Crowdsourcing Historical Tabular Data – 1961 Census of England and Wales

Christian Clausner
School of Science, Engineering & Environment
University of Salford
United Kingdom
c.clausner@primaresearch.org

Justin Hayes
School of Science, Engineering & Environment
University of Salford
United Kingdom
j.hayes@primaresearch.org

Apostolos Antonacopoulos
School of Science, Engineering & Environment
University of Salford
United Kingdom
a.antonacopoulos@primaresearch.

ABSTRACT

This paper describes how crowdsourcing can be incorporated as an integral part of a comprehensive technical workflow to identify, extract and validate data from large volumes of printed tabular statistics, and transform them into operable digital datasets using current structural and descriptive standards. The recently completed digitisation project for the 1961 Census of England and Wales (commissioned by the UK's Office for National Statistics) is used to provide details on data processing, crowdsourcing platform and tasks, crowd interaction, and validation of results. The multi-modal approach employed was very successful, delivering far more complete and validated data than automated processes alone could produce (due to the challenging nature of the source material).

KEYWORDS

Crowdsourcing, Historical, Tables, Census, Digitisation

1 Introduction

Despite its importance and usefulness, statistical data in historical documents is largely untouched due to the increased recognition difficulty when compared to textual content. The extraction of numerical information is a considerably more complex problem than creating a computer readable/editable form of the printed numbers on a page. Numerical information in tabular form expresses values of row/column relationships which must also be represented for the data to be useful. For instance, while in the case of text one may search for a particular keyword, in the case of numerical data one will need to search for the value of a relationship between variables (e.g. how many houses had a fixed bath in a given town) – there will not be a use case for searching for a given number e.g."173" in the document. Therefore, the analysis and recognition of the overall structure and meaning (columns, rows, semantics of data cells) requires specialised software solutions (recognition workflows, see [1] for example).

Compounding the difficulty of this higher-level table recognition problem is the fact that the numbers OCR (Optical Character Recognition) will produce cannot be readily validated as in the case of text (where dictionaries of allowed words exist) – at face value any number could be correct. However, users of numerical data have significantly higher expectations in terms accuracy. Textual errors can be forgiven, numerical errors less so – there is a high threshold for accuracy required for results to be statistically reliable and useful.

Crowdsourcing (paid or volunteer-based) is a solution often suggested as a possibility for either completely manual text entry (small-scale projects) or for OCR post-correction (see next section). Following document analysis and recognition, crowdsourcing can also be used to validate numerical information. For a small-scale project this could be done for all the data (depending on the capacity of the crowd). For large-scale datasets (most common real-world cases), however, the crowd must be used selectively, prioritising the most challenging items. The latter requires the design and application of a decision-making process as to which data items to crowdsource.

In this paper, we describe the crowdsourcing approach (and resulting insights) that was devised and employed as part of a complete workflow for digitising historical census data. The established Zooniverse platform [2] was used for crowd-based processing.

The next section briefly summarises crowdsourcing approaches, focussed on the correction and transcription of information on documents. Section 3 provides a brief overview of the 1961 Census digitisation project. Section 4 describes the part of the pipeline that deals with crowd-related processing. Section 5 provides details on the project website, results, and statistics. Sections 6 and 7 contain further insights and concluding remarks.

2 Crowdsourcing

Spinks et al. [3] discuss task workflow design in volunteer-based crowdsourcing. They argue that users prefer variety in data and autonomy in performing tasks. Simpler tasks typically lead to greater volume of results. In addition, interfaces that are more direct can lead to better result quality.

Traditional crowdsourcing platforms for (narrative) text (e.g. Trove [4], Digital Proofreaders / Gutenberg Project [5], TypeWright [6]), use contributors to work on whole blocks or on text lines (as identified by OCR engines). Platforms that focus more on field-based data (records or certificates) typically use a combination of manually specifying the boundaries of a field on the image and entering the text in a predefined form (e.g. software tools used by FamilySearch [7], ancestry.co.uk/.com [8]; see Fig. 1).

