Geoparsing History: Locating Commodities in Ten Million Pages of Nineteenth-Century Sources

Abstract

In the *Trading Consequences* project, historians, computational linguists and computer scientists collaborated to develop a text mining system that extracts information from a vast amount of digitized published English-language sources from the "long nineteenth century" (1789 to 1914).

The project focused on identifying relationships within the texts between commodities, geographical locations and dates. We explain the methodology, uses and the limitations of applying digital humanities techniques to historical research and argue that interdisciplinary approaches are critically important in addressing the technical challenges that arise. We believe that collaborative teamwork of the kind described here has considerable potential to produce further advances in the large-scale analysis of historical documents.
Key Words: Digital history, Text Mining, Geoparsing, Commodities, British world

The late nineteenth century experienced vastly expanding international trade, creating an early form of economic globalization. Between 1850 and 1913, the scale of trade increased tenfold (Darwin, 114). Much of this commerce involved natural commodities, often shipped from different parts of the globe to European and North American factories and markets. Economic historian Edward Barbier refers to the latter part of the period as "the 'Golden Age' of Resource-Based Development" (Barbier, 2). Historians have grappled with both macro and the micro features of this process: some scholars focus on the broad economic, political and social impacts of these expanding connections while others explore particular commodity trades that linked distant parts of the world. Funded by the

Digging into Data program, the *Trading Consequences* project aims to assist both types of focus, while at the same time permitting a third emphasis, an examination of the discourse surrounding commodity trade. While historians tend to study, and rightly so, the successful and well documented trades, to understand the importance of the trading mindset, it can be useful to focus on other proposed, rejected, and failed trading possibilities as well.

The *Trading Consequences* project facilitates exploration of commodity trades in the nineteenth-century British economic world through text mining and geo-parsing millions of pages of documents and the development of visual cartographic and textual interfaces that enable research with the results. A collaboration between historians, computational linguists and computer scientists, this project uses previously scanned printed historical sources and identifies geographical relationships for hundreds of commodities. This article explains the techniques and assumptions of the project, some of the challenges of using printed historical sources for text mining, and the potential of large-scale distant reading made possible through this approach. The discussion will proceed in two sections: In the first place, we describe the project by presenting the methodologies and the assumptions we adopted along with some of the difficulties of dealing with the vast quantity of printed historical material. Secondly, we present some early historical research findings made possible by the project.

A. METHODOLOGIES AND CHALLENGES

1. CORPUS

Assembling a corpus is a key step in a humanities text mining project, just as it is in most types of historical research. At an early stage we identified the British House of Commons Parliamentary Papers and Early Canadiana Online as the two principal sources for the project. The House of Commons papers include a wealth of government reports on nineteenth-century Britain and the wider British world. Early Canadiana Online provides a wider array of material including government documents alongside periodicals and literary texts. We later integrated the Confidential Prints collection, which includes reports, documents and correspondence from British diplomats based in Africa, Latin America, the Middle East and North America. Finally, we learned about a small but highly significant collection at Kew Gardens. Their Directors' Correspondence project digitized thousands of letters sent to the Royal Botanical Gardens during the nineteenth century. The original letters were generally handwritten and therefore could not be processed with Optical Character Recognition (OCR) software (see below); instead, their project team composed short electronic descriptions of each one. We mined the descriptions, the titles and any relevant metadata of the letters, which proved very effective. We selected these four collections for a number of reasons. We looked for very large collections that would provide the quantity of data needed to test the effectiveness of our text mining methods. With our

historical interests in the environmental and economic history of commodities, we knew government documents would be useful sources.[1]

Table 1 shows that these collections make up a total of 227,928 documents. Together, the processed documents contain a total of 6,955,547,159 word tokens. All documents within each collection were used in this project, with one exception: in the Early Canadiana Online data, we filtered out three sub-collections which we did not regard as containing sufficiently relevant information on commodity trading (including English Canadian Literature, Jesuit Relations and Aboriginal Studies) as well as all non-English-language documents. The language filtering still left some non-English text in the corpus in the case of bilingual Canadian government documents, but it removed a large proportion of documents written completely in a different language. Processing such documents would have required multilingual or language independent approaches to the text mining tools employed in Trading Consequences which due to the short project timescale was not feasible but which is something that can be addressed in future work.

Insert Table 1 HERE

2. TEXT MINING

Using this large corpus of historical sources, text mining in our project had the goal of identifying relationships between commodities, places and times contained within the texts. Essentially, this process involves teaching the

computer to identify such logical relationships in the same way a human reader would recognize them. The advantage of text mining is of course the speed and consistency with which a computer can process vast amounts of text.

An overview of the *Trading Consequences* text mining system is shown in Figure 1. It comprises a number of generic linguistic "pre-processing" steps, followed by the identification of commodities, locations, dates, quantities and the extraction of relationships between them. Because components of the system are arranged so that the output from one provides the input to the next, this architecture is often described as a "pipeline."

Insert Figure 1 HERE.

The corpus data arrived in a variety of formats from the providers, and we first converted each document to a consistent XML (eXtensible Markup Language) format. XML is an expressive and flexible markup language for natural language processing tasks. It lends itself as a format to a pipeline architecture since the linguistic information computed by successive components in the pipeline can be added incrementally as layers of XML annotation (Mikheev et al. 1999).

Once the corpus was converted into uniform XML, the process identified different types of structural, syntactic and linguistic information present in the corpus — these steps are referred to as "linguistic pre-processing" in Figure 1.

They included tokenization, sentence-boundary detection, part-of-speech tagging, lemmatization and chunking. Tokenization involves establishing the boundaries of individual words, while lemmatization identifies the base form of a specific word (e.g. "oat" is the lemma of the plural noun "oats" and of the adjective "oaty"). Each processing step provides information that enables subsequent steps to perform better. For example, part-of-speech tagging helps to inform the stage where we identify commodity terms. By only treating nouns as candidates, we can avoid false positives when a potential commodity term is used as a verb, such as "*grease* the machine" or "*salt* the fish." Sentence-boundary information was important as it was used to restrict the recognition of relationships between locations and commodities.

After pre-processing, the system tagged various types of terms, including mentions of commodities, locations, organizations, persons, diseases and disasters, temporal information, amounts and units. This kind of annotation is called Named Entity Recognition (NER). In the case of the commodities, this process involved matching nouns or noun phrases in the corpus to the lexicon of possible commodity terms created for the project and tagging them as positive matches given a series of context features. This was not a simple keyword search. We preferred longest string matching, so "coconut oil" was identified as a single commodity instead of two: "coconut" and "oil." The NER step created an

annotation for each of the place names, commodity terms, dates and other entities identified in each document and added them into the XML markup.

The *Trading Consequences* database includes tables with all the locations and commodities found in the corpus and the text snippets they occurred in, along with tables identifying the relations between commodities and locations. Relations in the database were also assigned a date, which was the date of publication of the document. It is therefore possible to query the database for all locations mentioned in the same sentence as "coconut" between 1840 and 1880 or all the commodities related to "Ceylon" in 1875. In addition to the interactive visualizations, we can export this information into Geographic Information Systems software to map the text-mined results and visually explore the results.

