Trading Consequences: A Case Study of Combining Text Mining & Visualisation to Facilitate Document Exploration

- Hinrichs,Uta
  SACHI, University of St. Andrews
  uh3@st-andrews.ac.uk
- Alex,Beatrice
  ILCC, School of Informatics, University of Edinburgh
  balex@staffmail.ed.ac.uk
- Clifford,Jim
  Department of History, University of Saskatchewan
  jim.clifford@usask.ca
- Quigley,Aaron
  SACHI, University of St. Andrews
  aquigley@st-andrews.ac.uk

Summary

Trading Consequences is an interdisciplinary research project between historians, computational linguists and visualization specialists. We use text mining and visualisations to explore the growth of the global commodity trade in the nineteenth century. Feedback from a group of environmental historians during a workshop provided essential information to adapt advanced text mining and visualisation techniques to historical research. Expert feedback is an essential tool for effective interdisciplinary research in the digital humanities.

1. Introduction

This paper reports on interdisciplinary work carried out as part of Trading Consequences [1], a two-year Digging into Data project [2]. The focus of the project is to mine large quantities of historical documents, extract information on commodity trading in the nineteenth century British World and visualise the mined output in dynamic and interesting ways, thereby bringing archives alive in ways that authors of original documents would have never imagined. The Trading Consequences interface is aimed at historians studying commodities and their environmental consequences. Their studies have tended to focus on a manageable number of commodities (e.g. William Cronon's research on beef, lumber and wheat [3]). The Trading Consequences Project aims at identifying global trends in commodity trading for many different natural resources, raw materials or lightly processed goods by correlating information extracted for one commodity with that of others or showing all commodities relevant to particular locations and dates.

In this paper, we first present an overview of this collaborative project that involved environmental historians, text mining, database experts and visualization researchers. We then report on lessons learned from a workshop where we collected feedback from historians and geographers after they interacted with the interface prototype in a series of exercises. This feedback informed the further adaptation of the underlying technologies for historical research.

2. Trading Consequences

The Trading Consequences system encompasses three main technical components: a text mining system, a database and a web-based user interface with dynamic visualisations (Fig. 1). The data analysed using this system is comprised up of several nineteenth century British and Canadian text collections [4]. These sources amount to over 11 million pages and over 7 billion analysed word tokens.

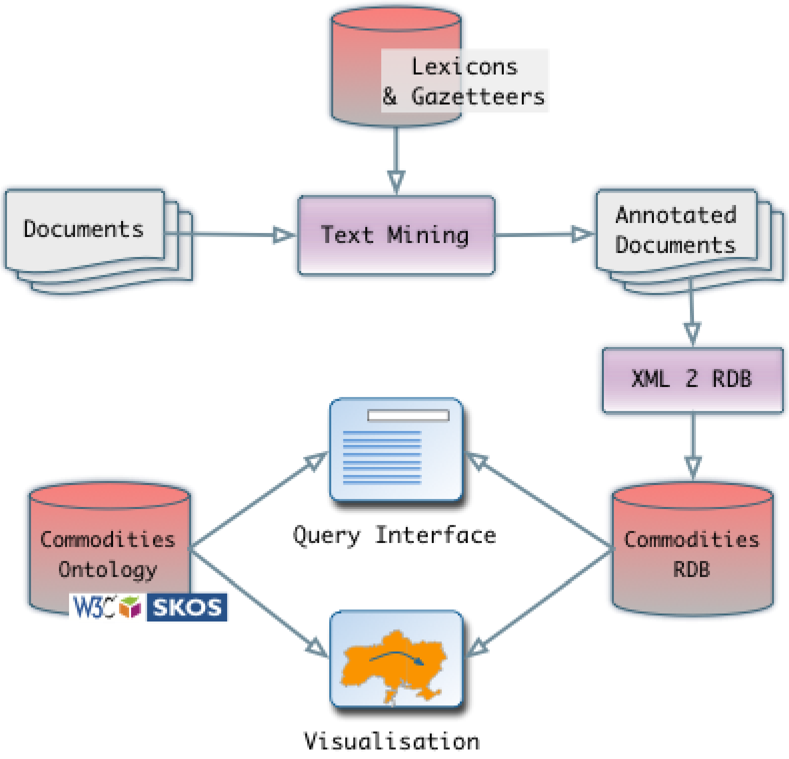


Fig. 1: System architecture.