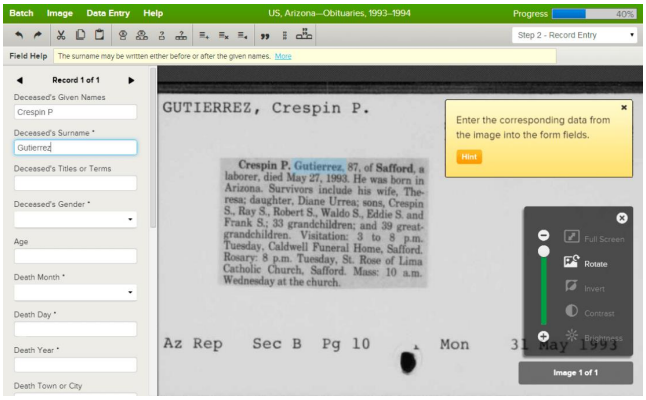

Figure 1: FamilySearch Indexing interface (familysearch.org).

Processing complete tables is particularly challenging and can be split into two main tasks: table structure recognition and text recognition. In some projects (e.g. Weather Rescue [9], Castaway [10]), the user is asked to perform both tasks. Usually, this leads to very complex workflows. Other projects (see Southern Weather Discovery [11], for example) perform the structure recognition beforehand (often manually) and present the user with smaller sections of a table for entering the associated text.

3 The 1961 Census Digitisation Project

The digitisation project for the 1961 Census for England and Wales [1][12] was conducted by the authors for the UK's Office for National Statistics (ONS) [13]. The goal was to produce a complete integrated digital dataset for publication by ONS, containing as much information as could possibly be extracted and validated from a set of approximately 140,000 digital images of computer printouts (via microfilm) of tables from the unpublished and almost entirely inaccessible 1961 Census Small Area Statistics. Although a fully automated processing pipeline was developed and applied, the recognition accuracy (about 98%) was not sufficient for census data. Crowdsourcing was employed for number-recognition only, presenting the users with only one table cell at a time.

Specific challenges included image quality and other issues such as:

- Inconsistent scan quality (illumination, warping, skew, scaling, placement).
- Faint print, handwritten corrections.
- Microfilm scratches and general degradation.
- Missing parts, printing errors.
- Unorganised data (pages not in any particular order).
- Dense tables, sometimes with no separation between columns.

4 Crowdsourcing Integration within the Overall Workflow

The 1961 Census Small Area Statistics consist of about 70,000 pages (as part of the larger image set), printed in seven different fixed page layouts (containing several tables each), repeated for different geographical areas. Fig. 2 shows an example page with 13 tables.

As part of the 1961 Census Digitisation Project, a processing pipeline was developed that includes OCR and template-based table recognition (more information in [1]). Text recognition I based on ABBYY FineReader Engine and Tesseract OCR. Template matching is used to find the position of the tables within an image. Figure x shows another example page with the aligned table templates.

The templates contain detailed information on all data cells (IDs, data types, parent table). This and additional external information can be used to validate extracted values. Once values have been extracted and associated with specific cells in the statistical tables, equivalencies within the data are exploited by carrying out a large set of arithmetic and logical comparisons (e.g. values across a row with a row total) to validate the values. The majority of table cells take part in at least one such comparison, and the validation enables the location of errors to be narrowed down to small groups of cells. These cells are then considered for crowdsourcing. The complete workflow is described next.

