mrs. Onley and two children and one nurse child. Same as No. 5. No. 7. Miss Noble, a sist...
...he intention of the Directors not to allow any nurse, who has been near the patients with puerperal ...
...1 Cigar Stripper, 1 Twine Spinner, 1 Carman, 1 Nurse, 1 Tailoress, 1 Straw Bonnet-maker, 1 Cooper, 1...
...tution. The mother of the deceased lady, and a nurse, were attacked with the same complaint and reco...
...er have been registered, one of which was of a nurse in the London Fever Hospital. It is to be obser...
... particularly the illegitimate, are put out to nurse at the earliest possible period, at a very smal...
...sease from a house where she was employed as a nurse girl and infected the remainder. The disease wa...
...r, 1 Stay-maker, 1 Dress-maker, 1 Charwoman, 1 Nurse, 1 Twine-Spinner, 1 Green-grocer, 1 Pipe-Maker,...
... over sound sleep on the part of the mother or nurse; and others are probably intentional. There wer...
...f only 7 feet 1 inch. The cubic space for each nurse averages 657 feet only, viz., from 487 feet to ...
...ations as wet-nurses, the babes are put out to nurse, or put out nce, increase of mora...
...en are likely to be left without the care of a nurse. Two of the cases of death from burns arose fro...
...eceased comprised 1 Seamstress, 1 Laundress, 1 Nurse, 1 Carman, 1 Brush-maker, 1 Twinespinner, 1 Cha...
... his wife, and seven children, together with a nurse-child, which had been sent there at 4s. a-week,...
...make those persons, who are in search of a wet-nurse, very careful in their inquiries respecting the...
... she would be unfitted for the duties of a wet-nurse. The cases of illness attended by the medical o...

Figure 1: Detail from visualisation of MOH reports data: 'nurse'. Each text snippet is horizontally centred on its date. The snippets come from separate documents, unless a bracket to their left indicates multiple are from the same document. Hovering over a snippet with the pointer reveals the report year and London area. Each snippet is hyperlinked through to the full document on the Wellcome Library website.

Figure 3: Zoomed out visualisation of MOH reports data: 'heroin'.

Figure 4: Zoomed out visualisation of MOH reports data: 'blitz'.

As an example, a user might be interested to explore the roles of nurses over time through these reports. By searching for the keyword 'nurse', a visualisation is generated of every instance of the keyword across the documents with a snippet of its surrounding text. The snippets are visualised at legible size and arranged by date (horizontally centred by date and vertically ordered from oldest to newest, see Figure 1). A horizontal date axis runs across the top of the visualisation. Each snippet is hyperlinked through to the full text for that document, and brackets to the snippets' left indicate when more than one snippet originates from the same document.

In this way, a researcher can explore: what sorts of things are being said in the texts about nurses at particular times? Is the context of the keyword in the narrative of the documents similar over time? Are there patterns in shifting descriptions or associations?

The focus of this tool is not primarily quantitative, but, since the snippets are arranged by date, a user can get a feel for their volume and patterns of occurrence from the overall shape. The snippets generally arrange in a slope, which is more vertical for a higher occurrence rate. Any kinks, gaps and gradient changes of the slope indicate changing frequencies of occurrence at particular times. For instance, the severe turn at the bottom of the visualisation generated when searching for 'putrid' reveals a sharp drop in frequency of use of this term 1920-onwards (see Figure 2). (In later iterations, it may be worth limiting visualised snippets to one per document, as otherwise this interferes with using the slope as an indicator of the document count for keyword matches)

Interview Evaluations

Each semi-structured interview was roughly 30 minutes long and conducted in person. The historian interviewees, recruited through contacts at the Wellcome Library and Institute of Historical Research, were a mixture of ages, gender, degree of seniority and of familiarity using digital tools/coding.

In each interview, I demonstrated the prototype with a few prepared example queries. I then invited interviewees to use the tool freely themselves. The historians often chose keywords relating to their current research topics or that they anticipated might produce interesting results.

The interview structure included questions about their general impression of the tool and the value of specific design features. Interviewees were asked to discuss how they might use the tool in their own research.

By analysing the aggregate keyword-context across the documents mapped by time and the overall visualisation shape, qualitative trends across these documents can be revealed. For example, the visualisation for 'heroin' features a strong column to the right, indicating a sudden surge in instances of the term 1960-onwards (see Figure 3). The text snippets reveal discussion around heroin suddenly shifts from 1960-onwards to drug addiction and abuse.

Similarly, the visualisation for 'blitz' forms a column of snippets from 1940 (see Figure 4), where it refers to the Nazi air raids on Britain starting that year. The text snippets expose the word's gradual adoption into English language from this time; it first appears in quotation marks, but over time instances without quotation marks increase and become the norm. By the 50s, the term is even used metaphorically.

The tool's front-end is built in JavaScript and D3.js [10]. Its back-end is an Elasticsearch [11] index.

Evaluation

To better understand how this timeline tool design can serve humanities research, I conducted 7 semi-structured interviews with scholars. These feedback sessions were an opportunity to evaluate the visualisation design and solicit suggestions for improvements and refinements.

Discussion and Further Research

The scholars confirmed they felt the visualisation is effective for overviewing commentary and would be most appropriate for exploratory work at the beginning of a research project. They also felt it would be quick and useful for organising and filtering source material – one commented it was like being able to "*look through all of them*" at the same time. The visualisation seemed particularly powerful in combination with a scholar's expertise with the corpus or knowledge of relevant framing events. A number also wished to try the interface with a fully customisable selection of digitised sources, crossing texts from different collections.

