Does Digitised Historical Text have to be mediOCRe?

Optical character recognition and text mining of historical documents

Beatrice Alex and Mike Bennett
University of Edinburgh
@bea_alex, balex@staffmail.ed.ac.uk

National Library of Scotland, Edinburgh, 29th of January 2020


TEXT MINING

THE UNIVERSITY *of* EDINBURGH
Edinburgh Futures Institute

THE ALAN TURING INSTITUTE

- Chancellor's Fellow and Turing Fellow at the Edinburgh Futures Institute
- Co-convener of the Data Science and Digital Humanities SIG at the Alan Turing Institute
- Head of the Edinburgh Language Technology Group
    - Information extraction, data linking and document classification
    - Developed and released the Edinburgh Geoparser

- Collaborations with domain experts in many disciplines:
  - History: Trading Consequences (Digging Into Data)
  - Literature: Palimpsest (AHRC Big Data)
  - Botany: BotaniTours (Smart Tourism)
  - Social Sciences: Text Mining Careers (CIF, Turing)
  - Healthcare: Text mining EHRs (MRC Pathfinder), PrepDoc (EIT Digital)
  - Plague.TXT (CIF)

Overview of today’s talk

- Plague Dot Text
- The Team
- Background
- Related work
- Data
- OCR quality
- Digitisation process
- Information extraction and manual annotation
- Preliminary analysis
- Summary and next steps

Plague DOT Text

- Approached by Lukas Engelmann about analyzing the a corpus of reports about the Third Plague Pandemic (1894-1952)
- Obtained funding from the Challenge Investment Fund at CAHSS at Edinburgh
- Successful small-scale project with presentations at Digital Humanities 2019, a paper at the HistoInformatics workshop at the 23rd International Conference on Theory and Practice of Digital Libraries (TPDL 2019) and invited for a journal article (in preparation)
- Internal follow-on funding from LLC and STIS for further annotation
- Looking for a larger grant to complete the text mining work and conduct historical/epidemiological research on the output

Funded by the Challenge Investment Fund 2018/19 at the College of Arts, Humanities and Social Sciences, University of Edinburgh

THE UNIVERSITY of EDINBURGH
School of Social and Political Science

THE UNIVERSITY of EDINBURGH
School of Literatures, Languages and Cultures

THE UNIVERSITY of EDINBURGH
informatics

Background

The Third Plague Pandemic (1894-1952)

- 1894: Alexandre Yersin identified pathogen
- Catastrophic outbreaks in India, Manchuria (~12 million dead)
- Outbreaks in almost every port city around the world (Echenberg 2007, Engelmann 2018)
- Rat-Flea-Human model proposed ~1900, widely accepted by 1910
- 1960, worldwide casualties dropped below 200 per year

Historical Use Case

- Epidemiologists used to write narrative accounts of outbreaks to understand the drivers of an epidemic. This genre has been widely overlooked in the historiography of ‘formal epidemiology’ (Morabia 2004).
- We aim to structure these narratives, to identify concepts and to extract epidemiological data from written accounts.
- Expected Data: geo-data, bibliometrics, descriptions of plague, accounts of environmental drivers, interventions, individual case files, treatments, laboratory analysis ...

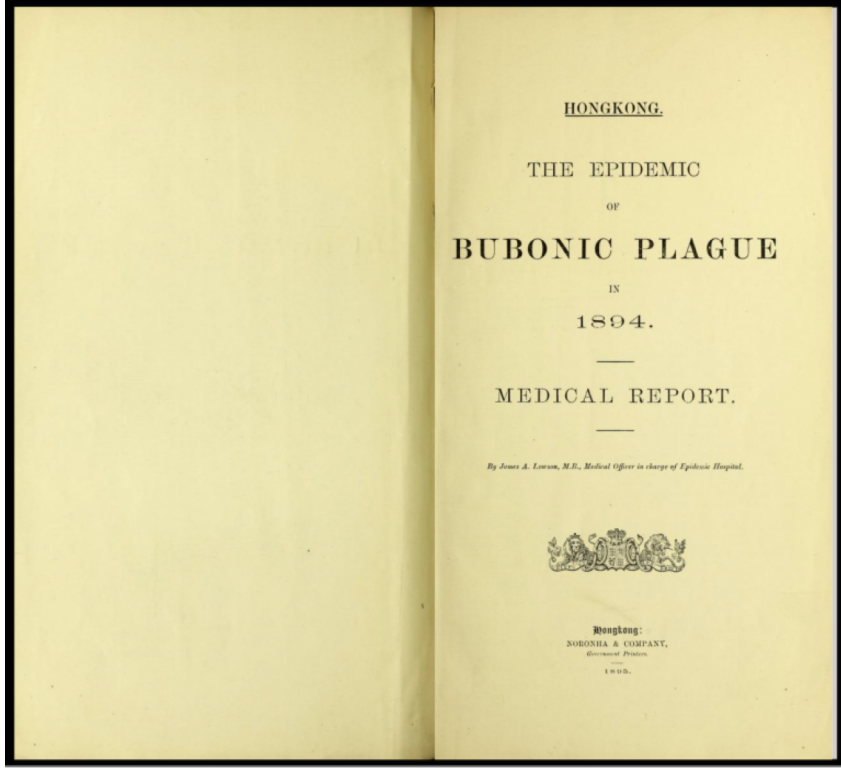
Plague in Hong Kong (1894), Hawaii (1900) and Liverpool (1913): https://www.repository.cam.ac.uk/handle/1810/275905

Related Work

- Analysis of reports during the Third Plague Pandemic involved largely manual collection of data (e.g. statistics across reports for mortality rates).
- Derived data used to reconstruct transmission trees from localised outbreaks (Dean et al. 2019) or to study potential sources of transmission to Europe (Branmanti et al. 2019).
- Krauer, University of Oslo, PhD project on developing mathematical models that simulate the spread of plague in preindustrial Europe. Digitised data from historical books and publications to be text mined and geoparsed for Old German.
- HistSearch (Petterson et al. 2016) is a web-based prototype tool for automatic processing of historical text (POS tagging and parsing).
- Trading Consequences and Litlong.org interfaces to historical/literary text

Data

- The third plague pandemic has been documented in over 100 outbreak reports for most major cities around the world
- Focus on English reports initially
- Many of them have been digitised, converted to text via optical character recognition (OCR)
- Available via the Internet Archive and the UK Medical Heritage Library
- Needed to explore the data to decide what is possible to do automatically


HONGKONG.

THE EPIDEMIC

OF

BUBONIC PLAGUE

IN

1894.

MEDICAL REPORT.

By James A. Lowson, M.B., Medical Officer in charge of Epidemic Hospital.

Hongkong:

NORONHA & COMPANY,

Government Printers

1895.

