

# Scoping the Medical Officer of Health Reports for Statistical Data Recovery

## Project Report

Justin Hayes  
Christian Clausner  
Apostolos Antonacopoulos

---

### Overview

This report provides details of the work carried out and the corresponding findings of a scoping study to explore the technical potential and user requirements for recovery and integration of operable statistical data from information contained in digitised tables captured via OCR from approximately 65,000 Medical Officer of Health Reports (MOH Reports) covering the UK from 1848 to 1978 (see <https://wellcomelibrary.org/moh/>).

The main components of the study were:

- A [high-level scoping exercise](#)
- A [review of current work and best practice](#)
- An [online user survey, followed up by an informal meeting with researchers](#)
- A [SWOT analysis](#)
- An [alternative OCR and templating exercise](#)
- A [data recovery and integration exercise](#)
- A [recommendation on the Minimum Viable Product for proposals for further work](#)

### 1 High-level scoping exercise

A high-level scoping exercise was carried out to explore Exploration and characterise the tables and statistical data present within the MOH Reports.

XML files containing structured text versions of individual tables from MOH reports currently available for London from the Wellcome website at <https://wellcomelibrary.org/moh/about-the-reports/using-the-report-data/> identify most, but not all, of the tables (approximately 270,000 in all) contained in the reports. Table text is differentiated and structured into captions, column headers and row text (from which row headers can be extracted, usually in the first or second columns). This allows these components of the tables to be analysed separately.

The illustration below shows an example of a table containing information on Causes of Death by Age from Shoreditch for 1955.

- **Caption** (dedicated descriptive table caption)
- **Column headers**
- **Row headers**

CAUSES OF DEATH, WITH AGE DISTRIBUTION

CAUSES OF DEATH	A G E S								S E X		Total deaths due to each cause.
	Under 1	1 to 5	5 to 15	15 to 25	25 to 45	45 to 65	65 to 75	Over 75	Males	Females	
1. Tuberculosis, respiratory	1	1	1	1	1	1	1	1	11	3	14
2. Tuberculosis, other forms	1	1	1	1	1	1	1	1	1	1	1
3. Syphilitic disease	1	1	1	1	1	1	1	1	1	1	1
4. Diphtheria	1	1	1	1	1	1	1	1	1	1	1
5. Whooping cough	1	1	1	1	1	1	1	1	1	1	1
6. Meningococcal infections	1	1	1	1	1	1	1	1	1	1	1
7. Acute poliomyelitis	1	1	1	1	1	1	1	1	1	1	1

Selected rows from the corresponding XML file are shown below.

```

<caption>
  <p>CAUSES OF DEATH, WITH AGE DISTRIBUTION.</p>
</caption>
...
<thead>
  <tr>
    <th rowspan="2">CAUSES OF DEATH</th>
    <th colspan="8">AGES</th>
    <th colspan="2">SEX</th>
    <th rowspan="2">Total deaths due to each cause.</th>
  </tr>
  <tr>
    <th>Under 1</th>
    <th>1 to 5</th>
    <th>5 to 15</th>
    <th>15 to 25</th>
    <th>25 to 45</th>
    <th>45 to 65</th>
    <th>65 to 75</th>
    <th>Over 75</th>
    <th>Males</th>
    <th>Females</th>
  </tr>
</thead>
...
<tbody>
  <tr>
    <td>1. Tuberculosis, respiratory</td>
    <td>-</td>
    <td>-</td>
    <td>-</td>
    <td>-</td>
    <td>1</td>
    <td>6</td>
    <td>7</td>
    <td>-</td>
    <td>11</td>
    <td>3</td>
    <td>14</td>
  </tr>
  ...

```

A MySQL database was created, and data from the XML files for all tables were aggregated and imported into a series of database tables to enable global operations to be carried out on it.

## 1.1 Table Clustering and Textual Analysis

Following initial investigations, it was agreed with Wellcome to use information from the XML files created by Wellcome to perform automated textual analysis across all tables in the reports, rather than a target a small subset of tables for intensive manual analysis as originally planned,

as this would provide more comprehensive scoping of the complete set of tables across all reports.

Textual similarity analyses were performed across texts from captions, column headers and row headers from all tables to identify similar patterns of text. Multiple analyses were run with parameters modified to allow varying degrees of ‘fuzziness’ in comparison by ignoring minor textual differences due to differences in capitalisation, spellings, etc. The example below shows some examples of similar text fragments grouped in this way.

“1. INSPECTIONS FOR PURPOSES OF PROVISIONS AS TO HEALTH. Including inspections made by Sanitary Inspectors.”  
 “1.—Inspections for purposes of provisions as to health Including inspections made by Sanitary Inspectors.”  
 “1. INSPECTION for purposes of provisions as to health (including inspections made by Sanitary Inspectors).”  
 “1. - Inspections for purposes of provisions as to health (including inspections made by Sanitary Inspectors).”  
 “1. I nspections for purposes of provisions as to health (including inspections made by Sanitary Inspectors).”  
 “1.Inspectios for purpses of provisions as to heaths(including inspections made by Sanitary Inspectors):-”

Groups (clusters) of tables with strong textual similarities across combinations of captions, column headers and row headers were identified. The following are examples of such groupings:

### **Example 1: Analysis of table captions with common column headers**

Below is a list of the number of tables (frequency within the complete dataset) identified with similar captions and the following common column headers:

*CAUSE OF DEATH.| Under 1 week.| 1 to 2 weeks.| 2 to 3 weeks.| 3 to 4 weeks.| Total under 4 weeks....(continues)*

<u>Frequency</u>	<u>Table Caption</u>
400	Net Deaths from stated causes at various Ages under One Year of Age
250	Net Deaths from stated Causes in Weeks and Months under One Year of Age
90	Net Deaths from stated causes at various Ages under 1 Year of Age

### **Example 2: Analysis of column headers with common caption and common row headers**

Number of tables (frequency within the complete dataset) identified with given column headers and the following common captions:

*Causes of Death as given by the Registrar-General*

and the following common column headers:

*Tuberculosis (Respiratory)| 2. Other forms of Tuberculosis| —| 3. Syphilitic Disease...(continues)*

<u>Frequency</u>	<u>Table Column Header</u>
350	Causes of Death  Males  Females  Total

## **1.2 Ontology**

From results of the textual similarity analyses of captions, column headers and row headers described above, an initial ontology of broad topic categories was developed as a tool to enable classification and characterisation of the information content of tables with the most commonly occurring groupings.

The ontology has the following structure:

**Demographics**  
 Age

Sex
Births
Deaths
Causes of death
Infant death
<b>Ailments</b>
Diseases
Infectious diseases
Notifiable diseases
Immunisations
<b>Environmental</b>
Inspections
Food
Conditions
Meteorological
<b>Financial</b>
<b>Legal</b>

## 1.3 Findings

### 1.3.1 Key table topics

The ontology described above covers the most frequently occurring tables. There are also several tables (not found on all reports) containing information on less common topics.

The table below is an overview of the results of topic identification within tables. Topics are derived from either the table caption, from the column headers, or from the row headers. Each proposed topic is related to an approximated count of individual tables across all reports in the dataset (containing a total of 270,000 individual tables). Due to the nature and brevity of this high-level scoping exercise the counts are likely to be incomplete, but they provide an initial overview of likely numbers and proportions.