A wider issue that emerged was the role of simple visualisation design for interpretability and transparency in this context. The historians responded positively to the simplicity of this design and how close it stays to the original texts: "*I really like the simplicity of your visualisation actually. You haven't tried to be flashy and sparkly; it's just 'this is the information'*". Trust and transparency in the processes by which results are returned was very important for these users: "*when it come to academic research … what I conclude from your tool feeds into my reputation*".

This issue of trust in the interface and its connection with simple, easy to understand design was an important concern for users. This sort of issue hints at the role designers can play in the development of tools for digital humanities research and in exploring how such tools are perceived and experienced.

Before running more in-depth evaluations with this prototype, it will be necessary to explore options for dealing with its scalability issues; currently if a large number of snippets are returned for a keyword, the visualisation becomes unwieldy. The author is also interested to extend exploration of some of the issues raised in this paper to different digitised collections and different types of cultural heritage data, or instance image data.

References

1. Stephen Boyd Davis, Emma Bevan, and Aleksei Kudikov. 2013. Just In Time: defining historical chronographics. In *Electronic Visualisation in Arts and Culture*. Springer, London, 243-257.
2. Stephen Boyd Davis and Florian Kräutli. 2015. The Idea and Image of Historical Time: Interactions between Design and Digital Humanities. In *Visible Language*, 49, 3, 100-119.
3. Stephen Boyd Davis, Olivia Vane, and Florian Kräutli. 2016. Using data visualisation to tell stories about collections. In *Proceedings of Electronic Visualisation and the Arts 2016,* British Computer Society, London, 221-228.
4. Megan R Brett. 2012. Topic modeling: a basic introduction. In *Journal of digital humanities* 2, 1, 12-16.
5. Anne Burdick. 2009. Design without designers. In *Conference on the Future of Art and Design Education in the 21st Century*, University of Brighton.
6. Christopher Collins, Fernanda B. Viégas, and Martin Wattenberg. 2009. Parallel tag clouds to explore and analyze faceted text corpora. In *Visual Analytics Science and Technology* (VAST '09), *IEEE Symposium on*, IEEE, 91-98.
7. Adam Crymble. 2013. Review of paper machines, produced by Chris Johnson-Roberson and Jo Guldi. In *Journal of Digital Humanities*, 2, 2-1.
8. Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. 2011. Textflow: Towards better understanding of evolving topics in text. In *IEEE transactions on visualization and computer graphics* 17, 12, 2412-2421.
9. Weiwei Cui, Shixia Liu, Zhuofeng Wu, and Hao Wei. 2014. How hierarchical topics evolve in large text corpora. In *IEEE transactions on visualization and computer graphics* 20, 12, 2281-2290.
10. D3.js. Retrieved 9 March, 2018 from https://d3js.org/
11. Elasticsearch. Retrieved 9 March, 2018 from https://www.elastic.co/products/elasticsearch
12. Susan Havre, Beth Hetzler, and Lucy Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *Information visualization* (InfoVis '00)*, IEEE symposium on*, IEEE, 115-123.
13. Florian Heimerl, Steffen Lohmann, Simon Lange, and Thomas Ertl. 2014. Word cloud explorer: Text analytics based on word clouds. In *System Sciences (*HICSS '14*), 47th Hawaii International Conference on*, IEEE, 1833-1842.
14. Uta Hinrichs, Beatrice Alex, Jim Clifford, Andrew Watson, Aaron Quigley, Ewan Klein, and Colin M. Coates. 2015. Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. In *Digital Scholarship in the Humanities*, 30, suppl_1, i50-i75.
15. Florian Kräutli. 2016. *Visualising cultural data: exploring digital collections through timeline visualisations*. Ph.D Dissertation. Royal College of Art, London, UK.
16. Florian Kräutli and Stephen Boyd Davis. 2013. Known Unknowns: representing uncertainty in historical time. In *Proceedings of Electronic Visualisation and the Arts 2013,* British Computer Society, London, 61-68.
17. Shixia Liu, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. 2009. Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (CIKM '09). ACM, New York, NY, USA, 543-552. DOI=http://dx.doi.org/10.1145/1645953.1646023
18. Paper Machines. 2013. Retrieved 12 March, 2018 from http://papermachines.org/

19. Benjamin M Schmidt. 2012. Words alone: Dismantling topic models in the humanities. In *Journal of Digital Humanities*, 2, 1, 49-65.
20. Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou. 2016. Text mining the history of medicine. In *PloS one*, 11, 1, e0144717.
21. Frank van Ham, Martin Wattenberg, and Fernanda B. Viégas. 2009. Mapping text with phrase nets. In *IEEE transactions on visualization and computer graphics* 15, 6.
22. Martin Wattenberg and Fernanda B. Viégas. 2008. The word tree, an interactive visual concordance. In *IEEE transactions on visualization and computer graphics*, 14, 6.
23. Wellcome Library. Medical Officer of Health reports. Retrieved 12 March, 2018 from http://wellcomelibrary.org/moh/

Acknowledgements

Thanks go to my supervisors Stephen Boyd Davis and John Stevens, Diana Tanase for her technical guidance, and the Wellcome Library (who digitised and made available these historical texts) for their help. Olivia's research is funded by Arts and Humanities Research Council award AH/L503782/1 through the London Doctoral Centre for Design.