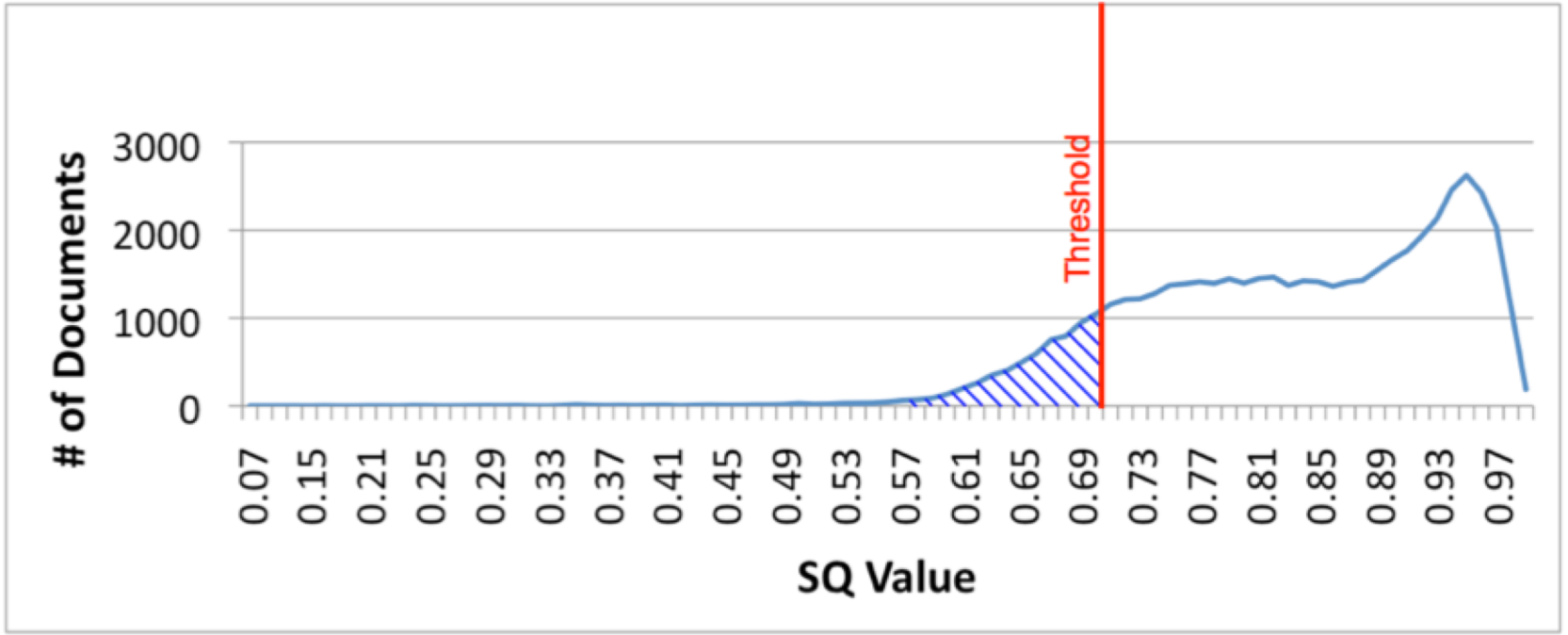
Optical Character Recognition

- OCR quality of available data did not look very good.
- But text quality is very important for natural language processing tasks (Hauser et al. 2017, Lopresti 2008, Gotscharek et al. 2011, Alex et al. 2012)
- Plus historians are not always aware about the percentage of data they miss out on using keyword search of historical text collection.

```
I have tlie lioiiour to foi'ward herewith foi- your information a Heport upon
the Kpideiiiic of Plague in HoDgkoug in 1894, so far as it concerns tlie medical
work which I carried out under your directions.

I regret extremely that several important matters – including the e[)idemiology
of the disease – which I could have wished to discuss at some length, have been
touched upon very superficially, or passed over altogether in this Report. I will
ask you to accept as an excuse for my shortcomings in these respects the following
facts of which you are, I believe, already cognizant : –
```

Text Quality

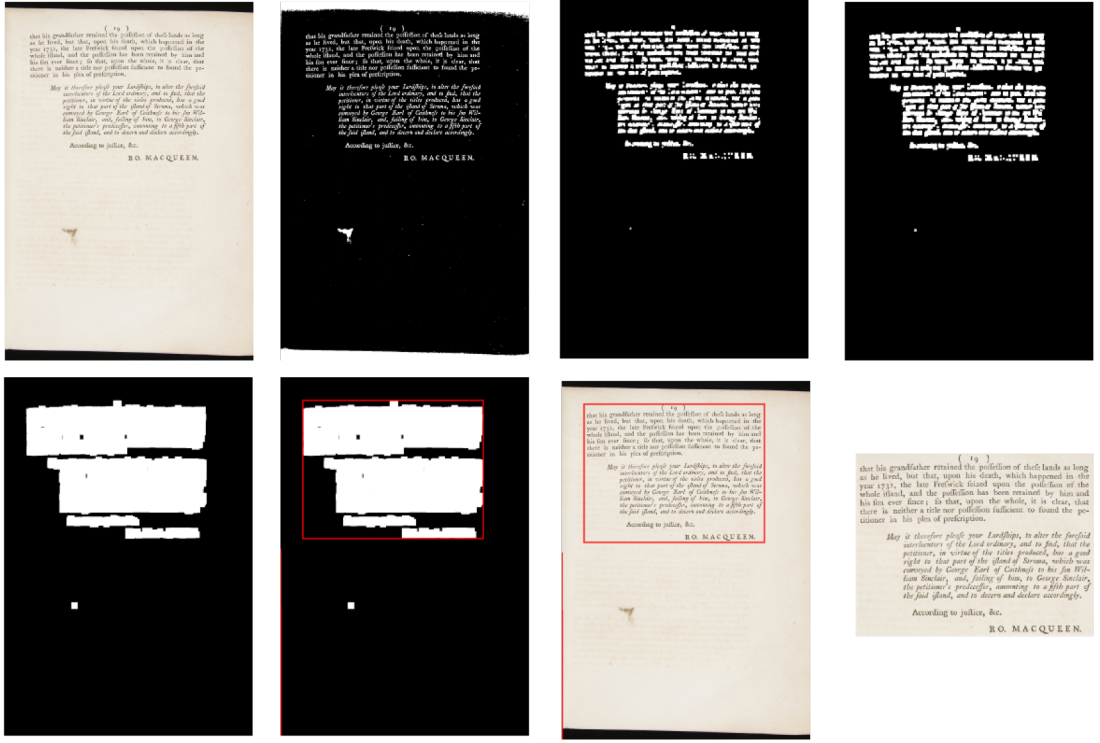


Quality score distribution over the English language portion of Early Canadiana Online (55,313 documents).

Alex and Burns, DATeCH, 2014

Optical Character Recognition

- Using computer vision techniques, we identify likely textual areas in report images, and produce an effective crop, to provide the OCR engine with less extraneous data (Fu et al. 2007, see also http://libraryblogs.is.ed.ac.uk/librarylabs/2017/06/23/automated-item-data-extraction/
- OCR is performed using Tesseract, trained specifically for typeface styles and document layouts common to the time period of the reports.



Custom trained Tesseract running on ECDF Eleanor cloud service

```
<TextLine ID="line_25" HPOS="80" VPOS="2252" WIDTH="1711" HEIGHT="47">
    <String ID="string_246" HPOS="80" VPOS="2252" WIDTH="58" HEIGHT="36" WC="0.96" CONTENT="the"/>
    <String ID="string_247" HPOS="158" VPOS="2252" WIDTH="184" HEIGHT="36" WC="0.28" CONTENT="condition"/>
    <String ID="string_248" HPOS="363" VPOS="2252" WIDTH="40" HEIGHT="36" WC="0.96" CONTENT="of"/>
    <String ID="string_249" HPOS="417" VPOS="2252" WIDTH="81" HEIGHT="35" WC="0.96" CONTENT="filth"/>
    <String ID="string_250" HPOS="519" VPOS="2252" WIDTH="37" HEIGHT="35" WC="0.96" CONTENT="in"/>
    <String ID="string_251" HPOS="577" VPOS="2252" WIDTH="118" HEIGHT="35" WC="0.79" CONTENT="whieh"/>
    <String ID="string_252" HPOS="719" VPOS="2252" WIDTH="12" HEIGHT="34" WC="0.45" CONTENT="I"/>
    <String ID="string_253" HPOS="754" VPOS="2252" WIDTH="114" HEIGHT="36" WC="0.94" CONTENT="found"/>
    <String ID="string_254" HPOS="889" VPOS="2253" WIDTH="61" HEIGHT="36" WC="0.96" CONTENT="the"/>
    <String ID="string_255" HPOS="971" VPOS="2254" WIDTH="142" HEIGHT="43" WC="0.96" CONTENT="houses,"/>
    <String ID="string_256" HPOS="1135" VPOS="2253" WIDTH="75" HEIGHT="35" WC="0.95" CONTENT="also"/>
    <String ID="string_257" HPOS="1230" VPOS="2258" WIDTH="129" HEIGHT="30" WC="0.96" CONTENT="streets"/>
    <String ID="string_258" HPOS="1384" VPOS="2252" WIDTH="68" HEIGHT="35" WC="0.95" CONTENT="and"/>
    <String ID="string_259" HPOS="1475" VPOS="2252" WIDTH="112" HEIGHT="47" WC="0.93" CONTENT="alleys"/>
    <String ID="string_260" HPOS="1621" VPOS="2252" WIDTH="34" HEIGHT="34" WC="0.74" CONTENT="in"/>
    <String ID="string_261" HPOS="1687" VPOS="2252" WIDTH="104" HEIGHT="36" WC="0.74" CONTENT="other"/>
</TextLine>
```

OCR challenges in historic documents

- With historical documents, the scanned page images are often of suboptimal quality for OCR
- Most OCR engines are designed to work best with pages of printed text on a white background
- As well as poorly printed letters or recto/verso bleeding, historical documents often contain a lot of “noise” (smudges, marks, damage, etc)
- Two main ways to tackle this issue:
  - 1) Improve the quality of the image before running the OCR process
  - 2) Train the OCR engine to work better with historical texts

Improving the image

- Many simple image processing functions can be chained together to prepare an image for OCR
- Most OCR engines will do this automatically, but they tend to be more conservative and less well geared for older documents
  - For example, binarisation of images can be done with several algorithms, some of which are better suited for less “clean” documents (O.D. Trier, T. Taxt 1995)
- By performing this processing before passing the image to the OCR engine, we maintain greater control over the process and we can overcome many of these issues

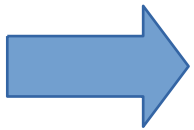
Improving the image

- The most important step is to binarise the image (that is, to convert it so that it only contains pure black and white pixels)
- This is done by converting the image to grayscale and then using a thresholding algorithm to decide which “strengths” of gray become black and which become white
- Because of the variance present in older documents, an adaptive approach works best, using a sliding window to make a judgement on each pixel dependent on the level of gray around it (Niblack 1986).
- This is especially helpful for yellowing pages as it can make them as if they were on white paper

VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.



