Unearthing the Recent Past: Digitising and Understanding Statistical Information from Census Tables[†]

Christian Clausner[1], Justin Hayes[2], Apostolos Antonacopoulos[1], and Stefan Pletschacher[1]

(1)
PRImA Research Lab
The University of Salford
United Kingdom
www.primaresearch.org

(2)
Jisc
56 Oxford Street
Manchester, M1 6EU, UK
Justin.Hayes@jisc.ac.uk

ABSTRACT
Censuses comprise a wealth of information at a large (national) scale that allow governments (who commission them) and the public to have a detailed snapshot of how people live (geographical distribution and characteristics). In addition to underpinning socio-economic research, the study of historical Census statistics provides a unique opportunity to understand several characteristics in a country and its heritage. This paper presents an overview of a complete account of the background, challenges, implemented pre-processing, recognition and post-processing pipeline, and the information-rich results obtained through a pilot digitisation project on the 1961 Census of England and Wales (the first time computers were used to process data and output very detailed information, a vital part of which is only available in the form of degraded historical computer printouts). The experience gained and the resulting methodology can also be used for digitising and understanding tabular information in a large variety of application scenarios.

Categories and Subject Descriptors
I.7.5 [**Document Capture**]: Language Constructs and Features – *Document analysis, Optical character recognition.*

General Terms
Algorithms, Management, Performance, Reliability, Experimentation.

Keywords
Digitisation, Tabular data, Printed documents, Census, Historical, Cultural Heritage, Preprocessing, Post-processing, Recognition.

DATeCH2017, June 1–2, 2017, Göttingen, Germany. © 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5265-9/17/06 …$15.00.
http://dx.doi.org/10.1145/3078081.3078106

[†] This work was funded in part by the UK Office for National Statistics.

1. INTRODUCTION
National censuses are primarily conducted to acquire information about the geographical distribution and characteristics of the population required to inform government spending and policy decisions. Census data represents an invaluable resource for researchers and other interested parties. Historical census data reveals a wealth of information on factors influencing culture and the heritage of a country. However, while more recent census results are available in digital form, older material exists only in the form of paper, microfilm, or image files (scans).

The majority of census data is presented across several tables, each corresponding (in the UK) to a county and its constituent local authorities. The tables on this rather broad area data were printed and published in book form. The 1961 Census saw the introduction of computers to store and aggregate the manually acquired (by the appointed Census Enumerators) data. The new technology enabled more fine-grained reporting of statistics down to small areas (e.g. about 100 households). Those Small Area Statistics (SAS) tables were sent as computer printouts on request to local authorities. Unfortunately, only 1 or 2 complete copies of this SAS data exist now (in scanned microfilm form of the original printouts) – all digital data has been lost.

The recently concluded Census 1961 Feasibility Study was conducted to ascertain whether the complete 1961 Census data collection can be digitised and the information extracted and made available online in a highly versatile form similar to the newer Censuses.

The study was conducted in two parts by the authors in cooperation with the Office for National Statistics (ONS) [1] from September 2015 to December 2016. The feasibility was tested by designing a digitisation pipeline, applying state-of-the-art page recognition systems, importing extracted fields into a database, applying sophisticated post-processing and quality assurance techniques and evaluating the results. The main questions to be answered were: What is the best way of digitising the material to maximise the quality of the output and is the quality high enough to satisfy the requirements of a trustworthy Census 1961 database with public access?

A prototype of a fully-functional pipeline was developed, including: image preprocessing, page analysis and recognition,

post-processing, and data export. Each individual part of the pipeline was evaluated individually by testing a range of different analysis and recognition approaches on a representative data sample. Well-established performance evaluation metrics [10] were used to precisely measure the impact of variations in the workflow on different types of data (image quality, page content etc.). In addition, the accuracy of the extracted tabular data was evaluated using model-intrinsic rules such as sums of values along table columns and/or rows and across different levels of geography.

In this paper, we present a template-based table recognition approach that was developed and used within the study. This enabled the Census data in bitmap image form to be transformed to a format that can be fed into an existing database structure. Specific table cells could be identified and the content recognised with high precision.