2.1 Text Mining

The text mining (TM) tools are developed by the Language Technology Group at the University of Edinburgh. We adapted an existing pipeline built on LT-XML2 and LT-TTT2 to process historic text *[5]*. The TM component is made up of a series of linguistic processing steps which build up the linguistic properties of the language in a given text. A pre-processing stage includes tokenisation, sentence-splitting, part-of-speech tagging and lemmatisation to determine words and sentences, identify their syntax and compute canonical forms of word tokens. All of this information aids down-stream TM processes. The next steps are named entity recognition and grounding. This means that mentions of locations, commodities and dates are automatically identified in the text and grounded to unique identifiers in existing knowledge databases. We ground location mentions to GeoNames identifiers and their corresponding latitude/longitude values *[6]*. We use an adapted version of the Edinburgh Geoparser for this geo-referencing process *[7]*. Commodity mentions are grounded to DBpedia *[8]* concepts in a semi-automatically constructed commodity lexicon developed in this project *[9]*. Finally, date mentions are grounded to year, month and date attributes. The last TM step identifies relations between commodity, date and location mentions to identify the relevance of commodities in space and time. The extracted and enriched TM output is stored in a relational PostGreSQL database set up and hosted by EDINA *[10]* for subsequent querying and visualisation.

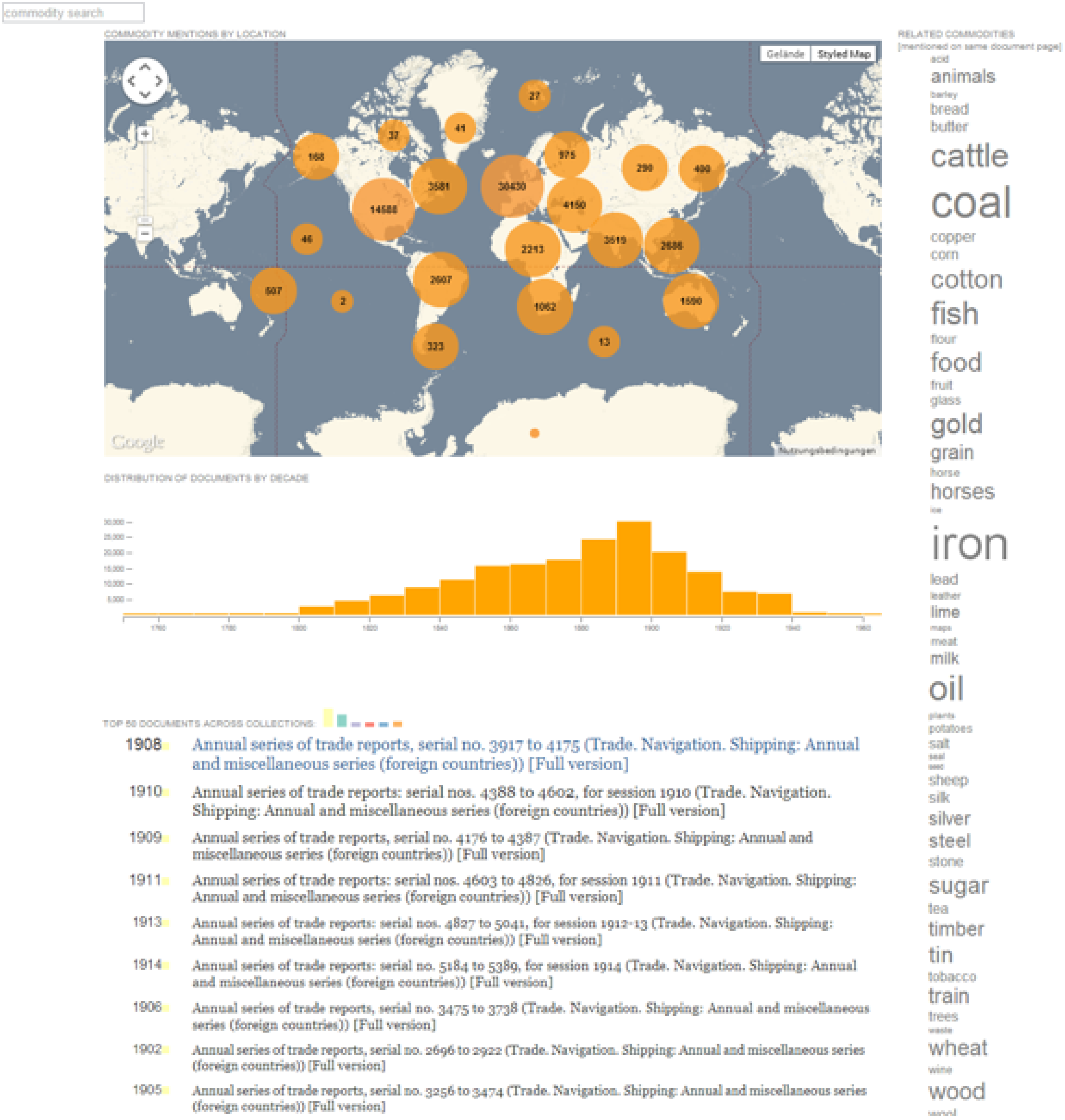
2.2 Interactive Visualisations

The strength of information visualisation is to make abstract concepts and relations within data visible and explorable *[11]*. In the context of Trading Consequences we aim at providing visualisations of the mined data to

1. enable open-ended explorations of the document corpus beyond target search *[12]*, i.e., supporting visual querying along spatial, temporal, and conceptual dimensions, and
