

Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration

Uta Hinrichs

SACHI Group, School of Computer Science, University of St Andrews, UK

Beatrice Alex

ILCC, School of Informatics, University of Edinburgh, UK

Jim Clifford

Department of History, University of Saskatchewan, Canada

Andrew Watson

Robarts Centre for Canadian Studies, York University, Canada

Aaron Quigley

SACHI Group, School of Computer Science, University of St Andrews, UK

Ewan Klein

ILCC, School of Informatics, University of Edinburgh, UK

Colin M. Coates

Robarts Centre for Canadian Studies, York University, Canada

Abstract

Large-scale digitization efforts and the availability of computational methods, including text mining and information visualization, have enabled new approaches to historical research. However, we lack case studies of how these methods can be applied in practice and what their potential impact may be. Trading Consequences is an interdisciplinary research project between environmental historians, computational linguists, and visualization specialists. It combines text mining and information visualization alongside traditional research methods in environmental history to explore commodity trade in the 19th century from a global perspective. Along with a unique data corpus, this project developed three visual interfaces to enable the exploration and analysis of four historical document collections, consisting of approximately 200,000 documents and 11 million pages related to commodity trading. In this article, we discuss the

Correspondence:

Uta Hinrichs, SACHI Group,
School of Computer Science,
University of St Andrews, St.
Andrews, Scotland,
UK.

E-mail:

uh3@st-andrews.ac.uk

potential and limitations of our approach based on feedback from historians we elicited over the course of this project. Informing the design of such tools in the larger context of digital humanities projects, our findings show that visualization-based interfaces are a valuable starting point to large-scale explorations in historical research. Besides providing multiple visual perspectives on the document collection to highlight general patterns, it is important to provide a context in which these patterns occur and offer analytical tools for more in-depth investigations.

1 Introduction

Trading Consequences¹ is a 2-year interdisciplinary project funded under the ‘Digging into Data’² initiative. The project, conducted by researchers in environmental history, computational linguistics, information visualization, and database systems, focuses on the exploration of commodity trade in the 19th-century British world. Information on commodities was computationally extracted from over 200,000 historical documents. This data was then visualized to enable large-scale open-ended explorations in ways that authors of the original documents would have never imagined.

Traditionally, historians studying commodities and their environmental consequences have tended to focus on a limited number of commodities (see, e.g. William Cronon’s influential research on beef, lumber and wheat; Cronon, 1992). In contrast, Trading Consequences aims to identify global trends in commodity trading for 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. The main contribution of the Trading Consequences project is the exploration of computational methods—text mining (TM) and information visualization—to help to identify and illustrate the significance of specific commodities in relation to particular places and time to facilitate in environmental history research.

This article describes our approach of facilitating the investigation of high-level trends in the collocation of commodities and place names in text, based on a large corpus of 19th-century (mostly) government documents, while enabling detailed analysis of

trends on a sentence-level. We particularly focus on the visualization-based tools that were developed to make the text-mined data explorable. We discuss design considerations and resulting functionalities but also reactions from historians within and outside our research team. As computational tools are increasingly becoming a vital part of humanities’ research methods, it is important to consider how they are utilized and experienced, and, ultimately, if and how they can facilitate new discoveries. Our work provides insights on the role that information visualization, beyond the ubiquitous map display, can play in environmental history research. We first present an overview of our general approach, followed by a description of the initial visualization prototype we developed and the feedback provided by a group of historians from the Network in Canadian History & Environment (NiCHE).³ We then present the second visualization prototype we developed based on this feedback and which historians in our team tested for several months. We end with a critical discussion of our approach to facilitate the analysis of large document corpora as part of historical research and outline considerations that can inform the design of exploratory, visual interfaces to support research in the humanities in general.

2 Trading Consequences— Underlying Sources and Technical Approach

As part of Trading Consequences, we computationally analysed digital documents from four major British and Canadian text collections relevant to trade in the 19th century.⁴ The British House of

Commons Parliamentary Papers (ProQuest) and Early Canadiana Online (Canadiana.org) are some of the largest digitized collections of British and Canadian historical documents with a combined total of more than 10 million pages. They are a major source for more conventional historical research and the most obvious collection for a text-mining project interested in British and Canadian history. The other two major collections, Confidential Prints (Adam Matthews) and Kew Gardens Directors Correspondence, are significantly smaller, but focus on global correspondence, which was of particular interest to this project. We only included English documents—processing documents in other languages was beyond the scope of this project. While the resulting document collection is limited, it still consists of over 200,000 documents, 11 million pages, and over 7 billion analysed word tokens. The digital documents form the basis of our text-mining approach which relies heavily on the use of lexicons and gazetteers. The mined information is stored in a relational database, which serves as the backend to three exploratory, visualization-based interfaces. An overview of the technical approach of Trading Consequences is shown in Fig. 1 and described in the remainder of this section.

2.1 Text mining

The TM tools were developed by the Language Technology Group at the University of Edinburgh.⁵ We adapted an existing text-mining pipeline built on in-house tools⁶ to process historical text. As mentioned above, we only processed English documents within a highly multilingual text collection (e.g. Early Canadiana Online). If a text's metadata did not include its language origin, we computed it automatically.⁷ For collections that we knew contained mostly English language content, we processed all documents.

A large amount of the input text was of poor quality as a result of optical character recognition (OCR)—the digitization process used to convert scanned images to electronic text. We devoted some research to rating the quality of the text documents automatically and to identify an appropriate threshold above which documents were considered

as sufficiently high quality for TM (Alex and Burns, 2014). However, this was carried out after completion of the last iteration of text-mining processing, and OCR accuracy-based filtering was not applied in practice in this project.

The TM component includes a series of processing steps that build up the linguistic information in a given text. A preprocessing stage includes tokenization, sentence-splitting, part-of-speech tagging, lemmatization and chunking to determine words and sentences, identification of their syntax, computation of canonical forms of word tokens, and recognition of verb and noun phrases within sentences. The next steps are named-entity recognition and grounding: mentions of locations, commodities, and dates are automatically identified in the text and grounded to unique identifiers in existing knowledge databases. For example, we ground location mentions to GeoNames⁸ identifiers, together with their corresponding latitude/longitude coordinates and feature types. We use an adapted version of the Edinburgh Geoparser for this geo-referencing process (Grover *et al.*, 2010; Alex *et al.*, 2015). The final TM step identifies relations between commodity, date, and location mentions. Here, a commodity–location relation is recognized if both appear within the same sentence. The resulting TM output is stored in XML format.⁹

At the start of the project, we did not have software for recognizing and grounding commodity mentions in text, and there was no comprehensive list of commodities relevant to the 19th century available. We therefore created such a resource semi-automatically, starting with a small seed set of commodities manually collected from archival sources, which we expanded automatically from several hundred commodities to a lexicon of over 20,000 entries linked to DBpedia¹⁰ concepts and categories. The historians then manually edited this resource: missing commodities were identified by means of contextual bigram analysis of development data. This manual error analysis resulted in an increase in both precision and recall of the mined commodity output (Klein *et al.*, 2014a, Klein *et al.*, 2014b). The output of the improved text-mining system forms the basis for the visualizations that are accessible today.