2. RELATED WORK

Table recognition from document images is commonly considered a two-stage process: table detection and table structure recognition [2]. During the detection phase entities that correspond to a certain table model are identified and segmented from the rest of the image. Structure recognition is then targeted at recovering the actual table content by analysing and decomposing such entities following the assumed model [3], [4].

Most table recognition systems employ generic models based on certain rules and/or features for describing the characteristics of what is considered a table. Over the years, several methods have been proposed following different approaches related to the two main stages from above and further broken down with regard to how observations are obtained (measurements, features), transformations (ways to emphasise features) and inference (decision if/how a certain model fits) [5].

Nevertheless, table recognition still remains a very challenging topic. Especially generic table recognisers have to strike a balance between accommodating various types of tables and achieving a high-enough accuracy. Conversely, scenarios in which the input material contains only a limited number of fixed table layouts can greatly benefit from specifically trained systems. The case in which the semantics and locations of all data cells are known can also be seen as a form recognition problem [6]. According to the nature of the approach, such systems will always have to be tailored to the material.

The largest part of the Census 1961 data consists of fixed table layouts. Those can be processed using templates that model the precise table structure. Other problems, such as inconsistently scanned images, geometric distortions, and poor image quality, still pose a considerable challenge. The remainder of the Census data also contains more complex content with more variable content (e.g. unknown number of table rows), although none of which require purely generic table recognition approaches.

Existing table recognition methods, such as implemented in ABBYY FineReader Engine 11 [7], produce a result with a table structure and cell content, but with very inconsistent quality (according to the authors' experiments on the data concerned). Most of the Census data is densely packed (to save paper) and only with narrow whitespace separators.

Furthermore, even if a recognition method correctly identified the content of a table cell (i.e. its correct numeric value) the relation between this recognised cell content and the table model (labelled cells) still needs to be established.

In the following, the complete workflow from image scans to extracted and accumulated Census 1961 data is described. The proposed approach transforms the semantically meaningless values identified in the images by OCR into higher-level information by associating each number with a unique cell identifier that references the specific combinations of characteristics that the number quantifies. This enables the numbers and their associated cell identifiers to be extracted from the images, in an automated way, as self-describing, atomic packages of data, which are then combined in a single structured and operable dataset.

3. DATASET

The full 1961 Census data consists of approximately 140,000 scanned pages. From these, a representative subset of about 9,000 pages was selected for the pilot study. The majority of the material consists of different types of tables that were either typeset (accumulative reports) or computer printouts (Small Area Statistics – SAS). The scans are characterised by a wide range of image quality with various production and scanning related issues and artefacts. Figure 1 shows three examples.

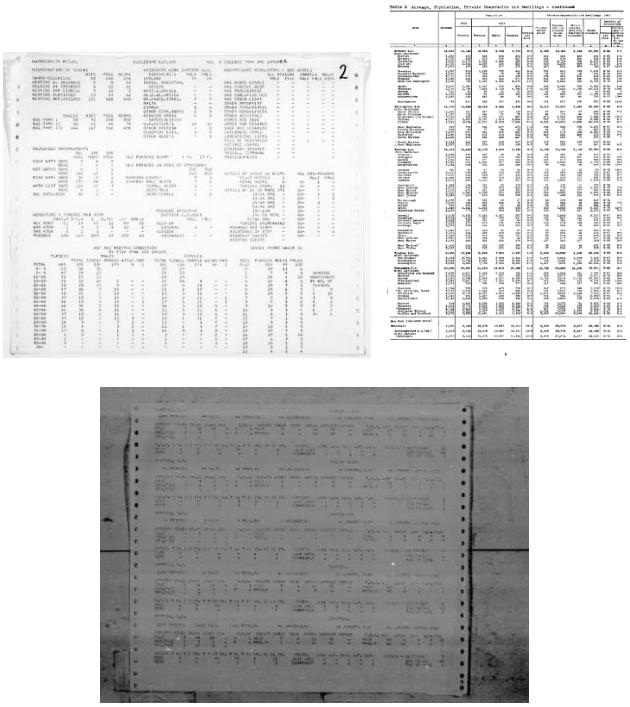

Figure 1. Examples of Census 1961 scans.