2. highlighting trends within a range of document data, for instance, relations between different commodity types.

While the first approach facilitates the discovery of related documents in ways that common text-based search interfaces cannot, the second approach can lead to new insights or research questions based on collection sizes that exceed possibilities of traditional research methods in the humanities. Our visualizations are web-based to make them easily accessible worldwide (see [1]). The implementation is based on JavaScript (D3.js and jQuery) and PHP.

In this paper we briefly describe the Trading Consequences visualisation tool and how it was experienced by environmental historians as part of a workshop. Inspired by [13], our visualisation consists of three interlinked representations (Fig. 2): a map showing the geographic context in which commodities were mentioned, a vertical tag cloud showing the 50 most frequently mentioned commodities, and a bar chart representing the temporal distribution of documents within the collection. A ranked document list provides direct access to the relevant articles.



1908 Annual series of trade reports, serial no. 3917 to 4175 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

1910 Annual series of trade reports: serial nos. 4388 to 4602, for session 1910 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

1909 Annual series of trade reports, serial no. 4176 to 4387 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

1911 Annual series of trade reports: serial nos. 4603 to 4826, for session 1911 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

1913 Annual series of trade reports: serial nos. 4827 to 5041, for session 1912-13 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

1914 Annual series of trade reports: serial no. 5184 to 5389, for session 1914 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

