Plague Dot Text: Text mining and annotation of outbreak reports of the Third Plague Pandemic (1894-1952)

Arlene Casey[1], Mike Bennett[2], Richard Tobin[1], Claire Grover[1], Lukas Engelmann[3], and Beatrice Alex[1,4]

[1] Institute for Language, Cognition and Computation, School of Informatics
[2] University of Edinburgh Library
[3] Science Technology and Innovation Studies, School of Social and Political Science
[4] Edinburgh Futures Institute, School of Literatures, Languages and Cultures
University of Edinburgh, Edinburgh, UK

Abstract

The design of models that govern diseases in their relation to population is built on information and data gathered from past outbreaks. However, epidemic outbreaks are never captured in statistical data alone but are communicated by narratives, built on empirical observations. Outbreak reports discuss correlations between populations, locations and the disease to infer insights into causes, vectors and potential interventions. The problem with these narratives is usually the lack of consistent structure that allows for exploration of their collection as a whole. Our interdisciplinary research investigates more than 100 reports from the third plague pandemic (1894-1952) evaluating ways of building a corpus to extract and structure information through text mining and manual annotation. In this paper we discuss the progress of our exploratory project, how we enhance optical character recognition (OCR) methods to improve text capture, our approach to structure the narratives and identify relevant entities in the reports. The structured corpus is made available via Solr enabling search and analysis across the whole collection for future research dedicated e.g. to the identification of concepts. The corpus will enable researchers to analyse the reports collectively and allows for deep insights into the global epidemiological consideration of plague in the early 20th century.

1 Introduction

The Third Plague Pandemic (1894 - 1950) spread along sea trade routes affecting almost every port in the world and almost all inhabited countries, killing millions of people in the late nineteenth and early twentieth centuries [8, 7]. However, as outbreaks differed in severity, mortality and longevity, questions emerged at the time of how to identify the common drivers of the epidemic. After the Pasteurian Alexandre Yersin had successfully identified the epidemic's pathogen in 1894, *yersinia pestis* [19], the attention of epidemiologists and medical officers turned to the specific local conditions to understand the conditions under which the presence of plague bacteria turned into an epidemic. These observations were regularly transferred into reports, written to deliver a comprehensive account of the aspects deemed important by the respective author. These reports are the underlying data set for ongoing work in the *Plague.TXT* project which is conducted by an interdisciplinary team of medical historians, computer scientists and computer linguists. While each historical report was written as a standalone document relating to the spread of disease in a particular city, the goal of our work is to

Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

bring these reports together as one systematically structured collection of knowledge, preparing an annotated corpus made accessible to the wider research community and for follow-on analysis of narrative structures and concepts.

The pandemic reports offer deep insights into the ways in which epidemiological knowledge about plague was articulated at the time of the pandemic. While the reports contain a wealth of statistics and tabulated data their main value is found in articulated viewpoints about the causes for a plague epidemic, about the attribution of responsibility to populations, locations or climate conditions as well as about evaluating various measurements of control. The corpus thus constitutes an archive, from which future analysis will discern *concepts*, with which plague has been shaped into an object of knowledge in modern epidemiology. Further, the corpus will allow inferences to be made on the history of narrative epidemiology, a genre that has been widely overlooked in the historiography of 'formal epidemiology' [15].

Analysis of reports produced during the third plague has been done before, involving mainly manual collation of data such as collecting statistics across reports for mortality rates. The derived data has been used to reconstruct transmission trees from localised outbreaks [6], and to study potential sources and transmission across Europe [5]. In this work we are focused on collating the knowledge available across our report collection to consider comparable conventions within the entire corpus. This approach is challenging though as each report differs in the presentation, ordering and style of content. Despite their differences in style, these communicative reports are intended for the same audience of government officials and fellow epidemiologists and will present their arguments comparably. We build a schema structure to label the information contained in individual reports such that similar discourse segments can be linked and studied across the collection to enable for example comparative analysis across time and places. In our structuring and annotation efforts, we apply genre analysis as our methodological approach, treating each report as a communicative event about a specific plague outbreak. Whilst these arguments may differ in style, our work has shown that they can be isolated, annotated and mapped under one structure.

