

## User collaboration in mass digitisation of textual materials

Aly Conteh

British Library, London, United Kingdom

[Aly.conteh@bl.uk](mailto:Aly.conteh@bl.uk)

Asaf Tzadok

IBM Israel – Science and Technology Ltd, Haifa, Israel

[asaf@il.ibm.com](mailto:asaf@il.ibm.com)

### Abstract

By utilising web-based collaboration tools, institutions can engage users in the building of historical printed text resources created by mass digitisation projects. The paper presents the drivers for developing such tools and identifies the benefits that can be derived by both the user community and cultural heritage institutions. The perceived risks, such as errors introduced by the users or whether users will engage with resources in this way, will be set out. The paper will present the lessons that can be learnt from existing activities, such as the National Library of Australia's newspaper website, which supports collaborative correction of Optical Character Recognition (OCR) output.

The user collaboration tools being created by the IMPACT Project (Improving Access to Text, <http://www.impact-project.eu>), a large-scale integrating project funded by the European Commission as part of the Seventh Framework Programme (FP7), will be detailed. A primary aim of IMPACT is to develop tools that help improve OCR results for historical printed texts, specifically those works published before the industrial production of books in the middle of the 19th century.

While technological improvements to image processing and OCR engine technology are key to improving access to historic text, engaging the user community also has an important role to play. Utilising the user community can aid in achieving the levels of accuracy currently found in born digital materials. Improving OCR results to this level is key to producing resources that support better resource discovery and enable greater performance when applying text mining and accessibility tools to the extracted text. The IMPACT project will specifically develop a tool that supports collaborative correction and validation of OCR results and to allow user involvement in building historical dictionaries that can be used to validate word recognition. The technologies use the characteristics of human perception as a basis for error detection.

**Keywords:** Digitisation, OCR, User Collaboration, IMPACT

## 1. Introduction

### The Challenge

The digitisation of historical text resources and use of sophisticated software tools to translate the images of text into machine-readable text has transformed the way researchers engage with these types of resources. CENL (Conference for European National Librarians) surveyed its members in 2008, revealing an expectation of a 350% increase in the digitisation of historical books and newspapers between 2006 and 2012, which would make them the most popular type of material being digitised[1]. The benefits of OCR (Optical Character Recognition) in the digitisation workflow were recognised but the experience at the British Library in their project to digitise 19<sup>th</sup> Century newspapers indicated that there were issues in the quality of the OCR text with, on average, over 20% of the text on a page not being correctly recognised [2]. Many factors influence poor performance, such as quality of the original material, storage practices, and the fonts and languages used. Current OCR is tuned to processing modern printed text and so there is a need to improve the performance of OCR tools when dealing with historical texts.

Achieving 100% percent or even the 99.9% accuracy that is usually specified through completely automated solutions will therefore be difficult to achieve for historical text. Indeed, when accuracies of this level are required then the preferred solution is to get human operators to re-key the data. While this produces good results, it is not scaleable and when institutions are digitising millions of pages the costs are prohibitively high. The cognitive ability of the humans make them ideally suited to task of recognising text that computers cannot. So there is a need to harness that power in a way that can scale to support projects that are digitising millions of pages of text. The ideal solution is to partner the strengths of the computer and humans to achieve our goals for accuracy.

## 2. The Power to be Harnessed

The advent of technologies which carry the moniker Web 2.0 technologies, the interactive web, has meant users are not just presented with information but play a full part in the creation, enhancement and semantic mark-up of information. Amazon, Wikipedia, Youtube and Facebook all provide evidence that the web community has fully embraced this paradigm, indeed Flickr recently announced that the 4 billionth photo had been uploaded [3].

The ability to harness the millions of users who interact with web-based cultural heritage resources to fill that gaps that OCR software leaves behind is attractive, but there remain questions as to whether users will engage to the same extent in collaborative correction.

The concept of user collaboration in the clean-up of OCR text is not a new one. Distributed Proofreaders have used a volunteer force of over 4,000 people to correct the OCR text of over 16,000 titles in a nine-year period [4]. While this approach

demonstrates that users are willing to engage in such activity for the benefit of the community, it is not a mainstream activity routinely deployed in the digitisation workflow of cultural heritage institutions. There is a need to significantly increase user throughput to match the levels of digitisation that is currently under way.