3. COMMODITY LEXICON

To program the computer to extract pertinent information from the texts, we supplied purpose-built lexicons for the NER process to utilize. Of key and specific importance to our project was the list of commodities to be recognized in the texts. Although historians have intensively studied resources like sugar, cotton, timber, coffee and so on, we hypothesized that the "long tail" of less studied commodities was relevant to an understanding of nineteenth-century global trade. However, we lacked an authoritative, comprehensive list of such substances. Moreover, even historians with expertise in nineteenth-century trade might be unlikely to know whether, say, the plant species *Caesalpinia coriaria* or

divi-divi (the seed pods of which are high in the tannins used in leather tanning) had economic significance in the nineteenth century. This meant that it would be impractical to ask experts to annotate a body of training material that could be used for supervised machine learning (a commonly used approach to text mining). We decided instead to address this issue by "bootstrapping" a commodity lexicon: starting from a small initial list of commodities found in archival customs records, we constructed a much larger lexicon by mining Wikipedia for semantically related terms and by adding predictable variants such as plural and hyphenated forms (Klein, Alex, and Clifford 2014). This lexicon of commodity names serves as the basis of the project.

4. CHALLENGES IN USING HISTORICAL DOCUMENTS

Trading Consequences was predicated on the availability of a large volume of electronic text created by scanning paper or microfilm. However, much of the digitization was carried out over a decade ago, using software that was significantly inferior to the current state-of-the-art. The low quality of the resulting electronic text is an enormous challenge to text processing techniques, aggravated further when the corpus consists of historical texts.

There are two basic steps in the digitization process.[2] First, a digital image is produced using a scanner or digital camera, and the metadata, including title, author, publisher, and date of publication, are normally entered by hand. This process results in a digital image of the original source that can be read onscreen

or printed. However this image is not machine-readable as a text, but is just an array of pixels. Since there has been no identification of characters, there is also no way of identifying or searching for words. Consequently, the second step is to process every scanned image with OCR software. This process identifies the characters, punctuation and spaces in each image and adds a machine-readable text layer for each page in a digitized document. In some case, the OCR output format contains image coordinates for each piece of text recognized. More advanced systems automatically identify all of the images and tables in the source material. Unfortunately, the quality of OCR is inconsistent, which can negatively affect the performance of text mining processes (Lopresti 2009, Alex and Burns 2014).

OCR software normally works well on a clean modern document or book. Scanned nineteenth-century sources, unfortunately, present numerous challenges. First of all, a dirty or damaged document confuses the software and often results in non-textual marks being recognized as characters. Second, if the printing is slightly out of kilter, if the margins between columns are too small or if the font is irregular, the software will make numerous errors. The problems compound when the digital copies were scanned from microfilm instead of the original documents. The OCR software used to process most of our corpus also failed to identify the structure of tables, making large sections of government documents less than optimal for text mining, an issue we examine in further detail below.[3]

We were able to address some common problems with historical texts. End-of-line hyphenations splitting words in two and the long “s” (ſ) which OCR software frequently mistakes for “f” both created errors that we could correct automatically (Alex et al. 2012). To fix end-of-line hyphenation, the last token on each line ending with a hyphen was considered a candidate for joining with the first token of the next line. Tokens were joined if, after removing the hyphen and concatenating them, the result was either a word that appeared in the dictionary or was a word that appeared elsewhere in the document. Thus, “cin- chona” would be concatenated to “cinchona”. To switch false “f” characters to “s”, we first created a lexicon from a corpus of correct text. For each word in the corpus, we generated all the possible misspellings caused by the “ſ” to “f” confusion. Words in the OCR output were corrected if there was a match in the lexicon and if their corrected form occurred more frequently in the corpus of correct text. Thus, “falt” would be replaced with “salt.”

The government documents of the period present additional challenges for text mining. The corpus includes large numbers of tables, and indeed the statistical information in these tables is often of great interest to economic historians. Many tables are merged with surrounding text, while others are found in documents with little textual context surrounding them. For example, each edition of the *Annual statement of the trade and navigation of the United Kingdom with foreign countries and British possessions* contains hundreds of

pages of tables. The sentence-boundary detection function in the linguistic preprocessing step of the system is not designed to process the OCR output which does not contain table markup. This can lead to artificial sentences ranging from very short to very long collections of words, numbers, punctuation and noise. The text snippet below is an example of how the system concatenated all information in one table into a single "sentence" containing numerous place names, commodities, a lot of punctuation and many OCR errors. This 65 word excerpt is a sample from a very long 210 word "sentence" found in our results:

> *GLASS— continued : „ Manufactures, Unenumerated " (I" L) D«nm&rlc (lncludltiR I'nrOo IslnntU) Netherlands Franco Spain Italy Turkey Egypt Japan (Including Konnoaa) .... United States ot America .... Braril Argentine Ilopubllo Other Forolgn Countries Total to Forolgn Countries Capo ot Good Hopo British India Auttrolln Kow Zealand Canada Other British l»ogao»«Ion« ToUlto Urttlah VomomIoiib •; •; TOTAIi •; •; •; •; GLUE, SIZE, and GELATINE…*[4]

Checking with the original image we find a standard table where it is clear that "glass," "glue" and "gelatine" are not actually related to all the locations found in the table, but the text mining system extracts numerous relationships from this example which it treats as one sentence.[5] While most human readers find it easy to interpret how the information in table cells relates to the headers, automatic identification of the structure of tables and interpretation of their content are not

trivial tasks due to the number of possible formats and orientations. Therefore, we needed to examine how the information mined from the tables affects the results overall. Using the text mining results, we can identify table-heavy documents with unusually high numbers of place names and commodities. It is therefore relatively easy to distinguish between pages with tables containing commodity trade information and those which do not. To experiment, we identified all the documents with titles that include the phrases "Trade and navigation" and the "Annual statement of the trade of the United Kingdom with foreign countries and British possessions." We then used a database query to extract all the data for "copper" twice: the first time including the table-heavy documents (Figure 2) and the second time excluding results from these documents (Figure 3). The maps below show two overlapping layers with the black layer including all the documents and the light grey layer using the filter. It is possible to see that the results from these tables associate a large number of country-level geographical locations with the word "copper". In most cases these are correct. For example, "Egypt" and "Chile" are prominent in the results that exclude large numbers of tables and are even more prominent with the tables included. Chile was a major source of copper, and Egypt was a market for British copper exports. However, it is important for us to recognize the influence a small number of table heavy documents have on the database of extracted relationships.

Insert Figure 2 HERE.

Insert Figure 3 HERE.

The second map for "copper" (Figure 3), focused on Europe, confirms that the tables mostly amplify the results extracted from the wider corpus.[6] We were encouraged by the fact that there are no prominent examples of locations where the results from the tables are not mirrored in the results from the rest of the corpus, although "Spain" does seem to be more amplified than other locations. "Saumur" in France is the only notable round black dot where there are no examples in the filtered results. This is a case where the commodity "Saumur Wine" has been falsely identified as a location and being related to the other commodities found in the tables.