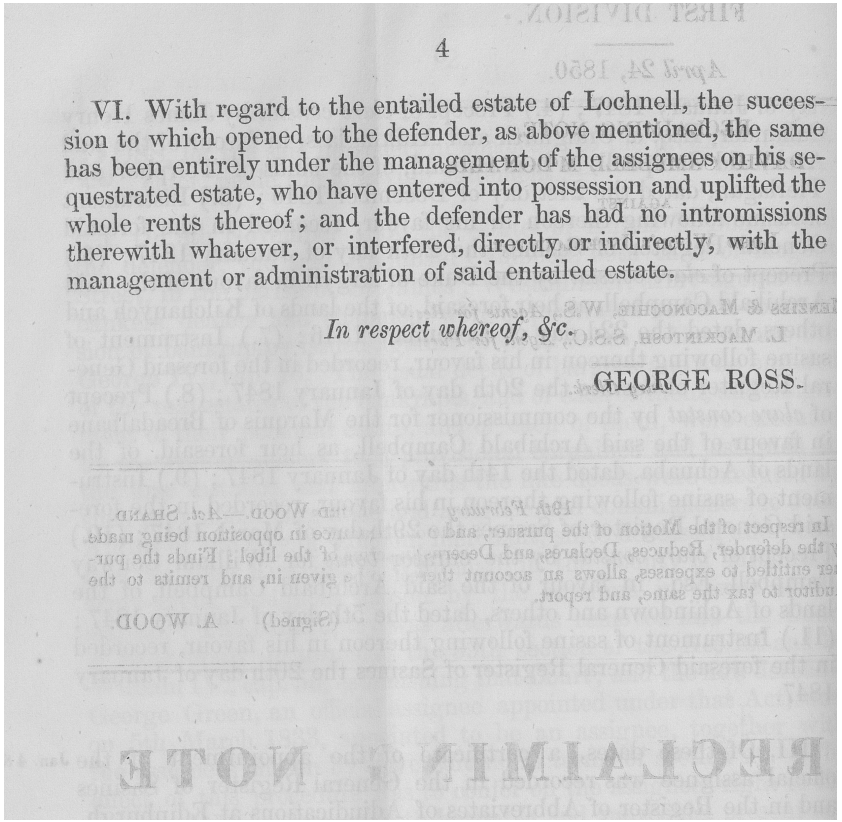
4

VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.

85% Threshold

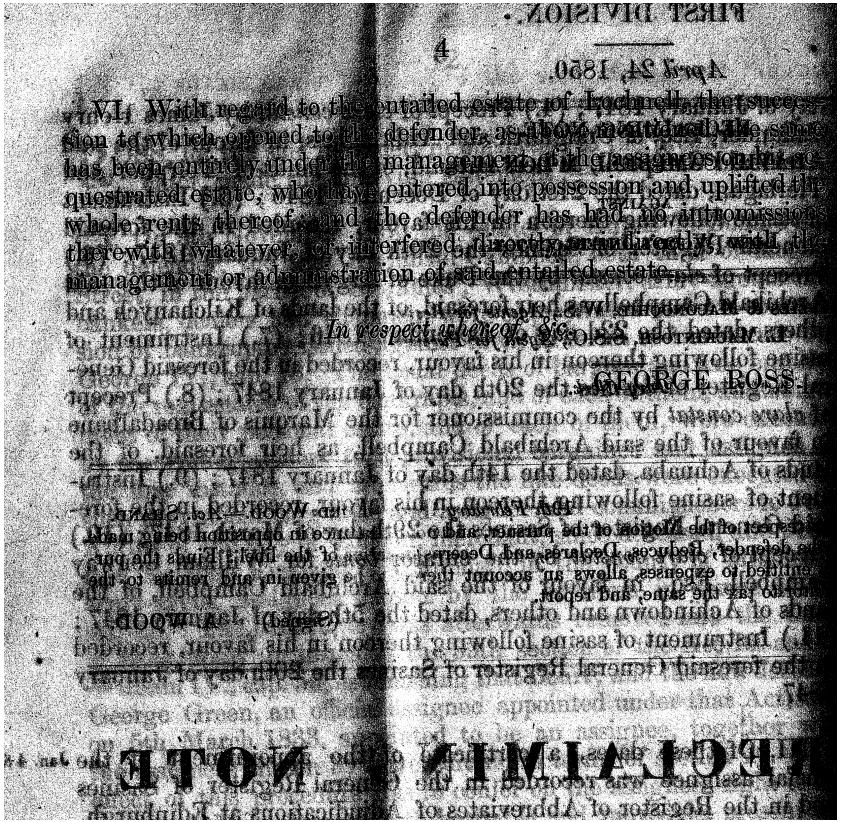


4

VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.



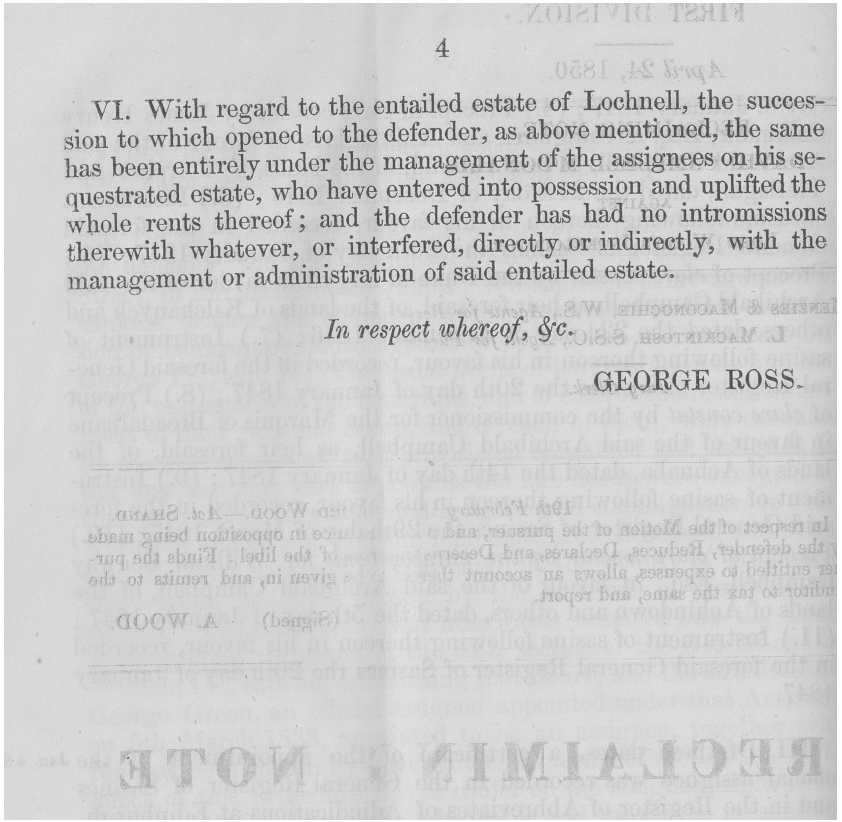
4

VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.

