Creating a Complete Workflow for Digitising Historical Census Documents: Considerations and Evaluation[†]

Christian Clausner, Justin Hayes, Apostolos Antonacopoulos and Stefan Pletschacher

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

ABSTRACT

The 1961 Census of England and Wales was the first UK census to make use of computers. However, only bound volumes and microfilm copies of printouts remain, locking a wealth of information in a form that is practically unusable for research. In this paper, we describe process of creating the digitisation workflow that was developed as part of a pilot study for the Office for National Statistics. The emphasis of the paper is on the issues originating from the historical nature of the material and how they were resolved. The steps described include image pre-processing, OCR setup, table recognition, post-processing, data ingestion, crowdsourcing, and quality assurance. Evaluation methods and results are presented for all steps.

Categories and Subject Descriptors

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

Keywords

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

1. INTRODUCTION

The main objectives of national censuses are to acquire information about the geographical distribution and characteristics of the population and to inform government spending and policy decisions. Historical census data reveals information on factors influencing culture and the heritage of a country. However, while more recent census results are available in fully searchable digital formats, older material exists only in the form of paper, microfilm, or scans of those physical items.

Most of the data in the 1961 Census of England and Wales is presented across several tables, each corresponding to a county and its constituent local authorities. Those tables were printed and published in book form. The introduction of computers enabled also more fine-grained Small Area Statistics (SAS), which were sent as computer printouts to local authorities (on request). Only one or two complete copies of this data survived (scanned microfilm of the original printouts) – all digital data has been lost.

The recently concluded Census 1961 Feasibility Study [1] was conducted in cooperation with the Office for National Statistics (ONS) [2]. Its aim was to ascertain whether the complete 1961 collection can be digitised, the information extracted, and made available online in a highly versatile form like the newer Censuses. 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, and evaluating the results.

Accurately capturing the content of the census tables was a central step. Table recognition from document images is commonly split into table detection and table structure recognition [3]. For detection, entities that correspond to a table model are identified and segmented from the rest of the image. Structure recognition then tries to recover the table content by analysing and decomposing such entities following a model [4], [5].

Most table recognition systems employ generic models based on certain rules and/or features for describing the characteristics of what is considered a table. Several methods have been proposed following different approaches related to the two main stages from above and further broken down according to how observations are obtained (measurements, features), transformations (ways to emphasise features) and inference (decision if/how a certain model fits) [6].

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 resembles a form recognition problem [7]. Typically, such systems are tailored specifically to the material.

The largest part of the Census 1961 data consists of known table layouts which can be processed using templates that model the precise table structure. Content-unrelated 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 tables with more variable content (e.g. unknown number of table rows).

Existing and readily available table recognition methods, such as implemented in ABBYY FineReader [8], produce results with general table structure and cell content, but with very inconsistent quality (as experiments for the initial feasibility study [1] showed). Most of the Census data is densely packed (to save paper) with narrow whitespace separators. Furthermore, even if a recognition method correctly identifies 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.

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

A template-based table recognition method was developed within the pilot, complemented by processing steps to compensate for issues originating from the historical nature of the material. The complete pipeline includes: image pre-processing, page analysis and recognition, template matching, post-processing, and data export. Well-established performance evaluation metrics [11] were used to precisely measure the impact of variations in the workflow on different input data (image quality, page content etc.). 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.

2. CENSUS DOCUMENTS

The available 1961 Census document set comprises about 140,000 scanned pages. From these, a representative subset of 9,000 pages was selected for the pilot study. Most of the material consists of different types of tables that were either typeset or computer-printed. The scans show a wide variation in image quality with various production and scanning related issues and artefacts. Figure 1 shows three examples and four snippets highlighting common issues. The set is a mix of bitonal and greyscale images with resolutions between 300 and 400 PPI. Unfortunately, JPEG compression was used on most of the images and compression artefacts are visible (although the impact of this on OCR could not be measured as rescanning was not possible).