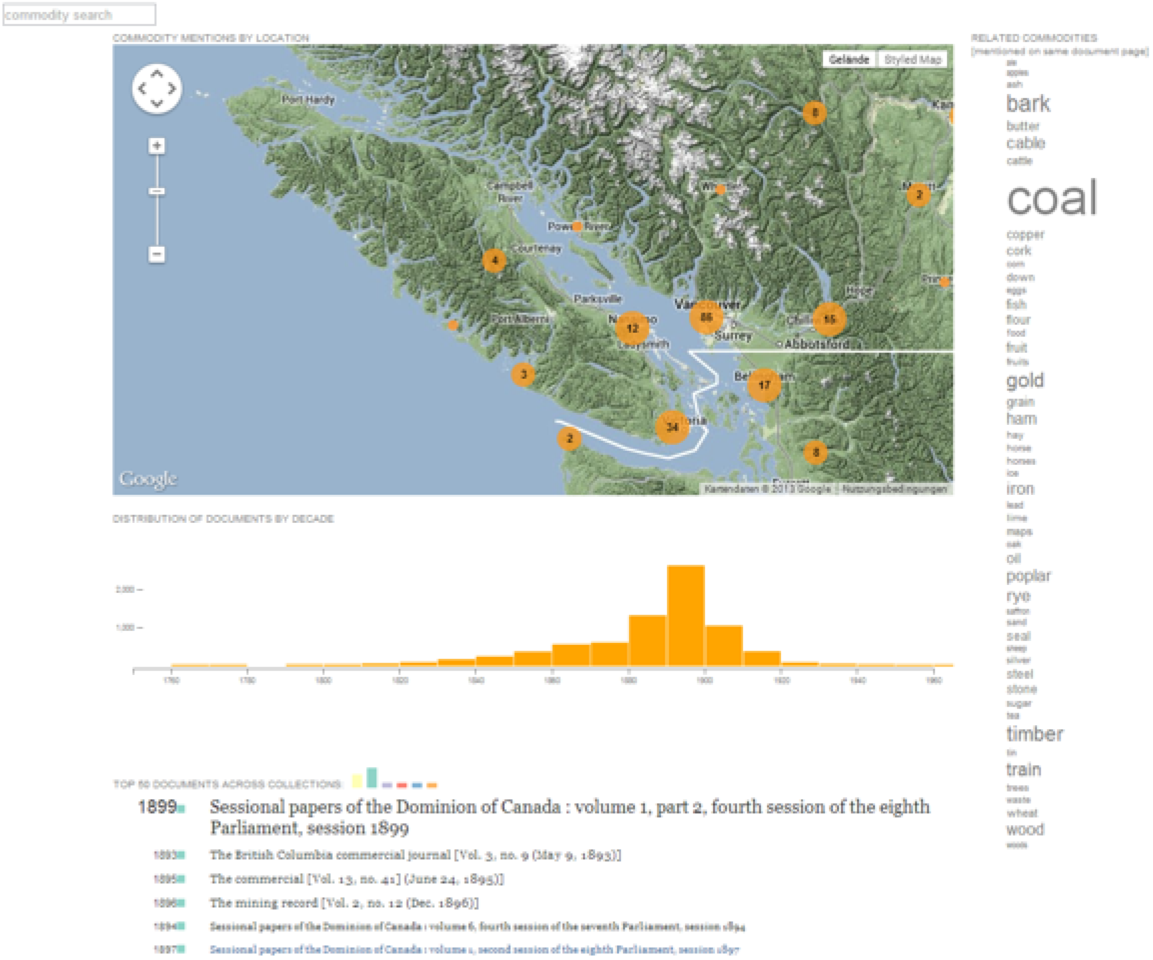
1906 Annual series of trade reports, serial no. 3475 to 3738 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

1902 Annual series of trade reports, serial no. 2696 to 2922 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

1905 Annual series of trade reports, serial no. 3256 to 3474 (Trade. Navigation. Shipping: Annual and miscellaneous series (foreign countries)) [Full version]

Fig. 2: Interlinked visualisations provide an overview of the document collection.

Interaction with one visualisation acts as a filtering mechanism and adjusts the data shown in the other visualisations. For instance, zooming into the map adjusts the tag cloud to only include commodities mentioned in relation to visible locations; the bar chart only shows documents that include these commodity/location mentions (Fig. 3). Particular time frames can be selected to further filter the document corpus; the other visualisations are updated accordingly (Fig. 4).



1899 Sessional papers of the Dominion of Canada : volume 1, part 2, fourth session of the eighth Parliament, session 1899