From caption	Table Count (approx.)	From column headers	Table Count (approx.)	From row headers	Table count (approx.)
Mortality / Cause of Death	2530	Financial	3670	Inspections	4010
General statistics / demographics	1900	Inspections	4670	Infectious diseases	1450
Infectious Diseases / Notifiable Diseases	1720	Food	580	Cause of death	730
Inspections / conditions	4360	Infant deaths	1280	Births	710
Minor ailments, dental, etc.	710	Cause of death	1060	Meteorological	2810
Financial	470	Diseases / conditions of children	430		
Food	330	Infectious Diseases	750		
Births	240	General / vital statistics	1520		

Meteorological	100				
Legal	190				
Immunisation	60				

### 1.3.2 Geographic areas generally reported on

As would be expected, most tables report information for district areas, but many tables also report for areas from smaller geographies such as sub-districts and wards. Some tables also contain information for larger geographies in order to draw regional and national comparisons.

### 1.3.3 Typical table/data structures

There is considerable variety of information content, and of physical structure in the arrangement of captions, columns and rows in tables across the reports. However, there is a great deal of similarity in the structures of certain tables with more commonly occurring combinations of variables (Cause of Death by Age and Sex, for example) which use standardised, externally defined classifications.

### 1.3.4 Potential for QA based on duplication within data

Many of the tables with more commonly occurring combinations of variables (Cause of Death by Age and Sex, for example) contain row and column totals. These can be used to validate values (extracted through OCR/manual correction) for individual categories across rows and down columns through automated summation and comparison with the column and row totals within a database. Values that fail these comparisons would be candidates for correction, which could be accomplished using a crowd-sourced approach similar to the one developed by the authors and currently in operation to correct errors in numerical data within the various tables in the 1961 Census Small Area Statistics (England and Wales) extracted from scanned pages supplied by the Office for National Statistics.

### 1.3.5 Level of consistency across locations and over time

There is wide variety in the information content and structures of tables across locations and time. However, there are tables with some combinations of variables (Cause of Death by Age and Sex, for example) that are present with a good level of consistency in structures across many locations and many years.

## 1.4 Conclusions

Results from the high-level scoping exercise suggest that it is possible to identify broad categories of similar information content across tables in the MOH reports using automated textual analysis, and to use them to develop a descriptive ontology to classify and characterise combinations of information content for a significant proportion of the tables.

Further work could extend and develop the descriptive ontology and create a searchable index to the existing tables in the MOH reports that could operate alongside the existing web access to enable users to quickly find pages containing tables with particular combinations of information.

## 2 Current Practices in Digitisation and Representation of Historical Numerical Data

A review of current work and best practice in the field of historical data recovery was carried out to identify reported examples of research and/or operational work with similar aims (bulk

recovery of information from multiple historical sources and integration as globally operable digital data), and summarise methods, subject materials and outputs.

A search through current publicly available information and a consultation with a small number of key stakeholders (researchers and content holders) were carried out to establish different types of digitisation and current approaches to making the digitised numerical data (in different forms) available for consumption/re-use. While this was not a large-scale nor exhaustive search, it is realistically indicative of the current state of the art.

The value of numerical information is widely recognised, as is also the realisation that current digitisation pipelines and OCR do not deal with this information in a satisfactory way. Numerical information is present mostly in tabular form and, while OCR can very often recognise individual numbers/digits quite well, this is far from sufficient. This is for two reasons:

1. Numerical information needs to be absolutely accurate to the original – OCR errors are not tolerated by users in the same way as the occasional error in text is.
2. Current OCR makes several errors with respect to preserving page layout, especially spacing in tables. Errors in table cell spacing are detrimental to the semantics of the number content of cells, i.e. suddenly a numerical entry may wrongly refer to a different table column altogether.

At the very lowest end of provision, users can simply view numerical tables (in image form) contained on pages within **scanned volumes which are indexed at the item level** with the usual library record metadata – examples are the Public Health Reports from the US National Library of Medicine [1] and various medical etc. reports at the Hathi Trust Digital Library [2]. Since the content is not searchable, users search for relevant items(s) and browse through the images of each item, page by page, to identify the tables they are interested in. Page images can be downloaded.

The first level of improvement comes with **indexing the individual scanned pages containing tables** with keywords extracted from the table headers – through OCR (corrected or not) or manually keyed. An example is the historical statistics available from Statistics Sweden – the national statistics organisation of Sweden [3], where the numerical content of the tables is not made available online due to the unreliability of the OCR results obtained. A report on the corresponding large-scale digitisation effort [4] offers more details on the difficulties of OCRing historical tables. Another example is the Online Historical Population Reports (histpop) [5] maintained by the UK Data Service. In addition to item-level indexing (whole reports/books), raw text OCR results are available for some of the scanned pages and used for indexing and searching. Some of the items available at the Hathi Trust Digital Library [2] also have associated raw (uncorrected) OCR results which afford some usefulness with keyword search.

The next level of making numerical information available is through extracting the data (manually corrected OCR or manual entry) and **re-creating the visual representation of tables in an encoded form**. The numerical information of each individual table can thus be copied by the user and pasted onto a spreadsheet etc. for further analysis. Examples of this, in the simplest form with tables represented in HTML, are the Digitised Collections of Statistics New Zealand [6] and the results of the German project Digital Reich Statistics (for some of the tables) [7]. The Medical Officer of Health (MOH) reports at the Wellcome Library [8] are the best example in this category, with individual tables recognised, corrected, indexed and visually reconstructed for viewing on the web, and the corresponding numerical information available for download in a variety of formats.

Finally, in the ultimate form that is most useful to researchers, numerical information is transcribed from tables, cross-referenced and standardised in a purpose-made database and is fully searchable (faceted search). It should be noted that the resources required (funding and time of field experts) to achieve this are very significant and, correspondingly, information is not currently available at large scale. The prime example in this category is Project Tycho® at the University of Pittsburgh [9] where completely transcribed data on disease counts from a variety of primary sources (scanned reports such as those from [1] and [2]) are available and visually presented as graphs.

## 3 User Consultation

### 3.1 User Responses

A focussed online survey ([see appendix A for full survey questions and results](#)) was created containing a set of questions developed in collaboration with Wellcome, and circulated to a select group of people identified from suggestions from Wellcome, as well as from other contacts with interests in using information from the MOH reports. In addition, an informal meeting was held with researchers from the Centre for the History of Science, Technology and Medicine at the University of Manchester to discuss the survey results and any wider perspectives.

In total, 15 responses to the survey were received. The information collected showed that respondents had a mixture of levels of awareness of the MOH reports. Those who had already used the reports had mainly done so for a variety of academic research purposes. Respondents were keen to make use of information on various topics contained in the reports, particularly information relating to

- Basic demographics,
- Mortality and cause of death,
- Ailments, and
- Fertility.

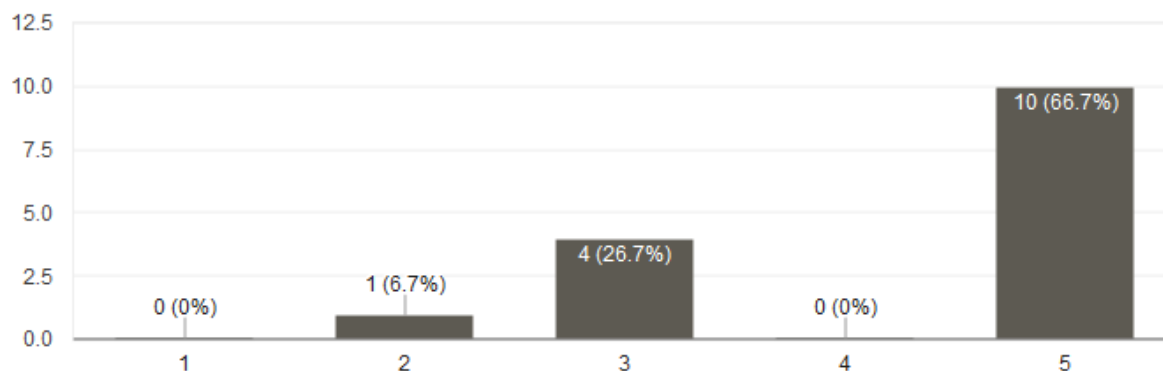
Respondents generally found the current access functionalities very useful, and made positive comments about them, along with some suggestions for improvements (shared with Wellcome).