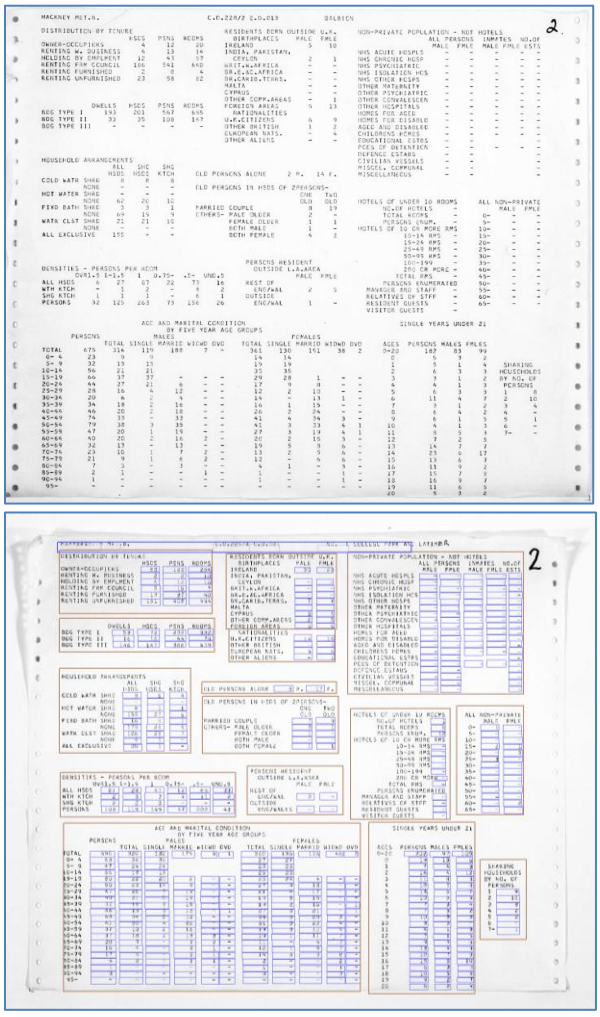

Figure 2: Example pages with Small Area Statistics (top: original image; bottom: matched template as overlay)

Fig. 3 provides an overview of the steps related to crowd-based processing. The workflow is as follows:

1) OCR and Template Matching provide the table structure and text content (simplified, the complete workflow includes multiple processing steps and branches).
2) Where the template matching confidence is low, a series of manual steps are performed:
    a. Export of all low-confidence pages to a single PDF with original images and template overlay.
    b. Visual check for template matching errors. If no error, continue with step 3.
    c. Manual template alignment using an interactive tool (Aletheia [14]).
3) Validation of extracted numerical table content. If no data disagreements, continue with step 6. Where there are validation errors or where no validation can be performed (due to lack of data redundancy):
4) Create image snippets of identified cells (with a bit of the surrounding area) and an overlay of the cell boundaries (see Fig. 4).
5) Upload to Zooniverse platform (required information, such as cell Id, encoded in filenames). Transcription of numerical content by volunteers (3 volunteers per cell). Download of results. Where users identified misaligned cell boundaries (feedback as #misaligned tag), go to step 2. Otherwise revalidate the values in step 3.
6) Data ingest into result database.

The workflow has repetitive elements. If the validation still fails after transcription by humans, another round of crowdsourcing is performed. Cells that cannot be completed in this manner need to be checked by an expert (see examples in Fig. 5).
The crowdsourcing of the 1961 Census was a big success. Results and statistics are discussed next.