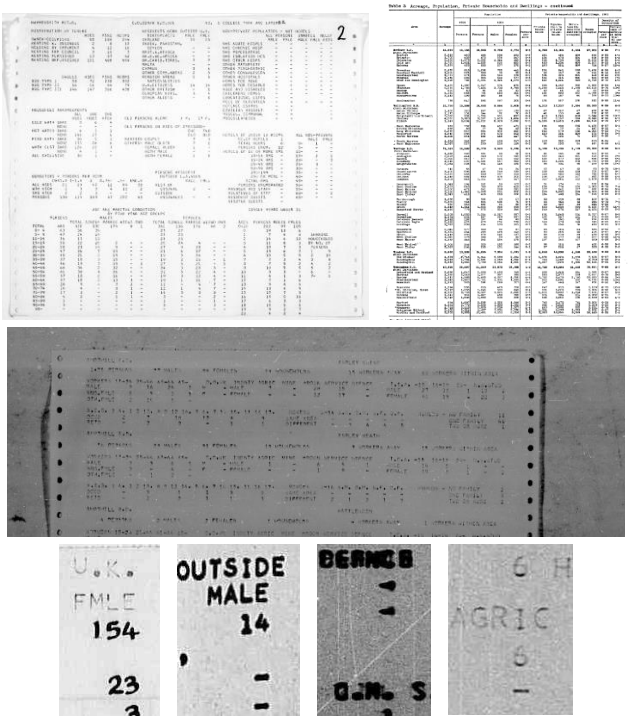

Figure 1. Example images (pages and details) of the 1961 Census.

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 considering printing/image 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 step was carried out, including: splitting multi-page documents into single-page image files, visual inspection, and conversion to TIFF images. In order to establish a baseline, a random sample of 1,000 images was tagged using 40 different keywords describing the condition of the material. The keywords include artefacts and characteristics related to following categories: production problems, ageing related issues, and problems originating from reproduction or scanning (see also [10]). Table 1 lists the most common conditions.

Table 1 - 30 most common image/content conditions

Keyword	Pages out of 1,000	Keyword	Pages out of 1,000
Punch holes	909	Filled-in characters	54
Annotations	906	Rotated text	41
Blurred characters	888	Binarisation artefacts	39
Uneven illumination	818	Non-straight text lines	28
Broken characters	507	Warped paper	27
Paper clips visible	357	Low contrast	25
Scanner background vis.	216	Out of focus	14
Scratches (microfilm)	202	Touching chars (vert.)	10
Skew	179	Noise from scanner	8
Faint characters	140	Page curl	7
Show-through	115	Folds	6
Salt-and-pepper noise	68	Handwritten (mostly)	5
Handwritten correction	59	Tears	4
Stains	58	Holes	4
Touching chars (hor.)	57	Missing parts	2

For evaluation purposes, detailed ground truth for tables and text content was produced for 60 images. This was carried out using the Aletheia Document Analysis System [9] (see Figure 2). In order to arrive at the required accuracy it took on average two hours to complete one page. Where useful (more efficient), pre-produced data (OCR results) from ABBYY FineReader was corrected, otherwise all data was entered from scratch. All ground truth is available in PAGE XML [11], a well-established data format representing both physical and logical document page content.

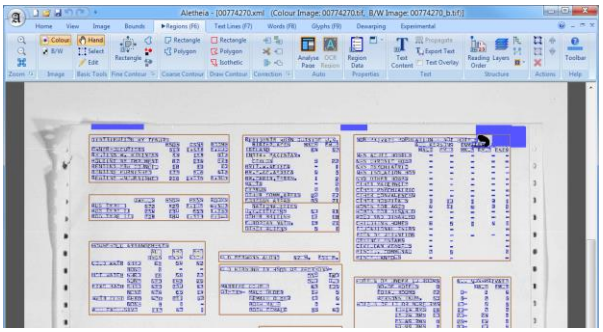

Figure 2. Ground truth in the Aletheia Document Analysis System.

3. INFORMATION 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, followed by data aggregation and quality assurance in the next section.

3.1 Designing the Processing Pipeline

As part of the pilot study, a processing pipeline was designed, implemented, and applied to the dataset. Figure 3 shows an overview of the digitisation pipeline. The overall goal is to extract all table information contained in image files (scans) and export it as comma-separated values (CSV) that can be fed into a database. The pipeline framework connecting the individual processing steps was implemented using the Microsoft PowerShell scripting language.