75% Threshold

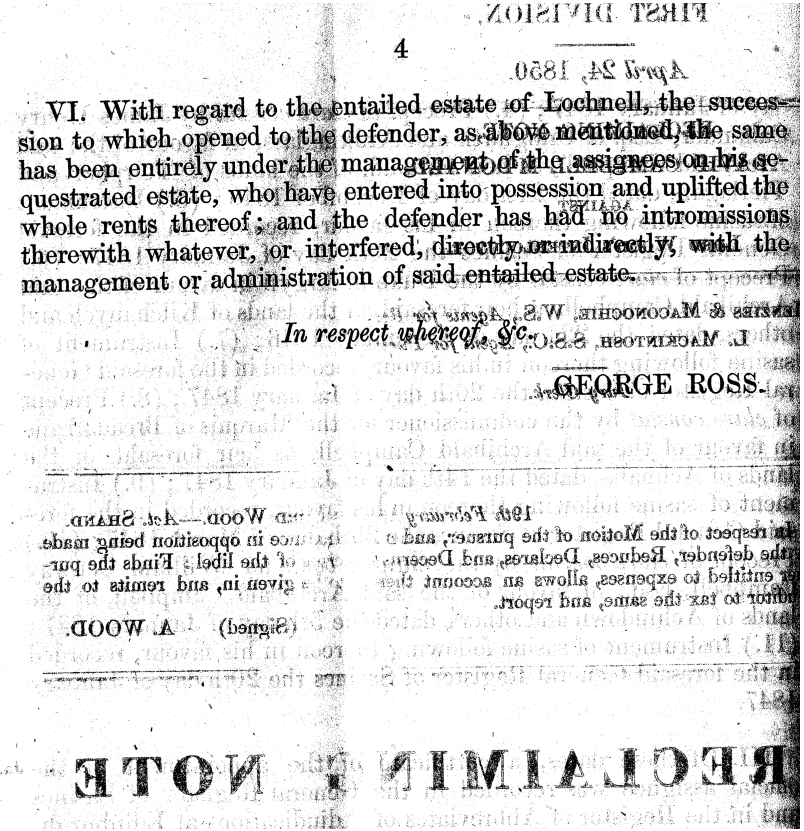


4

VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.



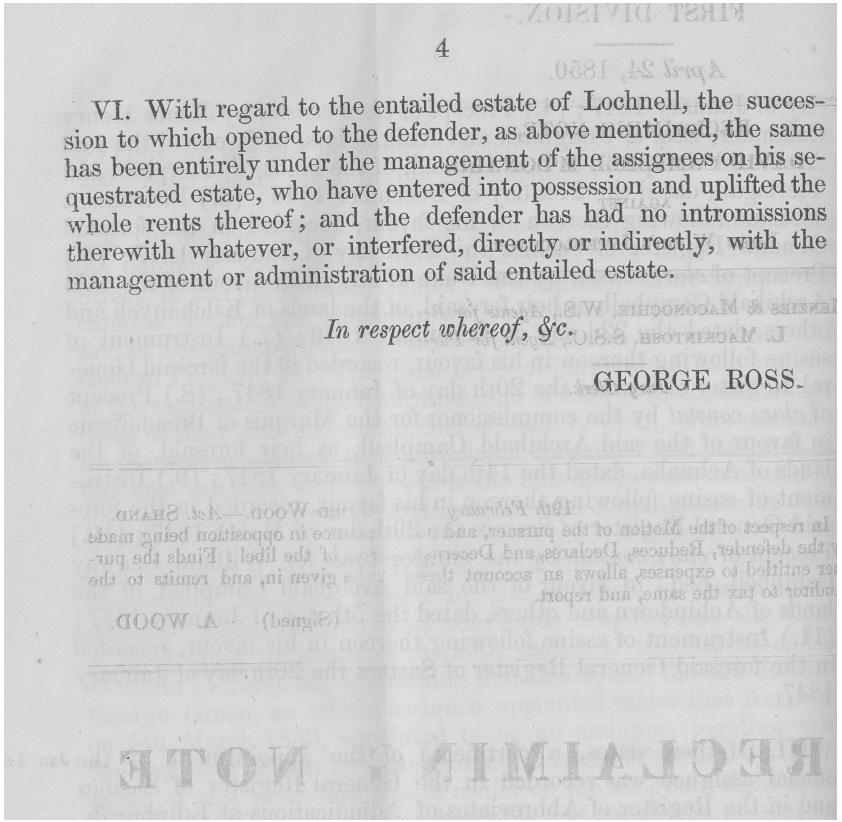
4

VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.

50% Threshold



VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.

4

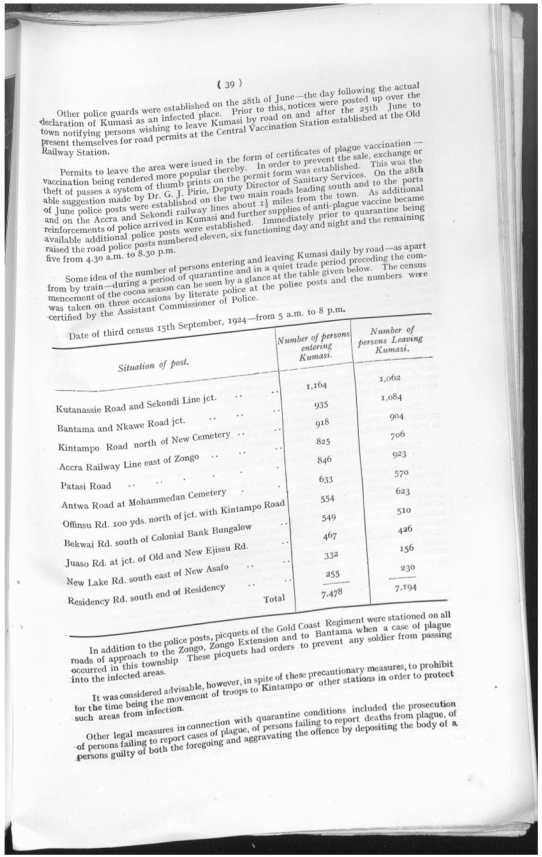
VI. With regard to the entailed estate of Lochnell, the succession to which opened to the defender, as above mentioned, the same has been entirely under the management of the assignees on his sequestrated estate, who have entered into possession and uplifted the whole rents thereof; and the defender has had no intromissions therewith whatever, or interfered, directly or indirectly, with the management or administration of said entailed estate.

In respect whereof, &c.

GEORGE ROSS.

Further steps – page flattening and deskewing

- Large volumes and conservation issues can mean that some volumes produce images with very warped pages.
- Correcting for this makes a significant difference to the quality of the OCR output.
- Algorithm treats page as a series of cylinders (Fu et al. 2007)



(39)

Other police guards were established on the 28th of June—the day following the actual declaration of Kumasi as an infected place. Prior to this, notices were posted up over the town notifying persons wishing to leave Kumasi by road on and after the 25th June to present themselves for road permits at the Central Vaccination Station established at the Old Railway Station.

Permits to leave the area were issued in the form of certificates of plague vaccination — vaccination being rendered more popular thereby. In order to prevent the sale, exchange or theft of passes a system of thumb prints on the permit form was established. This was the able suggestion made by Dr. G. J. Pirie, Deputy Director of Sanitary Services. On the 28th of June police posts were established on the two main roads leading south and to the ports and on the Accra and Sekondi railway lines about 1½ miles from the town. As additional reinforcements of police arrived in Kumasi and further supplies of anti-plague vaccine became available additional police posts were established. Immediately prior to quarantine being raised the road police posts numbered eleven, six functioning day and night and the remaining five from 4.30 a.m. to 8.30 p.m.

Some idea of the number of persons entering and leaving Kumasi daily by road—as apart from by train—during a period of quarantine and in a quiet trade period preceding the commencement of the cocoa season can be seen by a glance at the table given below. The census was taken on three occasions by literate police at the police posts and the numbers were certified by the Assistant Commissioner of Police.