Our contribution is the development of a systematically structured corpus, which we capture through annotation, to assimilate similar discourse segments such as causes or treatments across the reports. In addition, we develop an interactive search interface to our collection eliminating the need for manual read and search activities. This search tool in combination with our structured schema allows follow-on research to conduct automated exploration of a rich source about the conceptual thinking at the time on the plague pandemic, to better understand the historical epistemology of epidemiology and to thus provide valuable lessons about dealing with contemporary global spread of disease.

In the following sections we give an overview of our pilot study describing the data collection, the challenges presented by OCR and improvements we have made to the original digitised reports. Following this we describe our annotation process including our schema to structure the reports to extract information. We discuss our combination of manual annotation and automated text mining techniques that support the retrieval and structuring of information from the reports. We discuss aspects of our search interface, enabled through Solr ,which we use to make the collection available online. Finally we give some examples of potential use of this interface.

2 Data

The Third Plague Pandemic was documented in over 100 outbreak reports for most major cities around the world. Many of them have been digitised, converted to text via optical character

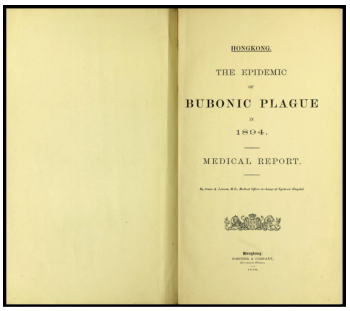

Figure 1: Bubonic plague report for Hong Kong, 1895.

Counts	Total	Min	Max	Mean	Stddev
Sent	229,043	32	17,635	2,245.5	3,713.6
Word	4,443,485	1,091	396,898	43,563.6	74,621.0

Table 1: Number of sentences and words in the collection of English plague reports, as well as corresponding counts for the smallest document (Min) and the largest document (Max), the average (Mean) and standard deviation (Stddev).

recognition (OCR) and are available via the Internet Archive and the UK Medical Heritage Library. See Figure 1 for an example of such a report covering the Hong Kong outbreak which was published in 1895 and is accessible with open access on Internet Archive.[1]

We treat all relevant reports for which we have a scan as one collection. While the majority of reports in this set (102) are written in English, there are further reports in French, Spanish, Portuguese and other languages which we excluded from the analysis at this stage.

Table 1 provides an overview of the data set in terms of counts of sentences and words in the collection and illustrates the variety of documents in this collection. To derive these counts we used automatic tokenisation and sentence detection over the raw OCR output which is part of the text mining pipeline described in section 3. While the smallest document is only 32 sentences long containing 1,091 word tokens, the largest report contains almost 400,000 word tokens. The collection contains 38 documents with up to 5K words each, 15 reports with between 5K and 10K words each, 32 documents with between 10K and 100K words each and 17 documents with 100K or more words each. In total, the reports amount to over 4.4 million word tokens and over 229K sentences.

2.1 OCR Improvements

When initially inspecting this digitised historical data, we realised that some of the OCR was of inadequate quality. We therefore spent time during the first part of the project on improving the OCR quality of the reports.

Using computer vision techniques, we processed the report images to remove warping artefacts [9]. We then identified likely textual areas in report images, and produced an effective crop, to provide the OCR engine with less extraneous data.[2] OCR is then performed using Tesseract[3], trained specifically for typeface styles and document layouts common to the time period of the reports.