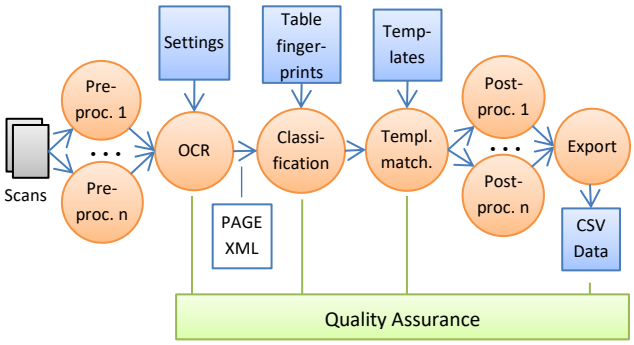

Figure 3. Digitisation pipeline.

3.1.1 Deciding on Pre-processing and OCR Setup

A combination of different image pre-processing steps and OCR setups were tested. The tools used were: ImageMagick [13], PRImA Image Tool (by the authors), ABBYY FineReader Engine 11 [8], and Tesseract 3.04 [14]. Table 2 shows the tested methods and setups. Scripts were used to run various combinations (several hundred) of the methods/steps on all pages for which ground truth was available.

All results were evaluated based on the OCR result quality (character recognition rate, see [12]). The final pipeline setup was then chosen individually for different types of pages with the help of an improvement matrix which compares the success after pre-processing with the baseline of no pre-processing (Figure 4).

Table 2 – Tested preprocessing steps and OCR setups

Tool	Tested Step / Setting
PRImA Image Tool	Dilation (greyscale or bitonal)
	Erosion (greyscale or bitonal)
	Sauvola binarisation
ImageMagick	Despeckle
	Equalize
	Contrast enhancement
	Enhance
	Sharpen
FineReader Engine	Normal font OCR
	Typewriter font OCR
	Low resolution OCR
Tesseract OCR	Standard OCR

Processing	Improvement of recognition rate relative to original images (no pre-processing)					
	Ampthill RD	Index of Place Names	Kent Film 68	Folke-stone MB	SAS Film Listings	Random Selection 1
Dilation	+2.5	-1.8	-0.4	-2.1	+3.8	-4.6
Erosion	+3.4	+0.4	+1.7	+1.1	-39.6	+2.7
Dilation + Erosion	+5.2	-0.4	-0.1	+1.6	+1.3	+2.0
Sauvola025	+16.0	0.0	+0.7	-0.2	+1.8	-2.3
Despeckle	+4.0	0.0	-2.5	+0.6	+2.1	-1.8
Equalize	+13.9	0.0	-79.8	-92.6	-7.3	-28.6
Contrast	+3.6	0.0	+0.7	0.0	0.0	-0.8
Enhance	+2.6	0.0	+1.2	+0.6	+0.5	+2.4
Sharpen	+1.6	0.0	-1.7	+0.4	+0.7	+1.2
Despeckle + Sharpen	-0.3	0.0	+1.7	+0.9	+0.7	+0.4
Enhance + Sauvola025	+14.0	0.0	-0.5	+0.2	+2.1	-3.6
Contrast 2x	+13.2	0.0	-0.2	+1.0	+0.3	-3.4
Enhance + Contrast	+7.7	0.0	+1.0	-0.5	+0.6	-2.1
Enhance + Dilation + Erosion	+7.5	-0.4	+0.3	+0.8	+1.9	+1.2

Figure 4. Pre-processing improvement matrix (pre-processing steps in different rows, image subsets in different columns, change in percent in comparison to baseline).

FineReader performed significantly better than Tesseract (especially in capturing the majority of page content) and was chosen as primary OCR engine. Tesseract is still used in post-processing. Since, in addition to the text content, detailed page layout data is required for the table recognition, the OCR engines were accessed via their respective API (application programming interface) and results were exported to a suitable file format (here, PAGE XML [11]).