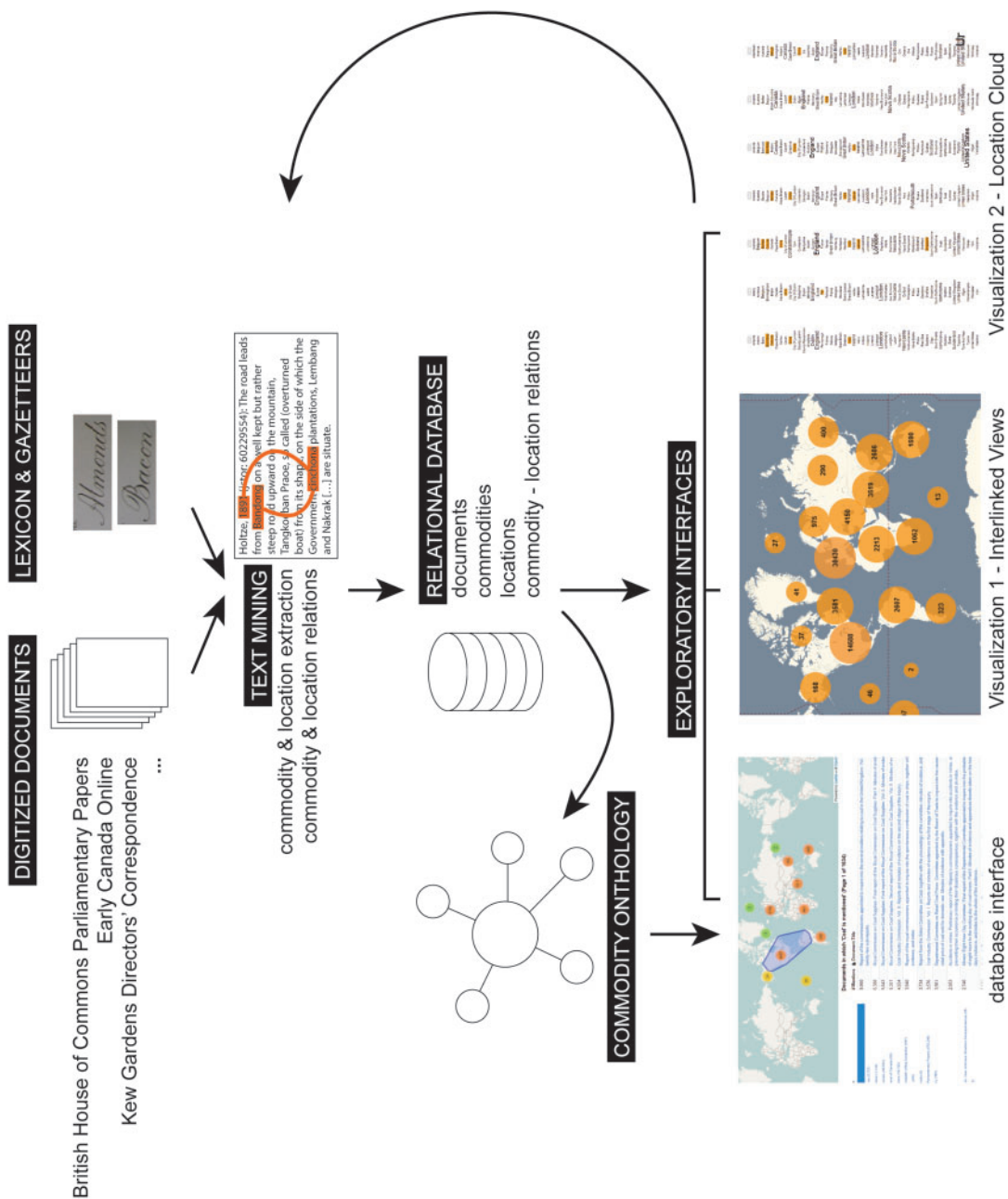


Fig. 1 Technical approach of trading consequences

2.2 Relational database

The TM output is stored in a relational PostgreSQL¹¹ database set up and hosted by EDINA¹² for subsequent querying and visualization. PostgreSQL is an open-source object-relational database system, with the advantage of supporting geospatial queries through PostGIS. The database consists of four main tables that contain information about each historical document, the extracted commodities, the extracted locations, and the commodity–location relations.

The Documents table stores information about 200,871 documents—document title, author, publication year, collection, and a URL to the original digitized document. The Commodity Mentions table stores all mentions of each commodity (currently 28,595,550), reproducing exactly the spelling of the commodity in the text. Each commodity mention is linked to the corresponding document, the page identifier, sentence from which it was extracted, and the corresponding DBpedia concepts and categories. The latter helps to categorize, e.g. commodity mentions of different spellings that refer to the same commodity.

All Location Mentions are stored along with their corresponding document, sentence, page identifier, latitude/longitude coordinates, and GeoNames identifier (see Section 2.1). There are currently 74,744,515 location mentions stored in the database, corresponding to 2,275,186 unique locations identified in the corpus. The TM component identified a total of 13,969,659 Commodity–Location Relations across the entire document corpus which are stored in an additional table. Each commodity–location pair is linked to the corresponding commodity/location mentions tables via identifiers to enrich them with additional information such as the corresponding document, sentence, commodity concept, and latitude and longitude. There are some limitations to our text-mining approach. For instance, while the text-mining tools automatically identified 595,121 unique commodities, this large number is the result of incorrect identifications: sequences of commodities found in tables (e.g. ‘Wheaten Bread Horned Cattle Sheep’) were misidentified as a single unique commodity. Such inaccuracies add noise to our data set and, in some cases, also limit

the accuracy of our exploration tools. Further improvements of the text-mining tools will address this issue. However, the vast majority of misidentified commodities appear less than ten times in the corpus, and many were not connected to a location, which is why this issue does not have a major effect on the tools’ overall functionality. In fact, most aspects of the exploration tools focus on patterns where individual commodities and locations collocated hundreds or thousands of times.

As the numbers above show, the data resulting from the TM output is large and stored in a format unfamiliar to a majority of historians. While the database is available to scholars directly and it is possible to retrieve and filter data via database queries, this still leaves researchers with thousands of rows of data, which are effectively impossible to read and analyse manually or through conventional tools such as Excel. It was our goal to make the mined data accessible and explorable to historians across the world, without requiring knowledge of database queries.

Geographic Information Systems (GIS), familiar to a growing minority of historians, provide methods of visually mapping the type of data we produced. For instance, one could query the database for all location mentions related to ‘coal’ and input these data into GIS software. Even temporal aspects could be explored, using the Timeline features recently introduced into ArcGIS.¹³ This, however, still requires writing SQL queries and experience with GIS software, and the time involved in transferring the database and becoming familiar with its content hinders an exploratory approach to the data.

We therefore developed a number of visualization-based web interfaces that show the different aspects of our mined data in textual and visual ways and, as part of this, enable interactive explorations, without requiring prior knowledge of database systems or GIS tools.

2.3 Database interface

We first built a database interface that supports standard text-based search to query for commodities and locations of interest.¹⁴ A search for a commodity brings up a view that lists its corresponding DBpedia categories (see Fig. 2A) and provides a zoomable

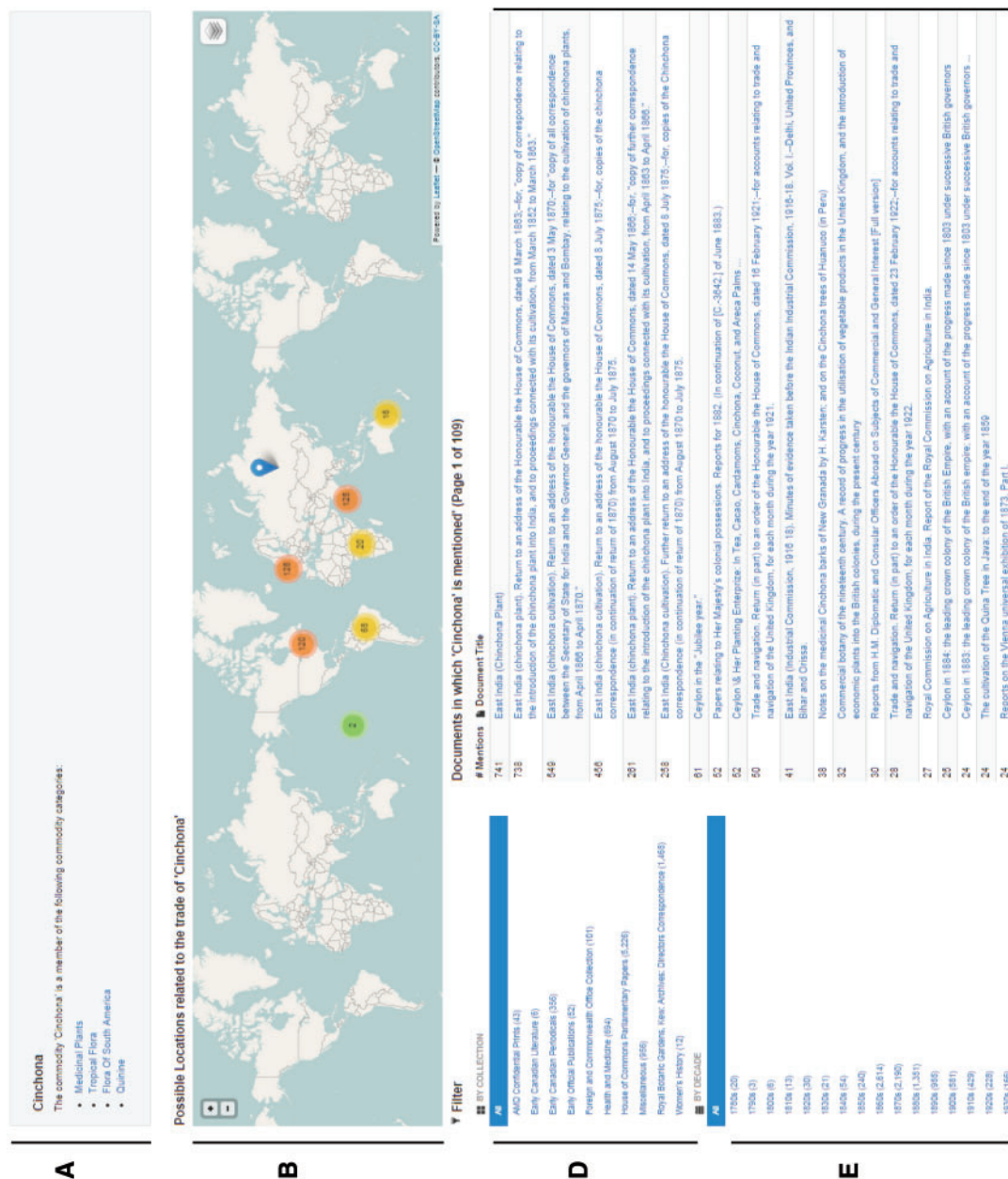


Fig. 2 Trading Consequences database interface. Here, the commodity “Cinchona” has been entered. A: corresponding DBpedia categories; B: geographic locations mentioned in the context of “Cinchona”; C: corresponding document list; D: corresponding collections; E: corresponding decades extracted from the documents’ publication years

Search by:

Commodity

Location

Eighth report from the Select Committee on Sugar and Coffee Planting; together with an appendix.