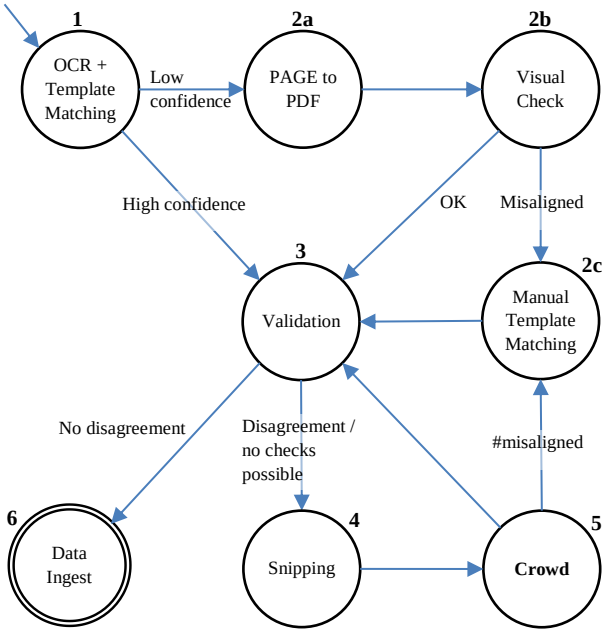

Figure 3: Overview of workflow that includes crowdsourcing of selected data

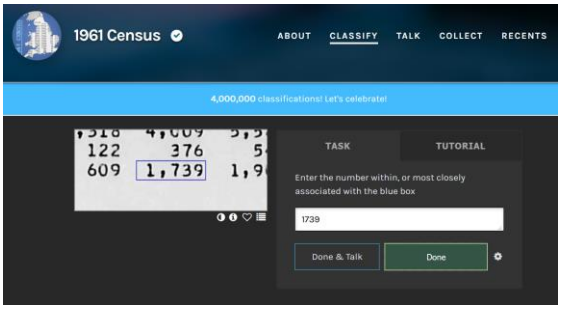

Figure 4: Image snippet of table cell as shown on Zooniverse platform.

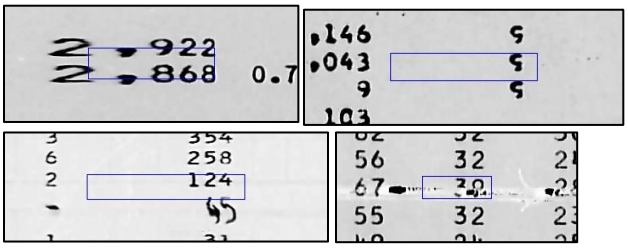

Figure 5: Examples of difficult or unclear cells.

5 The 1961 Census Zooniverse Project

This section starts with an overview of the project's web presence on Zooniverse [15]. Then, the project outcome is presented in form of results and statistics.

Zooniverse [16] is an online platform claiming to be the "world's largest and most popular platform for people-powered research". In contrast to commercial solutions (e.g. Amazon Mechanical Turk) it uses a philanthropic approach with free projects and unpaid volunteers.

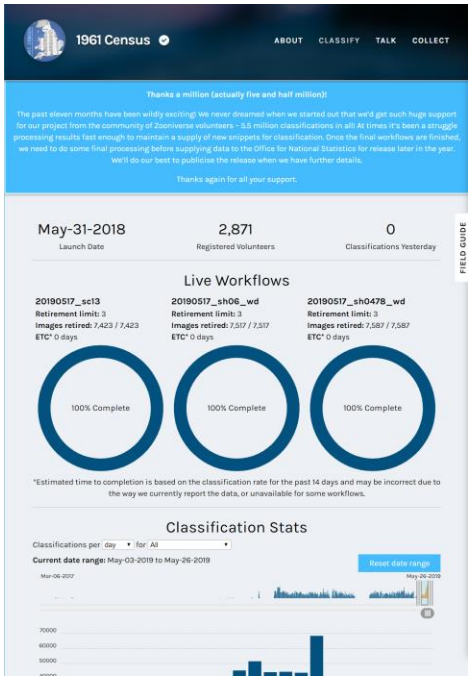

Figure 6: 1961 Census Zooniverse project page.

5.1 Project Details

Zooniverse [16], an open platform, allows anyone to create a project with an initial limit of 10,000 subjects (i.e. uploaded images). However, projects will only become visible to the public after a review process. If the project is of philanthropic nature, Zooniverse is also likely to increase the subject limit.

Data and processes are organised as:

- Subjects: The images to be presented to the crowd.
- Subject sets: Collections of images.
- Workflows: Tasks for the crowd and for selected subject sets.
- Classifications: Completed user tasks for subjects.
- Retirement count: Number of classifications by different users to retire a subject (retired subjects will not be presented to more users). For the Census project a retirement count of 3 was selected (best balance of fast turnaround and reliability).
- Workflow exports: Download of classifications and metadata in JSON format.