The following excerpts from the survey results illustrate responses to the questions that were specific to the potential for retrieval and integration of quantitative data from tables:

**(Question 6) Are you interested in using quantitative data from MOH report tables?**

67% scored 5 out of 5 on a scale of 1 (no interest) to 5 (very interested).

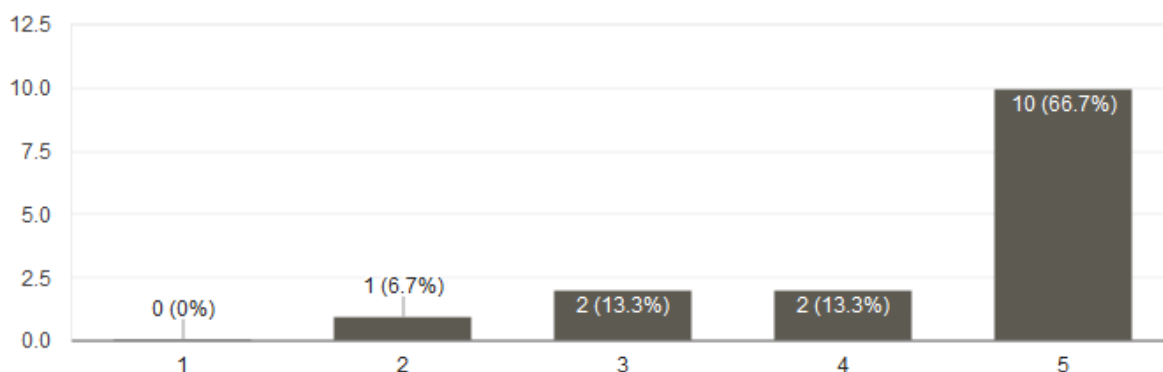
93% scored 3 out of 5 or above on the same scale.



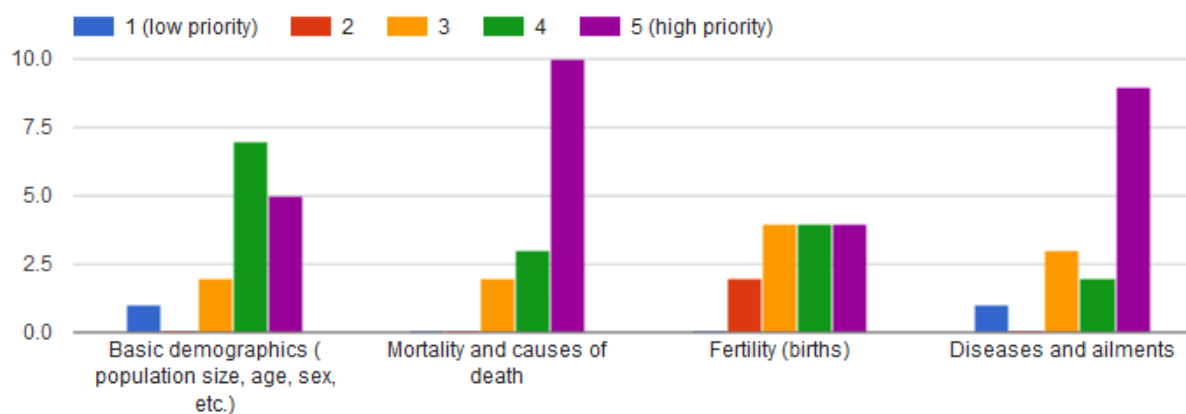
**(Question 7) Would having integrated data make this easier for you?**

67% scored 5 out of 5 on a scale of 1 (no easier) to 5 (much easier).

93% scored 3 out of 5 or above on the same scale.



**(Question 8) Which of the following topics would you most like to have integrated quantitative data from MOH report tables about?**





As mentioned earlier in this section, an informal meeting was also held with staff from the Centre for the History of Science, Technology and Medicine at the University of Manchester. In addition to confirming the results of the survey, participants suggested that there would be a wide range of different audiences from different disciplines with different interests in using data retrieved from the MOH tables (eg historians of various kinds, geographers, demographers, medical researchers). Interests would vary from those interested in finding and using individual tables containing data of interest in the context of their containing report to researchers (eg epidemiologists) more interested in comparative analyses of large subsets of data spanning many years and/or different areas.

## 3.2 Findings

### 3.2.1 Who are the audiences?

There are likely to be a variety of audiences from different professional research disciplines including historians and researchers in medicine, epidemiology, public health, sociology, geography, economics, etc., as well as individual researchers such as genealogists who would be interested in using different parts of the data retrieved from MOH tables in different ways due to the wide range of information the tables contain.

### 3.2.2 What data are they after (geographical, linguistic, terms, etc.)?

This will vary according to the nature of the research. Integration of retrieved values to an external data model will be important to provide a framework for objective comparison of values from tables from different years and locations.

### 3.2.3 How deep/broad (broad brush, specific tables, whole tables, individual values)

Again, this will vary according to the nature of the research. Some researchers (eg historians) are likely to be interested in finding and using individual tables in the context of the corresponding MOH reports, and will want to be able to view relevant report pages. Other researchers (eg epidemiologists) will be interested in using integrated and structured data about particular topics retrieved from all of the MOH tables for large scale comparative analyses. They will be more interested in the quality and descriptive accuracy of the values as a dataset, and aren't likely to want to view individual report pages.

### 3.2.4 How do they want to discover it (to get to the above)

Multiple routes of access will be important. Some users will want to explore the report pages and tables and, once they have found particular reports or tables of interest, may then also want to explore integrated data with similar characteristics. Other users will want to approach from exploration and query of an integrated dataset of retrieved values, and may (or may not) then also want to view corresponding report pages.

### 3.2.5 How do people want their data served (formats, standards, APIs, download functionality)

A variety of access mechanisms will be important to meet the requirements of different users. Users will generally want to discover and obtain information as quickly and easily as possible in relation to the complexity of their requirements. Evidence from other data dissemination applications suggests that the majority of users will want simple file downloads in generic formats once they have found information of interest. Online analysis and visualisations are also of growing interest. APIs will appeal to a much smaller audience of more technical users, but can enable the data to reach much wider audiences through the development of secondary dissemination applications.

## 4 SWOT Analysis

A SWOT analysis informed by outcomes from the high-level scoping exercise and the user survey was carried out between the project team and Wellcome Library staff to explore and assess the suitability of tabular statistical data already captured from the MOH Reports for integration and conversion to forms that will meet the requirements of potential users, and the value that would be provided by doing so.

SWOT analysis of current position in relation to Wellcome MOH resource.