The largest part of the material contains tables with a fixed layout, where the number of columns and rows, heading text, and spacing are identical (not taking into account distortions) for each instance. More complicated layouts include pages with unknown combinations of tables and tables with variable row count and/or different abbreviations used in the headings and row/column labels.

To enable experiments and evaluation, an initial data preparation was carried out, including: splitting multi-page documents into single-page image files, visual inspection, conversion to TIFF images, and binarisation.

In order to measure digitisation results, reference data is needed. Ground truth can be seen as the ideal result of a page recognition method, or, in other words, the result a perfect (error-free) OCR

system would produce. Because the production of ground truth is very labour-intensive, only a relatively small set of document images could be ground truthed within the context of this study. It was therefore crucial to select the images carefully in order to capture a representative selection. 60 images were chosen, from all relevant subsets. In average, it took in the order of two hours per page to create ground truth. The production was carried out with the Aletheia Document Analysis System [8] (see Fig. 2). Where useful, pre-produced data (OCR results) from ABBYY FineReader Engine 11 was corrected, otherwise the ground truth was created from scratch. Both page layout and text content were transcribed. The output format is PAGE XML [9], a well-established data format representing both physical and logical document page content.

4. DATA EXTRACTION

The digitisation workflow consists of two major parts: (1) the recognition and information extraction pipeline and (2) a stage for data aggregation and quality assurance. This section describes the processing pipeline and its evaluation and the next section describes the data aggregation and quality assurance.

4.1 Pipeline

As part of the pilot study, a processing pipeline was designed and all essential parts were implemented and applied to the aforementioned dataset. Figure 2 shows an overview of this digitisation pipeline. The target is to extract the table information from image files (scans) and export it as comma-separated values (CSV) that can be fed into a database. The pipeline framework was implemented using the Microsoft PowerShell scripting language.

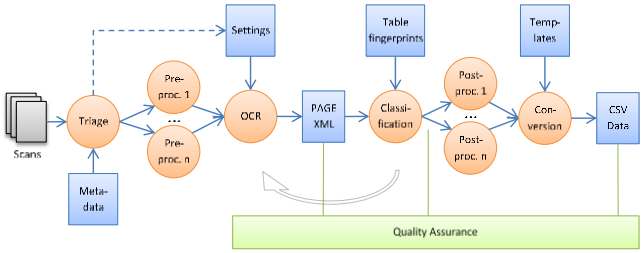

Figure 2. Census digitisation pipeline.

The preprocessing step performs one or multiple (alternative) image operations to improve the input to OCR and ultimately its results. Input and output are therefore image files. It should be noted that the improvement can only be measured objectively based on the output of the subsequent steps. Since not every preprocessing approach works equally well on each type of image, a triage step can be used to select the best preprocessing method for the image at hand. The decision process is based on metadata, which can either be readily available or calculated on-the-fly (e.g. feature extraction).

OCR is arguably the most important part of the pipeline. It should be noted that, page analysis and recognition is the preferable term since the process includes page segmentation (into regions, text lines, words and glyphs), region classification (e.g. text, picture, chart, table etc.) and content recognition (including optical character recognition (OCR)). Nevertheless, an OCR engine is generally understood to perform all of the above tasks.

The results of an OCR system (page layout objects and text content) need to be exported to a detailed and versatile file format (here, PAGE XML) that enables further (automated) processing of the data. The OCR engine itself is interchangeable, as long as export functionality to PAGE XML is available. For the pilot study the latest versions of two state-of the-art systems were applied and evaluated (see Section 4.2).

The recognition process of an OCR engine can be influenced through various settings. Choosing an appropriate setup can have a considerable positive impact on the quality of the OCR results. The triage component described earlier can also be used to select settings that work well for the type of image that is currently being processed.