The creation and maintenance of a project page is comparable to a website that is administrated via a content management system. There are public-facing pages and management pages:

- Public:
    a. Home: Project overview, active workflows etc.
    b. Stats: Statistics on active workflows and the project activity (see Fig. 6).
    c. About: Project details with sub-pages: Research, The Team, Results, FAQ.
    d. Tutorial, Field Guide: Step-by-step guides for the classification tasks.
    e. Classify: Classification page (where the workflow tasks are performed).
    f. Talk: Social section with discussions and user feedback.
    g. Collect, Recents: User-specific pages.
- Management (private, called "Lab"):
    a. Content management: Editing the public pages, including styling and layout.
    b. Workflows: For creating workflows and tasks.
    c. Subject sets: For creating sets and uploading data. Large subject sets can also be uploaded via a Python script.
    d. Data exports: For download of classifications of completed workflows and other project data.

Once active, the project is visible on the Zooniverse main site and volunteers can start the work. If problems occur, a project can be hidden, at which point it is only visible to invited users.

5.2 Results and Statistics

The Zooniverse project for the 1961 Census was active from July 2018 to May 2019. During that period, over 2,800 volunteers performed more than 5 million classifications (a classification is a single task in Zooniverse terminology). Fig. 7 shows the classifications per month.

Several thousand Talk messages were recorded (user feedback on Zooniverse related to specific image snippets or about the project in general). The volunteers were encouraged to use specific hashtags (e.g. #misaligned) to inform the researchers about problems.

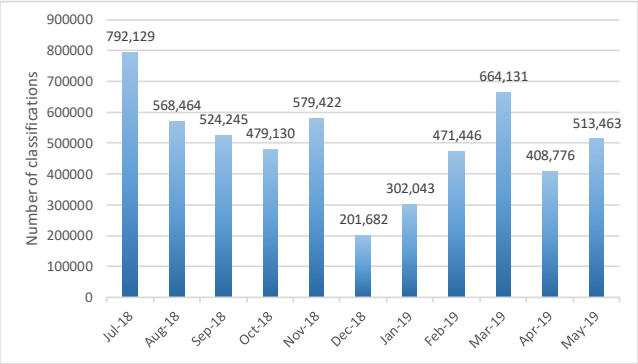

Figure 7: Number of classifications (transcriptions) per month.

The activity per individual user peaked at over 400,000 classifications performed by one volunteer. The classification count then falls approximately exponentially (a usual phenomenon in crowdsourcing). Fig. 8 shows the top ten individual users and the classification graph of the top 200 users.
The snippet creation and upload were limited (by technical constraints) to approximately 500 snippets per hour. This was a bottleneck, slowing down the correction process at peak times. Overall, there was no shortage in volunteering work. The main limitations originated at the researchers' end.