Training was done across a range of truth data, covering period documents obtained from the IMPACT Project Datasets[4], documents from Project Gutenberg prepared for OCR training[5], internal ground-truth data compiled as part of the Scottish Session Papers project at the

[1] https://archive.org/details/b24398287
[2] http://libraryblogs.is.ed.ac.uk/librarylabs/2017/06/23/automated-item-data-extraction/
[3] https://opensource.google.com/projects/tesseract
[4] https://www.digitisation.eu/tools-resources/image-and-ground-truth-resources/
[5] https://github.com/PedroBarcha/old-books-dataset

Zones	Description
Title-matter	Title page
Preface	Preface information
Content-page	Content page information
Introduction	State of the epidemic at the time of the production of the report, summary of key features, evaluation of significance of the epidemic
Disease history	General points on the history of the epidemic, origin of outbreak
Outbreak history	Geographical and chronological overview of local outbreak. What happened this place this year
Local conditions	Descriptions of key elements that are considered noteworthy, something that has contributed or impacted the outbreak
Causes	Causes identified by the author e.g. usually points of origin, specific local conditions or descriptions of import
Measures	List of the measures e.g. undertaken to curb the outbreak, sanitary improvements, quarantines, disinfection or fumigation and rat catching
Clinical appearances	Description of the disease appearance, its usual course and its mortality
Laboratory	Description of bacteriological analysis, human lab work
Treatment	Description of the treatment given to patients
Cases	List of individual cases, usually with age, gender, occupation, course of disease, and time and dates of infection and death
Statistics	Contains many lists or tables of statistics such as deaths
Epizootics	Contains information solely about animals, experiments or discussions
Appendix	Labelled appendix
Conclusion	Conclusion

Table 2: Zones

University of Edinburgh[6] and typeface datasets designed for Digital Humanities collections.[7]

While we have not yet formally evaluated the improvements made to the OCR, observation of the new OCR output shows clear improvements in text quality. This is important as it affects the quality of downstream text mining steps. Previous research and experiments have found that errors in OCRed text have a negative cascading effect on natural language processing or information retrieval tasks [12, 14, 10, 1]. In future work, we would like to conduct a formal evaluation comparing the two versions of OCRed text to quantify the quality improvement.

3 Annotation

This section describes the schema we implemented to structure the information contained within the reports and automatic and manual annotation applied to our collection of plague reports. We first processed them using a text mining pipeline which we adapted and enhanced specifically for this data. The text mining output annotations are then corrected and enriched during a manual annotation phase which is still ongoing. Each report that has undergone manual annotation then undergoes further automatic geo-resolution and date normalisation to normalise and disambiguate different mentions of location names and dates in the text.

[6] https://www.projects.ed.ac.uk/project/luc020/brief/overview
[7] https://github.com/jbest/typeface-corpus

Entity Type	Entity Mentions
person	Professor Zabolotny, Professor Kitasato, Dr. Yersin, M. Haffkine
location	India, Bombay, City of Bombay, San Francisco, Venice
geographic-feature	house, hospital, port, store, street
plague-ontology-term	plague, bubo, bacilli, pneumonia, hemorrhages, vomiting
date	1898, March 1897, 4th February 1897, the beginning of June, next day
date-range	1900-1907, July 1898 to March 1899, since September 1896
time	midnight, noon, 8 a.m., 4:30 p.m.
duration	ten days, months, a week, 48 hours, winter, a long time
distance	20 miles, 100 yards, six miles, 30 feet
population/group of people	Chinese, Europeans, Indian, Russian, Asiatics, coolies, villagers
percent	8%, 25 per cent, ten per cent

Table 3: Entity types and examples of entity mentions in the plague reports.

3.1 Developing a Schematic Structure for the Reports

Whilst our corpus provides a rich source of material the collation of report narratives into a structured format for information retrieval is not straightforward. The authors approach the narrative with different styles making the application of a schema to support extraction challenging. As discussed in the Introduction our methodological approach is to treat each report as a communicative event and we hypothesise that the reports - despite their variation of styles - as they are intended for the same audience of fellow epidemiologists and government officials will present and structure their arguments comparably. This hypothesis is based on the works of Swales and Bhatia [18, 4] who propose that authors achieve their argumentative structure through steps and moves. Whilst the sequence of arguments may differ in order of presentation, nonetheless they can identified and mapped to a schematic structure. Our schematic structure is represented by the zoning schema presented in Table 2.

Annotating text with zones creates structure within our reports, allowing us to collate similar discourse text segments, such as those that discuss treatments or local conditions. Identifying similar discourse segments allows for targeted analysis which can reveal knowledge and thinking on these zones and how this knowledge developed over time. Our zoning schema was created from studying a subsection of reports, section titles and three rounds of pilot annotation on a subsection of documents.