> ...**Copper** - Rice „ Meal •; Rivets, Iron and Steel •; Rockingham Ware Rock Phosphate - Rods, Iron, Wrought •; „ Steel „ Wire, Iron and St«««l •; Ropo, Hempen, i&c. Rosin Rum Saccharin .... Sacks Saddlery .... Sailcloth .... Sailing Ships •; •; Salmon, Canned •; Salt Saltcako .... Saltpetre .... Sarmnes, Canned •; Satin .BroadsLulft Sauces •; **Saumur** (Wine)...[7]

A more frequent problem is when a commodity is related to the wrong place. In the following example "copper" is linked to "Russia."

> Total of COPPER and Brass - EXPORTATIONS in the MONTH of JANUARY Lead: Pig-, Rolled, Sheet, Piping, Tubing and Lead Shot " -

To Russia - - - Tons France United States - China and Hong Kong British India - Australia - - - Other Countries - Total - - - 186 9."[8]

However, when we check the scanned document, it is clear that the phrase "Total of Copper and Brass" relates to locations found on the preceding page, and "Russia" should only be linked to "lead."[9] As the data does not differentiate tables from running text, such false positive errors are likely to appear frequently within the textual sections corresponding to the information in tables. In the future, text mining projects will benefit from algorithms which can identify tables within OCR'ed documents and weigh their structure appropriately in order to extract the information within them more accurately.

5. GEOPARSING AND GEO-TAGGING ISSUES

A key goal of the project was to apply the geoparsing expertise at the University of Edinburgh to historical texts. Geoparsing is a subfield of text mining that involves identifying geographic information in texts by first using NER. In a second process called toponym resolution, the place-name mentions found in the text are then disambiguated by linking them to their corresponding gazetteer entries, thereby assigning latitude/longitude coordinates and, depending on the gazetteer, other relevant information stored for the entry such as a country code or location type (e.g., town, province, state, country). Geoparsing is particularly attractive for place-based historical research, as it expands the sources available for historical Geographic Information Systems (GIS) from maps and statistical

tables to the text found in most digitized historical documents (Gregory and Hardie 2011). Geoparsing makes it possible to automatically recognize the places discussed in text and explore their relationship with other significant terms. This may include identifying the places mentioned in the context of a single word, such as "cholera" to explore the spatial history of disease in a large corpus of texts, or it may involve extracting a wide range of relationships between places and a list of other words pertinent to a subfield of history (Gregory and Geddes 2014; Murrieta-Flores et al. 2014). In *Trading Consequences* we made use of the Edinburgh Geoparser, which we adapted to historical text (Alex et al. 2015; Grover et al. 2010).[10] In this case, the geoparser's parameters were set so as to deal with texts written about commodities and locations anywhere in the worlds.

The geographic component of the NER process relied on an existing gazetteer of place names. We required a location gazetteer containing an entry for each location along with latitude/longitude coordinates and extra information about the location, including location type and population size if applicable. For that reason and because it has world coverage, we chose the gazetteer GeoNames for *Trading Consequences*.[11] However, even this excellent resource did not solve all problems. We illustrate the geoparsing process and some challenges it created with the following example sentence found in a 1923 edition of the *Coconut planter's manual*:

> "Originally **coconut** cultivation was confined to the coast and to sea level under the impression that proximity to the sea was a sine qua non; but this theory having been exploded, we **now** find it carried on round **Kandy**, **Peradeniya**, **Gampola**...." (Ferguson and Drieberg 1923).

Our text mining system tries to find grounding identifiers for all named entity mentions that it has tagged. Aside from locations, it also normalizes dates to a structured and machine-readable year-month-day format.[12] It also links commodities to DBpedia URLs, which in turn can be linked back to Wikipedia pages for the commodity.[13] The commodity, location and date entities found in the text are therefore enriched with additional information via normalization and grounding

1. commodity terms:
    a. coconut, concept: http://dbpedia.org/resource/Coconut
2. dates (year of publication or in text):
    a. 1923 (year: 1923), now (relative date)
3. location mentions:
    a. Kandy, latitude: 7.2622, longitude: 80.5841, country: Sri Lanka
    b. Peradeniya, lat: 7.1643, long: 80.5696, country: Sri Lanka
    c. Gambola, lat: 6.9895, long: 81.0557, country: Sri Lanka

The aim of the toponym resolution step is to select the correct latitude and longitude for a place name mention. The Edinburgh Geoparser uses contextual

information in the document to perform this task. It matches a toponym found in the text to all possible gazetteer entries with that name, as well as alternative names and abbreviations of the toponym. Where there are multiple candidates for one toponym, the geoparser ranks them to determine the most plausible interpretation in context. In addition to population size, we also assume that a textual document generally has a degree of geographic coherence to it so that, for example, the interpretation of "Paris" in a text will be influenced by whether the text also mentions "Texas" or "France." We model this by assigning a geo-coherence score to each candidate reflecting how close it is to all the other place name interpretations in the document. In the above example, the interpretation of Kandy in Sri Lanka was correctly ranked higher than the location in modern-day Uzbekistan as a result of population and geographic coherence with other Sri Lankan place names mentioned in the text. The information of the top-ranked candidate was then added to the markup of the location mention in the document (Grover et al. 2010).

In assigning geographical identifications, historical texts present specific challenges as illustrated with the following excerpt:

> 1 do not, however, entirely despair of the successful cultivation of American plants in the drier and cleared parts of north-west India, where irrigation may be employed to save a crop ; for it is only by the aid of irrigation that it succeeds in Egypt, and I found no diffi¬ culty in

> cultivating the American cotton at Saharunpore ; and Colonel Colvin successfully introduced it into many villages along the Delhi Canal.[14]

The geoparser correctly identified and resolved the country name "India" but did not identify the Indian location "Saharunpore" because it is not contained in the location gazetteers applied as part of the NER step. In this instance, "Saharunpore" is not an ORC error but is instead one of many examples of a place name where the transliteration into English changed during the course of the nineteenth century.[15] The modern variant Saharanpur is contained in the location gazetteers and therefore most likely would have been resolved correctly. Place names like "British North America," the "British West India Islands," the "Straits Settlements," "French West African Possessions" and the "Oil River Protectorates" are just a few examples of common nineteenth-century geographic terms that are not found in the modern digital gazetteer GeoNames used by our project. In some cases, such as British West India Islands, the geoparser incorrectly grounds the location to West India, while in other cases, such as the Straits Settlements, it cannot identify a location in the first place. We began developing a list of key locations that we could add as a supplemental gazetteer. However, nineteenth-century spelling fluctuated and the English conventions changed for numerous place names which means that this issue extends beyond a handful of locations. The only way to truly solve the problem is to develop a global historical gazetteer (Southall, Aucott, and Westwood 2014).

Other consistent errors resulted from text mining such a disparate collection of historical documents. Toponym resolution accuracy improves considerably when the region or country of most of the place names in the text is known. The geoparser can be set to give preference to locations within a specific locality if it is known that a text focusses on a specific area. For example, if most of the place names within a text are located within the United Kingdom or North America, the algorithm will have an easier task distinguishing between the "Surrey" outside of London and the "Surrey" outside of Vancouver. *Trading Consequences* could not exploit this feature because the aim was to better understand the globalization of commodity trading. We did improve our results by checking whether a candidate coincides with an entry in a Ports gazetteer that we constructed specifically for *Trading Consequences*, given that nineteenth-century trade was largely conducted by ship. Privileging countries also helped increase the accuracy of the geoparser at this global scale. We also discovered through experimentation that the optimal number of GeoNames candidates to consider for the geo-resolution of the *Trading Consequences* data is 15 (Alex et al. 2015).