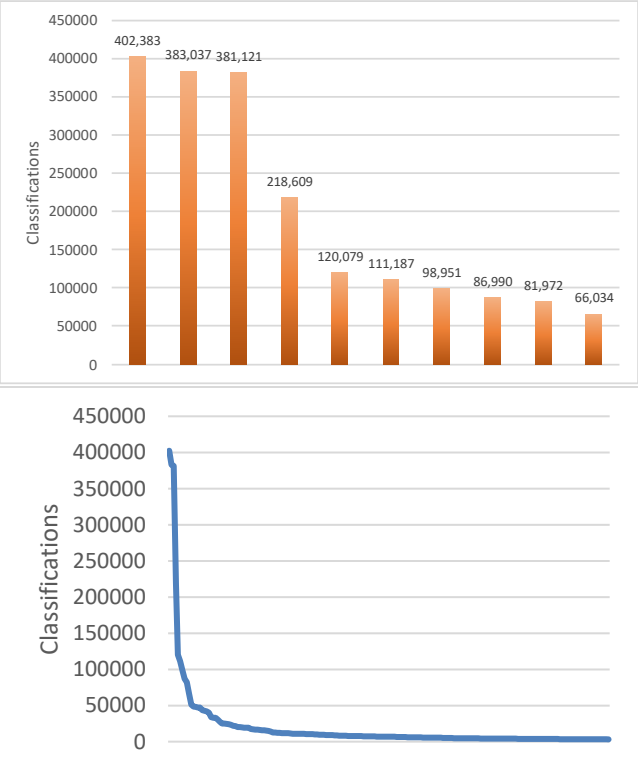

Figure 8: Volunteering distribution by classification count. (top: top 10 volunteers; bottom: top 200 volunteers).

6 Discussion

Thanks to the volunteering efforts and the support by Zooniverse, the 1961 Census project was completed within the planned project duration. All main data subsets were delivered in full. This required manual correction by an expert for less than 0.5 percent of the table cells (after crowdsourcing). The accuracy of the automated recognition (table alignment and OCR) was over 95%.
Before the public launch, convincing enough volunteers to work on the census project was a big concern. But as it turned out, the concern was unnecessary. In fact, the number of users and the number of individual classifications were surprisingly high. No promotion was required to complete the project. This is most likely due to six reasons:
(1) Zooniverse has a large existing base of volunteers and a website structure that highlights new projects but also projects that need particular help (e.g. due to inactivity).
(2) Feedback suggests that many users are interested in historical projects such as the 1961 Census.
(3) The simpler the crowd tasks the more volunteers are to be expected (see Fig. 9). This "Micro-tasking" approach requires careful consideration (What is essential to present to the users? What is clutter? How to break down complex workflows?) and extensive pre-processing. Less complex tasks attract a kind of volunteer that uses the work for relaxation and distraction (there was regular user feedback asking for more data at times where we could not keep up).
(4) User engagement is crucial to retain the most active users. This includes filling and updating all project pages and replying to forum questions. During the active period, about one person-hour per day was devoted to such project maintenance. Very little active promotion was carried out because enough users were attracted to the project. However, in brief periods of elevated need (approaching deadlines), significantly more user activity was achieved by increasing the engagement (pro-active messaging, stand-out banners and logos etc.). External promotion (outside Zooniverse) can provide a short boost in activity but can be short-lived (users do not necessarily stay active, as experienced during this project).
(5) Consistency and speed in uploading new data is key to keep the momentum. After a gap (where no subjects are available for classification) the user activity is restored very gradually. This could be noticed the most after a three-week Christmas break. Problems need to be resolved quickly to not lose volunteers to other Zooniverse projects.
(6) Power users, the most active volunteers, are important and should get special attention (direct messages etc.). For this project, the top 40 users (out of 2,800) were responsible for completing 50% of the work.

Although the general experience with Zooniverse was very positive, there were a few stumbling stones. At peak times, the data upload can fail repeatedly. This can only be dealt with by keeping a reserve of subject sets. Malicious users can cause problems by entering wrong or empty values. This can be identified by looking at the classification speed and/or results. If

values are entered at an inhuman rate or if empty values are submitted repeatedly, malicious intent can be assumed. Users cannot be blocked directly, but the Zooniverse team is very approachable and can help by making certain users' classification results not count towards the retirement of a subject.

As with any platform of a complexity as Zooniverse's, software bugs are to be expected. Within the Census project problems were encountered, but the Zooniverse developers were very responsive and fixes were provided quickly.

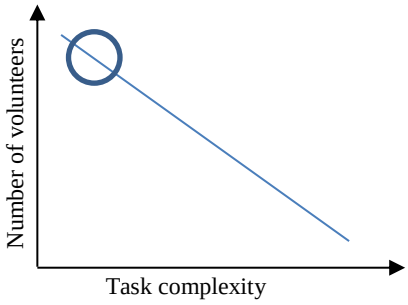

Figure 9: Relation between task complexity and number of volunteers (target area of 1961 Census project highlighted).