Collection

House of Commons Parliamentary Papers

Author

Not known

Publication Year

1847

Web address

http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&res_dat=xri.hcpp&rft_dat=xri.hcpp.fulltext:1847-024979

Sentences in which 'Sugar' is mentioned' (Page 1 of 75)

Fig. 3 Database interface—document view

heatmap displaying location clusters related to the selected commodity term (see Fig. 2B).

In addition, a list of documents is shown that contains the selected commodity term (see Fig. 2C) and which can be filtered by collection and decade (see Fig. 2D and E). Selecting a document brings up a view with more details about related commodity/location mentions within the document, their immediate context, and links to their digital sources (see Fig. 3).

The database interface also supports location queries. Fig. 4, for example, shows the view resulting from a query for 'Ceylon (Sri Lanka)', which highlights over 60,000 commodity mentions, nearly 56,000 found in the HCPP collection, peaking at approximately 12,000 mentions early in the 20th century. The text-mining results generally correspond with the island's major 19th-century exports, and tea and coffee rank second and third in the list.

The database interface employs a familiar search style, as seen in library catalogues. Research shows that such interfaces efficiently support targeted search where the search goal is clearly defined (Marchionini, 2006). However, more open-ended inquiries that characterize approaches in historical research are not well supported through list-based results, especially since these can contain hundreds to thousands of documents. We therefore designed alternative interfaces that draw from information visualization techniques, aiming to facilitate both targeted and open-ended exploration of the Trading Consequences data.

2.4 Visualization approach

Information visualization as 'the use of computer-supported, interactive, visual representations of abstract data to amplify cognition' (Card *et al.*, 1999, p. 7) enables the discovery and exploration



Fig. 4 Database interface—results of a location query on “Ceylon (Sri Lanka)”

of patterns and relations within data that would otherwise be difficult or impossible to recognize. In the context of Trading Consequences, our strategy was to provide information visualizations of the mined data to:

- (1) Highlight trends across the historical documents, in particular, relations between different types of commodities and the discourse around them in the context of geographic location and time.
- (2) Enable the interactive and iterative exploration of the document collection, supporting both targeted and open-ended search approaches.

Overall, our goal was to facilitate the development of new insights and/or research questions based on collection sizes that exceed possibilities of traditional humanities' research methods, and to promote data probing without requiring well-defined queries.

We decided early on to implement all visualizations as web-based tools to make them easily accessible and sharable by historians worldwide.¹⁵ The tools are implemented using PHP and JavaScript (D3.js¹⁶ and jQuery¹⁷); no installation of additional software is required. The visualizations have been optimized for and tested in the freely available Google Chrome¹⁸ Browser. In the following sections, we describe two visualization tools we developed as part of Trading Consequences. We outline how historians within and outside of our research team used and experienced these tools and how the visualizations can facilitate environmental history research in general.

3 Interlinked Views Visualization

The first visualization tool we developed, Interlinked Views¹⁹, aims at representing the three major attributes of our mined data: the temporal document distribution, the variety of commodities mentioned within these documents, and the spatial relationships extracted by the TM pipeline. Linking different types of visualizations leverages the advantages of different visualization techniques and has

been previously applied, for example, in the context of news articles (Dörk *et al.*, 2008).

3.1 Design and functionality

Interlinked Views consists of three interlinked information visualizations (see Fig. 5). A map visualization provides an overview of the geographic context of commodity mentions (see Fig. 5A). A vertical tag cloud shows the top fifty most frequently mentioned commodities in alphabetical order (see Fig. 5B). A bar chart presents the temporal distribution of historical documents across the source collections (see Fig. 5C). A ranked document list provides direct access to the relevant articles (see Fig. 5D), and a horizontal chart shows the distribution of documents across the collections (see Fig. 5E).

Interaction with one visualization acts as a filtering mechanism of the data shown in the others. For instance, zooming into the map adjusts the tag cloud to include only commodities mentioned in relation to visible locations, and the bar chart only shows documents that include these commodity/location mentions. Particular time frames can be selected to further filter the document corpus; the other visualizations are updated accordingly (see Fig. 6).

Historians can specify commodities of interest, either by textual query (see Fig. 5F) or by selecting commodities from the tag cloud. All visualizations adjust, with the tag cloud showing commodities related to the selected ones (i.e. commodities that are mentioned on the same document page with the selected commodity). An additional line chart presents the frequency of mentions of the selected commodities across time (see Fig. 7).

Selecting a document title from the list loads the corresponding digitized version of its source document in a new browser tab.

3.2 Feedback from historians

To gain expert feedback on our Interlinked Views, we conducted a half-day workshop at the NiCHE Summer School 2013 in Nanaimo on Vancouver Island, Canada²⁰ (see Fig. 8). Over twenty environmental historians and geographers at varying career stages participated in our workshop (PhD students, postdoctoral researchers, and tenured professors).

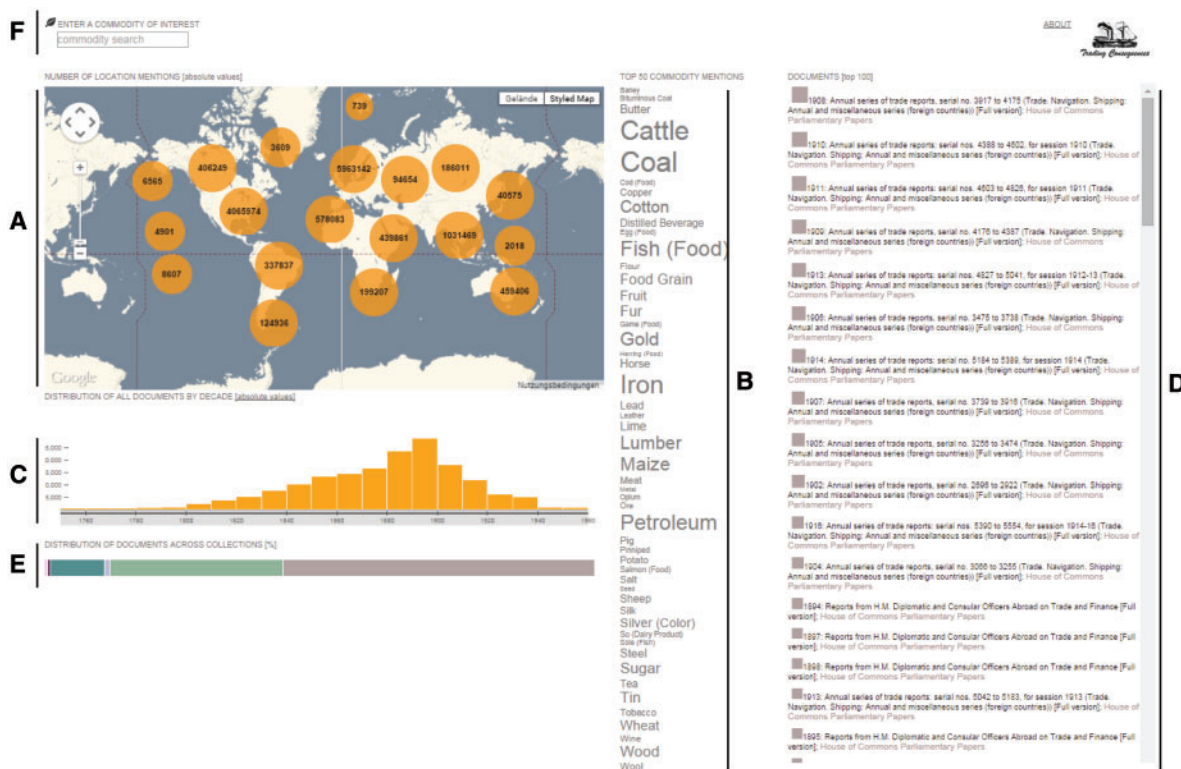


Fig. 5 Interlinked Views visualization with A: a map to provide a geographic context of location mentions, B: a tag cloud showing the most frequently mentioned commodity terms, C: a bar chart showing the temporal distribution of corresponding documents, D: the corresponding document list, and E: the distribution of these documents across the collections. A text field (F) can be used to specify commodities of interest

While all participants use computational tools as part of their research (e.g. digital search interfaces, databases, or GIS tools), none of them had used this visualization before.