For the final step of feeding the census data into a database, the OCR results need to be converted from PAGE XML format to a predefined table layout in CSV format. The census tables have a known layout which is modelled by a table template (also in PAGE XML format). The OCR result can be transferred to a template, as long as the type of the currently processed tables is known or can be determined on the fly. To this end, within the digitisation pipeline a table classification step is used to uniquely identify the type(s) of table(s) that are contained in the current page.

Before the text content can be transferred, the template needs to be aligned with the OCR result because the scanning was not performed in a precisely controlled way, leading to table positions that vary considerably. In addition, slight geometric variations (scaling, skew) need to be compensated. A matching algorithm was designed and implemented in a new tool called PRImA Layout Aligner. It takes advantage of the fact that the tables in the Census data have a certain proportion of fixed text content that is repeated in each image of the same type (headings, column and row headers, notes etc.). By sliding the template over an OCR result and calculating how well the fixed content of the template matches the OCR result at the current position, the best matching position can be determined (Figure 3). The match score is based on the character recognition rate that is also used for the performance evaluation of the OCR results (see [10]). The matching is therefore based on the polygonal layout data of glyph objects (not image data).

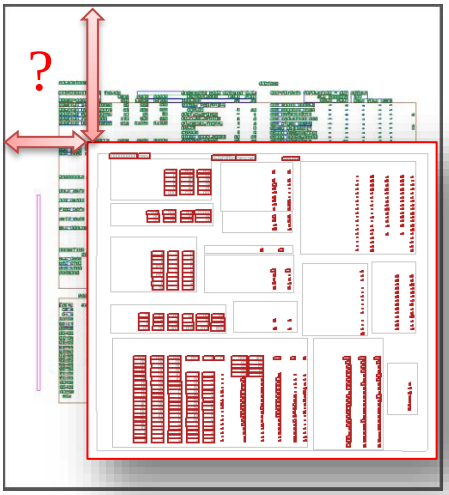

Figure 3. Illustration of template matching.

The actual alignment process is carried out by testing all possible positions of the template within the OCR result. For improved efficiency, this is done in two stages: (1) Rough

estimation of the location using a sliding step with equal to the average glyph width and (2) detailed calculation of the best match in the neighbourhood of the estimation using a sliding step width of 1 pixel.

If multiple table templates can be found on a single page, the matching process is performed for all templates and the templates are then used in the order from best match to worst. Overlap of templates in thereby not allowed. Similarly, if specific templates can occur multiple times on a page, this fact can be defined in the processing pipeline and the matching will be carried out accordingly.

Once the ideal offset is known, the template can be filled with the text from the OCR result (text transferal). This is done by copying each glyph object (a layout object with shape description, location and contained text character) of the OCR result to the word object in the template it overlaps most. If a glyph overlaps no template word, it is disregarded. After all glyphs have been processed, the final text is propagated to the region level (by composing words from the glyphs, text lines from the words and regions from the lines). The result is a copy of the template with the text of the OCR result filled into the cell regions which are labelled with the predefined IDs.

Another software tool (Table Exporter) was implemented to realise the final conversion from the filled-in template (PAGE XML file) to the desired table format.

Post-processing can be used to try to improve the results of the core digitisation pipeline. As with preprocessing, it is not strictly required but the overall results can be enhanced incrementally by identifying shortcomings and applying certain correction rules or reprocessing parts of the data. Methods tested within the study include: re-OCRing empty table cells and invalid results, character replacement and removal, and specialised OCR (restricted to recognise digits only, for example).

4.2 Evaluation

The output of OCR engines can be evaluated by comparing it against the ground truth. A requirement is that both pieces of data (OCR result and ground truth) are available in the same data format. For this study the PAGE XML format was used, which stores detailed information about location, shape and content of page layout objects (including but not limited to: regions, text lines, words and glyphs).

Two sets of text-based performance measures were used to establish a quality baseline for two state-of-the-art OCR engines: ABBYY FineReader Engine 11 (commercial) [7] and Tesseract 3.04 (open source) [11]. The first set of measures is character-based and describes the recognition rate. A rate of 100% thereby means that all characters have been found and identified correctly by the OCR engine. In order to be able to focus on the important pieces of data (in the context of this study), three variations of this measure have been implemented: (1) Character recognition rate excluding "replacement" characters (which are markers for unreadable text), (2) Recognition rate for digits only (characters "0" to "9"), and (3) Recognition rate for numerical characters (digits plus "-", "+", "(" etc.). This has been implemented as an extension to an existing layout-based evaluation tool [10]. The second set of measures uses the "Bag of Words" approach [12], mentioned earlier.