Once the system had grounded the named entity mentions, it carried out relation extraction to identify relationships between commodities and places. In *Trading Consequences*, we specified that a relationship holds between entities only if they occurred in the same sentence. Clearly this is a vast oversimplification, but it was effective as an approximation. Consequently, this

step looked for sentences containing one or more place names and one or more commodities and put them into commodity-location relations.

We used the text mining results to build interactive visualizations to help historians perform open-ended exploratory research. To facilitate the web visualizations and to make the results easier to query, we extracted the core text mining results out of the XML files into a single PostgreSQL database with a PostGIS extension.[16]

6. PERFORMANCE EVALUATIONS

With the text mining complete and available to explore through a number of public web visualization research tools, historians began assessing the results (Lawson 2014). The interactive websites, which we discuss in a separate publication, significantly reduce the technical barriers to exploring the database and make it possible for historians to use our results for heuristic research (Hinrichs et al. 2016).[17]

The visualizations make our text-mined results available to all historians interested in commodity histories. At the same time, perhaps ironically, the visualizations serve, for some researchers, to magnify specific errors. Over the course of the collaboration, we identified a number of issues with the text mining and geo-parsing elements of the project. Some of these challenges were specific to historical sources and research, and the collaboration between team members from different disciplinary backgrounds proved very useful in identifying and

addressing them. In some cases, we were able to fix problems fairly easily. Others remain difficult to resolve, though we contend that specific problems do not outweigh the benefits of such broad-based research.

Colleagues who have used the on-line tools were quick to notice, for example, OCR errors which confuse “time” or “line” for “lime” and give the small fruit and calcium-rich mineral an undue prominence in the results. A bug in the geoparser located the vast majority of the instances of “Italy” in the corpus with the longitude and latitude of a small Texan town, instead of the country.[1] It is simple to reassign all of the cases of Italy grounded in Texas to the correct geography for Italy in Europe and to filter out obvious false positives caused by OCR errors.

False positives are fairly easy to recognize, and in an iterative process among team members, we have been able to address them through specific fixes. The bigger challenges are the false negatives, where the text mining tool failed to identify place names or commodities in the text. False negatives may stem from poor OCR quality, a historical place name missing from the GeoNames gazetteer or a commodity term not included in the lexicon. These missing results are not displayed in our visualizations and therefore are difficult to spot and correct.

To evaluate this problem, we used intrinsic evaluation to assess the quality of our text-mined results. Computational linguists and natural language processing researchers do this by testing their system output against a gold

standard of human-annotated documents. They use the gold standard to calculate the scores based on the number of false positives, false negatives and correct identifications and produce tables with statistical results that allow them to compare the performance of one text mining system against other similar projects or variations of their own system. They can therefore also test whether changes made to a system prototype improve or reduce performance. In the case of *Trading Consequences*, this method of evaluation enabled us, for example, to determine the number of commodities or locations in the gold standard that were correctly identified by the system (true positives), those which were missed (false negatives) and any erroneous commodities or locations that were extracted by the system (false positives). Using these counts, NER performance on the gold standard can then be computed in terms of precision and recall and balanced F-score.[18] Given a sufficiently large gold standard, the results can be used as estimates for the text mining performance on datasets which are similar to those in the gold standard.

Creating a gold standard for commodity recognition posed its own difficulties. The tasks we are trying to accomplish through text mining are not straightforward for humans to perform, and the knowledge and judgments of the two annotators created significantly different results working with the same material. We concluded this after measuring inter-annotator agreement (IAA) for a double-annotated sub part of the data in the same way as the system is evaluated

using balanced F-score. The IAA F-score for commodities when comparing the double-annotated sub-section of the gold standard was 0.72. The IAA F-score for annotating locations was 0.82, showing that the annotators found it easier to agree on these designations.

We are not aware of other gold standard annotation efforts which have determined IAA for commodity annotations. However, location-specific agreement figures usually tend to be higher on contemporary texts, particularly if the annotators underwent an extended training period (e.g. 0.97 for résumé data, see Alex et al. 2010). It is likely that the quality of the data affected our scores. In fact, we showed that in the entire gold standard data 14.9% of all locations contain OCR errors (Alex and Burns 2014). Some people are better at recognizing OCR errors and identifying actual locations than others. This explains why the IAA figures for locations are much lower compared to those obtained when annotating contemporary text. The same holds true for commodities but to a lesser extent as only 9.1% of commodities in gold data contain OCR errors.

The random sample also shaped the results as they happened to include pages with the Latin nomenclature for numerous plants. The list of commodities developed for this project did not include these terms, as they were not major commodities, and these were not the words we expected to be used in trade. One of the annotators made a different judgment call and tagged these words as commodities, decreasing the IAA score.

These results make it clear that even highly trained environmental historians are unlikely to agree on all parameters of what to consider a commodity. The IAA figures also put the performance of the text mining system into perspective. If humans find it difficult to agree on a task then it is likely to be more challenging for a computer as well since the latter is evaluated against human annotations. The final system F-scores for commodities and locations were 0.62 and 0.61 respectively. While these results are lower than we would except for text mining methods applied to contemporary sources, the pipeline still created useful results for historical research, particularly because we applied it to a very large corpus. It is also notable that most documents of particular interest to historians repeat locations and commodities with some frequency, increasing the chance that the pipeline correctly identified relevant documents, even when poor OCR prevented it from capturing all the relevant commodity-location relationships.

We only employed one annotator to carry out the annotation for the toponym resolution step but believed it would be an easier task to agree on as it simply involves placing a location pin on the correct spot on the map. When measuring the accuracy of the toponym resolution step of the geoparser we obtained a high score, with 85 percent of the samples mapped within five kilometers of the annotator's selection.

We improved the text mining tools as a result of experimentation with the gold standard carried out by the developers of the tools and constant feedback from the historians who identified obvious errors in the output. We believe that this combination of interdisciplinary collaboration is the key to creating better software and tailoring tools and technology to their users. The process of error analysis continued beyond the end of 2013 when the text mining development specific to *Trading Consequences* was completed, as the historians continued to work with the resulting database and visualizations. The Edinburgh text mining team who continue to develop and improve the geoparser value this feedback as it enabled them to improve their technology for future applications.

B: Research with the Tool

The *Trading Consequences* database provides a powerful way to research the growing commodity trade across the long nineteenth century. Text mining coupled with purpose-built and dynamic web research tools significantly increase the efficiency of historians as they can explore macro-level trends and identify specific pages in documents that address the relationships between commodities and places of interest. The text-mined results allow us to explore and research the history of commodities in new ways. The database and visualization tools allow historians to sort the date-place-commodity relationships to explore patterns spatially and chronologically. These visualizations serve as research tools to prompt historians to explore the relationships between geography and

commodities and develop new hypotheses, while at the same time identifying source material to help answer those questions. With the ability to "read" millions of pages, the speed with which historians may explore trends and source material is greatly increased. This section discusses the research that is possible at three different scales: major (well-known) commodity flows, minor commodity trades and the discourse on commerce.