Additional improvements were made by restricting the target character set of the OCR engines. All census tables are limited to standard Latin characters, digits, and punctuation marks. By prohibiting all other characters OCR performs notably better. Some tables are also limited to upper case letters, reducing the target character space even further.

3.1.2 Table Recognition

To be able to select the correct table template for a given image, every page needs to be classified first (since the dataset was only partially structured and the type(s) of tables within a page is not known beforehand). A text-based method was implemented that uses the static text content of the tables (e.g. headers) as "fingerprints" and compares them to the recognised text of the page at hand using a bag-of-words evaluation measure (as implemented in [15]).

Table detection and recognition is done by aligning a previously created template (the table model containing all table cells as polygons with metadata) with the OCR result. A match score based on character bounding boxes is used instead of pixel-based matching, providing robustness against small layout variations. A matching algorithm was designed and implemented in a new tool called PRImA Layout Aligner.

Due to variations during scanning or capturing on microfilm, tables within a page usually vary slightly with respect to scale, aspect ratio, and rotation. An automatic scaling detection and correction step was implemented to compensate for size differences. This is based on distances between unique words which can be found in both the template and the OCR result. FineReader's deskewing feature was used to correct the skew angle, if any.

The actual alignment process is carried out by testing all possible positions of the template within the OCR result (Figure 5). For efficiency, this is done in two stages: (1) Rough estimation of the location using a sliding step width equal to the average glyph width and (2) detailed calculation of the best match in the neighbourhood of the estimation using a one-pixel sliding step.

If multiple table templates can be found on a single page (the same table or set of tables repeated multiple times), 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 is thereby not allowed.

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. The result is a copy of the template with the text of the OCR result filled into the cell regions which are labelled with predefined IDs.

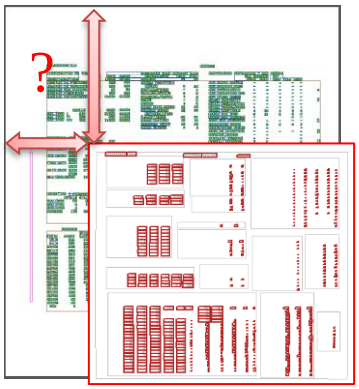

Figure 5. Illustration of template matching.

3.1.3 *Post-processing, confidence, and export*

Metadata in the table templates includes the type of cell content (text, Integer number etc.). This can be used to validate the recognised content against the expected content. A rule-based post-processing step is applied to the recognised data in order to auto-correct certain mistakes by using statistical information (e.g. upper case I can be replaced by the digit one if the cell is known to contain a number).

If this type of automated correction is not possible, the whole cell is OCRed with a secondary method (Tesseract). This is done by creating an image snippet of the respective table cell and sending it to Tesseract with the option of recognising a single text line. The target character set can be reduced even further in case of cells with numerical content (digits and punctuations).

Several steps in the pipeline (OCR, page classification, and template matching) produce confidence values. They can be used as indicators for problems where manual intervention may be necessary. Having indicators throughout the pipeline helps to find issues early in the process and avoids cumbersome searches when errors are found in the final stages of data processing.

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 (comma-separated values).

3.2 Evaluation

The output of OCR can be evaluated by comparing it against the ground truth. A requirement is that both pieces of data 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/characters).

Two sets of text-based performance measures were used to establish a quality baseline for the two state-of-the-art OCR engines: ABBYY FineReader Engine 11 (commercial) [8] and Tesseract 3.04 (open source) [14]. The first set of measures is character-based and describes the recognition rate. It is a very precise measure, made possible by having ground truth glyphs with their location on the page and the assigned character. Each glyph of the OCR result can then be matched against a glyph of the ground truth. 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 [12].

The second set of measures uses the "Bag of Words" approach [15], mentioned earlier. It measures how many of the ground truth words were recognised regardless of their position and how many wrong words were added.

Figure 6 shows a comparison between FineReader and Tesseract OCR. FineReader has a clear advantage in all but two subsets. Especially for the subset representing the largest amount of pages (Small Area Statistics, Kent Film 68), FineReader outperforms the open source engine by over 5%.

Figure 7 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 good enough to be used as secondary (alternative) OCR during post-processing.