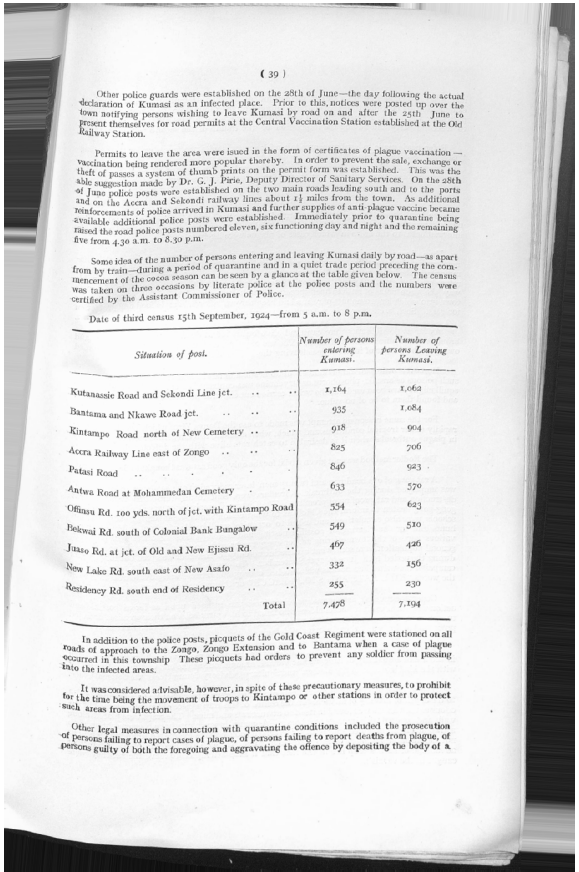
Date of third census 15th September, 1924—from 5 a.m. to 8 p.m.

Situation of post.	*Number of persons entering Kumasi.*	*Number of persons Leaving Kumasi.*
Kutanassie Road and Sekondi Line jct.	1,164	1,062
Bantama and Nkawe Road jct.	935	1,084
Kintampo Road north of New Cemetery	918	904
Accra Railway Line east of Zongo	825	706
Patasi Road	846	923
Antwa Road at Mohammedan Cemetery . .	633	570
Offinsu Rd. 100 yds. north of jct. with Kintampo Road ..	554	623
Bekwai Rd. south of Colonial Bank Bungalow ..	549	510
Juaso Rd. at jct. of Old and New Ejisu Rd. ..	467	426
New Lake Rd. south east of New Asafo	332	156
Residency Rd. south end of Residency	255	230
Total	7,478	7,194

In addition to the police posts, picquets of the Gold Coast Regiment were stationed on all roads of approach to the Zongo, Zongo Extension and to Bantama when a case of plague occurred in this township These picquets had orders to prevent any soldier from passing into the infected areas.

It was considered advisable, however, in spite of these precautionary measures, to prohibit for the time being the movement of troops to Kintampo or other stations in order to protect such areas from infection.

Other legal measures in connection with quarantine conditions included the prosecution of persons failing to report cases of plague, of persons failing to report deaths from plague, of persons guilty of both the foregoing and aggravating the offence by depositing the body of a



(39)

Other police guards were established on the 28th of June—the day following the actual declaration of Kumasi as an infected place. Prior to this, notices were posted up over the town notifying persons wishing to leave Kumasi by road on and after the 25th June to present themselves for road permits at the Central Vaccination Station established at the Old Railway Station.

Permits to leave the area were issued in the form of certificates of plague vaccination — vaccination being rendered more popular thereby. In order to prevent the sale, exchange or theft of passes a system of thumb prints on the permit form was established. This was the able suggestion made by Dr. G. J. Pirie, Deputy Director of Sanitary Services. On the 28th of June police posts were established on the two main roads leading south and to the ports and on the Accra and Sekondi railway lines about 1½ miles from the town. As additional reinforcements of police arrived in Kumasi and further supplies of anti-plague vaccine became available additional police posts were established. Immediately prior to quarantine being raised the road police posts numbered eleven, six functioning day and night and the remaining five from 4.30 a.m. to 8.30 p.m.

Some idea of the number of persons entering and leaving Kumasi daily by road—as apart from by train—during a period of quarantine and in a quiet trade period preceding the commencement of the cocoa season can be seen by a glance at the table given below. The census was taken on three occasions by literate police at the police posts and the numbers were certified by the Assistant Commissioner of Police.

Date of third census 15th September, 1924—from 5 a.m. to 8 p.m.

Situation of post.	*Number of persons entering Kumasi.*	*Number of persons Leaving Kumasi.*
Kutanassie Road and Sekondi Line jct.	1,164	1,062
Bantama and Nkawe Road jct.	935	1,084
Kintampo Road north of New Cemetery	918	904
Accra Railway Line east of Zongo	825	706
Patasi Road	846	923
Antwa Road at Mohammedan Cemetery . .	633	570
Offinsu Rd. 100 yds. north of jct. with Kintampo Road ..	554	623
Bekwai Rd. south of Colonial Bank Bungalow ..	549	510
Juaso Rd. at jct. of Old and New Ejisu Rd. ..	467	426
New Lake Rd. south east of New Asafo	332	156
Residency Rd. south end of Residency	255	230
Total	7,478	7,194

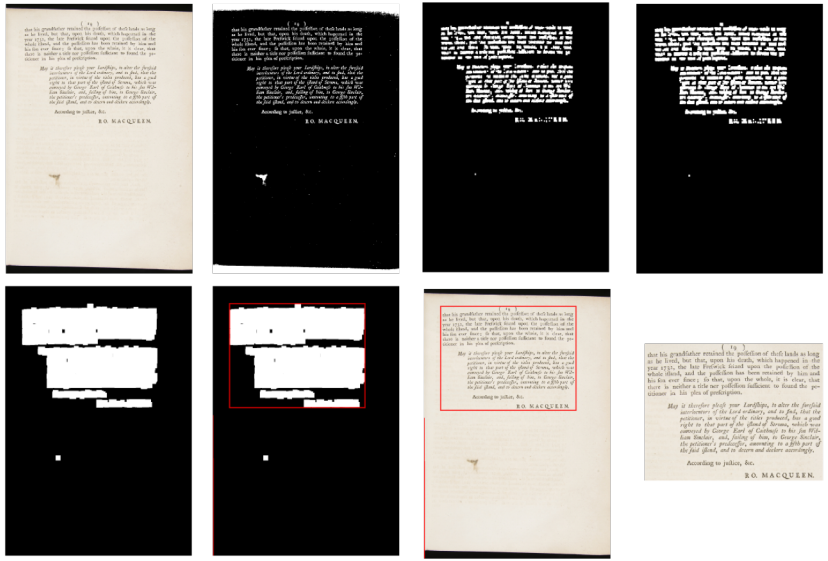
In addition to the police posts, picquets of the Gold Coast Regiment were stationed on all roads of approach to the Zongo, Zongo Extension and to Bantama when a case of plague occurred in this township These picquets had orders to prevent any soldier from passing into the infected areas.

It was considered advisable, however, in spite of these precautionary measures, to prohibit for the time being the movement of troops to Kintampo or other stations in order to protect such areas from infection.

Other legal measures in connection with quarantine conditions included the prosecution of persons failing to report cases of plague, of persons failing to report deaths from plague, of persons guilty of both the foregoing and aggravating the offence by depositing the body of a

Finding text areas in the image

- Another way to improve the quality of the OCR output is to remove extraneous parts of the image, those without any text content
- We do this using a similar chain of simple image processing techniques, to produce a mask of areas of the image containing text



Training the OCR engine

- One major advantage of using an open-source non-proprietary OCR engine such as Tesseract is the ability to have full control over the process by which the engine recognises text within the image
- There are open source datasets for many languages provided both by the Tesseract project and also as outputs from GLAM research projects
- For the Plague.TXT project, we compiled a set of training data geared specifically towards late 19 / early 20 century printing styles
- This included data from the IMPACT Project (http://www.digitisation.eu) as well as typeface specific training data
- More generally, this approach can help with dealing with issues such as variant lettering in older documents (Long-S!!) as well as older typefaces not well recognised by modern-geared software

ABBYY FineReader versus Tesseract

- Approx. half of the corpus is available on Internet Archive as text OCRed using ABBYY FineReader 11.0
- Eyeballing some of the reports showed OCR quality was not that great in places
- Mike’s work using Tesseract lead to output that looked much better quality at least for documents with bad original OCR
- We need to do a formal evaluation using a gold standard to confirm that
- Previous work has found Tesseract to work better for some test sets than others when compared to FineReader (Heliński et al. 2012, *Report on the comparison of Tesseract and ABBYY FineReader OCR engines*)

Available OCR

I have tlie lioiiour to foi'ward herewith foi- your information a Heport upon the Kpideiiiiic of Plague in HoDgkoug in 1894, so far as it concerns tlie medical work which I carried out under your directions. I regret extremely that several important matters — including the e[)idemiology of the disease — which I could have wished to discuss at some length, have been touched upon very superficially, or passed over altogether in this Report. I will ask you to accept as an excuse for my shortcomings in these respects the following facts of which you are, I believe, already cognizant:

Improved OCR

J have the honour to forward herewith for your information a Report upon the Epidemic of Plague in Hongkong in 1894, so far as it concerns the medical work which I carried ont under your directions. I regret extremely that several important matters — including the epidemiology of the disease — which I could have wished to discuss at some length, have been touched upon very superficially, or passed over altogether in this Report [will ask you to accept as an exeuse for my shortcomings in these respects the following facts of which you are, I believe, already cognizant:

Available OCR

That the Litrines are a source of))ropao^atino: the infection as described by Dr. LowsoN there is no doubt, and |)roc)f is afforded by the dates of the closiiiii; of the surrounding liouses. I found on inquiry that during- the end of jMay and the beginning of June, when the jn-evailing winds were from the east and north, the houses to the west and south of the latrines were closed and afterwards, when the prevailing winds were fi'om the south and west, the houses to the north and east of the latrines were closed, being fonnd infected and more than three deaths having occurred in each of them. Mr. RA^[made elaborate phins of the City of Victoria showing where the plague existed, and the proportion of houses in each district that were infected.