To be able to examine and evaluate a variety of preprocessing methods efficiently, a framework of scripts, evaluation tools, and analysis approaches was created. That way, experiments could be set up and modified easily with as little manual labour as possible. A cascade of scripts processes a subset of the census data using all possible combinations of pre-processing steps, OCR engine settings, evaluation tools, and evaluation settings. The data is accumulated and transferred into an interactive spreadsheet that can be used for detailed comparative analysis. Figure 4 shows a comparison of the pipeline using no pre-processing and default OCR settings vs. the best pre-processing OCR setup (determined by experiments). Tesseract performs worse than FineReader (86.6% vs. 97.6% digit recognition accuracy) but it is still used as secondary (alternative) OCR during post-processing if FineReader fails for a specific table cell.

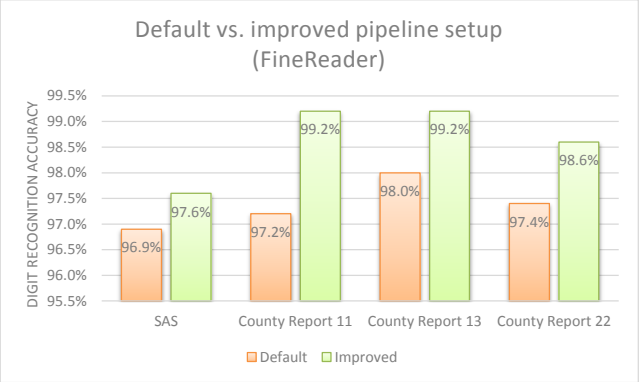

Figure 4. Digit recognition accuracy for different subsets and setups (ABBYY FineReader)

Table data in CSV format represent the final output of the digitisation pipeline of the census data within the pilot study. Errors in the data can originate from:

1. Mistakes in the original print (e.g. mixed up table layouts and/or actual typos).
2. OCR errors.
3. Errors in table type classification (a table was misclassified as another type than it actually is).
4. Errors in the pre-specified templates (templates do not match the actual table layout or cell IDs have been mislabelled).
5. Template matching / alignment errors (due to geometric distortions in the scan or bad OCR results for instance).
6. Errors in the transferral from OCR result to the template (too much, too little or wrong data cell content was transferred).
7. Problems with CSV export (e.g. data content - such as commas - not escaped properly, interfering with the CSV format; encoding / special characters).

An evaluation of the whole pipeline can be done by using data analysis based on intrinsic rules in the table models (sums across rows and columns, for example). Using stored intermediate results of the digitisation pipeline and processing reports, errors can be traced back to their source and can be corrected if possible. The data accumulation and analysis is explained in the next section.

5. DATA IMPORT AND VALIDATION

This section describes the final stage of the census digitisation in which the extracted raw data is fed into a database with a logical model of the Census. The model allows for detailed quality assurance - a crucial part of the workflow since the limited quality of the image data leads to imperfect recognition results. Being able to discover and pinpoint problems is the basis to achieve reliable Census information at the end of the digitisation effort.

5.1 Logical Model

The initial scoping of the image set enabled a logical model to be constructed in a database that incorporates and integrates the geographies (Figure 5) and characteristics described by the data together with relationships between them. The model provides a clear picture of data that can be expected in the outputs from OCR processing, and so is useful for assessing their completeness. It also provides a framework for receiving and storing the data and metadata in the outputs in a way that makes them accessible and operable for quality assurance as well as future dissemination and analysis.