1. MAJOR COMMODITIES

With the *Trading Consequences* interfaces, we can map hundreds of thousands of relationships related to thousands of commodities that spread throughout the British world. Trade statistics identify where British industry sourced raw materials, but they generally focus on national or regional levels, showing the quantities of goods imported from, for example, British West African possessions or United States Eastern Ports. In contrast, text mining can find mentions of goods in sentences discussing towns, cities, provinces, and countries. To take one example, the trans-Atlantic cattle trade is fairly well known in the literature. In 1891, over 100,000 live cattle, valued at almost $9 million, were shipped from Canada to Britain (Foran 2011: 266). Figure 4 shows how we can identify periods of time where the commodity appeared with greater frequency in the corpus. The increased frequency of "cattle" in the corpus in the 1860s, reflects a European rinderpest crisis featured in a number of reports on the "cattle plague" published in 1866 and 1868. The Commodity Search tool identifies key

documents discussing cattle in the 1890s focused on transatlantic trade.[19] The spike in 1891, for example, is the result of a report on the transatlantic cattle trade and references to cattle exports in the Canadian Sessional Papers (government department reports) at a time of increasing concern about disease in Canadian animals.[20]

Insert Figure 4 HERE.

The Location Cloud Visualization, purpose-built for *Trading Consequences*, provides a powerful option for identifying interesting patterns in the data. The visualization shows the changing normalized frequency in different locations in commodity-location relations decade by decade. Clicking on any of the place names in the location clouds organized by decade provides a link to a list of documents ranked by the relevance of the commodity and place name. This visualization is similar to the graph showing the changing frequency of cattle across the corpus but with the focus on the geoparsing results. Figure 5 shows the prominence of England and Ireland in references to "cattle" in the 1860s and 1870s (the 1860s are the seventh column), while Canada and the United States stand out at the end of the century (the 1890s are the tenth column). The visualization, however, also includes references to many locations, including the Channel Islands, Dublin, Hamburg, Leeds, and Rotterdam for the 1860s and Alberta, Chicago, Glasgow, Manitoba, Toronto and Winnipeg for the 1890s. By clicking on individual locations it is possible to quickly see that Antwerp was

discussed in sentences discussing the cattle plague and prohibition, while Winnipeg and Alberta appear in sentences that discuss shipping cattle to Liverpool and Great Britain. This visualization provides a powerful tool to explore the relationship between major commodities and different localities.

Insert Figure 5 HERE.

2. MINOR COMMODITIES

Trading Consequences interfaces complement traditional archival research by allowing us to explore specific commodity trade histories on a very detailed level. In the case of a relatively minor commodity used in leather tanning, the Cloud Visualization tool proved of critical help in identifying the documents necessary to answer a research question regarding a curious anomaly in nineteenth-century British trade statistics. The seedpod of a tropical tree called the divi-divi was a minor source of vegetable tannins used by leather tanneries in London. The annual trade statistics record only imports of divi-divi to the United Kingdom from 1855-1861 and 1865-1866. Since all other types of vegetable tannins appeared in later decades, an obvious research question emerged: did divi-divi cease to be traded after the 1860s, or were the trade tables simply adjusted to include divi-divi as part of a larger aggregated category of commodities?

A simple internet search on divi-divi revealed the tannin was derived from a tropical tree native to the Caribbean coast of South and Central America. Entering the commodity term ‘divi-divi’ into the Cloud Visualization tool turned

up an array of locations related to this commodity in the corpus. Isolating only returns from South America, however, generated a more useful number of locations. The larger the font of the locations listed in the Cloud Visualization, the more mentions of the commodity in relation to that location. Venezuela appeared prominently in the cloud. Clicking on the 'Venezuela' location under the first decade of the twentieth century brings up a list of snippets of text containing 'divi-divi' and 'Venezuela'. In 1907, the Diplomatic and Consular Reports of the British Foreign Office in Caracas recorded 3,630 tons of divi-divi exported from Venezuela.[21] These passages clarify that divi-divi was being traded during the first decade of the twentieth century. Thus, *Trading Consequences* identified a critical document necessary in answering the research question above. By following the annual series of consular reports from Venezuela back into the 1890s, it becomes apparent that divi-divi remained an export commodity during the late nineteenth century.

More qualitative information is also discernible from the snippets of text identified by *Trading Consequences* through the Cloud Visualization tool. The 1901 Diplomatic and Consular Reports, for example, mention that, "Business in divi-divi was prejudiced both by the poor harvest and by political troubles in Colombia and in Venezuela, which had the effect of raising prices and restricting the demand of the tanning industry for this article.[22] Although this does not

provide a complete picture of the story, this report indicates that divi-divi had become an unreliable resource for British industry.

With the help of *Trading Consequences*, identifying these documents was much less difficult and time-consuming, making it rather easy to answer the question: what happened to divi-divi when it disappeared from British trade statistics in the 1860s? The value of exploring the database instead of simply searching the collections themselves is twofold. First, it draws on texts compiled from numerous different collections simultaneously. And second, thanks to the built-in commodity-place relational algorithm, the project has the advantage over simply keyword-searching of the parent collections by narrowing down the number of relevant documents much more quickly. Thus, the database makes it possible to conduct meaningful research, not only for the most important commodities of the nineteenth century, but for many minor commodities as well.

3. DISCOURSE

As the divi-divi example demonstrates, trade statistics privilege successful supply chains and provide little information on the many failed attempts to establish agriculture and other extractive industries in unsuitable territories. For instance, British trade statistics record insignificant quantities of raw cotton imported from Sub-Saharan Africa during the second half of the nineteenth century. The United States, India and Egypt provided the vast majority of British raw cotton imports. The text-mined database, however, includes many sentences

featuring African place names and "cotton." Many of the sentences discuss trading cotton cloth in Africa, but a considerable number discuss the suitability of Africa for cotton cultivation. The following sentence, for example, found in a collection of papers related to the cultivation of cotton in Africa published in 1857, discusses the prospect for developing cotton exports from west Africa: "From the above observations, which are founded, not on theory or hypothesis, but on personal experience and extensive inquiry on the spot, it is fair to expect that a moderate and increasing supply of cotton may be obtained, from certain districts on the west coast of Africa."[23] Figure 6 shows the trade statistics, recorded by weight, for all commodities related to cotton in 1871, which include raw cotton, cotton seeds, and cotton rags. The second map, Figure 7, displays an intensity heat map and a sample of labelled points from the text-mined data from 1870 to 1872. The only recorded imports from Africa were a little more than four thousand tons of raw cotton from British possessions in South Africa. It is possible that a smaller amount of cotton might have been imported from West Africa and grouped with other small producers in the Other Countries category which cannot be visualized in this map.[24]

Insert Figure 6 HERE.

Insert Figure 7 HERE.

Pursuing this topic further, the text-mined results from 1870 to 1872 include ten locations in Sub-Saharan Africa, which we can examine individually.