3.2 Automatic Annotation and Text Mining

To process the plague reports, we used the Edinburgh Geoparser [11], a text mining pipeline which has been previously applied to other types of historical text [2]. This tool is made up of a series of processing components. It takes an input raw text and performs standard text pre-processing on documents in XML format, including tokenisation, sentence detection, lemmatisation, part-of-speech tagging and chunking as well as named entity recognition. Before tokenising the text, we also applied a script to repair broken words which were split in the input text as a result of end-of-line hyphenation [3].

We adapted the Edinburgh Geoparser by expanding the list of types of entities it recognises in text, including `geographic-feature`, `plague-ontology-term` and `population/group` etc.[8]

[8]Note that our goal was to emulate descriptions used by the report authors at the time, mirroring concepts of race and ethnicity that were often implicated in the construction of epidemiological arguments. Some examples of the `population/group` entities show that these are often derogatory and considered offensive today. They are

The full list of entity types extracted from the plague reports and examples are presented in Table 3. Date entity normalisation and geo-resolution are also applied once the manual annotation (described in the next section) for a document is completed. This is to ensure that the corrected text mining output is disambiguated, including manual corrections of spelling mistakes occurring in entity mentions.

3.3 Manual Annotation

Manual annotation was necessary for a number of reasons. Whilst some zones could be identified automatically from section titles we found that this was often hampered by spelling errors due to OCR issues arising from typeface styles and title placements in margins. In addition, depending on author narrative styles some zones could be found nested within sections with no titles. This created the need to manually annotate zones. The automatic recognition of named entities (see Section 3.2) was partially successful but also suffered from spelling errors and OCR issues. In addition, as more reports are being annotated new entity mentions are identified. Thus manual correction of erroneous and correction of spurious entity mentions was required.

Manual annotation is conducted using Brat[9], a web-based text annotation tool. After the text was processed automatically as described above, it was converted from XML into Brat format to be able to correct the text mining output and add zone annotations.[10] Figure 2 shows an excerpt of an example report being annotated in Brat. Entities such as date, location or geographic feature listed in Table 3 can be seen highlighted in the text. The start of an outbreak history zone is also marked at the beginning of the excerpt.

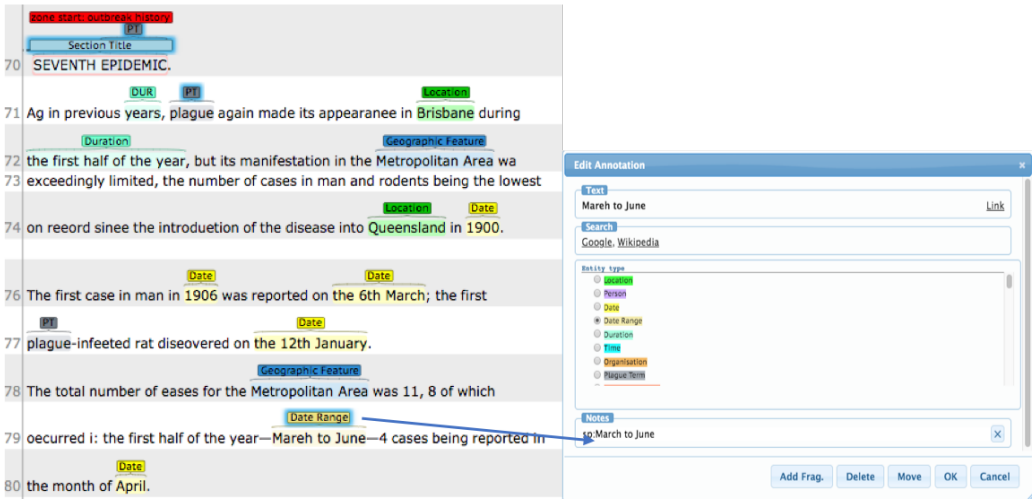

Figure 2: Brat annotation tool.

strictly understood to be of value only for the illumination of historical discourse.