During the workshop, we first introduced the general background and goals of Trading Consequences. This was followed by a short introduction of the Interlinked Views visualization, including its range of features. Participants were then asked to explore the visualization in groups of two. Each group had their own desktop computer with the tool launched in a browser, showing the initial overview screen (see Fig. 5). We chose to have participants work together in groups to promote discussions during their explorations.

While participants were generally encouraged to explore the visualizations based on their own

(research) interests, we suggested four open-ended exercises to initiate engagement with the different features of the visualization: exploring the commodities (1) ‘cinchona’ and (2) ‘cheese’, (3) zooming into a location of interest, and (4) exploring a particular time period. The exercises were introduced one-by-one, and participants were given 10 min to explore corresponding features. Each exploration session was followed by an open-ended discussion among workshop participants about their discoveries and impressions of the tool. The workshop concluded with a general discussion about the visualization and the use of computational methods as a part of environmental history research. Some historians immediately started to focus their exploration on the Vancouver Island area, where the workshop took place. Others experimented with

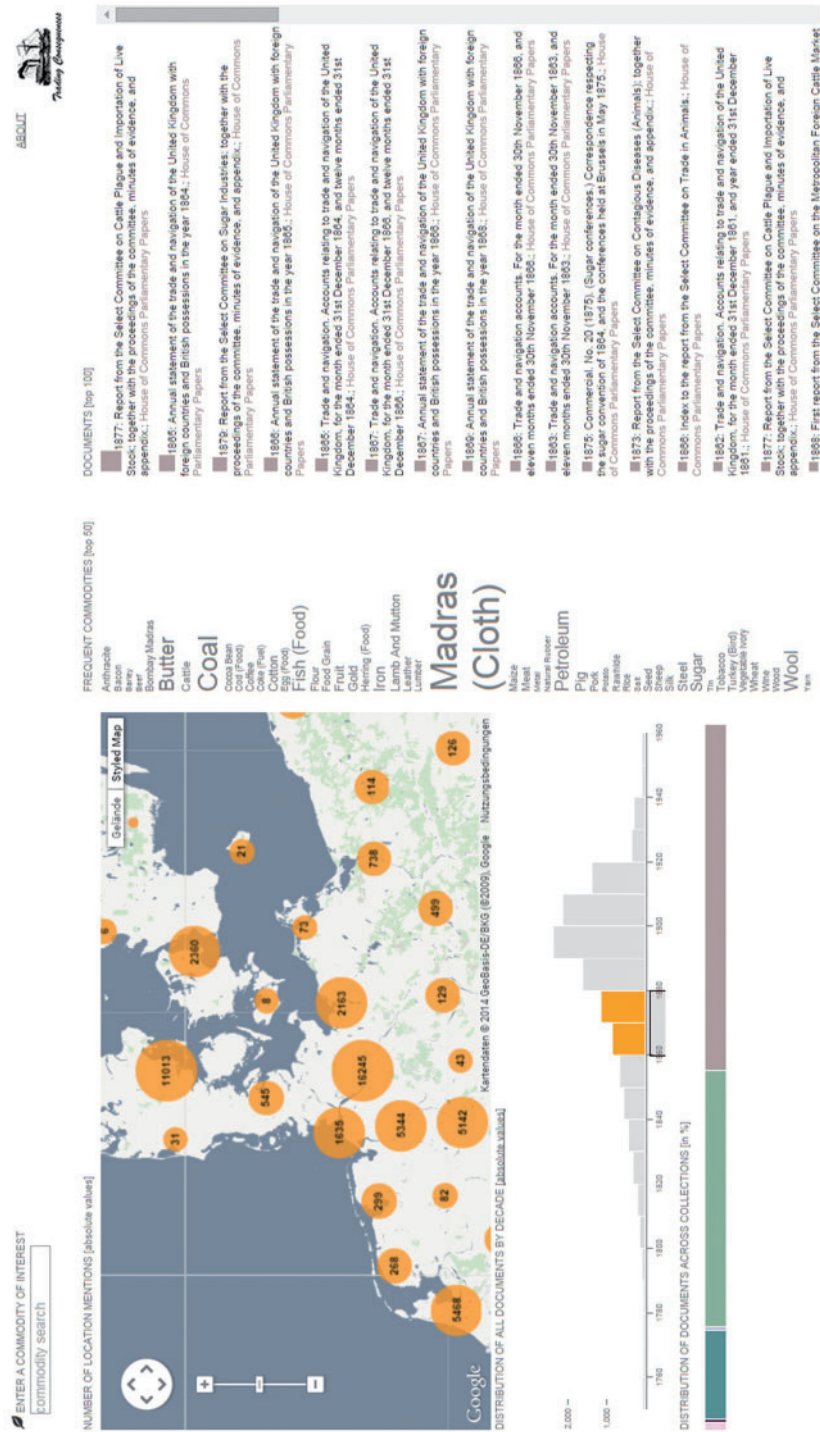


Fig. 6 Adjusting one visualization, e.g., zooming in on the map or selecting a time period of interest, acts as a data filter on the others

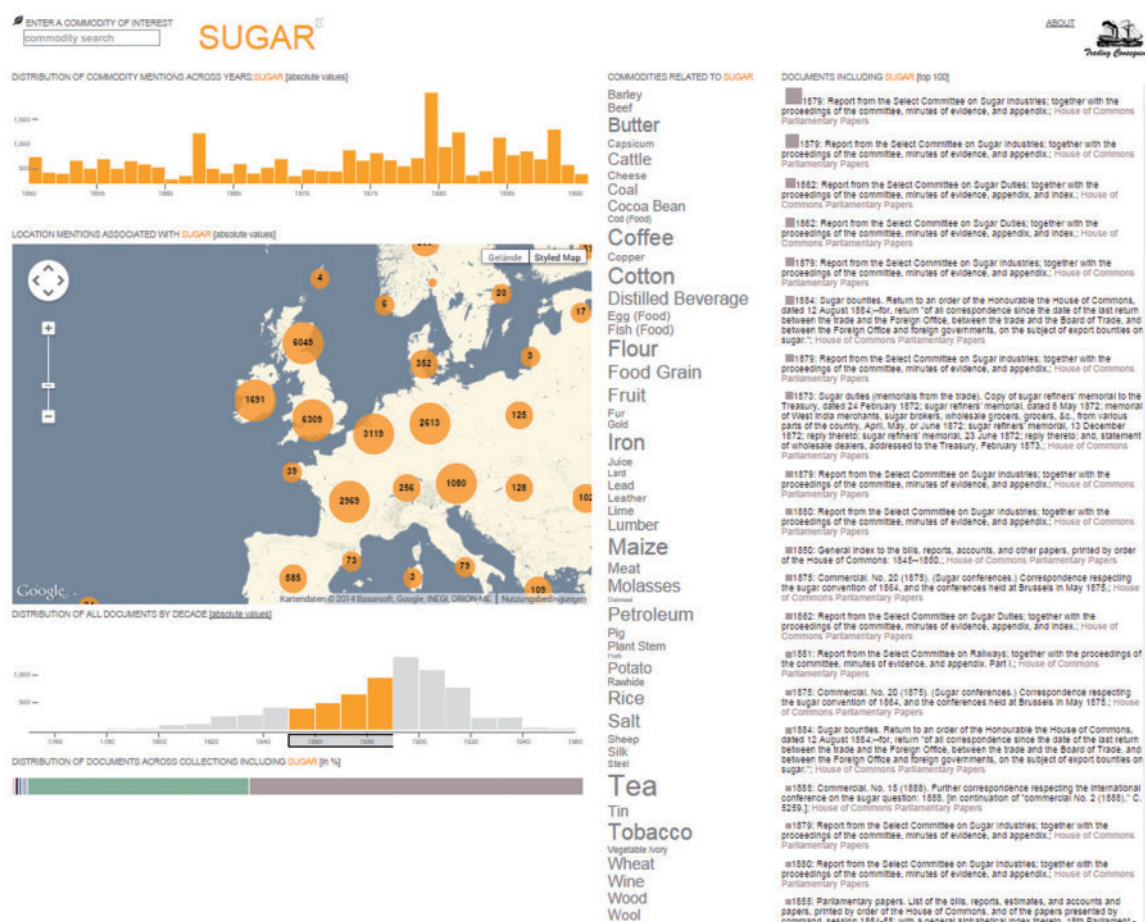


Fig. 7 Selecting a commodity (in this case “Sugar”) further adjusts all visualizations

commodities and locations related to their own research. In general, these first exploration periods were about confirming familiar facts to assess the capabilities of the visualization and the trustworthiness of the underlying data. The historians quickly understood the general purpose and high-level functionality of the visualizations 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 individual commodity/location mentions?). Observing changes in the visualizations while adjusting parameters improved understanding, but we reflect that clear

labelling and tooltips are crucial for visualizations 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 visualizations a valid research tool. Workshop participants found the meta-level overviews of the visualizations 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 investigations, this shows that



Fig. 8 Gathering feedback from workshop participants

information visualization has the potential to support exploration and insight in the context of historical research.