"Volga," located in West Africa, comes from a sentence discussing cotton districts upriver from Lake Volga.[25] The "Madagascar" reference discusses the quality of "native industries," including cotton cloth, in the capital region.[26] Some instances reflect toponym resolution errors. However, one reference to "Zanzibar" points towards a very interesting discussion of the economic potential of Zanzibar and its hinterlands:

> Zanzibar would become a second Singapore or Kurrachee for that part of the world, more especially now the Suez Canal is opened... According to the accounts of the recent discoveries of Dr. Livingstone and others, we have in the interior of that part of Africa a country equal in resources to any part of India, and I believe more healthy as a rule; the sea-board and the rivers are unhealthy, but when you get some distance from the coast you rise to a lovely table land, and it is a country which, from what I saw, and from what I know from other men who have travelled there, is second in beauty to hardly any in the world, and it is also a most productive country. Iron abounds in all directions; in fact the Portuguese get all their iron from there. Coal is to be found; lead I have seen myself in large quantities, and cotton can be grown of any extent. I have seen very large quantities of cotton there.[27]

This extract comes from a document entitled "Report from the Select Committee on Slave Trade (East Coast of Africa)," which is the 72nd most relevant document identified when using the British House of Commons Parliamentary Papers keyword search interface for "Cotton and Africa." Since cotton only appears 25 times in this document, it would rank very low if we searched for "Cotton" alone. Exploring unexpected locations in text-mined results leads us to find and read documents that we might otherwise miss. As with the divi-divi example, exploring the text-mined results alongside the trade statistics

provides a more complete understanding of how the British government understood the geography of cotton in the early 1870s. They knew that the United States, Egypt and India supplied the vast majority of Britain's cotton, but remained attentive to other possible suppliers and "native industries" as well as the conflicts resulting from cheap Manchester cotton flooding local markets.

In a second example, we analyzed a report from a commission on potential trade between British North America and the Caribbean islands, British Guiana and Brazil published the year before Canadian Confederation in 1867, a text that has been cited by other historians, but is not well known for the light it sheds on the nineteenth-century colonial economy.[28] For the commissioners, great potential existed for increasing trade between British North America and the British and Dutch colonies and independent states in the Caribbean and Latin America. The northern locations could supply softwood timber, fish, and flour in exchange for the fruits, coffee, and hardwoods of the warmer climates. In other words, the commodities of one region could complement those from a different climate. The commissioners insisted on the relative ease of shipping: the distance from Halifax, Nova Scotia, to the Brazilian ports was actually shorter than from New York to Brazil.

If one focused merely on the statistics of trade from British North America to the southern locations at the time of the first commission, fish and timber predominated in quantity and value. Through text mining the document, we see

many minor commodities such as beans, bran and hay, often identified to specific locations in the British North American colonies. This attention to regional detail on the part of the authors shows that part of the purpose of the commission was to convince reluctant colonists in the northern colonies that Canadian Confederation offered positive economic benefits. The commissioners optimistically argued that the region offered "a market for the entire present surplus of our principal staples."[29] Text mining illustrates the range of items that the commissioners thought could enhance hemispheric commerce as well as commodities that actually constituted current exchange. Nonetheless, these close trading ties did not develop to any great extent; a similar commission released a report in 1910 that reached almost identical conclusions about the promise of trade between the regions.[30] Therefore, *Trading Consequences* points researchers to unrealized prospects, something a mere compilation of trade statistics would not. It captures the discourse of commodity trade which relfected many aspirations far beyond the extensive trading relations that came to exist.

Conclusion

Trading Consequences built a research tool to explore the environmental and economic consequences of the globalizing commodity trade during the nineteenth century. Our approach enabled us to work with a very large and diverse corpus of primary sources to map the changing geography of commodities

in our corpus on a global scale over the course of the long nineteenth century (circa 1789 to 1914). The project demonstrated the value of historians and digital humanists collaborating with text mining experts to adapt existing software to analyze historical sources on a larger scale than previously possible. The difficulties that arise from applying standard text mining techniques to historical sources make it unwise to rely on off-the-shelf text mining packages.[31]

As it is difficult to sustain web interfaces long-term, we make the database available for researchers to search it directly or to build their own interfaces to it, and potentially correct it via crowd sourcing.[32] It is possible to interact with the database directly using Structured Query Language (SQL), allowing for even more fine-grained examination or searches focused on particular documents within the corpus. For example, we used a SQL query to produce the maps comparing the results for "copper" including and excluding tables.[33] All these approaches combine a distant-reading analysis, allowing researchers to explore millions of pages of documents, with the ability to identify which documents and pages are the most relevant for our research questions. After identifying documents, we can download, read and scrutinize the sources using standard historical research methods. Text mining combined with effective visualizations allows for a useful hybrid between digital and more traditional research methods.

Solutions exist to many of the problems encountered while developing the *Trading Consequences* project. We are confident that continued cooperation

between humanists and computational linguists will further improve methods for adapting named-entity extraction and geoparsing historical documents. No historical methodologies are perfect. Historians miss important information with keyword searches and archival research in ways that are analogous to the difficulty we have encountered extracting everything correctly using text mining methods. In some respects, there is a tendency to hold computational tools to a higher standard than human efforts. But, as we recognize some of the errors in the *Trading Consequences* results, it is important to acknowledge the fallibilities of all historical methodologies. Even with the challenges of low-quality OCR, the difficulties of creating a list of terms used to describe commodities in the nineteenth century, and the problems involved in geoparsing historical place names, our project demonstrates that the ability to process millions of pages and identify geographic trends for commodities significantly improves historical research, particularly when combined with existing methods. By working with the database we learn to see past the errors and quickly test results to see if common false positives influence trends in the data. We can filter out some of the documents with large numbers of tables, fix the common geoparsing errors to improve our database, and produce very functional results. As we learn to research with text-mined results we increasingly find interesting documents that discuss commodities in their geographic context beyond the obvious "tea from China" trends. It is also worth noting that the common critique that text mining

tends to confirm well known patterns can be flipped around. When the visualizations display expected relationships between major commodities and locations, the results suggest that text mining works. When we see predictable trends in the visualizations, we can have greater confidence in the usefulness of other, less obvious relationships.

For historians who work on the environmental and economic history of nineteenth-century resource extraction, commodification, colonial trade, industrialization, and consumer cultures, our text mining results provide powerful tools for visualizing broad patterns contained within a very large corpus and identifying specific documents for further study. Taken together, projects such as *Trading Consequences* and *Spatial Humanities: Texts, GIS & Places* demonstrate that geospatial text mining has the capacity to dramatically enhance research methods.[34] The traditional approach to reviewing primary sources (one set of documents at a time, one page at a time) means research questions are shaped slowly as each new bit of information is explored and incorporated into changing thinking. The geospatial text mining tools enable historians to explore and incorporate information from many sets of documents at the same time, creating the ability to refine and check research questions more quickly. Moreover, the ability to move between geospatial scales means historians can pursue global history at the same time as they investigate microhistories. Questions asked and explored from one scale can be used to pursue the same issues at another scale.

Similarly, the ability to adjust the time period under investigation allows exploration of issues over longer or shorter periods.