Figure 8 shows a comparison between the general character recognition rate and the digit recognition rate. For the most important subsets (Kent), the digit recognition surpasses the general character recognition rate. This is encouraging since the main data that is to be extracted from the images are numbers.

Some of the images (e.g. Ampthill) are of particularly poor quality. Even the manual ground truth production was a challenge and automated processing is unlikely to produce usable results. Fortunately, this seems to be limited to one microfilm (apparently the first one to be produced) and manual transcription seems a viable option.

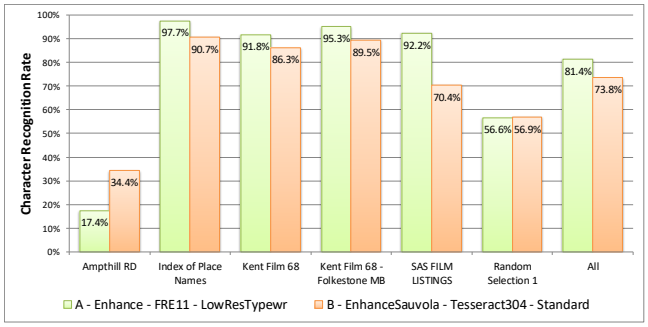

Figure 6. Character recognition accuracy for FineReader (left bars) and Tesseract (right bars) for different subsets

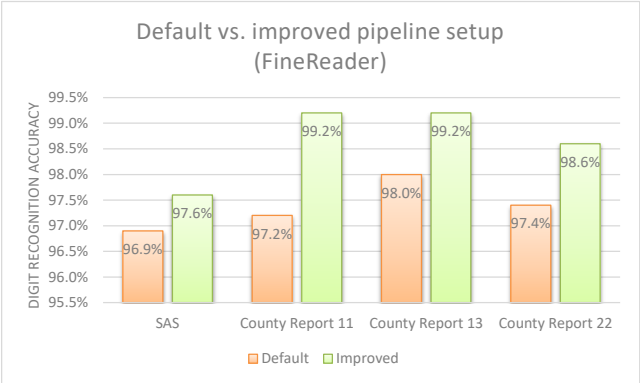

Figure 7. Digit recognition accuracy for different subsets and setups (ABBYY FineReader)

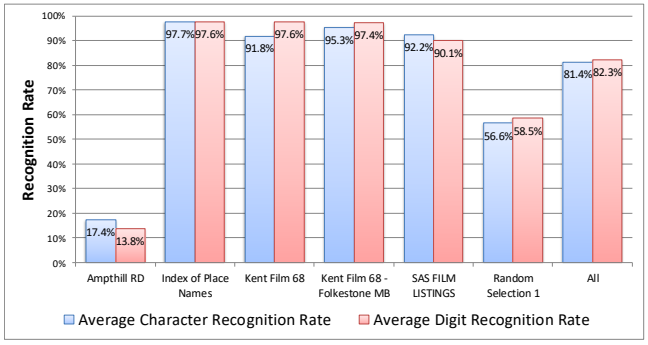

Figure 8. Character vs. digit recognition (FineReader)

Table data in CSV format represents the final output of the digitisation pipeline. Errors in the data can originate from:

1. Mistakes in the original print.
2. OCR errors.
3. Errors in table type classification (wrong type detected).
4. Errors in the pre-specified templates.
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 cell content was transferred).
7. Problems with CSV export (e.g. decoding issues).

Table 3 shows the evaluation results of a few random samples. Extracted table cells have therein been checked for correctness and errors logged. The success rate is the number of correct cells divided by the total number of cells. Most errors are caused by OCR misrecognition and a few by problems during text transfer from OCR result to the template (due to misalignment).

Table 3 – Evaluation of whole pipeline

Sample set	Cells checked	Cells wrong	Accuracy
SAS Battersea	967	16	98.3%
SAS Camberwell	967	8	99.2%
SAS Deptford	967	10	99.0%
SAS Fulham	967	14	98.6%
SAS Hammersmith	967	6	99.4%
SAS Lambeth	967	13	98.7%
SAS Lewisham	967	18	98.1%
SAS Wards & Parish.	1407	93	93.4%
SAS Local Authorit.	1072	30	97.2%
County Reports	494	9	98.2%
	9742	**217**	**97.8%**

An evaluation of the whole pipeline without ground truth can be done by leveraging intrinsic rules in the table models via automated data analysis (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 step is explained in the next section.