7 Conclusion and Future Work

There is no doubt that the 1961 Census digitisation project would not have been a complete success without the help of the volunteers and the Zooniverse team. Crowdsourcing alternatives exist (e.g. Amazon Mechanical Turk with paid workers), but the large userbase and the philanthropic approach made the project and Zooniverse a perfect symbiosis.

The concept of micro-tasking attracts more users but requires more work on the research team's side. This helped to make 1961 Census one of the most active projects on Zooniverse for the entire duration (based on classification count).

The data validation is central to identifying cells that need user action. The complete Small Area Statistics data contained more than 15 million values, which could not have been crowdsourced in the short project timescale.

The extracted tabular data was delivered to the Office for National Statistics and will be published soon.

Future work is going to include more census data. If the amount of subjects is too big for the free Zooniverse offering, an independent instance of the Zooniverse platform can be created (the Zooniverse system is open source) and linked in to the main website to gain access to volunteers and the Talk system.

The crowd tasks could be extended slightly (without overburdening users), for instance to allow multiple alternatives where cell contents are hard to read. Also, crowd-corrected cells could be used for OCR training, which was only done as a side experiment during the census project (leading to small but measurable improvements).

REFERENCES

[1] C. Clausner, J. Hayes, A. Antonacopoulos, S. Pletschacher. 2017. Creating a Complete Workflow for Digitising Historical Census Documents: Considerations and Evaluation. In *Proceedings of the 2017 Workshop on Historical Document Imaging and Processing (HIP2017), Kyoto, Japan, November 2017, pp. 83-88*. https://doi.org/10.1145/3151509.3151525

[2] Zooniverse crowdsourcing platform. https://www.zooniverse.org. Last access 09/06/2019.

[3] James Sprinks, Jessica Wardlaw, Robert Houghton, Steven Bamford, Jeremy Morley. 2017. Task Workflow Design and its impact on performance and volunteers' subjective preference in Virtual Citizen Science. In *International Journal of Human-Computer Studies, Volume 104, August 2017, Pages 50-63*. https://doi.org/10.1016/j.ijhcs.2017.03.003

[4] Trove. National Library of Australia. https://trove.nla.gov.au. Last access 09/06/2019.

[5] Digital Proofreaders. Distributed Proofreaders Foundation. https://www.pgdp.net. Last access 09/06/2019.

[6] TypeWrigth. 18thConnect. http://www.18thconnect.org/typewright/documents. Last access 09/06/2019.

[7] FamilySearch. https://www.familysearch.org. Last access 09/06/2019.

[8] Ancestry. https://www.ancestry.com. Last access 09/06/2019.

[9] Weather Rescue. University of Reading. https://www.zooniverse.org/projects/edh/weather-rescue. Last access 09/06/2019.

[10] Castaway. https://www.zooniverse.org/projects/zhcreech/castaway. Last access 09/06/2019.

[11] Southern Weather Discovery. https://www.zooniverse.org/projects/drewdeepsouth/southern-weather-discovery. Last access 09/06/2019.

[12] C. Clausner, J. Hayes, A. Antonacopoulos, S. Pletschacher. 2017. In *Proceedings of Second International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2017), Goettingen, Germany, 01 - 02 June 2017*. https://doi.org/10.1145/3078081.3078106

[13] Office for National Statistics, United Kingdom. https://www.ons.gov.uk/. Last access 09/06/2019.

[14] C. Clausner, S. Pletschacher, A. Antonacopoulos. 2011. Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, September 2011, pp. 48-52*. https://doi.org/10.1109/ICDAR.2011.19

[15] 1961 Census. University of Salford, UK. https://www.zooniverse.org/projects/dataliberation/1961-census. Last accessed 09/06/2019.

[16] Zooniverse. https://www.zooniverse.org. Last accessed 09/06/2019.