Developing a feedback mechanism to report errors might allow a future project to harness a crowd of humanities researchers, students and the general public to identify toponym resolution errors and missed place names. Combining the methods developed for the *Trading Consequences* system with supervised machine learning for particularly challenging and ambiguous words may help solve some of the more common false positives. Historians can also make a major contribution by developing a global historical gazetteer, with the spelling variations for place names, historical population data and temporally specific geographic data (including latitude/longitude and boundary polygons). Historians can also work with computer scientists to develop better linked data structures. Instead of relying on the categories mined from Wikipedia, historians could help build structured lexicons of historical words and phrases for text mining a range of different topics. Economic and environmental historians could collaborate on improving our lexicon of commodity terms, while medical historians could develop a lexicon of disease and illness terms, and military historians could develop a lexicon of words related to warfare. Identifying and solving bugs in the text mining code is a considerably easier task than developing history-specific data structures that will facilitate more nuanced, sophisticated and accurate historical text mining in the future. Since algorithms developed to mine the

internet will not work on historical texts in a straightforward fashion, historians should not simply wait for software developers to refine their code. They should contribute their considerable knowledge of context and content to such team projects to improve the process of incorporating historical texts into the digital humanities.

Reference List

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh geoparser for Historical georeferencing. *International Journal of Humanities and Arts Computing*, *9*(1), 15–35. http://doi.org/10.3366/ijhac.2015.0136

Alex, B. and Burns, J. (2014). Estimating and Rating the Quality of Optically Character Recognised Text. In *Proceedings of DATeCH 2014*, Madrid, Spain.

Alex, B., Clifford, J., & Hinrichs, U. (2013, May 2). Bringing Kew's archive alive. Retrieved from http://www.kew.org/discover/blogs/bringing-kews-archive-alive

Alex, B., Grover, C., Klein, E., & Tobin, R. (2012). Digitised historical text: does it have to be mediOCRe? In *Proceedings of KONVENS 2012 (First International Workshop on Language Technology for Historical Text(s))*. Vienna, Austria. Retrieved from http://www.oegai.at/konvens2012/proceedings/59_alex12w/59_alex12w.pdf

Alex, B., Grover, C., Shen, R. and Kabadjov, M. (2010). Agile Corpus Annotation in Practice: An Overview of Manual and Automatic Annotation of CVs. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*, Uppsala, Sweden.

Barbier, Edward B. (2011) *Scarcity and Frontiers: How Economies have developed through natural resource exploitation*, Cambridge: Cambridge University Press.

Brown, T., Baldridge, J., Esteva, M., & Weijia Xu. (2012). The substantial words are in the ground and sea: computationally linking text and geography. *Texas*

Studies in Literature & Language, *54*(3), 324–339. http://doi.org/10.7560/TSLL54303

Coates, C., Klein, E., Quigley, A., Alex, B., Clifford, J., Hinrichs, U., … Grover, C. (2014). Trading Consequences | exploring the trading of commodities in the 19th Century. Retrieved January 26, 2014, from http://tradingconsequences.blogs.edina.ac.uk/

Cohen, D., Gibbs, F., Hitchcock, T., Rockwell, G., Sander, J., Shoemaker, R., … others. (2014). *Data mining with criminal intent white paper*. Retrieved from http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final.pdf

Cohen, D. J., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A. M., … Turkel, W. J. (2008). Interchange: the promise of digital history. *The Journal of American History*, *95*(2), 452–491. http://doi.org/10.2307/25095630

Darwin, John (2009), *The Empire Project: The Rise and Fall of the British World-System, 1830-1970*, Cambridge: Cambridge University Press.

Ferguson, John, and Drieberg, C. (1923) *Coconut Planter's Manual*. 5th ed. Columbo: Ceylon Observer Press https://archive.org/details/coconutplantersm00ferg.

Foran, Max. (2011) The Politics of Animal Health: The British Embargo on Canadian Cattle, 1892-1932 in Marchildon, Gregory P., ed., *Agricultural History* Regina: Canadian Plains Research Centre.

Gibbs, F. W., & Cohen, D. J. (2011). A conversation with data: prospecting Victorian words and ideas. *Victorian Studies*, *54*(1), 69–77.

Goldstone, A., & Underwood, T. (2014). The quiet transformations of literary studies: what thirteen thousand scholars could tell us. *New Literary History*, *45*(3), 359–384.

Graham, S., Milligan, I., & Weingart, S. (Forthcoming). *The historian's macroscope: big digital history (open access draft)*. Imperial College Press. Retrieved from http://www.themacroscope.org/

Gregory, I. N., & Geddes, A. (Eds.). (2014). *Toward spatial humanities: historical GIS and spatial history*. Bloomington: Indiana University Press.

Gregory, I. N., & Hardie, A. (2011). Visual GISting: bringing together corpus linguistics and geographical information systems. *Literary and Linguistic Computing*, *26*(3), 297–314. http://doi.org/10.1093/llc/fqr022

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875–3889. http://doi.org/10.1098/rsta.2010.0149

Guldi, J., & Armitage, D. (2014). *The history manifesto*. Cambridge: Cambridge University Press. Retrieved from http://historymanifesto.cambridge.org/

Hinichs, B., Alex, A., Clifford, J., Watson, A., Quigley, A., Klein, E., Coates, C. (2016) Trading Consequences: A Case Study of Combining Text Mining & Visualisation to Facilitate Document Exploration. Digital Scholarship in the Humanities; DH2014 Special Issue.

Hitchcock, T. (2013a). Confronting the digital: or how academic history writing lost the plot. *Cultural and Social History*, *10*(1), 9–23. http://doi.org/10.2752/147800413X13515292098070.

Hitchcock, T. (2013b, December 9). Big data for dead people: digital readings and the conundrums of positivism. Retrieved from http://historyonics.blogspot.co.uk/2013/12/big-data-for-dead-people-digital.html.

Jockers, M. L. (2013). *Macroanalysis: digital methods and literary history*. Urbana: University of Illinois Press.

Klein, E., Alex, B., & Clifford, J. (2014). Bootstrapping a historical commodities lexicon with SKOS and DBpedia. In *Proceedings of LaTech 2014 at EACL 2014*. Gothenburg, Sweden. Retrieved from http://www.aclweb.org/anthology/W/W14/W14-0603.pdf.

Lawson, K. M. (2014, April 8). Exploring Trading Consequences. Retrieved from http://chronicle.com/blogs/profhacker/exploring-trading-consequences/56415.

Lopresti, D. (2009). Optical Character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, *12*(3), 141–151.

Mikheev, A., Grover C. and Moens, M. (1999). XML tools and architecture for Named Entity recognition. *Journal of Markup Languages: Theory and Practice, 1*(3), 89–113.

Milligan, I. (2013). Illusionary order: online databases, optical character recognition, and Canadian history, 1997–2010. *The Canadian Historical Review*, *94*(4), 540–569.

Moretti, F. (2007). *Graphs, maps, trees: abstract models for literary history*. London: Verso.

Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A., & Rayson, P. (2014). Automatically analyzing large texts in a GIS Environment: the registrar general's reports and cholera in the 19th century. *Transactions in GIS*, n/a–n/a. http://doi.org/10.1111/tgis.12106.

Southall, H., Aucott, P., & Westwood, J. (2014). PastPlace – the global gazetteer from the people who brought you "A Vision of Britain through Time." In *UK Archives Discovery Forum*. London. Retrieved from http://eprints.port.ac.uk/15626/.

Turkel, W. J. (2006, April 5). Methodology for the infinite archive. Retrieved from http://digitalhistoryhacks.blogspot.de/2006/04/methodology-for-infinite-archive.html.