STRENGTHS	WEAKNESSES
<ul style="list-style-type: none"><li>○ Digitised MOH material is a unique and information-rich resource with very significant potential value.</li><li>○ Scanning and OCR have greatly improved discovery and usability of the qualitative, narrative component of the MOH information.</li><li>○ Numerical information available as individual tables in several formats - equal to state-of-the-art.</li><li>○ Good user base.</li><li>○ MOH resource supports Wellcome Library's great reputation for information availability and innovation.</li></ul>	<ul style="list-style-type: none"><li>○ Users find it cumbersome to locate the numerical tables they need within the MOH reports.</li><li>○ Users find it cumbersome to use (collate data across time and geographies) the downloaded individual numerical tables.</li><li>○ Most users are researchers who are willing to invest significant time and effort to find and extract the numerical information. It seems that non-researchers are not readily engaging with the collection.</li><li>○ Information is at the moment available mostly for London only.</li></ul>

OPPORTUNITIES	THREATS
<ul style="list-style-type: none"> <li>○ Very useful content (remaining MOH reports) can be digitised and made available.</li> <li>○ New processing pipelines may deliver more/better information for the same or less cost. Improvements can be introduced in stages with different resource implications.</li> <li>○ Significant value for different user segments can be added to existing collection by making available numerical information in more accessible and usable forms.</li> <li>○ User base can be expanded by attracting individuals and organisations (additional market segments) requiring aggregated and higher-level numerical information from the MOH reports.</li> </ul>	<ul style="list-style-type: none"> <li>○ Other content holding institutions developing and deploying more advanced ways of delivering numerical information. Most institutions are expressing their significant interest to make better use of this type of data.</li> <li>○ Third party organisations may re-use the currently available information in the Wellcome library and produce a better (in terms of accessibility and usefulness) resource attracting the current user base away.</li> <li>○ Funding (internal and/or external) may not be available to realise all proposed stages of improving the accessibility and usefulness of numerical information.</li> </ul>

## 5 Alternative OCR and Templating Exercise

An alternative OCR and templating exercise was carried out, from the original printed texts of a sample of MOH Reports identified through the high-level scoping exercise, employing innovative techniques developed by the project team during work to extract numerical tabular data from the scans of the 1961 Census of England and Wales supplied by the Office for National Statistics. The goal of the exercise was to provide a measure of the quality of the existing tabular statistical data capture from the reports, and an indication of whether the quality could be improved if it was found to be a limiting factor in data recovery from the tables of statistical data already captured from MOH Reports by the Wellcome Library.

The existing XML files were found to have been produced from manual correction of ALTO output from OCR. The manual correction would have required a lot of resource. If a way can be found to automate the correction process it will save a lot of time and effort.

### 5.1 Finding images containing specific tables

The high-level scoping exercise identified common variables and categories present in the MOH report tables, and common combinations in multivariate tables. By examining multiple instances of tables containing equivalent information described by categories belonging to the same variables, it is possible to construct table ‘fingerprints’ containing sets of key words most often associated with particular combinations of variables, and especially words that are unique to particular variables. These fingerprints can be used to search for tables in unstructured text directly produced from OCR.

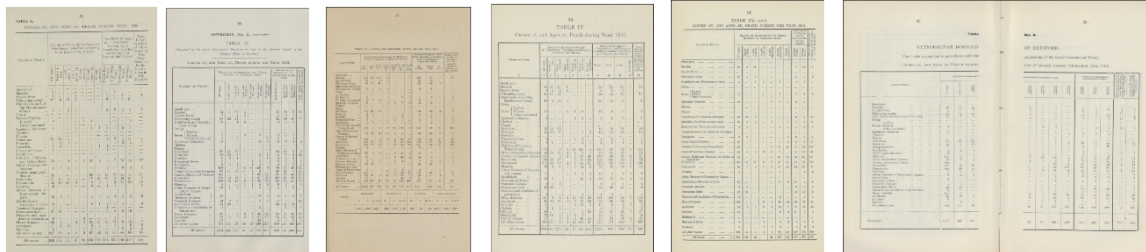
b18106316_13:	CAUSES OF, AND AGES AT, DEATH DURING THE YEAR 1910.
b18106651_103:	Causes of, and Ages at, Death during Year, 1910.
b18106857_101:	Causes of, and Ages at, Death during the Year 1910.
b18111099_37:	Causes of, and Ages at Death during the 52 weeks ending December 31 <sup>st</sup> , 1910.
b18111701_49:	Table IV,—Births and Corrected Deaths for the Year 1910.
b19783450_70:	CAUSES OF. AND AGES AT, DEATH DURING YEAR, 1910.

Causes of Death. Cause of Death.	Deaths in or belonging to whole District at subjoined Ages. Deaths in, or belonging to, Whole District at Subjoined Ages. Deaths at the subjoined ages of "Residents" whether occurring in or beyond the District.							(Localities)	Total Deaths in Public Institutions in the District. Total Deaths in Public Institutions in the District Total Deaths whether of Residents or Non-residents in Public Institutions in the District. Total Deaths whether of "Residents" or "Non-residents" in Public Institutions in the District. (similar)
	All Ages	Under 1 Year	1 and under 5	5 and under 15	15 and under 25	25 and under 65	65 and upwards	...	
1	2	3	4	5	6	7	8		13
Small-pox									
Measles									
Scarlet Fever									
Whooping-cough									
Diphtheria and Membranous Croup Diphtheria and Membranous Croup Diphtheria (including Membranous croup)									
Croup									
Fever (not otherwise specified)	Typhus								
	Enteric								
	Other continued								
Epidemic Influenza									
...									
All causes									

Common text fragments identified in tables from different districts

### 5.1.1 Example: Cause of Death Table

Six pages containing equivalent (but laid out differently) information on Cause of Death by Age from different London districts were identified (all for 1910). The table text fingerprint was then created by comparing the table headers and finding common words (and excluding words that appear only in some of the six tables).



[Acton](#)

[Battersea](#)

[Bermondsey](#)

[Bethnal Green](#)

[Chelsea](#)

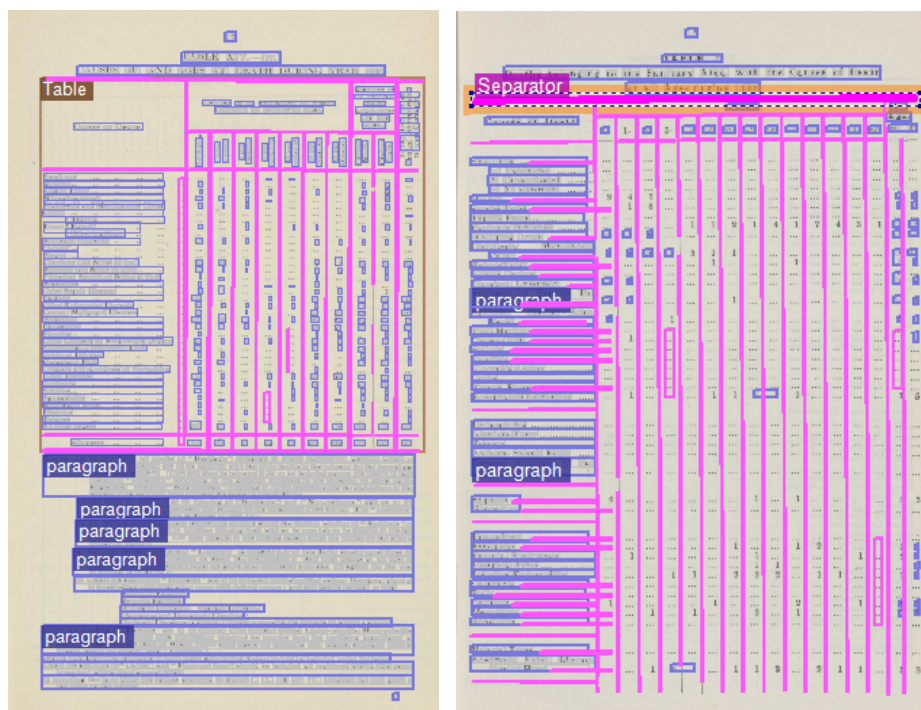
[Deptford](#)

Table and variable fingerprints can be used to search for and identify images containing tables with specific information content.

Experiments were conducted with table fingerprint searches for tables containing Cause of Death categories, and it was found that there was only a small reduction in accuracy of matching from raw OCR of images compared to XML from corrected ALTO files. A detailed description of the experiments is given in Appendix B.