1893 The British Columbia commercial journal [Vol. 3, no. 9 (May 9, 1893)]

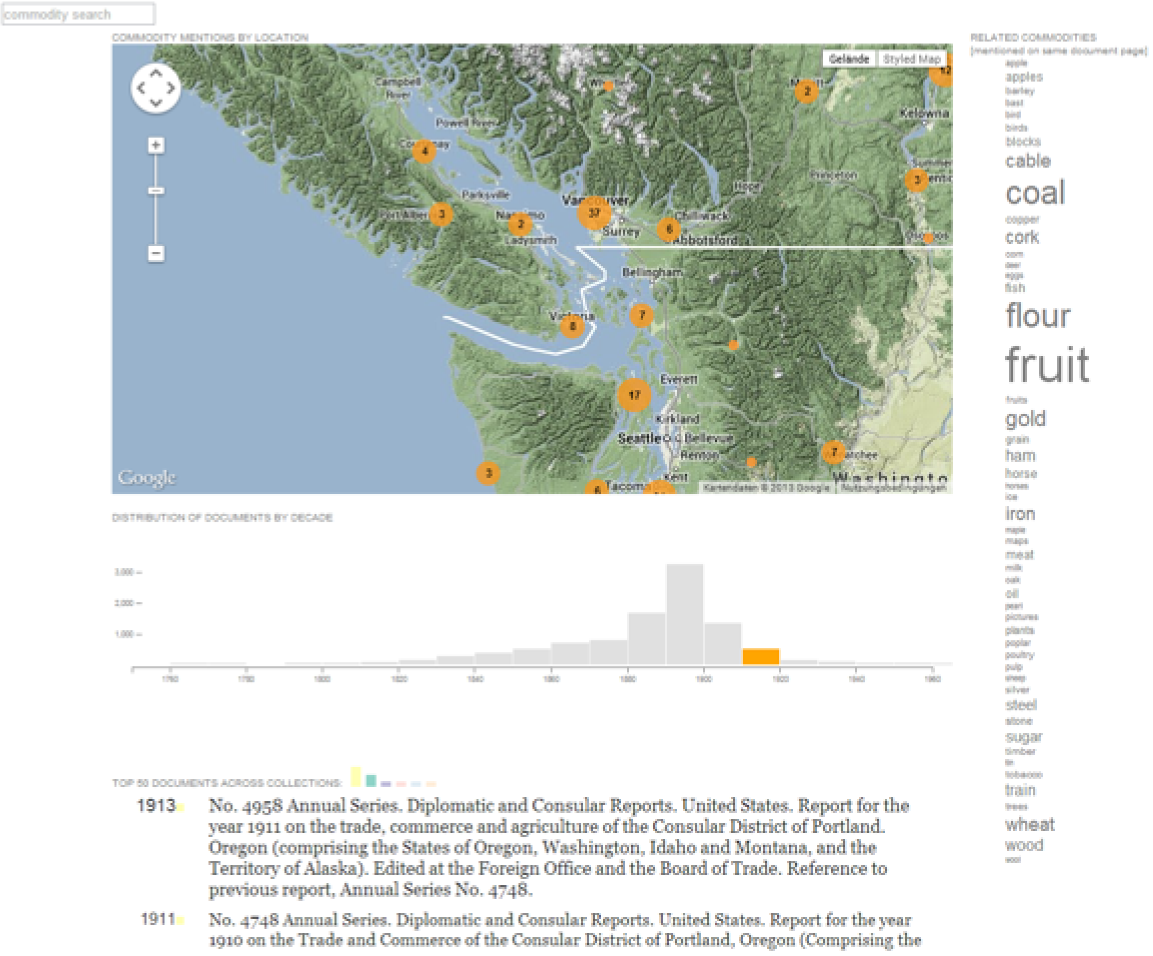
1895 The commercial [Vol. 13, no. 41] (June 24, 1895)]

1896 The mining record [Vol. 2, no. 12 (Dec. 1896)]

1894 Sessional papers of the Dominion of Canada : volume 6, fourth session of the seventh Parliament, session 1894

1897 Sessional papers of the Dominion of Canada : volume 1, second session of the eighth Parliament, session 1897

Fig. 3: Specifying the location adjusts the other visualisations.