---

[1] In addition to the main collections of documents in the corpus we included just over three hundred books and documents downloaded from ProQuest's British Periodicals collection, the Internet Archive and Jstor by one of the historians on the team, with a particular focus on Ceylon (Sri Lanka), palm oil, coconuts and soap production. This is a small sample of the material available from these three digital archives and not a significant component of the corpus.

[2] http://archive.org/scanning

[3] Advanced OCR software can now identify tables and extract the information into a spreadsheet, but it still struggles with the irregular layout of nineteenth-century printed tables and requires significant human intervention on a page-by-page level. It would also be very expensive to re-OCR already digitized collections to exploit the advances of this technology.

[4] http://tcqdev.edina.ac.uk/search/commodity/154-Tallow/document/193559.

[5] *Annual Statement of the Trade of the United Kingdom with Foreign Countries and British Possessions. 1911 Compared with the Four Preceding Years. Compiled in the Statistical Office of the Customs and Excise Department.* Volume I., Document Type: COMMAND PAPERS; ACCOUNTS AND PAPERS, 1912, 357, http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&res_dat=xri:hcpp&rft_dat=xri:hcpp:fulltext:1912-015254:357.

[6] It is important to note that we do not have a complete understanding of the frequency of tables throughout the rest of the corpus.

[7] http://tcqdev.edina.ac.uk/search/location/53646-Saumur/commodity/136-Copper/document/183168

[8] http://tcqdev.edina.ac.uk/search/location/476-Russia/commodity/136-Copper/document/127168

[9] Trade and Navigation Accounts. For the Month Ended 31st December 1869, and Year Ended 31st December 1869; and Customs Duties and Excise Accounts for Each of the Years 1867, 1868, and 1869., Document Type: HOUSE OF COMMONS PAPERS; ACCOUNTS AND PAPERS, 69 1868, 26, http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&res_dat=xri:hcpp-us&rft_dat=xri:hcpp:rec:1868-045715.

[10] This process is also called geotagging in some of the literature.

[11] http://www.geonames.org

[12] This results in a standard problem in historical databases where the computer requires more precise information than we necessarily have available. As a result the database ends up with the majority of dates normalized to January 1 in the year of publication as the default. It is possible to simply extract the year from these data fields when working with the data, but it is important to remain aware that databased information is often more precise than the historical records.

[13] In some cases we needed to create a new URL to ground commodities not found in DBpedia.

[14] http://tcqdev.edina.ac.uk/search/location/61-Egypt/commodity/176-Cotton/document/106916

[15] Google's Ngrams suggests Saharunpore was more the common spelling in the 1850s and that Saharanpur gained prominence in the 1880s:

https://books.google.com/ngrams/graph?content=Saharunpore%2CSaharanpur&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2CSaharunpore%3B%2Cc0%3B.t1%3B%2CSaharanpur%3B%2Cc0

[16] PostGIS includes support for spatial information.

[17] These visualizations are discussed in a forthcoming article in *Digital Scholarship in the Humanities.*

[18] Precision is computed as: P = true positives / true positives + false positives. Recall is computed as: R = true positive / true positives + false negatives. Balanced F-score (or F1) is the harmonic mean of precision and recall: F1 = 2 x (precision x recall) / (precision + recall).

[19] Trading Consequences Search Interface: http://tcqdev.edina.ac.uk/search/commodity/

[20]Simple web visualization developed by Jon Bath to show the frequency of commodities over the long nineteenth century: http://drcspatial.usask.ca/clifford/ngram/; Trading Consequences Commodity Search for Cattle in the 1860s: http://tcqdev.edina.ac.uk/search/commodity/160-Cattle/document/123642; Trading Consequences Commodity Search for Cattle in the 1890s: http://tcqdev.edina.ac.uk/search/commodity/160-Cattle/decade/1890.

[21] "Diplomatic and Consular Reports. No. 3772 Annual Series. Venezuela," *Report for the Year 1906 on the Trade and Commerce of the Consular District of*

Caracus, Presented to both Houses of Parliament by Command of His Majesty, May 1907. (London: Harrison and Sons, 1907), 11.

[22] "Diplomatic and Consular Reports. No. 2628 Annual Series. Germany," *Report for the Year 1900 on the Trade and Commerce of the Consular District of Hamburg, Presented to both Houses of Parliament by Command of His Majesty, June 1901.* (London: Harrison and Sons, 1901), 40.

[23] *Africa. Papers relating to the cultivation of cotton in Africa.* Document Type: COMMAND PAPERS; ACCOUNTS AND PAPERS, 1857, Session 2, http://gateway.proquest.com.cyber.usask.ca/openurl?url_ver=Z39.88-2004&res_dat=xri:hcpp-us&rft_dat=xri:hcpp:fulltext:1857-033656:7.

[24] *Annual Statement of the Trade of the United Kingdom with Foreign Countries and British Possessions for the Year 1871.*, Document Type: COMMAND PAPERS; ACCOUNTS AND PAPERS, 1872, http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&res_dat=xri:hcpp-us&rft_dat=xri:hcpp:rec:1872-048695.

[25] http://tcqdev.edina.ac.uk/search/location/9896-Volta/commodity/176-Cotton/document/129138

[26] http://tcqdev.edina.ac.uk/search/location/327-Madagascar/commodity/176-Cotton/document/129151

[27] http://tcqdev.edina.ac.uk/search/location/615-Zanzibar/commodity/176-Cotton/document/128499; Russell GURNEY, *Report from the Select Committee on Slave Trade (East Coast of Africa); Together with the Proceedings of the Committee, Minutes of Evidence, Appendix and Index.*, Document Type: HOUSE OF COMMONS PAPERS; REPORTS OF COMMITTEES, 1871, http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&res_dat=xri:hcpp-us&rft_dat=xri:hcpp:rec:1871-047068.

[28] *Report of the Commissioners from British North America Appointed to Enquire into the Trade of the West Indies, Mexico and Brazil* (1866) Ottawa: Hunter.

[29] Ibid., 168.

[30] *Report of the Royal Commission on Trade Relations between Canada and the West Indies* (1910) London: Eyre and Spittiswoode.

[31] At this point in time there are no easy and fully functioning geoparsing tools, but there are geotaggers included as a part of Named Entity Recognition that humanists can apply to a corpus. Paper Machines does include a simple geoparsing tool: http://papermachines.org/how-to-use-paper-machines/.

[32] The Trading Consequences database can be downloaded at: https://github.com/digtrade/digtrade

[33] The queries are posted on Github: https://github.com/jburnford/HGISlab_USASK/blob/master/trading%20consequences%20queries%20Bath%20and%20Clifford

[34] A cognate project, *Spatial Humanities: Texts, GIS & Places* is a collaborative and interdisciplinary endeavor that also uses the Edinburgh Geoparser for historical and literary history research. Their approach, which has produced impressive results and methodological advances, has selected smaller amounts of material than *Trading Consequences*, largely limited to a subset of the Lancaster Newsbooks Corpus with 870,000 words and the Registrar General's reports for England and Wales with two million words (Gregory and Hardie 2011; Murrieta-Flores et al. 2014). The benefit of this approach is that the researchers know a great deal about the contents of the corpus and can focus on the relatively limited geography of England and Wales or Europe. As a consequence, the project has been able to develop techniques that give detailed answers to specific research questions, such as the geography of discussions of cholera in the Registrar General reports.