## 5.2 Locating tables within images containing tables

ABBYY FineReader detects most tables automatically. Where it does not, tables can be identified from other features, such as large numbers of vertical separators to locate them and define their content. While exact numbers would require more extensive experiments, accumulations of vertical separators (like in the example below) should very reliably point to tabular content. A text-based analysis of OCR result, looking at the number of digits found, can further enhance the table detection accuracy.



FineReader results (left: table detected; right: no table region, but many separators (magenta))



### 5.3 Table Data Extraction

Once a table candidate page was identified, it can be tried to find relevant table columns and rows. A prototype algorithm was developed which tries to find predefined table headers within an uncorrected FineReader OCR result.

The input of the algorithm are all expected row and column header texts (ordered as they should appear). A word-based text matching and alignment is then used to find the headers on the given page. Based on the positions, the numerical content area can be identified too. The initial recognition works well and can be refined with little development effort. Using inner-table summations (e.g. row or column totals), recognition errors could be identified in an automated way. Corrections could then be targeted to specific problem areas.

17

TABLE XIV.—(iv.)  
CAUSES OF, AND AGES AT, DEATH DURING YEAR 1905.

CAUSES OF DEATH.	DEATHS IN OR BELONGING TO WHOLE DISTRICT AT RESPECTED AGES.					DEATHS IN OR BELONGING TO WHOLE DISTRICT AT RESPECTED AGES (AT ALL AGES).		TOTAL DEATHS IN OR BELONGING TO WHOLE DISTRICT AT RESPECTED AGES (AT ALL AGES).
	Under 5.	5 to 14.	15 to 24.	25 to 44.	45 to 64.	65 and over.	Total.	
Smallpox	1	1	1	1	1	1	6	6
Scarlet fever	1	1	1	1	1	1	6	6
Diphtheria	1	1	1	1	1	1	6	6
Whooping cough	1	1	1	1	1	1	6	6
Measles	1	1	1	1	1	1	6	6
Fever	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6
Typhus	1	1	1	1	1	1	6	6
Other continued	1	1	1	1	1	1	6	6
Typhoid	1	1	1	1	1	1	6	6

In order to integrate data and metadata retrieved from multiple tables containing information about similar characteristics a further challenge must be addressed:

- Development of global data models with structures (eg variables and categories) that are standardised, but contain the variation required to describe the full range of input data and metadata.

A data recovery and integration exercise was carried out using the Wellcome MOH XML files to explore the feasibility of, and effort required to create an *integrated and operable* information resource that could provide a back end for application development to deliver users with discovery, browse and query functions across the information contained in commonly occurring tables across different reports (across geographies and time). This would satisfy the majority of user requirements. The exercise focussed on tables containing information about Cause of Death by Age and Sex. These tables were chosen as most likely to deliver high value for effort because:

- The user survey identified information on mortality and cause of death as a high priority for users.
- Tables of Cause of Death by Age and Sex are some of the most commonly occurring tables across the reports, and so will provide data with most comprehensive coverage.
- The physical structures of Cause of Death tables are relatively consistent, and incorporate externally standardised classifications (for causes of death) that will make it easier to construct models that will accommodate changes in the classification over time.
- The existence of row and column totals within these tables provides a good basis for quality assurance based on comparing equivalent groups of values.

Cause of Death tables from six districts from 1910 (the same tables for which images are shown in section 5.1.1 above) were examined, and data and metadata were retrieved from them and integrated into a spreadsheet accompanying this report. The simple data filtering functionality (see arrows in table header row) demonstrate the benefits that recovery and transformation of information from images into operable digital data provide.

Development and variation in the classifications of Cause of Death used in the MOH reports over the period of the reports was also investigated. As well as the textual information available from the table header rows themselves, several external sources of information (such as the curious Wolfbane website at <http://www.wolfbane.com/icd/index.html>) were discovered. From these, together with input from domain researchers, it would be possible to construct the kind of ontologies encompassing variation in and relationships between the classifications over time that would be required for the global data models described above.

## 7 Recommendations on Minimum Viable Product and Additional Options for Further Work

Based on the findings of the high-level scoping exercise, the review of current work and best practice, the user consultation, the SWOT analysis, and the alternative OCR and table identification/recognition exercise, this section describes concrete recommendations, starting from a Minimum Viable Product (MVP) and progressing, in steps of increasing sophistication and effort through further stages of development of the information and service. The two different starting points – corrected OCR in XML structure for the London reports and scanned pages/raw OCR for the rest of England and Wales – have been taken into account in the following proposals.

As MVP, feedback from users indicates that a way of locating the tables they need (across all reports) for their work would add considerable value to the collection and its attractiveness for use. Having an index of relevant tables created based on user query terms would facilitate researchers' work considerably. Such an index, generated on the fly or as a set of pre-stored query results, can be presented in chronological order and/or according to geographical area. Having carried out this initial exercise, we are confident that this work can be done with minimal resources.

## 7.1 London MOH Reports already scanned and OCRed

### **Proposal 1.1: Create index of tables for online search for user-specified variables/categories within London MOH reports**

Individual tables will be described using a simple, high-level ontology (see example in Section 1.2 above) from analysis of existing London MOH XML. User-initiated search will then be able to produce results in the form of a list of links to pages containing the tables matching search keywords. Existing structured text for those tables can be downloaded via links. Tables in existing website could have 'see more tables like this for other areas, or for other years' links added.

### **Proposal 1.2: Create integrated data resource from London MOH tables for online search across locations and time**

Based on the XML versions of existing structured text, tables will be analysed and the data they contain will be described using a detailed ontology of variables and categories to facilitate integration within a global structure/database. Work will prioritise the most valuable (to users) and prevalent tables to ensure best returns from allocated effort. Online search can then query the database, and along with the integrated results provide links to relevant tables on the MOH report pages.

## 7.2 MOH Reports currently being scanned and OCRed for England and Wales, Scotland, Northern Ireland and the Colonies

### **Proposal 2.1: Create index of tables for online search for user-specified variables/categories within England and Wales MOH reports**

Effectively this will be an extension of Proposal 1.1 for tables across England and Wales but based on raw/uncorrected OCR outputs instead of the existing XML and corrected OCR. Tables will be described using the simple, high-level ontology from the London reports with further development for England and Wales if/where necessary. User-initiated search will then be able to produce results in the form of a list of links to pages containing the tables matching search keywords, although there will be no downloads of structured text for tables as in the London reports.

### **Proposal 2.2: Create structured text for tables from England and Wales MOH reports**

This will be an enhancement of Proposal 2.1 above to enable structured text downloads as in the case of the London MOH reports. Based on knowledge of table locations and variable/category content, table structures and data will be described and validated using the same detailed ontology developed for London MOH reports.

### **Proposal 2.3: Create integrated data resource from England and Wales MOH tables for online search across locations and time**

This will be an extension of Proposal 1.2 for tables across England and Wales, using the structured table data produced in Proposal 2.2 above.