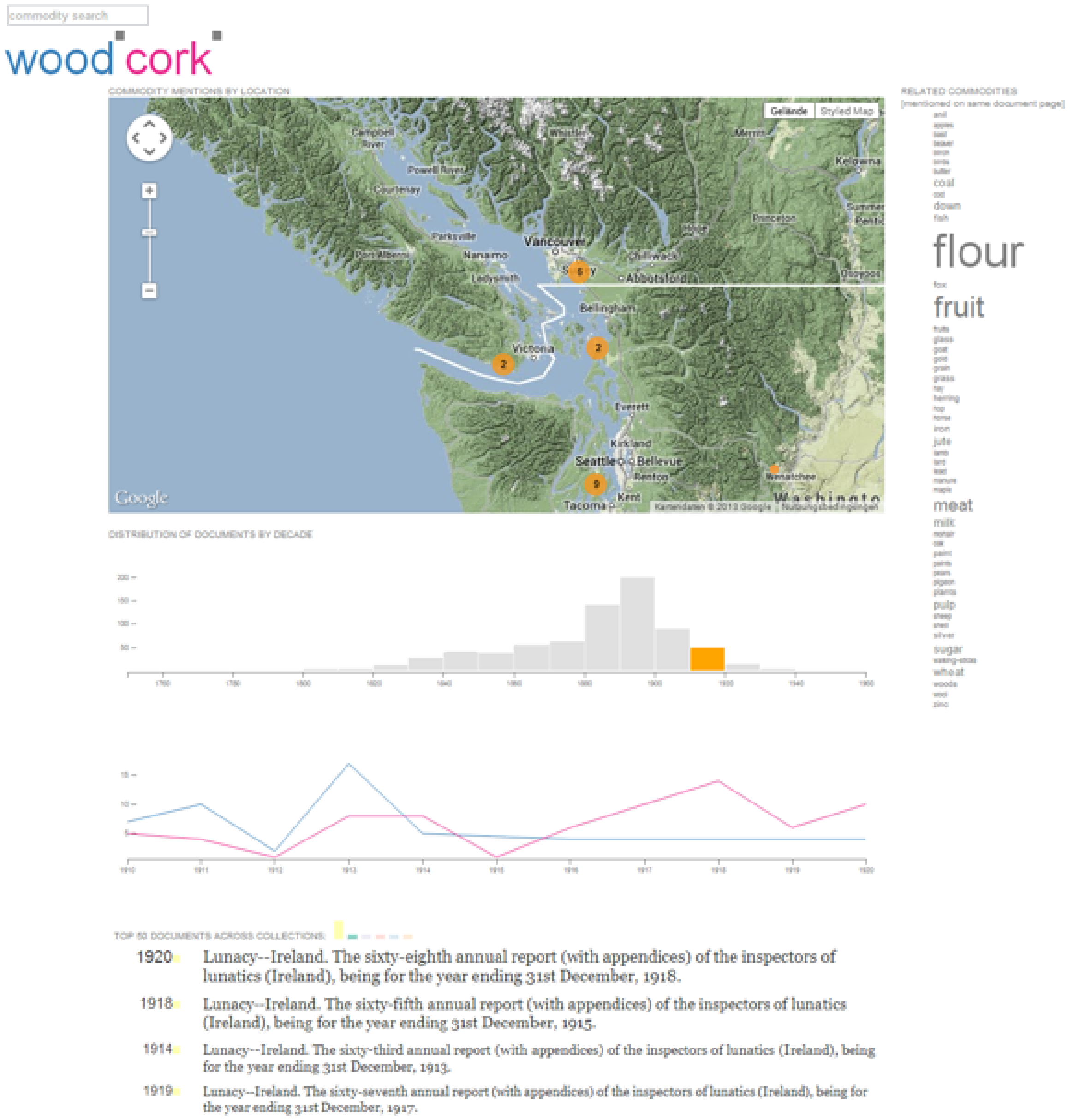


1913 No. 4958 Annual Series. Diplomatic and Consular Reports. United States. Report for the year 1911 on the trade, commerce and agriculture of the Consular District of Portland, Oregon (comprising the States of Oregon, Washington, Idaho and Montana, and the Territory of Alaska). Edited at the Foreign Office and the Board of Trade. Reference to previous report, Annual Series No. 4748.

1911 No. 4748 Annual Series. Diplomatic and Consular Reports. United States. Report for the year 1910 on the Trade and Commerce of the Consular District of Portland, Oregon (Comprising the

Fig. 4: Specifying a time frame adjusts commodities and locations shown in the tag cloud and map.

Lastly, historians can specify commodities of interest, either by textual query or by selecting commodities from the tag cloud. All visualisations adjust, with the tag cloud showing commodities related to the selected ones. An additional line chart presents the frequency of mentions of selected commodities across time (Fig. 5).



1920 Lunacy--Ireland. The sixty-eighth annual report (with appendices) of the inspectors of lunatics (Ireland), being for the year ending 31st December, 1918.

1918 Lunacy--Ireland. The sixty-fifth annual report (with appendices) of the inspectors of lunatics (Ireland), being for the year ending 31st December, 1915.

1914 Lunacy--Ireland. The sixty-third annual report (with appendices) of the inspectors of lunatics (Ireland), being for the year ending 31st December, 1913.

1919 Lunacy--Ireland. The sixty-seventh annual report (with appendices) of the inspectors of lunatics (Ireland), being for the year ending 31st December, 1917.

Fig. 5: Specifying particular commodities of interest further adjusts the visualisation.

3. Feedback from Historians

To gain expert feedback on our approach of combining text mining with visualisations to facilitate research in environmental history, we conducted a half-day workshop where we introduced our visualisation prototype to historians. The workshop was held at the Canadian History & Environment Summer School 2013 with over 20 environmental historians participating *[14]*. At the workshop, we asked historians to explore the visualisation tool in small groups (Fig. 6). To promote engagement with the different visualisations and to fuel discussions, the explorations were guided by a number of open-ended tasks, such as querying for commodities of interest or focusing on a geographic area.