Since exploration can result in long document lists, additional features that allow a quick exploration of the document content were suggested. We addressed this in an iteration of the Interlinked Views visualization by providing a glimpse of sentence snippets as they occur in a document, based on a selected commodity (see Fig. 9).

Other criticisms included the lack of feedback mechanisms that communicate, for example, the loading of data: a sometimes lengthy process. Furthermore, historians expressed the need for confidence values to provide a better understanding of the reliability of the data and the corresponding visualizations. Future research needs to explore these issues from both a TM and visualization perspective.

Some historians pointed out that the visualizations represent the rhetoric of commodity trading in the 19th century: they show a correspondence about commodities and locations, 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 visualizations

are based on to avoid misinterpretations. With the Location Cloud visualization, we aimed at addressing this particular challenge.

4 Location Cloud Visualization

The Location Cloud moves towards a text-based visualization approach, highlighting the rhetoric nature of our data, while still providing a strong focus on its underlying geographic and temporal features. Furthermore, we aimed to provide more analytic functionalities that could help quantitative comparisons of geographic references and their frequency over time. The design of the Location Cloud is inspired by previous approaches from the information visualization literature, in particular, by utilizing vertical tag clouds to visualize email content across time (Viégas *et al.*, 2006) and the concept of parallel tag clouds (Collins *et al.*, 2009).

4.1 Location Cloud—design and functionality

In the Location Cloud, all locations mentioned in relation to a selected commodity are plotted in stacks, horizontally ordered by decade from 1800 to

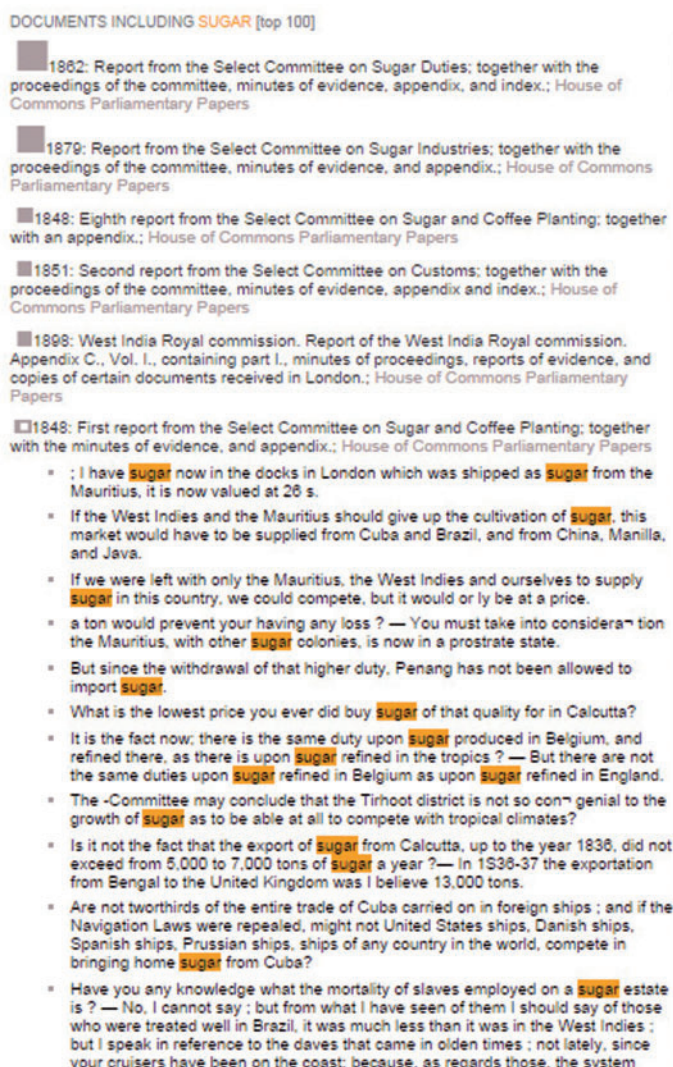


Fig. 9 Integrated sentence snippets in the document view

1920. The size of location terms corresponds to their relative frequency of mentions within the decade. To account for the varying distribution of documents across decades, we normalize the number of location mentions related to a specific commodity based on the total number of location–commodity relations within the particular decade. The Location Cloud in Fig. 10 shows that Mauritius, England/UK, and Cuba were prominent locations mentioned in relation to the commodity ‘sugar’. This type of text visualization has two advantages: Firstly, it does not

suggest a particular context of the data (e.g. export/import or production locations), but merely reports that a relation has been identified.

Secondly, in contrast to a map which can only provide an overview of commodity–location relations within one particular time frame in a single view, the Location Cloud provides details about temporal changes within the data. For instance, the distribution of mentioned locations becomes increasingly even in later decades as individual locations become less pronounced. This neatly reflects

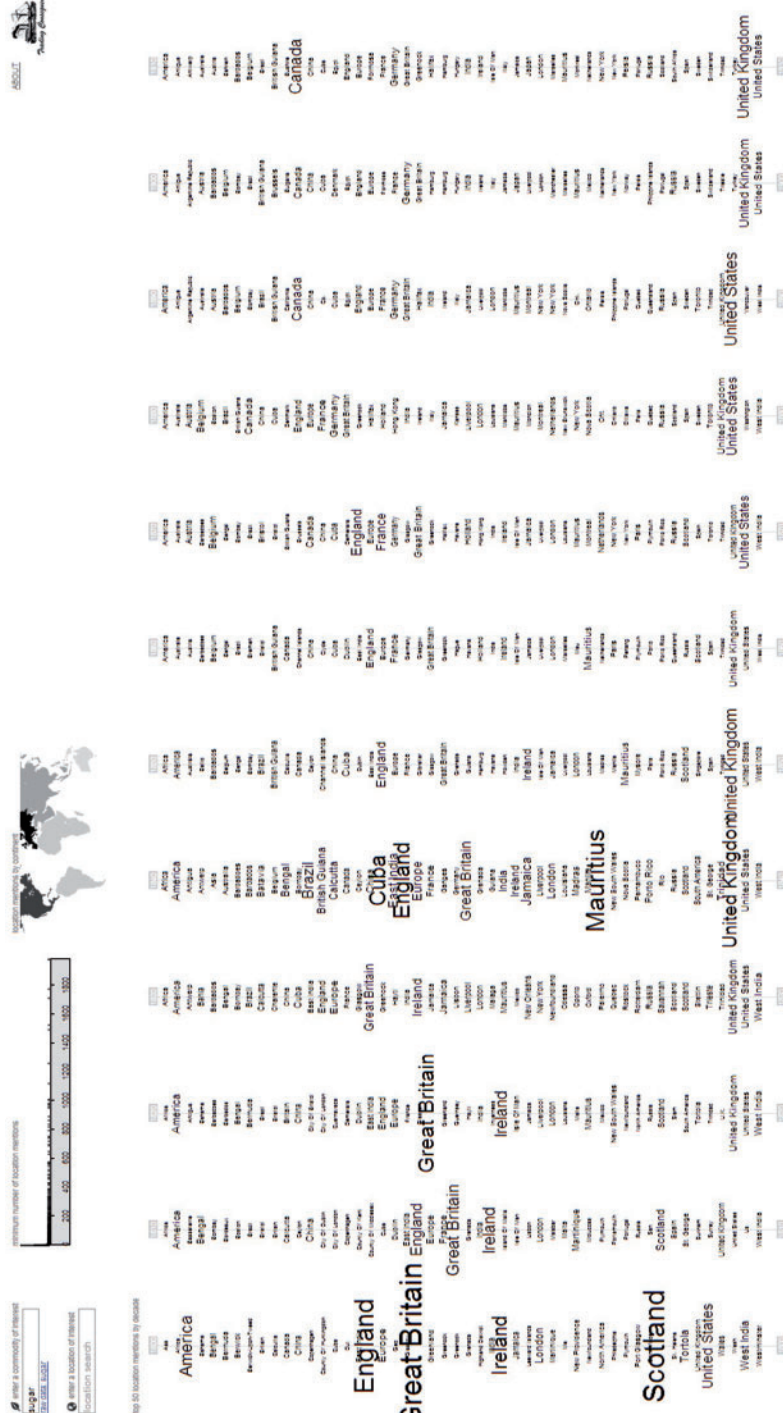


Fig. 10 Location Cloud for the commodity ‘sugar’

how trade globalization has increased the number of relevant locations over time and, in the case of sugar, the growing importance of sugar beets grown in Europe. Also, the frequency of particular locations changes over time. In later decades, locations such as Canada and the USA become more prominent. Location terms are ordered alphabetically to facilitate searching the lists for a location of interest. Depending on the selected commodity and decade, hundreds and thousands of associated location terms may need to be plotted. This can lead to overlapping location terms obscuring individual terms. To overcome this issue, we decided to show only the top sixty locations within each decade.