## References

- [1] Public Health Reports, US National Library of Medicine,  
<https://www.ncbi.nlm.nih.gov/pmc/journals/347/> (Last Accessed: 16/03/2018)
- [2] Hathi Trust Digital Library, <https://www.hathitrust.org> (Last Accessed: 16/03/2018)
- [3] Statistics Sweden – Historical Statistics, [http://www.scb.se/en /Finding-statistics/Historical-statistics/Some-facts-about-historical-statistics/](http://www.scb.se/en/Finding-statistics/Historical-statistics/Some-facts-about-historical-statistics/) (Last Accessed: 14/03/2018)
- [4] Digitisation of Bidrag till Sveriges officiella statistik (BiSOS) – Project Report,  
[http://www.rj.se/GlobalAssets/Slutredovisningar/2006/Rolf-Allan\\_Norrmosse\\_eng.pdf](http://www.rj.se/GlobalAssets/Slutredovisningar/2006/Rolf-Allan_Norrmosse_eng.pdf)  
(Last Accessed: 14/03/2018)
- [5] Online Historical Population Reports (histpop), <http://www.histpop.org/> (Last Accessed: 14/03/2018)
- [6] Statistics New Zealand, Digitised Collections,  
[http://archive.stats.govt.nz/browse\\_for\\_stats/snapshots-of-nz/digitised-collections.aspx](http://archive.stats.govt.nz/browse_for_stats/snapshots-of-nz/digitised-collections.aspx)  
(Last Accessed: 14/03/2018)
- [7] Digital Reich Statistics – Digitisation of the Statistics of the German Reich – Alte Folge – (1873-1883), <http://www.digitalereichsstatistik.de> (Last Accessed: 14/03/2018)
- [8] London's Pulse: Medical Officer of Health Reports (1848-1972),  
<https://wellcomelibrary.org/moh/> (Last Accessed: 16/03/2018)
- [9] Project Tycho®, <https://www.tycho.pitt.edu> (Last Accessed: 16/03/2018)

## Appendix A: Survey Question Text

### Improving access to the Medical Officer of Health Reports

Annual reports on a wide range of topics relating to public health were produced by the Medical Officers of Health (MOH) for each local authority district in England and Wales between the 1840s and 1970s. The Wellcome Library has recently digitised the full texts of Medical Officer of Health Reports for districts in Greater London and made them available online at <http://wellcomelibrary.org/moh/>, and work is underway to digitise the texts of reports for the rest of England, Wales, Scotland and Ireland.

In addition to narrative text, the reports contain large numbers of statistical tables containing quantitative data. 271,783 tables have been identified within the 6652 reports already digitised for Greater London, all of which are presented in tabular layouts distinct from narrative text within texts of individual reports online. An example of a table of Cause of Death by Age for the district of Acton for 1910 is at <https://goo.gl/8CpPnw>, with tabular data below the image viewer.

Wellcome is working to improve access to, and usability of, the wealth of valuable information contained in the reports. Please help us to do this by completing this short survey about current access functionalities and potential development related to the tabular data. The survey should take no longer than 10 minutes, and all questions are optional.

1. Are you already aware of the Medical Officer of Health Reports?

Please tick all responses that apply to you.

Tick all that apply.

- ☐ No, I wasn't aware of them
- ☐ I was aware of them, but haven't used them
- ☐ I've used narrative text from the reports
- ☐ I've used quantitative data from tables in the reports
- ☐ Other:

2. If you've used the reports, what has it been for?

Please give a brief description below, including links and references if relevant.

3. Would you like to use information from the Medical Officer of Health reports about the following topics?

The topics below are some of the most common across the reports, but there are many others. Please make requests under 'other'.

Tick all that apply.

- ☐ Basic demographics (population size, age, sex, etc.)
- ☐ Mortality and causes of death
- ☐ Fertility (births)
- ☐ Diseases and ailments
- ☐ Health inspections
- ☐ Workplace and housing inspections
- ☐ Food inspections
- ☐ Finance and expenditure
- ☐ Other:

4. How useful do you find the current access functionalities?

The current website is available at <https://goo.gl/yev7u8>.

Mark only one oval per row.

	1 (not useful)	2	3	4	5 (very useful)
Keyword search					
Filtering (by location or years)					
Browsing (by location or year ranges)					
Downloads of individual report and table texts (eg <a href="https://goo.gl/PcrbcU">https://goo.gl/PcrbcU</a> )					
Bulk downloads of all report and table texts ( <a href="https://goo.gl/DJ7u2n">https://goo.gl/DJ7u2n</a> )					

5. What do you like and/or dislike about the current access functionalities, and how might they be improved?

Please try to be as specific as possible.

### Integration of tabular data

The digitisation and tabulation of quantitative data contained in tables within the reports has made them much easier to use for individual districts, but it is still difficult and time-consuming to locate and combine equivalent data from tables in multiple reports across districts and years to enable large-scale comparative analyses (for example, trends in deaths from cholera across all districts in London between 1850 and 1900). As a result, research to date has focussed mainly on analysis of the narrative text of reports. A scoping study is exploring the technical feasibility of integrating equivalent quantitative data from tables across multiple reports into structured and operable data to make it much easier to access and use. The study also aims to assess the potential value of, and demand for such integrated data. A simple example of this kind of integration of equivalent data on Cause of Death by Age from tables in reports from six different districts in 1910 into a single, queryable dataset contained in a spreadsheet can be seen at <https://goo.gl/yiAvSK>. The spreadsheet can be downloaded to enable filtering by column.

6. Are you interested in using quantitative data from MOH report tables?

Mark only one oval.

1	2	3	4	5
No interest				Very interested

7. Would having integrated data make this easier for you?

Mark only one oval.

1	2	3	4	5
No easier				Much easier

8. Which of the following topics would you most like to have integrated quantitative data from MOH report tables about?

These topics have been identified as relatively common and consistent in tables across multiple reports.

Mark only one oval per row.

	1 (low priority)	2	3	4	5 (high priority)
Basic demographics (population size, age, sex, etc.)					
Mortality and causes of death					
Fertility (births)					
Diseases and ailments					

### About you

It would help us to interpret responses to this survey if we have some understanding of the people who provide them. All responses below are optional.

Please describe your research interests

Please select the sector(s) most relevant to you

Tick all that apply.

- ☐ Academic
- ☐ Commercial
- ☐ Charity / not for profit
- ☐ Central government
- ☐ Local government
- ☐ Health
- ☐ Personal use
- ☐ Other:

Are you aware of similar information available from other sources?

If so, please provide some details (source, content, location, etc.)

### Your details

It might also be helpful to be able to contact respondents to follow up responses and/or to provide updates and gain feedback on future developments. If you're happy for us to do this, please provide your contact details below. Your details will not be shared or used for any other purpose.

Name

Email address

Do you have any further comments or questions?

## Appendix B: Text-based Table Detection

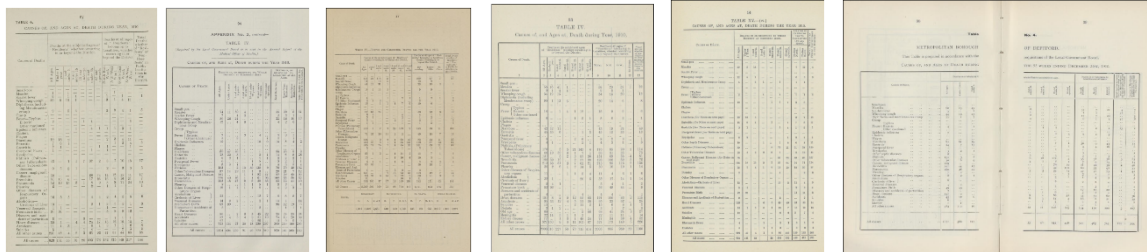
Experiments were conducted to determine if table extraction tasks could be automated for Medical Officer of Health Report images. The current table data was produced manually by correcting ALTO OCR outputs and extracting all tables to individual XML files. To reduce manual effort, information could be extracted from uncorrected OCR outputs in order to specifically target only those pages that contain tables of interest.

### B.1 Identification of Tables

In order to find candidates of pages containing a table of interest among many thousands of pages, a text matching approach was tested. The idea is to look for specific words (the table fingerprint) within the complete text content of a page. If a certain percentage of the words are found, the page is a candidate for containing the table of interest.

#### B.1.1 Example: Cause of Death Table

Six pages with equivalent data on Cause of Death by Age from different London districts were identified (all for 1910). The table text fingerprint was then created by comparing the table headers and finding common words (and excluding words that appear only in some of the six tables).