Improved OCR

That the latrines are a source of propagating the infection as described by Dr. Lowson there is no doubt, and proof is afforded by the dates of the closing of the surrounding houses. | Found on inquiry that during the end of May «and the beginning of June, when the prevailing winds were from the cast and north, the houses to the west and south of the latrines were closed and afterwards, when the prevailing winds were from the south and west, the houses to the north aud east of the latrines were closed, being found infected and more than three deaths having occurred in each of them. Mr. Ram made elaborate plans of the City of Victoria showing where the plague existed, and the proportion of houses in each district that were infected,

OCR Evaluation

- Current methods of evaluating OCR quality can be crude (“How many of these words can be found in a dictionary?”) and can be easily tripped up by things such as scientific terms, hyphenated words, tables of data, etc
- Development underway at UoE Digital Library on a toolkit for evaluating the quality of OCR renderings that can work for different document types
- Designed to use a series of filters to accept or reject individual tokens with the text allowing easy customisation depending on expected document content

OCR Evaluation

- Each filter focuses on one specific type of token or analysis method
- For example, a filter that finds web addresses
- Useful before a dictionary check where these tokens are likely to be marked as “bad” (“http://” doesn’t appear in many dictionaries!) and incorrectly lower the score
- Chainable filters mean that this check could be used or not depending on the document type. A 19 C Plague report is unlikely to have web addresses, but a scan of a modern thesis might

Data Stats

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).

- 38 reports with up to 5,000 words each, 15 reports with between 5,000 and 10,000 words, 32 documents with between 10,000 and 100K words each and 17 documents with 100K or more words each.
- Over 4.4 million word tokens and over 229K sentences.

Edinburgh Geoparser

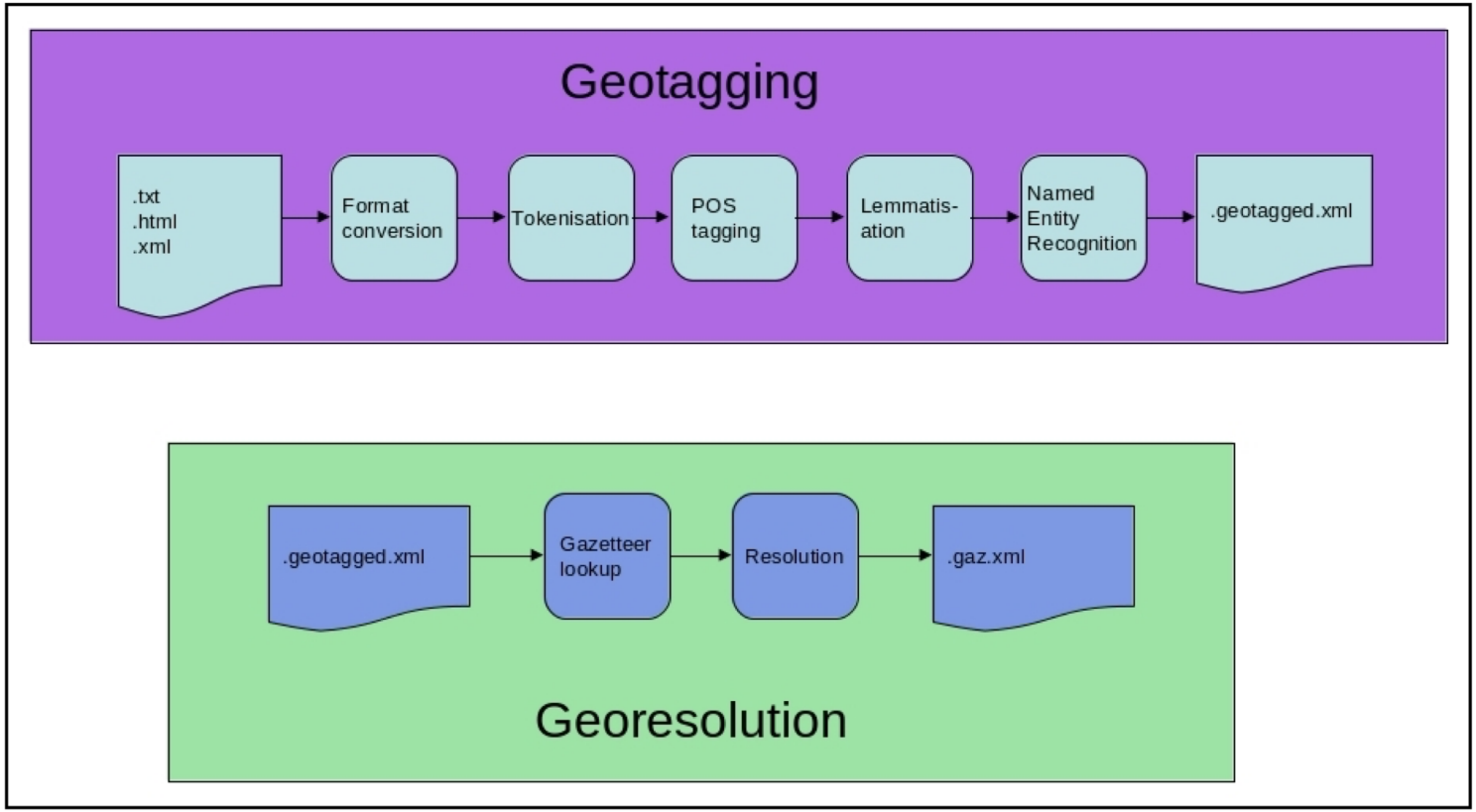
The Edinburgh Geoparser is a language processing tool designed to detect placename references in English text and ground them against an authoritative gazetteer so that they can be plotted on a map. It was developed by researchers at the Language Technology Group at the School of Informatics at the University of Edinburgh. It has been applied in a number of research projects, for example to geo-locate literature set in Edinburgh or to geo-reference historical documents on commodity trade in the 19th century.

Click on a lat/long to centre the map there.