[9] https://brat.nlplab.org/

[10] We have not yet conducted double annotation to determine inter-annotator agreement for this work but this is something we are planning to do in the future.

3.3.1 Zone Annotation

Zone annotation, as defined by our schema shown in Figure 2, is applied inclusive of a section title and can be nested. For example, zones of cases are often found inside treatment or clinical appearances zones. Footnote zones were added as these often break the flow of the text and make downstream natural language processing challenging. In addition, we added Header/Footer markup to be able to exclude headers and footers, e.g. the publisher name or report name repeated on each page, from further analysis or search.

Tables were a challenge for the OCR and unusable for the most part. When marking up tables, we also record its page number. Text within tables is currently ignored when ingesting the structured data to Solr (see Section 4). However, tables include a lot of valuable statistical information. In the next phase of the project we will investigate whether this information can be successfully extracted or whether it will need to be manually collated.

3.3.2 Entity Annotation

During manual annotation we instruct our annotators to correct any wrongly automated entities and add those that were missed. Any mis-spellings of entity mentions, mostly caused by the OCR process, are also corrected in the note field, as shown in Figure 2. The mis-spellings are used as part of our text cleaning process. The corrected forms are also used to geo-resolve place names and normalise dates. These final two processing steps of the Edinburgh Geoparser are carried out on each report once it has been manually annotated and converted back to XML.

4 Data Search Interface

Historians and humanities researchers are often faced with the daunting task of manual search through document collections to find information pertinent to their research interest. Additionally, the challenges of working with such text digitally require interdisciplinary collaboration. HistSearch [16], an on-line tool applied to historical texts, demonstrates how computational linguists and historians can work together to automate access to information extraction. One goal of the *Plague.TXT* project is to make our digital collection available as an on-line search and retrieval resource but in addition this collection should be accessible. This means being available to computational linguists for in-depth analysis as well as interactive search for humanities researchers. We do this with Apache Solr[11].

Solr is an open-source enterprise-search platform, widely used for digital collections. The features available through the Solr search interface make our collection accessible to a wide audience. For example, it offers faceting features that support grouping and organising data in multiple ways, whilst data interrogation can be achieved through its simple interface with term, query, range and data faceting. Solr also supports rich document handling with text analytic features and direct access to data in a variety of formats.[12]

We are currently customising and improving the filtering of the data for downstream analysis in Solr. Below we describe on-going filtering steps with Solr and provide examples to demonstrate a search interface customisation and to show analysis that can be done from data retrieved via the search interface.

[11] https://lucene.apache.org/solr/
[12] See the Solr website for further description of features.

4.1 Data Preparation and Filtering in Solr

The annotated data is prepared and imported to Solr using Python, with annotations created both automatically by the Geoparser and manually by the annotators mapped to appropriate data fields (e.g. date-range entities are mapped to a Date Range field[13]), enabling complex queries across the values expressed in the document text. Additionally, manual spelling corrections are used to replace the corresponding text in the OCR rendering prior to Solr ingestion, thus improving the accuracy of language-based queries and further textual analysis. We also implement lexicon-based entity recognition for entities that have been missed during the annotation and for additional entity types, e.g animals. Solr allows for storing and searching by geo-spatial coordinates and we import geo-coordinates associated with entities identified by the Geoparser. Geo-coordinates can be used to support interactive visualisations, as developed in the Trading and Consequences project [13] which visualises commodities through their geo-spatial history. In addition, this location information can be used in analysis such as transmission and spread, e.g. geo-referenced plague outbreak records have been used to show how major trade routes contributed to the spread of the plague [20]. Using 'Case Zones' we are currently assessing NLP techniques to extract case information into a more structured format for direct access to statistical information from hundreds of individual case descriptions.