[Acton](#)

[Battersea](#)

[Bermondsey](#)

[Bethnal Green](#)

[Chelsea](#)

[Deptford](#)

b18106316\_13: CAUSES OF, AND AGES AT, DEATH DURING THE YEAR 1910.  
b18106451\_103: Causes of, and Ages at, Death during Year, 1910.  
b18106857\_101: Causes of, and Ages at, Death during the Year 1910.  
b18111099\_37: Causes of, and Ages at Death during the 52 weeks ending December 31<sup>st</sup>, 1910.  
b18111701\_48: Table IV, —Births and Corrected Deaths for the Year 1910.  
b19783450\_70: CAUSES OF, AND AGES AT, DEATH DURING YEAR, 1910.

Causes of Death. Cause of Death.	Deaths in or belonging to whole District at subjoined Ages. Deaths in, or belonging to, Whole District at Subjoined Ages. Deaths at the subjoined ages of "Residents" whether occurring in or beyond the District.							(Lo coll ties )	Total Deaths in Public institutions in the District. Total Deaths in Public in-stitutions in the District. Total Deaths whether of "Residents" or "Non-residents" in Public in-stitutions in the District. (similar)
	All Ages	Under 1 Year	1 and under 5	5 and under 15	15 and under 25	25 and under 65	65 and upwards		
1	2	3	4	5	6	7	8		13
Small-pox									
Measles									
Scarlet Fever									
Whooping-cough									
Diphtheria and Membranous Croup									
Diphtheria and Membranous Croup									
Diphtheria (including Membranous croup)									
Croup									
Fever Typhus (not other)									
Enteric									
Other continued									
Epidemic Influenza									
...									
All causes									

Finding common words in tables from different districts

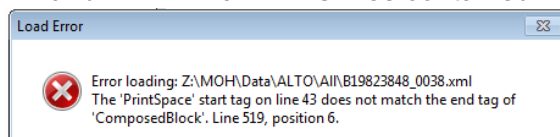
### B.1.2 ALTO XMLs

To test the table identification by text content, a realistic dataset was to be used. Since the original OCR results of the London MOH data was not available, the (corrected) ALTO XMLs were used. To this end, a PowerShell script was created to download all ALTO files containing at least one table. A large proportion of the pages also contain narrative text. In total, 206,359 ALTO XML files were collected.

#### *Problems with ALTO files*

Several problems were encountered during the download and use of the ALTO files.

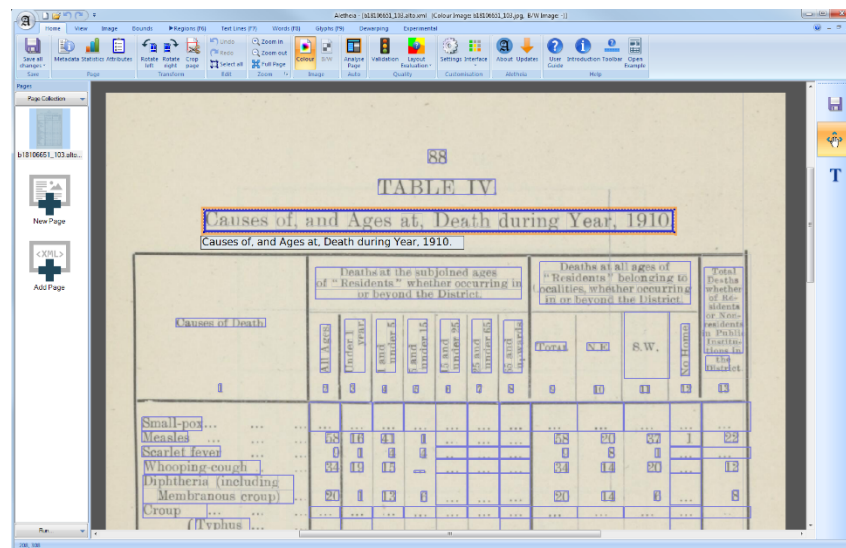
1. **ID format** – The page IDs were taken from the XML table files. The format is usually a B-number. In some cases, a page number was appended, leading to download failures in the batch download. The complete total file count was calculated to be 206,486. From those, 886 files could not be downloaded initially. 759 IDs were corrected and 127 files were omitted from the experiment.
2. **Invalid XML** – 102 ALTO files contained invalid XML content:



3. **Missing text orientation information** – Some table headers have 90 degree rotated text. However, that information is not contained in the ALTO files. Since text is stored as separate words, correctly composing a complete header from words is not possible without the rotation information (the result will be a jumbled up header with the wrong word order).
4. **Original image size** – The size and coordinates of the ALTO files are stored using a 1/10mm unit. For one example the height adds up to a page height of 21.5cm (8.5 inch). The corresponding JPG would then be precisely 300 PPI (even though the image metadata says 96 DPI). 21 cm seems a bit small (like A5), but possible.

#### *Inspection of the ALTO files*

The ALTO files can be visualised as overlay on the corresponding image using the Aletheia software. Looking at the word outlines and text content, it can be clearly seen that many OCR errors were manually corrected. Missing table cells were added and the text content transcribed or corrected. The word outlines were not corrected in many cases. A few occurrences of wrong text were spotted during the inspection (e.g. upper case letter i instead of a digit 1).



Aletheia Document Analysis System ([www.primaresearch.org](http://www.primaresearch.org))

Causes of Death	Deaths at the subjoined ages of "Residents" whether occurring in or beyond the District								Deaths at all ages of "Residents" belonging to localities whether occurring in or beyond the District				Total Deaths whether of Residents or Non-Residents in Public Institutions during the District
	All ages	Under 1 year	1 and under 5	5 and under 15	15 and under 25	25 and under 35	35 and under 55	55 and over	Total	N.E.	S.W.	No Home	
Small-pox...	...	...	...	...	...	...	...	...	...	...	...	...	...
Measles ...	68	10	13	1	...	...	...	...	68	20	37	1	28
Scarlet fever ...	0	1	1	1	...	...	...	...	0	1	1	...	3
Whooping-cough ...	34	13	13	...	...	...	...	...	34	13	20	...	12
Diphtheria (including Membranous croup)	20	1	13	1	...	...	...	...	20	1	1	...	8
Croup ...	...	...	...	...	...	...	...	...	...	...	...	...	...
(Typhus ...)	...	...	...	...	...	...	...	...	...	...	...	...	...

Selected word outlines of an ALTO file

### B.1.3 Identifying Candidate Pages

A text analysis tool, developed at the PRImA lab, was extended to aid with the table identification task. The tool requires a comma- or tab-separated table as input, hence the ALTO text contents were transferred to a single large TSV file (tab-separated values). The table has two columns – one for the filename and one for the text content (serialised). Since there is one row per file, the resulting table had 206,359 rows.

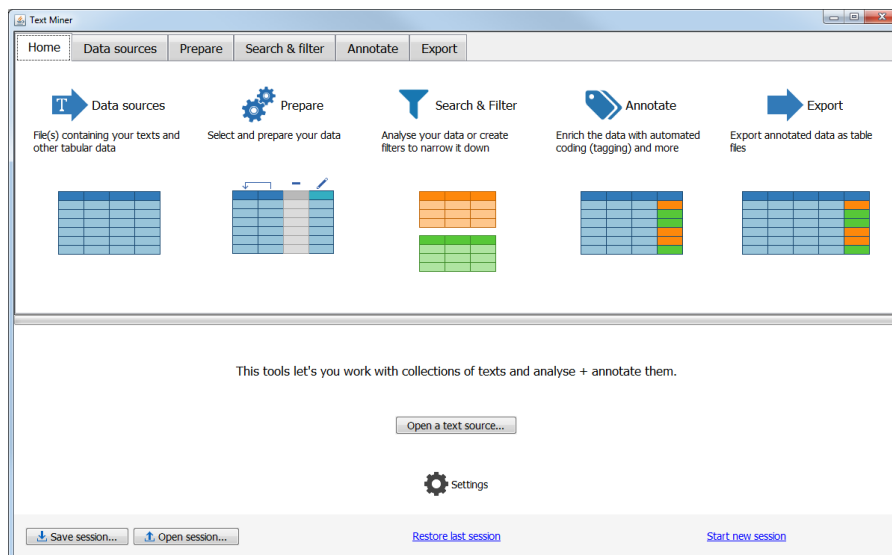
A text matching component was added to the tool, allowing to match a list of reference words (fingerprint) against all ALTO text contents. A match score (between 0% and 100%) is calculated and the table is sorted from best to worst match. The higher the match score of a page the more likely it is that the page contains the table of interest.