Some historians immediately started to focus on the Vancouver Island area where the workshop took place. Others experimented with commodities and locations related to their own research. In general, these first exploration periods were about verifying familiar facts to assess the capabilities of the visualisation and the trustworthiness of the underlying data. The historians quickly understood the general purpose and high-level functionality of the visualisations and were able to start their explorations immediately. There was some confusion, however, about lower level details. For instance, the meaning of the size and number of clusters in the map was unclear (e.g. do they represent number of documents, or number of commodity mentions?). Observing changes in the visualisations while adjusting parameters helped, but our observations highlight that clear labelling and tooltips are crucial for visualisations in the context of digital humanities, not only because these are a novel addition to traditional research methodologies, but also because they can be easily misinterpreted. The meaning of visual representations needs to be clear in order to make visualisations a valid research tool.

Fig. 6: User workshop at CHESS 2013.

Workshop participants found the meta-level overviews of the visualisations valuable as these can aggregate information about the document corpus beyond human capacity. In the short time of the workshop, historians made (sometimes surprising) discoveries that sparked their interest to conduct further research. While it is unclear if these discoveries withstand more detailed investigation (there is still some noise in the data), this shows that visualisation has the potential to support exploration and insight in the context of history research.

A large part of the discussions focussed on what kind of insights can be gathered from the visualisations. Some historians pointed out that the visualisations represent the *rhetoric* around commodity trading in the 19th century: they show where and when a *dialogue* about particular commodities took place, rather than providing information about the occurrence of commodities in certain locations. This raises the question of how we can clarify what kind of data the visualisations are based on to avoid misinterpretation.

4. Conclusion

In general, we received positive feedback about our approach of combining text mining and visualisation to help research processes in environmental history. Historians saw the largest potential in the amounts of data that can be considered for research but also in the open-ended character of the explorations that the visualisations support in contrast to common database search interfaces. Other types of visualisations were suggested to help analyse and discover relations and patterns in the data, something that we are currently developing.

Our future research will explore how our approach integrates into research processes in environmental history and how it can produce profound outcomes. This will involve controlled experiments including directed and open-ended tasks. We will also conduct long-term studies to evaluate the discoveries and limitations that historians encounter when using our tools. The wide-ranging feedback from the workshop was crucial in helping the computer science team members understand priorities and research methodologies of environmental historians. Expert feedback is an important component of interdisciplinary research in digital humanities.

References

1. Trading Consequences: http://tradingconsequences.blogs.edina.ac.uk/
2. Digging Into Data: http://www.diggingintodata.org/
3. William Cronon (1992). Nature's Metropolis: Chicago and the Great West. W.W. Norton, New York.
4. Data collections: http://tradingconsequences.blogs.edina.ac.uk/about/the-corpus/
5. LT-XML2: http://www.ltg.ed.ac.uk/software/ltxml2; LT-TTT2: http://www.ltg.ed.ac.uk/software/lt-ttt2
6. GeoNames: http://www.geonames.org/
7. Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball (2010). Use of the Edinburgh Geoparser for georeferencing digitised historical collections. Philosophical Transactions of the Royal Society A.
8. DBpedia: http://dbpedia.org. We accessed DBpedia via the SPARQL endpoint (http://dbpedia.org/OnlineAccess), most recently on 16/12/2013, corresponding to DBpedia version 3.9.
9. Ewan Klein, Beatrice Alex and Jim Clifford (2014). Bootstrapping a historical commodities lexicon with SKOS and DBpedia, In: Proceedings of the LaTeCH 2014 workshop at EACL 2014.
10. EDINA: Jisc-designated centre for digital expertise & online service delivery; http://edina.ac.uk/
11. Stuart K. Card, Jock D. Mackinlay, Ben Shneiderman (eds.) (1999). Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Publishers, Chapter 1, pp. 1-34.
12. Gary Marchionini (2006). Exploratory search: From finding to understanding. Communications of the ACM 49, 4, 41–46.
13. Marian Dörk, Sheelagh Carpendale, Christopher Collins and Carey Williamson (2008). VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. IEEE Transactions on Visualization and Computer Graphics, 14(6), pp. 1205-1212.
14. CHESS 2013: http://70.32.75.219/2013/04/12/cfp-canadian-history-and-environment-summer-school-2013-vancouver-island/