4.2 Use Case Example: Illustration of Interactive Search

Figure 3 shows one of our customised search interfaces. This allows users to search across the entire report collection displaying original image snippets from the reports containing the search term(s). Snippet search is supported by indexing OCR transcriptions from word-level ALTO-XML[14] in Solr and then by using Whiiif[15], an implementation of the IIIF Search API[16] designed to provide full-text search with granular, word-level annotation results to enable front-end highlighting. Further technical details about the Whiiif package can be found on the University of Edinburgh Library Labs blog[17]

4.3 Use Case Example: Finding Discussion Concepts in Causes Across Time Periods

Our search interface facilitates queries across the collection on content and meta-data as well as queries based on zones or entity types using facets such as date range. We are interested to see topic discussed in causes zones and if these differ between report time periods. Using Solr we search for cause zones published during the pandemic 1894-6 comparing these to cause zones in reports 1904 and beyond. We apply stop-word removal and removal of all non dictionary terms to the cause zone text. In future, we will make indexed versions of cleaned data in this format directly accessible from Solr. We use topic modelling (LDA with Gensim Python library[18]) to compare the cause zone text at the different time points, selecting two topics.

Results are presented in Table 4. The earlier reports show the discussion centering around environment aspects with focus on populations, conditions of living and buildings and how this might cause the spread. The second topic is linked to the concepts at that time period, about how the diseases may spread through the water system, with studies of ordinance maps of

[13] https://lucene.apache.org/solr/guide/8_1/working-with-dates.html#date-range-formatting
[14] http://www.loc.gov/standards/alto/
[15] https://github.com/mbennett-uoe/whiiif
[16] https://iiif.io/api/search/1.0/
[17] http://libraryblogs.is.ed.ac.uk/librarylabs/2019/07/03/introducing-whiiif/
[18] https://radimrehurek.com/gensim/index.html

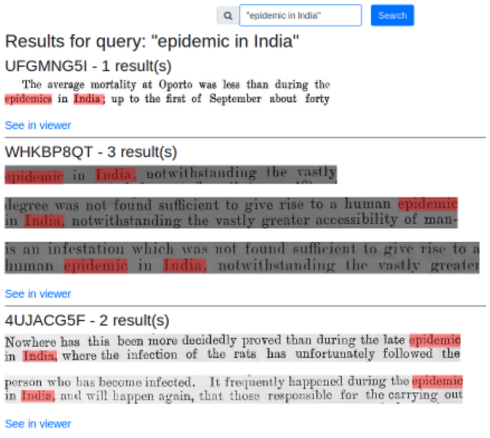

Figure 3: Snippet search example.

Topic/Date	keywords
(1) 1894-96	latrine, house, soil, street, find, case, time, plague, infection, opinion, condition, may, must, question, see
(2) 1894-96	house, people, ordinance, well, supply, cause, must, condition, drain, disease, pig, matter, area, water, provision
(1) 1904-07	plague, rat, case, infection, man, flea, may, infect, place, fact, evidence, disease, instance, produce, find
(2) 1904-07	year, month, temperature, epidemic, influence, season, infection, december, condition, may, june, prevalence, rat, follow, number

Table 4: Discussion topics from cause zones by time period

sewerage and water ways. Looking at the later reports we now see rats and fleas and infection are more prominent as a discussion topic but also season, temperature and weather form a topic being discussed as a causal factor.

5 Discussion, Conclusions and Future Work

In this paper we have presented the work done in the pilot stage of our *Plague.TXT* project. The work is the outcome of an interdisciplinary team working together to understand the nature and complexities of a historical text collection and the needs of the potential different types of users of this collection. A major contribution of this project is the development of an annotation schema to bring individual reports together as one corpus. This enables streamlined and efficient linking of knowledge and concepts used in the comprehension of the third plague pandemic covering the time period of the collection and provides the ability to analyse these reports as a one coherent corpus. By making this data accessible through the Solr search interface, we can share it with the research community in ways that cater for the needs of different field experts.

As a next step we will explore spelling normalisation. Diachronic and synchronic spelling variance is a known issue in historical documents [17] but in addition the OCR process also introduces mis-spellings. We will use existing methods for spelling normalisation and fuzzy string matching capabilities within Solr to correct for spelling variation introduced by OCR.