Hovering over a location highlights its appearance in the documents across decades (see Fig. 11). Locations that are not part of the 'top sixty' will appear at the bottom of the location list of their corresponding decade. For instance, Mauritius was not mentioned in relation to 'sugar' in the 1800s, but occurs in the 1810s, though not as part of the 'top sixty' locations.

The Location Cloud provides a range of features to evaluate and further explore apparent trends in more detail. For instance, locations can be filtered by continent and by frequency.

Fig. 12 shows all African locations that have been mentioned in relation to 'sugar' more than twenty times. This type of filtering allows historians to focus on particular continents and to filter out potentially irrelevant locations due to their low frequency within the document set. At the same time, it allows for a more detailed exploration of 'outlier' locations that have not been mentioned often, but may still be of interest.

Selecting a location of interest brings up a detailed view of the distribution of location mentions across individual years as well as sentence snippets that contain the mentions of the selected commodity and location within the selected decade (see Fig. 13).

If this detail view shows promise, a full document list can be brought up that contains all documents relevant to the selected commodity/location mention, including all corresponding sentences with the selected commodity and location mentions highlighted (see Fig. 14). Selecting a document will open its original digital copy in a new browser tab.

4.2 Location Cloud—internal evaluation

The historians on the Trading Consequences team provided ongoing feedback throughout the development of the Location Cloud and have been utilizing it as part of their research since early 2014. Their feedback confirmed that one of the strengths of this visualization is its ability to show change over time—a key goal of historical inquiry. In contrast, the Interlinked Views visualization allows the selection of particular decades, but it does not allow the side-by-side comparison of related location data without opening multiple browser tabs. Historians confirm that the Location Cloud facilitates the detection and exploration of the places associated with different commodities in the corpus across time.

4.2.1 Confirmation of known historical facts

Our team of historians found that, in many cases, the Location Cloud confirms expected trends. For instance, searching for the term 'wheat', shows that place names in Western Canada become more prominent by the end of the 19th century (see Fig. 15a): Wheat - Manitoba, b) Wheat - Ontario, and c) Wheat - Winnipeg). Furthermore, India is strongly associated with cinchona during the 1860s (see Fig. 16). The success of the visualization in highlighting known historical trends suggests the tool works and increases historians' confidence in researching the unexpected trends.

4.2.2 Unexpected trends and their background

In other cases, the visualization shows unexpected results, which highlight the power of TM coupled with exploratory visualizations. Nineteenth-century historians are confronted with a deluge of sources, and it is much easier to follow obvious trends and more difficult to find sources that provide a richer perspective. Trade statistics, which are easy to find and read, show the collapse of Russian tallow exports to the UK during the mid-1860s. It is surprising, therefore, to see the continued prominence of Russia through to the 1890s in the 'tallow' Location Cloud (see Fig. 17). This is where historians found the ability to link through to the actual documents particularly helpful. In some cases, the connection between Russia and tallow remains because the trade statistics tables kept Russia in place, even as



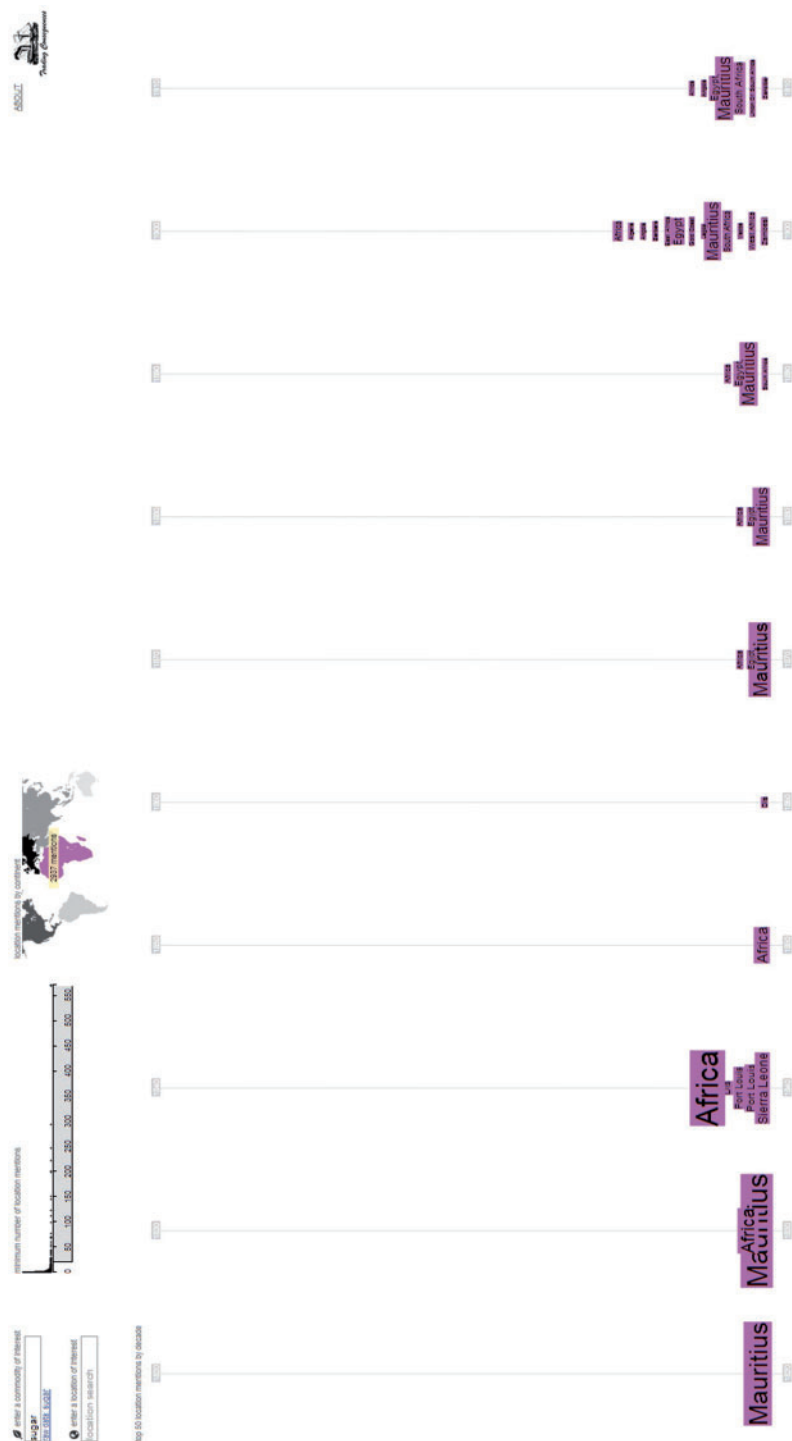


Fig. 12 Filtering the Location Cloud by continent and frequency of mentions shows how ‘sugar’ has been mentioned in relation to locations in Africa

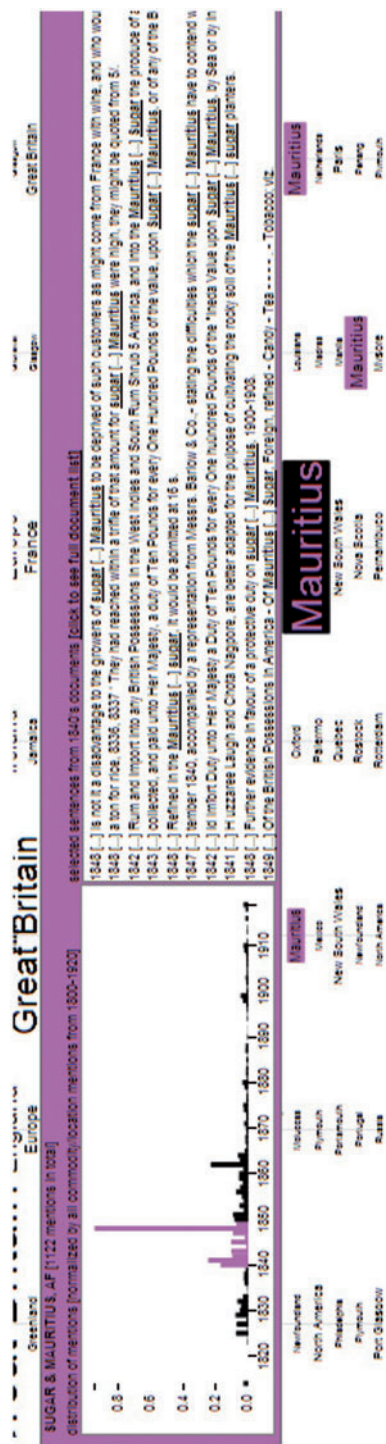


Fig. 13 Detail view for the location ‘Mauritius’ corresponding to the commodity ‘Sugar’