The matching returned the following results:

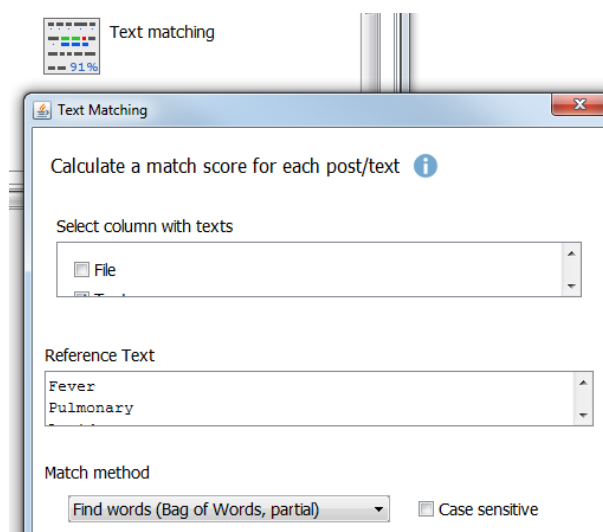
- 7 match scores over 90%
- About 300 match scores over 80%
- The first 10 matches contain relevant tables (not more tested)

File	Text
"B19970249_0040.alto.xml"	"TABLE IV.—Causes of, and Ages a..."
"B18043537_0010.alto.xml"	"TABLE X.—(iv.) Causes of, an..."

ALTO text contents as Tab-Separated Values



Text Analysis Tool ("Text Miner")



Text Matching Component

#### B.1.4 OCR with ABBYY FineReader

A few selected pages were re-OCR'd with ABBYY FineReader Engine 11 (FRE) in order to establish a baseline of the current state-of-the-art.

##### *Cause of Death tables 1910*

The six pages from earlier were processed with FineReader and PAGE XML (similar to ALTO) was produced. In addition, the pages were also processed with the popular open source OCR engine Tesseract, for comparison. The results can be seen in the following table. The percentages represent the word error rate (missed/wrong words, using Bag of Words measure). FineReader has a significantly higher success rate.

Note that the OCR was done using JPG images (lossy compression). Better-quality images might result in increased accuracy.



Image	FRE Normal	Tesseract
b18106316_13	8.22%	52.56%
b18106651_103	9.70%	52.72%
b18106857_101	9.76%	53.61%
b18111099_37	21.88%	42.46%
b18111701_49	7.13%	20.34%
b19783450_70	17.23%	19.21%
<b>Avg</b>	<b>12.32%</b>	<b>40.15%</b>

Further experiments were carried out to find out if image preprocessing or different OCR engine setups can improve the recognition results. While most approaches lead to improvements for individual pages, only FineReader's low-resolution mode generates slightly better results on average.

FRE Pre-processed	FRE LowRes	FRE Type-Writer	FRE Sharpen-ed	FRE Sharpened LowRes	FRE Enhanced	FRE De-speckled
10.60%	7.99%	10.19%	7.11%	8.29%	8.98%	10.03%
21.62%	12.56%	13.25%	10.57%	12.06%	11.88%	13.65%
11.52%	9.87%	20.40%	10.82%	10.35%	11.12%	10.42%
55.04%	16.49%	27.21%	50.28%	50.20%	19.10%	28.75%
6.91%	7.50%	16.41%	6.19%	6.21%	8.87%	7.23%
17.55%	17.60%	16.51%	13.28%	12.98%	14.06%	14.43%
<b>20.54%</b>	<b>12.00%</b>	<b>17.33%</b>	<b>16.37%</b>	<b>16.68%</b>	<b>12.34%</b>	<b>14.08%</b>

Image	FRE Normal	Tesseract	FRE Pre-processed	FRE LowRes	FRE TypeWriter	FRE Sharpened	FRE Sharpened LowRes	FRE Enhanced	FRE Despeckled
b18106316_13	8.22%	52.56%	10.60%	7.99%	10.19%	7.11%	8.29%	8.98%	10.03%
b18106651_103	9.70%	52.72%	21.62%	12.56%	13.25%	10.57%	12.06%	11.88%	13.65%
b18106857_101	9.76%	53.61%	11.52%	9.87%	20.40%	10.82%	10.35%	11.12%	10.42%
b18111099_37	21.88%	42.46%	55.04%	16.49%	27.21%	50.28%	50.20%	19.10%	28.75%
b18111701_49	7.13%	20.34%	6.91%	7.50%	16.41%	6.19%	6.21%	8.87%	7.23%
b19783450_70	17.23%	19.21%	17.55%	17.60%	16.51%	13.28%	12.98%	14.06%	14.43%

### Text Matching

To determine whether the table identification via text matching also works for the uncorrected OCR results (and not just for the corrected ALTO files), 20 pages were selected for processing with FineReader. These are the 10 best matches from the ALTO matching and 10 pages with decreasing match scores from 80% down to below 10%. In additions to the direct word matching introduced earlier (match score 1), a more forgiving match method was developed which allows for a certain degree of spelling mistakes (using an edit distance measure; match score 2). The following table contrasts the two different match scores for the 20 pages. The is good correlation between the ALTO and FineReader match scores, with match score 2 having less difference in average.

From this we can conclude that the FineReader results are good enough to allow table identification.

File	MOH ALTO Match Score 1	MOH ALTO Match Score 2	FineReader Match Score 1	FineReader Match Score 2
B18106316_0013.xml	98.48%	100%	96.97%	98.51%
B18106249_0014.xml	98.48%	100%	96.97%	98.51%
B18106304_0014.xml	95.45%	98.51%	95.45%	98.51%
B19970249_0040.xml	92.42%	94.03%	89.39%	92.54%
B19970237_0032.xml	92.42%	94.03%	89.39%	94.03%
B18044578_0016.xml	92.42%	94.03%	93.42%	92.54%
B19975922_0024.xml	90.91%	92.54%	87.88%	91.04%
B18043537_0010.xml	89.39%	89.55%	86.36%	88.06%
B18248950_0040.xml	89.39%	92.54%	86.36%	91.04%
b18106109_0024.xml	89.39%	92.54%	80.30%	85.07%
B18048304_0129.xml	84.85%	88.06%	78.79%	83.58%
B19970572_0048.xml	77.27%	86.57%	71.21%	80.60%
B19823265_0062.xml	75.76%	79.10%	75.76%	79.10%
b18120854_0057.xml	66.67%	71.64%	65.15%	70.15%
B18223011_0007.xml	53.03%	59.70%	50.00%	58.21%
B19822704_0099.xml	46.97%	52.24%	45.45%	50.75%
B18246540_0022.xml	37.88%	47.76%	37.88%	46.27%
B19823563_0092.xml	28.79%	41.79%	28.79%	41.79%
B19883432_0045.xml	18.18%	28.36%	16.67%	28.36%
B19785720_0022.xml	4.55%	16.42%	4.55%	17.91%