4. INFORMATION INGEST, VALIDATION AND FURTHER CORRECTION

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. Detected errors can then be corrected manually – either directly or via crowdsourcing.

The initial scoping of the image set enabled a logical model to be constructed in a database that incorporates and integrates the geographies 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.

4.1 Correction and Quality Assurance of OCR Output Values

It is possible to derive and compare multiple values for many of the population characteristics 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 such as (Males + Females).

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 by summing values from corresponding cells for smaller areas contained within them.

The within-area and geographical summation cell group comparisons were carried out programmatically on the values from each image in the OCR outputs in turn. Each of the values taking part receives a 'disagreement score' based on the level of agreement among groups of values that should be the same.

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 the corrected values. OCR values for the relatively small number of the largest (district) areas are processed first to provide 'true' corrected values as absolute, rather than relative targets for geographical summation comparisons, which significantly reduces noise in the resulting disagreement scores.

4.2 Crowdsourcing

Even though the overall pipeline table data recognition success rate is quite good considering the material (about 98%), the size of the dataset makes manual correction of the remaining 2% very costly (millions of table cells). But since this kind of data might incur public interest, crowdsourcing was explored and implemented within the pilot study.

The Zooniverse [17] platform was chosen since it is an established and popular crowdsourcing provider offering free hosting (up to a limit) and an intuitive app creation interface.

It was decided to make the correction task for users as simple as possible. Users are presented with a single table cell at a time. The cell is outlined over an image snippet which is extracted from a full page, including a bit of the surrounding area. The user is then asked to simply type the cell content that is associated with the outlined cell. Figure 9 shows the interface, in this case on a mobile device. Unnecessary punctuations can be left out.

Data is uploaded in batches, selected by the disagreement scores from the quality assurance step. Once finished, corrected data is fed back into the database and the QA step is repeated.

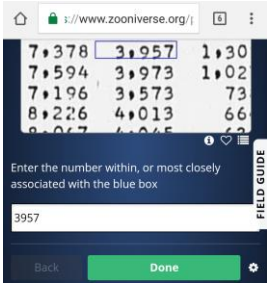

Figure 9. Crowdsourcing interface (mobile platform)

5. CONCLUSION AND FUTURE WORK

The feasibility study was considered a success and a follow-up project is planned to process the complete Census 1961 document set. Further improvements and more automation can be achieved and initial experiments are being carried out including OCR engine training and cloud-based processing.

The quality of the processing pipeline is sufficient to enable large-scale digitisation with limited resources. Data validation and error correction methods and strategies ensure high-quality Census data that extends the more recent datasets further into the past. The validation step is precise enough to even reveal errors in the original printouts – some of which have truncated data due to space limitations.

The results will be made publicly available.

Similar document collections exist (from the UK and abroad) and the workflow has also been evaluated on samples from those with a view to further digitisation projects.

REFERENCES

[1] Clausner, C., Hayes, J., Antonacopoulos, A., Pletschacher S. 2017. Unearthing the Recent Past: Digitising and Understanding Statistical Information from Census Tables. Proc 2nd Int Conf on Digital Access to Textual Cultural Heritage (DATeCH2017), Göttingen, Germany, June 2017.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] ABBYY FineReader Engine 11, www.abbyy.com/ocr-sdk

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

[10] C. Clausner, C. Papadopoulos, S. Pletschacher, A. Antonacopoulos. 2015. The ENP Image and Ground Truth Dataset of Historical Newspapers. Proc 13th Int Conf on Document Analysis and Recognition (ICDAR2015), Nancy, France, August 2015, pp. 931-935.

[11] 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.

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

[13] ImageMagick, www.imagemagick.org

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

[15] PRImA Text Evaluation Tool, University of Salford, UK, www.primaresearch.org/tools/PerformanceEvaluation

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

[17] Zooniverse crowdsourcing platform, www.zooniverse.org/