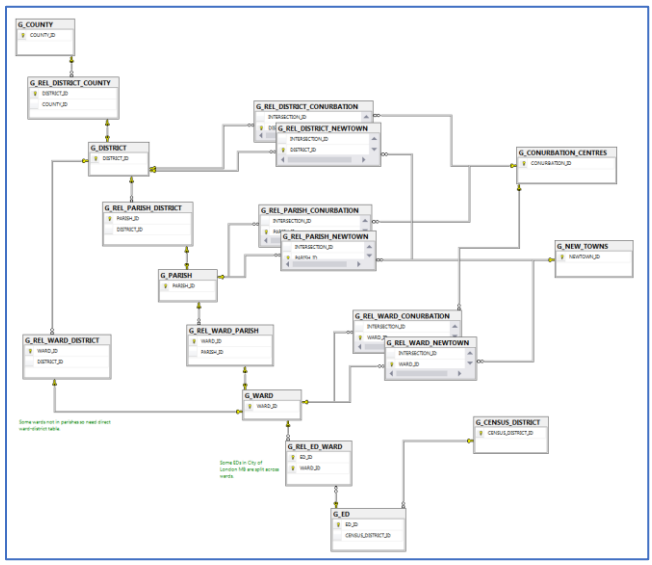

Figure 5. Geographical model for Census 1961 data. Printed tables were produced at different levels (from low to high): Enumeration districts (ED), Wards, Parishes, Local Authorities, and Counties.

5.2 Correction and Quality Assurance of OCR Output Values

It is possible to derive and compare multiple values for many of the population characteristics described in the 1961 SAS from different table cells, or combinations of cells for the same area. For instance, cells for All People appear in several tables, and values for this characteristic can also be generated by combining values for groups of cells containing sub-categories of All People, such as (Males + Females), or (Single + Married + Widowed + Divorced), etc. Of the 967 cells contained in the 18 tables of the 1961 SAS, 760 can be subjected to these within-area cell group comparisons, with each cell taking part in an average of 12 comparisons.

In addition to the within-area comparisons, it is also possible to derive multiple values for the characteristics represented by each table cell for larger areas (e.g. districts) by summing values from corresponding cells for smaller areas (e.g. wards and enumeration districts) contained within them. Each of the 967 cells in the 1961 SAS can take part in either 2 or 3 of these geographical summation cell group comparisons, depending upon geographical level.

The within-area and geographical summation cell group comparisons were carried out programmatically on the values from each image in the raw OCR outputs in turn. Every time a comparison is carried out, each of the values taking part receives a 'disagreement score' based on the level of agreement between groups of values that should have the same value (Equation 1).

$$\frac{\left(\frac{G}{V}-1\right)^2}{N} \qquad (1)$$

Where G is the number of 'comparison groups' that should have a value equivalent to a particular 'comparison characteristic' (e.g. "All people"), V is the number of check groups that share the value of the check group in which the value belongs, and N is the number of values within the comparison group in which the value belongs.

All values take part in at least two comparisons, and some values take part in many more. Disagreement scores from each comparison are summed to identify values which take part in comparisons as part of different groups in which disagreements persistently occur. High cumulative disagreement scores suggest that a value is likely to be the source of comparison errors. Values with the highest disagreement scores are selected for interactive correction (re-OCR or manual input). The raw OCR output values are then updated with corrected values, and the QA processing repeated iteratively. OCR values for the relatively small number of the largest (district) areas are processed first in order to provide 'true' corrected values as absolute, rather than relative targets for geographical summation comparisons, which significantly reduces noise in the resulting disagreement scores. Figure 6 shows a histogram for the disagreement scores for two London boroughs. The lowest scores (including 0 – full agreement) are the most frequent. Higher scores are much less frequent, but still considerable. It should be noted however, that a disagreement score greater than zero does not mean the corresponding table cell was misrecognised. It only means at least one cell of the corresponding comparison group is wrong and therefore there is a likelihood that the cell data is wrong and needs to be checked.