Manual annotation is time consuming and can be an error prone process. As we increase the number of reports annotated with zone markup, we also intend to investigate if text similarity measures can be used for automatic zone identification. We are also developing our methods to directly access the statistical information contained within case zones and within tables.

6 Acknowledgements

This work was funded by the Challenge Investment Fund 2018-19 from the College of Arts, Humanities and Social Sciences, University of Edinburgh.

References

[1] B. Alex and J. Burns. Estimating and rating the quality of optically character recognised text. In *In Proceedings 1st DATeCH*, pages 97–102, New York, NY, USA, 2014.

[2] B. Alex, K. Byrne, C. Grover, and R. Tobin. Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal for Humanities and Arts Computing*, 9(1):15–35, 2015.

[3] B. Alex, C. Grover, E. Klein, and R. Tobin. Digitised Historical Text: Does it have to be mediOCRe? In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 401–409, 2012.

[4] V. K. Bhatia. Analysing Genre: Language use in Professional Settings. *New Youk:Routledge*, 2014.

[5] B. Bramanti, K.R. Dean, L. Walle, and N.C. Stenseth. The third plague pandemic in europe. *Proceedings of the Royal Society B: Biological Sciences*, 286(1901):20182429, 2019.

[6] K.R. Dean, F. Krauer, and B.V. Schmid. Epidemiology of a bubonic plague outbreak in glasgow, scotland in 1900. *Royal Society Open Science*, 6(1):181695, 2019.

[7] M. J. Echenberg. *Plague Ports: The Global Urban Impact of Bubonic Plague,1894-1901*. New York University Press, New York, 2007.

[8] L. Engelmann. Mapping Early Epidemiology: Concepts of Causality in Reports of the Third Plague Pandemic 18941950. In E. T. Ewing and K. Randall, editors, *Viral Networks: Connecting Digital Humanities and Medical History*, pages 89–118. VT Publishing, 2018.

[9] B. Fu, M. Wu, R. Li, W. Li, Z. Xu, and C. Yang. A model-based book dewarping method using text line detection. In *Proc. CBDAR 2007*, pages 63–70, 2007.

[10] A. Gotscharek, U. Reffle, C. Ringlstetter, K. U. Schulz, and A. Neumann. Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *IJDAR*, 14(2):159–171, 2011.

[11] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball. Use of the Edinburgh Geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A*, 368(1925):3875–3889, 2010.

[12] A. Hauser, M. Heller, E. Leiss, K. U. Schulz, and C. Wanzeck. Information access to historical documents from the Early New High German period. In L. Burnard, M. Dobreva, N. Fuhr, and A. Lüdeling, editors, *Digital Historical Corpora- Architecture, Annotation, and Retrieval*, Dagstuhl, Germany, 2007.

[13] U. Hinrichs, B. Alex, J. Clifford, A. Watson, A. Quigley, E. Klein, and C.M. Coates. Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities*, 30(suppl_1):i50–i75, 10 2015.

[14] D. Lopresti. Measuring the impact of character recognition errors on downstream text analysis. In B.A. Yanikoglu and K. Berkner, editors, *Document Recognition and Retrieval*, volume 6815. SPIE, 2008.

[15] A. Morabia, editor. *A history of epidemiologic methods and concepts*. Birkhauser Verlag, Basel ; Boston, 2004. OCLC: 55534998.

[16] E. Pettersson, J. Lindström, B. Jacobsson, and R. Fiebranz. Histsearch – implementation and evaluation of a web-based tool for automatic information extraction from historical text. In *3rd HistoInformatics Workshop*, Krakow, Poland, 2016.

[17] M. Piotrowski. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157, 2012.

[18] J. Swales. Aspects of article introductions. *Language Studies Unit. University of Aston in Birmingham*, 1981.

[19] Yersin. La Peste Bubonique a Hong Kong. *Annales de l'Institut Pasteur*, pages 662–667, 1894. f667, 1.

[20] R. Yue, H.F. Lee, and C.Y.H. Wu. Trade routes and plague transmission in pre-industrial Europe. *Scientific reports*, 7(1):12973, 2017.