Edinburgh 55.952,-3.196 55.823,-3.093 55.950,-3.193 -37.068,

Edinburgh Geoparser

Grover et al. 2010, Alex et al. 2015

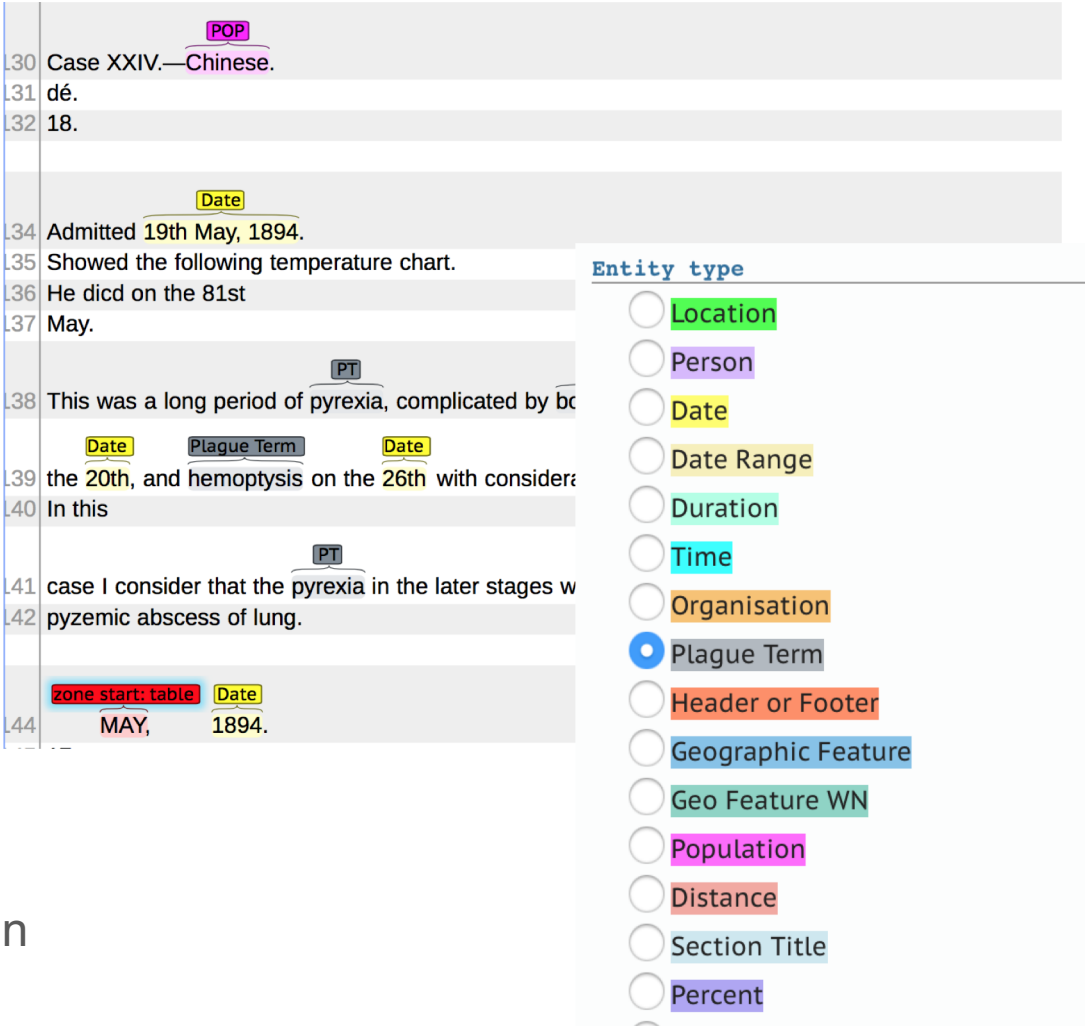


Entity Annotations

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

Information Extraction

- Adapted the Edinburgh Geoparser to this data
- Applied post-OCR text correction (soft-hyphen deletion)
- Extracted named entities, dates, date ranges, duration and time expression, geographic features, population, and plague related terminology etc.
- Output contains errors but it can assist with the analysis



Annotation

- We conducted an annotation pilot with feedback to improve the information extraction output and decided on additional manual annotation of zones in the text
- Derived a lexicon of plague related terms
- Decided on a way to correct OCR errors in entities manually
- Used the Brat annotation tool