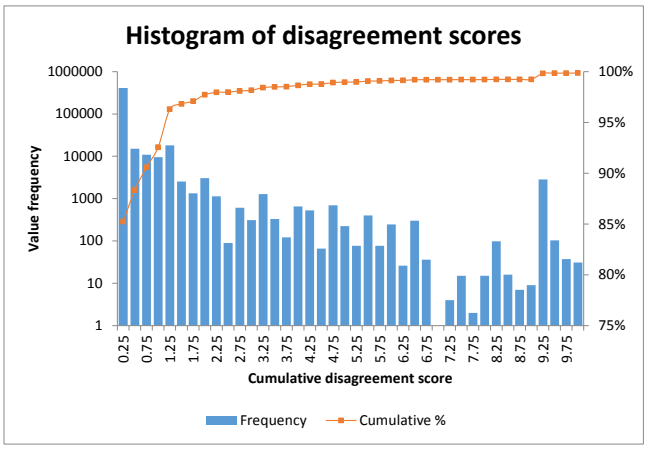

Figure 6. Histogram of cumulative disagreement scores from QA comparisons on raw OCR outputs of values for Hammersmith and Lewisham districts in London (logarithmic scale for frequency).

6. CONCLUSION AND FUTURE WORK

The pilot study showed that the quality of automated recognition is good enough to carry out the digitisation of the complete Census 1961 data. Problems in the processing pipeline can be detected and traced back to the source for manual correction, for instance via crowdsourcing.

A complete workflow was developed including pre-processing, OCR, table recognition, post-processing, data conversion, data integration, and quality assurance. All steps are automated already but need further work with respect to reliability and performance. Further improvements based on OCR training will be explored. Initial experiments show promising results.

A relatively large subset of the available Census material was already used for the pilot study and more data is being processed in an extension project in early 2017. The complete data is going to be processed within a larger follow-up project. More collections of similar data (Censuses and other datasets with tabular content) will be considered for further future studies and projects. The workflow is generic enough to be able to deal with a variety of tabular and form-like data sources.

The Census 1961 data will be made available online in an interactive platform (see [13]), once all data is processed and validated.

7. REFERENCES

[1] Office for National Statistics, United Kingdom, https://www.ons.gov.uk/

[2] Hu, J., Kashi, R.S., Lopresti, D., Wilfong, G.T. 2002. Evaluating the performance of table processing algorithms. *International Journal on Document Analysis and Recognition*, Volume 4, Issue 3 (March 2002), pp 140-153.

[3] Lopresti, D., Nagy, G. 1999. Automated Table Processing: An (Opinionated) Survey. *Proceedings of the 3rd International Workshop on Graphics Recognition* (Jaipur, India, 26–27 September 1999). pp 109-134.

[4] Costa e Silva, A., Jorge, A.M., Torgo, L. 2006. Design of an end-to-end method to extract information from tables. *International Journal of Document Analysis and Recognition (IJDAR)*, Volume 8, Issue 2 (June 2006), pp 144-171.

[5] Zanibbi, R., Blostein, D., Cordy, J.R. 2004. A survey of table recognition: Models, observations, transformations, and inferences. *Document Analysis and Recognition*, Volume 7, Issue 1 (March 2004), pp 1-16.

[6] Lopresti, D., Nagy, G. 2001. A Tabular Survey of Automated Table Processing. *Graphics Recognition Recent Advances*, Volume 1941 of the series Lecture Notes in Computer Science (April 2001), pp 93-120.

[7] ABBYY FineReader Engine 11, http://www.abbyy.com/ocr-sdk

[8] Clausner C., Pletschacher S., and Antonacopoulos A. 2011. Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments. *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011)* (Beijing, China, September 2011), pp. 48-52.

[9] Pletschacher S., and Antonacopoulos A. 2010. The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. *Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010)* (Istanbul, Turkey, August 23-26, 2010), IEEE-CS Press, pp. 257-260.

[10] Clausner C., Pletschacher S., and Antonacopoulos A. 2011. Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods. *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011)* (Beijing, China, September 2011), pp. 1404-1408.

[11] Tesseract OCR, https://github.com/tesseract-ocr

[12] PRImA Text Evaluation Tool, University of Salford, United Kingdom, http://www.primaresearch.org/tools/PerformanceEvaluation

[13] InFuse, UK Data Service, http://infuse.ukdataservice.ac.uk/