the quantity of exported tallow declined. In other cases, however, the documents actually discussed the lack of tallow exports and the significant reduction in the number of Russian cattle since the mid-19th century. The fact that British consular officials were still discussing the collapse of Russian tallow exports in the 1890s is historically significant, and it would have been hard to find these discussions without the Location Cloud visualization prompting the question and providing a link to the source. This example shows how the Location Cloud can facilitate exploring the relationship between the TM results and trade statistics. In future work, we need to focus on distinguishing between commodity–location relations with positive and negative linguistic contexts and more fine-grained attributes as shown in this example. The Russia tallow example is not a eureka moment where an unknown relationship between a major commodity and an important location is discovered, but this is rarely how we develop historical knowledge. Instead, it is an example of a historian with prior knowledge of the global tallow trade using the Location Cloud to find relevant new information about the long decline of the tallow industry on the Eurasian Steppe. Computational methods overlap with more traditional methods, and it is in combination that they provide new insights.

4.2.3 Additional features

While the Location Cloud shows a lot of promise, our in-team evaluation revealed some aspects that can be addressed through the introduction of additional features. The nature of the corpus, which includes significantly more documents towards the end of the 19th century than at the beginning, causes most commodities to follow a similar trend. To counteract this, we normalized the data as described above. This improved the results in that apparent trends have more merit in many of the cases the historians studied. However, for certain analyses, different normalizations or even no data normalization at all may be desirable. Future features could provide ways of changing the type of normalization on demand, fluidly adjusting font sizes of location mentions in the Location Cloud.

Documents from the 1840's mentioning SUGAR in the context of MAURITIUS, AF

[click collection square to see sentences]

- 1848: Eighth report from the Select Committee on Sugar and Coffee Planting; together with an appendix.; House of Commons Parliamentary Papers
- 1848: First report from the Select Committee on Sugar and Coffee Planting; together with the minutes of evidence, and appendix.; House of Commons Parliamentary Papers
- 1841: Hill coolies. Return to an order of the Honourable the House of Commons, dated 29 January 1841;--for, copies of a letter from the Secretary to the Government of India to the committee appointed to inquire respecting the exportation of hill coolies, dated the 1st day of August 1838;--of the report made by that committee; with the minutes of evidence and appendix;--of any minute recorded on that report by any member of the committee;--of the letters from the government of India to the Court of Directors of the East India Company, dated the 16th and 19th days of October 1840, on the same subject.; House of Commons Parliamentary Papers
- Supposing no protection were granted to **Sugar** grown in India, admitted in England, and that from any causes the production of **Sugar** in the other British Colonies, such as **Mauritius**, the West Indies, and British Guiana, were materially lessened, or ceased altogether, do you consider it at all certain that British India could produce **Sugar** for export, cheaper than the other **Sugar**-growing countries of the world, such as Siam, Cochin China, Brazil, Java and Manila)--- (should say, certainly not.
 - What was you told at the police? — Iurdo, a duffadar, shipped me; I was at the police before embarkation; I was asked at the police why I came here; I said I was brought by the duffadar for employment in Calcutta; I was asked if I would take Sivekye for it? **Mauritius**; I said I did not know where the **Mauritius** was, and what kind of place it was; the gentlemen of the police said it was two months' voyage on board ship; I agreed to go; I was told I was to be employed for five years in **Sugar** cultivation, at a pay of five rupees per month, get food and clothes, rice 14 chittacks, dholl three chittacks, ghee oh chittack, and nothing else, and every six months one blanket and one dhoty; I received 20 rupees, four months' advance, of which 12 rupees was taken from me by the duffadars, and I carried eight rupees with me.
 - and was employed in the **Sugar** cultivation; began work on the firing of the morning gun, and left it once at eight o'clock to breakfast; we began work one hour afterwards, and left it at three o'clock; we dressed and took our meal, and had no work at night; my master is a good man, he never beat me, but his blacks did so; we used to complain to the master, but he did not understand us; from the effect of the beating I fell sick, and was in the hospital for four months; we received rice and salt daily, but no fish nor any thing else; from the hospital I was sent to the chief police, where I was told that as I was disabled by the fall of a tree on my wrist, I had better return to my own country, and I was put on board ship, the name of which I do not know; I received two months' pay, eight rupees, at the **Mauritius**, which was taken away from me by the blacks; this was done at the police before the eye of the magistrate, who took no notice of it.
 - To proceed to **Mauritius** as general agricultural labourers for a term of five years, calculated from date of contract, and work on the estate or estates of or any of which he may have the agency, in the cultivation of **Sugar**-cane, manufacture of **Sugar**, digging of land, mending roads, clearing forest or jungle, care of cattle, and all such lawful work as shall be required of them.
 - I sent Mr. Hanley, the English superintendent, whom I had hitherto employed to learn the practice of **Sugar** planters at **Mauritius**, and as his return.
- 1848: Second report from the Select Committee on Sugar and Coffee Planting; together with the minutes of evidence, and appendix.; House of Commons Parliamentary Papers
- 1840: Report from the Select Committee on East India Produce; together with the minutes of evidence, an appendix, and index.; House of Commons Parliamentary Papers
- 1848: Mauritius. Return to an address of the Honourable the House of Commons, dated 26 November 1847;--for, "copies or extracts of correspondence between the Secretary of State and the Governor of Mauritius, with respect to the general condition of the colony, and the measures adopted to meet the difficulties to which it may be exposed by the recent commercial failures affecting it."--(In continuation of the papers ordered by the House of Commons to be printed, 8 August 1845, no. 641; 26 August 1846, no. 691, parts I. II. and III.; and 27 April 1847, no. 325).; House of Commons Parliamentary Papers
- 1842: British possessions abroad. A bill to amend the laws for the regulation of the trade of the British possessions abroad.; House of Commons Parliamentary Papers
- 1845: Trade of British possessions abroad. A bill to regulate the trade of British possessions abroad.; House of Commons Parliamentary Papers
- 1848: Third report from the Select Committee on Sugar and Coffee Planting; together with the minutes of evidence, and appendix.; House of Commons Parliamentary Papers
- 1843: Journal and proceedings of the House of Assembly, 1843; Early Canadiana Online
- 1847: Customs. Abstract of the net annual produce of the duties of customs on all articles imported into the United Kingdom, in the two years 1845 and 1846, &c.; House of Commons Parliamentary Papers

Fig. 14 Full document list, including sentence snippets corresponding to a selected commodity ('Sugar') and location ('Mauritius')

Furthermore, the Location Cloud shows individual location terms as they occur throughout the documents—terms such as 'Britain' and 'United Kingdom' with identical meaning, as well as 'Montreal', 'Quebec', and 'Canada' are treated in the same way. Future features could allow different

types of aggregations of location terms, e.g. by country or by equivalencies on demand.

The document list view currently appears in a separate browser tab which the historians experienced as slightly disruptive to the flow of explorations. The document list could be integrated into





Fig. 16 Cinchona—India



Fig. 17 Tallow—Russia

the main visualization view to enable fluid switching between the visualizations and corresponding documents (similar to the Interlinked Views).

Last but not least, in terms of quantitative analysis, the Location Cloud is just scratching the surface of possibilities. Future iterations will investigate further features that enable a fine-grained analysis of changes of location mentions over time. Overall, the Location Cloud is a highly promising approach to visualizing location/commodity relations over time. In fact, our historians requested an additional visualization that focuses on changes of commodity mentions across time, based on selected locations (an inverted version of the Location Cloud). This can show the changing prominence of different commodities for locations such as Canada or Singapore.

5 Discussion

As a case study, Trading Consequences provides rich examples and insights into how the combination of TM and information visualization facilitates historical research. In the following, we discuss the role of information visualization in such research and provide considerations to inform the design of visualization-based exploratory interfaces for humanities research.

5.1 Information visualizations as a starting point for historical research

At a glance, information visualizations can help historians to draw conclusions about the relationships between commodities and places, and how these change over time. Yet, to use these tools effectively, historians need to resist the urge to treat the visualizations as an end point. The entire social, economic, and environmental history of a specific commodity is not represented in the visualizations—what Trading Consequences provides is a starting point. The reasons for this are related to the nature of the data we have processed, which researchers should bear in mind while using our visualization tools.

First, Trading Consequences relies on a finite number of sources: the database comprises

location–commodity mentions contained in the written documents of a selected number of digitized primary source collections which, while large, are not comprehensive. As a result, depending on the commodity, some of the story is missing.