Document Zone Annotations

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

Visualisation of Zones

disease-history

statistics

measures

conclusion

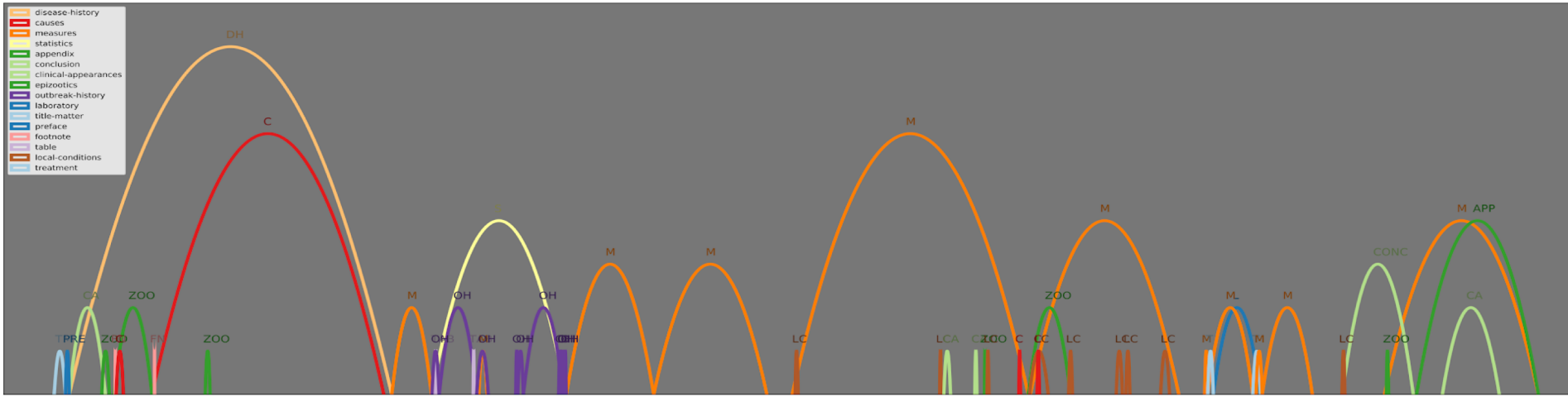
causes

outbreak-history

local-conditions

table

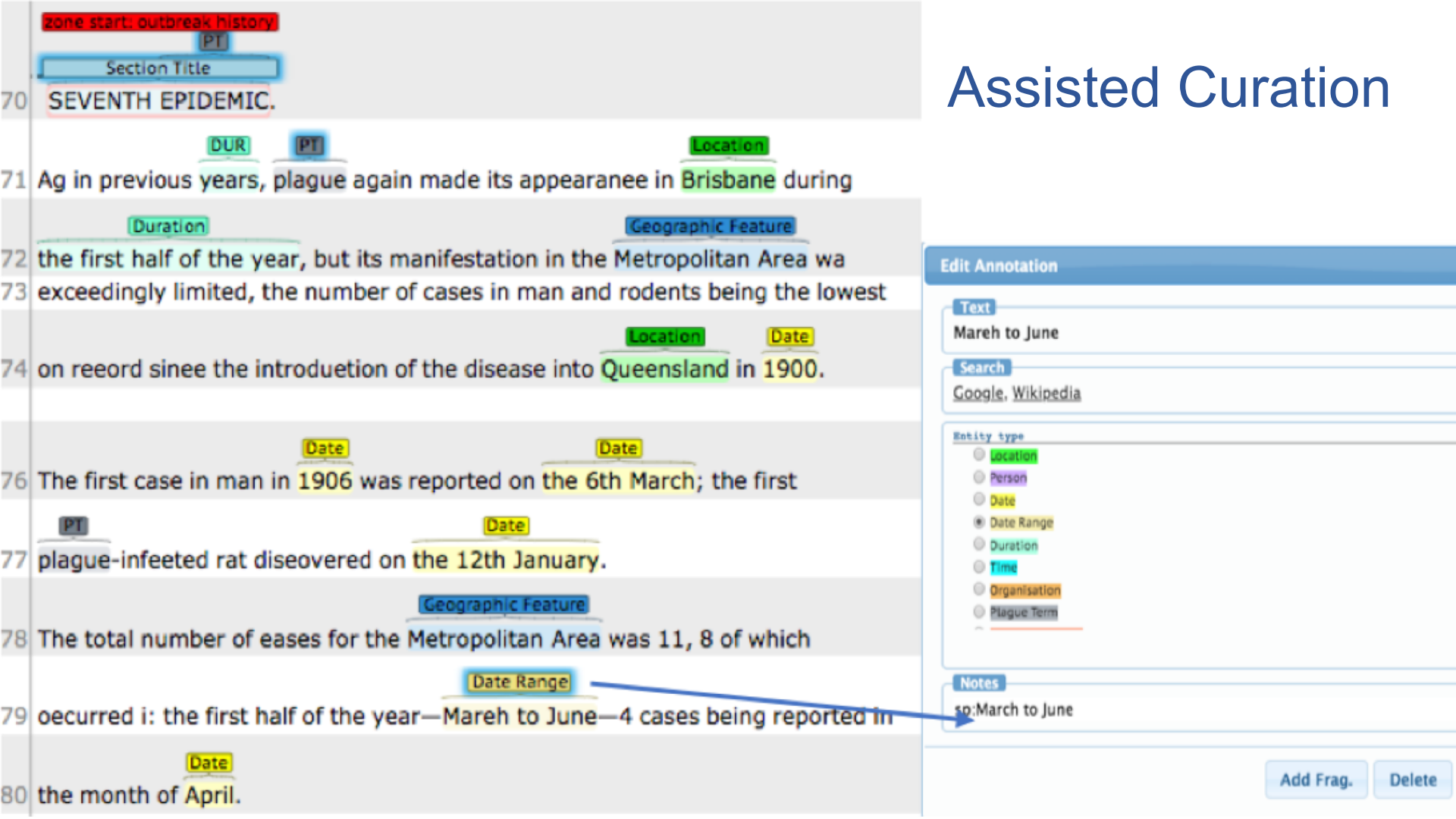
clinical-appearances



Pilot

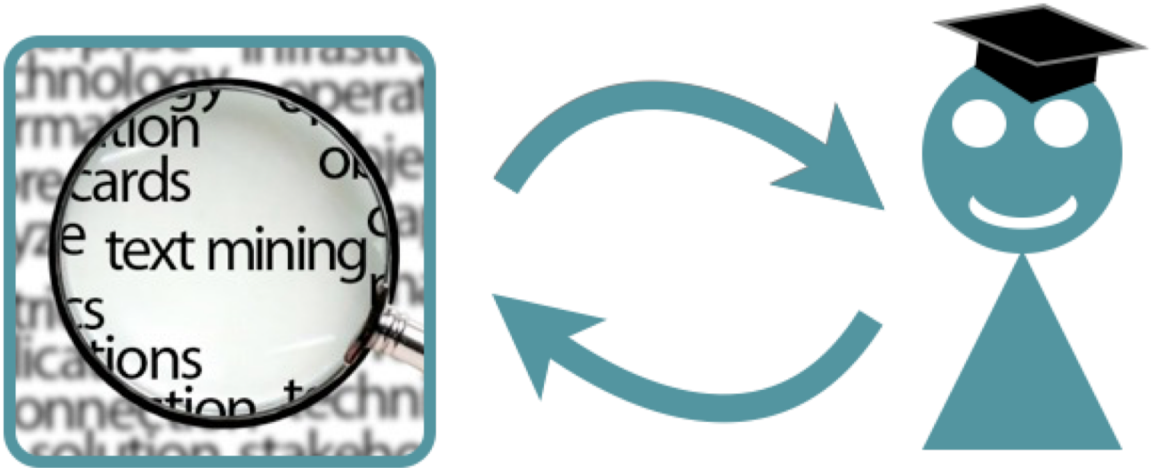
- We conducted a one-week long annotation sprint with three annotators.
- Annotation is still ongoing to mark up as many of the reports in the collection.
- Annotation involves correction of the information extraction output, targeted spelling correction and zone annotation.

Assisted Curation



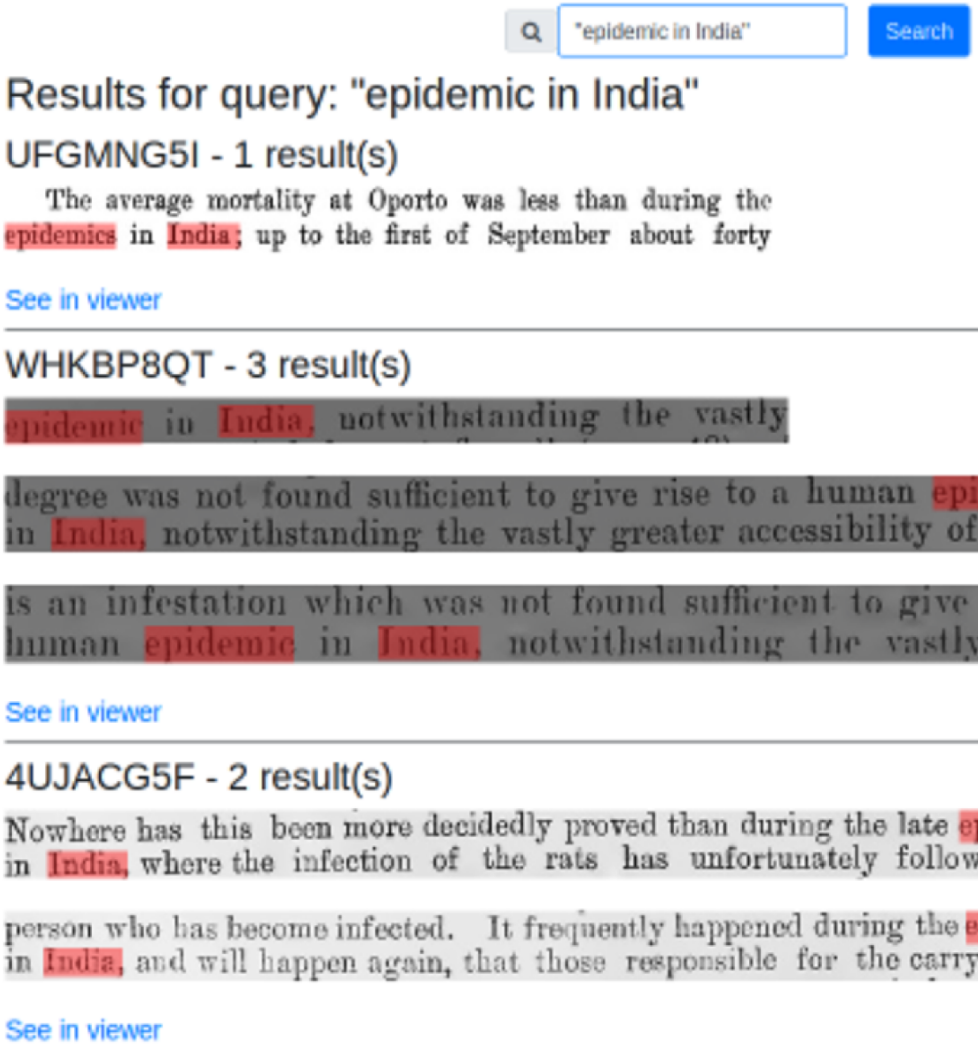
Development of Text Mining Methods

- Iterative development (meetings, prototyping, interviews, manual annotation, continuous feedback).



Use case 1: Search

- Using Solr
- Full text search across reports
- Filtering within annotated zones (e.g Search only within outbreak reports)
- Word stemming (e.g. a search for “vomiting” will also find past-tense “vomited”)
- Results served via custom IIIF Search service



Use case 2: Topic Modelling

- Combination of automatic & manual annotation as well as topic modelling allows for analysis of this data in interesting ways.
- Topics from zones of type **cause** by time period (earlier versus later reports).

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: Topics from cause zones by time period

Use case 3: Corpus Analysis

adjective + man|men		adjective + woman|women	
count	adjective	count	adjective
316	medical	21	old
27	young	12	married
22	sick	10	pregnant
19	old	9	young
13	medieal	8	chinese
9	poor	7	purdah
8	influential	5	native
7	other	4	other
7	infected	3	parturient
7	healthy	3	dead
6	intelligent	2	well-nourished
6	few	2	unfortunate
5	well-nourished	2	indian
5	twelve	2	few
5	scientific	1	weakly
4	white	1	sick
4	trained	1	several
4	several	1	russian
4	muscular	1	respectable
4	great	1	purdak

Table 5: Most frequent adjectives followed by the nouns *man* or *men* versus *woman* or *women*.

Summary and Next Steps

- Plague.TXT pilot: promising results, mostly required to prepare and mark-up the collection of plague outbreak reports
- Evaluation of OCR improvements and information extraction not feasible in pilot, and the epidemiological analysis remains to be done
- Annotation is still ongoing, once completed we will share any data we can share
- Looking to expand this work into a larger project
- Ideas for future work: Zone-specific analysis, automatic zone detection, OCR evaluation and improvement, spelling normalisation, epistemic network analysis, linking to newspaper, data visualisations

Thank you!

Beatrice Alex, Chancellor's Fellow, University of Edinburgh
Edinburgh Language Technology Group, @bea_alex

Mike Bennett, University of Edinburgh Library

Plague.TXT: https://www.ltg.ed.ac.uk/projects/plague-txt/
GitHub: https://github.com/Edinburgh-LTG/PlagueDotTxt
Edinburgh Geoparser: https://www.ltg.ed.ac.uk/software/geoparser/