Second, our corpus includes documents written in English by and for people living in Britain and British North America or Canada. British sources tend to privilege the perspective and context of places where commodities were marketed, processed, and consumed, and under-represents places where commodities originated.²¹ Furthermore, the corpus draws mainly, although not exclusively, from sources produced during the second half of the 19th century. The data available for decades between 1850 and 1900 is more plentiful and therefore more reliable than the data available for before 1850 and after 1900.

Finally, the text-mining output contains mistakes, resulting from the error-prone OCR process, which can only be partially addressed by fine-tuning the text-mining technology for processing historical texts. This noise does not significantly affect the larger trends presented by visualizations, but results become less reliable when drilling down into the data. For instance, ‘Madras’ is sometimes incorrectly identified as a type of cloth, rather than a location; ‘Italy’ is sometimes identified as a town in Texas; ‘Baltic’ as a town in Connecticut; or ‘British West India Islands’ as ‘Western India’. That said, the results produced when searching common historical document databases also come with a non-trivial level of noise. However, false negatives are rarely acknowledged by these interfaces, because the textual lists of search results do not reveal the missed pages where poor OCR caused the search to fail (Milligan, 2013). As our explorations show, information visualizations highlight this noise—historians in our workshop recognized irregularities within just minutes of exploration. While this can reduce trust in the underlying data, visualizations can also add transparency to the data, leading to a critical and responsible use of digital tools and interfaces as part of historical research. Furthermore, we found that producing visualizations of the mined data early on in the project facilitated improvement of the TM.

Text-based information visualization is always an abstraction of rich sources. As such, it cannot replace an in-depth analysis of original texts. Using the Trading Consequences visualization tools effectively means keeping these limitations in mind. Their benefit to historical enquiry is derived from the information contained within the digitized corpus and limited by the capacity of the technologies applied, not that it tells us the entire history of a specific commodity. Being clear eyed about these limitations does not take away from the powerful new tools created as part of Trading Consequences—historians simply need to remain alert to what the visualizations actually show and use more traditional research methods in conjunction to help ground any observations made with the tools.

5.2 Design considerations

The insights we gained from designing the three visual exploration tools and our discussions with more than twenty historians, from our team and beyond, have led to a number of considerations that may facilitate the design of visualization tools to support research in history and the humanities at large.

- **Offering Different Visual Perspectives on the Data.** Current search interfaces for document collections are typically dominated by text-based lists. In contrast, information visualizations can provide rich visual perspectives on the different aspects of the collection, without even requiring a textual query. In many cases, these visual perspectives can be enriched with interactive features to support the filtering of data.
- **Making the Context of Data Visible.** Large document collections include a range of topics and perspectives. It is therefore important to provide contextual details as part of the exploration process. Such a context can be provided by juxtaposing visualizations with snippets from the actual data sources (see, e.g. Figs 9 and 13). As our explorations show, disconnecting metadata-based visualizations from the actual documents can disrupt the fluidity of the exploration process and, in the worst case, lead to misinterpretations.

- **Quantitative Details.** In order to make visual exploration tools a valid addition to historical research methods, it is important to provide detailed and exact quantitative analytical views.

For instance, different types of normalizations can help to explore trends from various angles. Since researchers have different interests and requirements, it can be useful to support the download of corresponding data subsets, resulting from an exploration of the visualizations, to enable the researcher to analyse these data further using their preferred tools.

6 Conclusions

As part of Trading Consequences, we have explored how to combine different computational approaches—TM and information visualization—to facilitate research in environmental history. The three visualization-based exploration tools that we have presented enable the navigation of large-scale historical document collections that would otherwise be impossible. Furthermore, they highlight trends and relations within the data, which can facilitate discoveries and inform new research questions in environmental history.

The feedback from historians which we have gathered throughout the project, highlights the great potential of our approach of combining TM and information visualization to facilitate research in the humanities, but also point to limitations. Results produced by such tools therefore have to be considered as a starting point to humanities' research—they are a valuable addition, not a replacement of current research methods.

Acknowledgements

We would like to thank all CHESS'13 workshop participants and NiCHE members, in particular Josh MacFadyen, who provided invaluable feedback on our early prototypes and throughout this project. We would like to thank our partners who have provided the corpus of texts at the heart of Trading Consequences: Adam Matthew, Gale Cengage

Learning, JSTOR, Royal Botanic Gardens Kew, and ProQuest. We also thank Canadiana.org, Nipissing University, the Open University, University of British Columbia, and the University of Guelph for supporting this project.

Funding

This work was supported by the Digging Into Data programme.

References

- Alex, B. and Burns, J. (2014). *Estimating and Rating the Quality of Optically Character Recognised Text*, *Proceedings of DATECH 2014*, Madrid, Spain.
- Alex, B., Byrne, K., Grover, C., and Tobin, R. (2015). Adapting the Edinburgh Geoparser to historical georeferencing. *International Journal for Humanities and Arts Computing*, 9(1): 15–35.
- Card, S.K., Mackinlay, J.D., and Shneiderman, B. (eds) (1999). *Readings in Information Visualization: Using Vision to Think*. Burlington: Morgan Kaufmann Publishers, Chapter 1: Information Visualization, pp. 1–34.
- Collins, C., Viégas, F.B., and Wattenberg, M. (2009), *Parallel Tag Clouds to Explore and Analyze Facted Text Corpora*, *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, Atlantic City, NJ, pp. 91–8.
- Cronon, W. (1992), *Nature's Metropolis: Chicago and the Great West*. New York: W.W. Norton.
- Dörk, M., Carpendale, S., Collins, C., and Williamson, C. (2008), VisGets: coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6): 1205–12.
- Grover, G., Tobin, R., Byrne, K., *et al.* (2010), Use of the Edinburgh Geoparser for georeferencing digitised historical collections. *Philosophical Transactions of the Royal Society*, 368(1925), The Royal Society, 3875–89.
- Hitchcock, T. (2013). Big data for dead people: digital readings and the conundrums of positivism. *Historyonics*. <http://historyonics.blogspot.co.uk/2013/12/big-data-for-dead-people-digital.html> (accessed October 2013).
- Klein, E., Alex, B., and Clifford, J. (2014a). *Bootstrapping a Historical Commodities Lexicon with SKOS and DBpedia*, *Proceedings of the LaTeCH 2014 Workshop at EAEL*.
- Klein, E., Alex, B., Grover, C., *et al.* (2014b). *Digging into Data White Paper: Trading Consequences*. Research report, School of Informatics, Edinburgh, UK: University of Edinburgh; Multidisciplinary Studies Department, Glendon College, York University, Toronto, Canada; SACHI, School of Computer Science, University of St Andrews, UK; EDINA, University of Edinburgh, UK.
- Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4): 41–6.
- Milligan, I. (2013). Illusionary order: online databases, optical character recognition, and canadian history, 1997–2010. *The Canadian Historical Review*, 94: 540–9.
- Viégas, F. B., Golder, S., and Donath, J. (2006), *Visualizing Email Content: Portraying Relationships from Conversational Histories*, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY.

Notes

- Trading Consequences: <http://tradingconsequences.blogs.edina.ac.uk/>
- Digging into Data: <http://diggingintodata.org/>
- <http://niche-canada.org/>
- Data collections: <http://tradingconsequences.blogs.edina.ac.uk/about/the-corpus/>
- <http://www.ltg.ed.ac.uk/>
- LT-XML2: <http://www.ltg.ed.ac.uk/software/ltxml2>; LT-TTT2: <http://www.ltg.ed.ac.uk/software/ltttt2>
- TextCat: <http://odur.let.rug.nl/~vannoord/TextCat/>
- GeoNames: <http://www.geonames.org/>
- <http://tradingconsequences.blogs.edina.ac.uk/about/the-corpus/text-mining-document-example/>
- DBpedia: <http://wiki.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.
- <http://www.postgresql.org/>
- EDINA, the Jisc-designated centre for digital expertise & online service delivery: <http://edina.ac.uk/>
- <https://www.arcgis.com/features/index.html>
- <http://tcqdev.edina.ac.uk/search/commodity/>

- 15 All information visualizations as well as the database interface we describe in this paper are accessible on the Trading Consequences webpage: <http://tradingconsequences.blogs.edina.ac.uk/>
- 16 <http://d3js.org/>
- 17 <http://jquery.com/>
- 18 <https://www.google.ca/chrome/>
- 19 <http://tcqdev.edina.ac.uk/vis/tradConVis/>
- 20 CHESS 2013: <http://70.32.75.219/2013/04/12/cfp-canadian-history-and-environment-summer-school-2013-vancouver-island/>
- 21 Tim Hitchcock argues this is a major problem for the digital humanities more generally, as digital collections privilege an elite European and North American perspective (Hitchcock, 2013).