Two recent initiatives provide further demonstrable proof that user collaboration is a key tool to resolving the accuracy gap over purely computational approaches. Indeed a fusion of the computational approach of modern computer software allied with the cognitive power of the human brain can help us achieve the results we seek.

### Case Studies

reCAPTCHA [5], which has a tag line of “Stop spam, Read books”, is the result of a project undertaken by the School of Computer Science based at Carnegie Mellon University. CAPTCHA stands for Completely Automated Turing Test to Tell Computers and Humans Apart, and are text forms used by websites to prevent automated programs from accessing websites for malicious purposes. reCAPTCHA uses computer-unrecognised OCR results for these text forms, and thereby uses human interaction to improve the initial OCR results for historical documents.-There are no published statistics on the number of words that reCAPTCHA has corrected but over 200 million CAPTCHAs are solved every day [5] and almost one billion reCAPTCHAs were solved in 2007/2008[6].

In July 2008 the National Library of Australia released a public beta of the Australian Newspaper service [7] which supported public collaborative OCR text correction. It was the first service of its kind. It was a low-key launch with a desire to get feedback from early users to see what they required from such a service and how they would engage with the service. One part of this approach was to leave moderation of user changes to the community.

In the first 6 months of the project the usage of the service was as follows:

- 2,994 registered users
- 2.2 million lines of text corrected
- 104,000 articles corrected

The verdict of both the National Library of Australia and the site’s users on the first 6 months of the service is overwhelmingly positive. The issue of users making malicious changes did not surface and no vandalism of text was detected. Indeed, users validated the community moderation approach by explicitly stating that users should moderate each other, rather than the Library moderating, and have the ability to report and correct issues.

These two examples demonstrate that user communities are a power to be harnessed in improving the quality of OCR text.

### 3. The IMPACT approach to Collaborative Correction

IMPACT (Improving Access to Text, <http://www.impact-project.eu>) is a large-scale integrating project funded by the European Commission as part of the Seventh Framework Programme (FP7). The project commenced in January 2008 with the following objectives:

- Develop OCR software and technologies which exceed the accurateness of current software significantly.
- Provide a software system which will allow the realisation of new concepts of collaborative correction (in order to lower the costs for full featured full-text) by taking up and integrating Web2.0 concepts
- Develop language tools and lexica in order to provide access to historical texts independently of historical variants of a given language.
- Support adopters of these tools so that more European historical lexica can be built.
- Develop a number of smaller modules such as image enhancement and segmentation toolkits, functional parsers, etc. in order to support the automated text recognition and/or access to historical text.

It was recognised that while the project could seek to advance the state-of-the-art for language and OCR tools, there was a need to provide advanced solutions in engaging users to improve the word accuracy of digitised historical texts. The context for this is the i2010 vision of a European Digital Library: an ambitious plan for large scale digitisation projects that will transform Europe's printed heritage into digitally available resources.

Building on the experience of such initiatives as reCAPTCHA and the Australian National Library's Newspaper project, IMPACT will develop an alternative approach to user collaboration in the clean-up of OCR text which will increase the power of these types of tools.

The tools will allow for involvement of the general public in validation and correction of OCR results. These tools will be based on the SmartKey idea described below. This technology uses the characteristics of human perception as a basis for error detection. The result is a data acquisition procedure that is very efficient and virtually error-free.

The OCR engine recognition process concludes with a rating score for each character. These scores are further refined by the spell checker, which can either increase or decrease the probable success of individual characters. Based on such probabilities, all the characters are classified as Sure (characters with a high enough probability), Medium or Unsure (characters with a low enough probability). While Sure characters need no further verification and can be accepted automatically, and Unsure characters are sent for manual data entry, the Medium characters are sent for fast verification via "carpets".

In a "carpets session", all the Medium characters from different sources (possibly different pages, chapters or even books) are sorted in alphabetical order. For example, all the characters that were recognised as an 'A', but with a low score rate, are grouped together in a single "carpet". It has been shown to be very easy to identify the

few errors and thus automatically approve the other characters as valid ‘A’s. Hence, instead of keying in 100 characters, it is sufficient to point the mouse at few errors and the others will be automatically deduced. As a result, the validation process becomes very fast and effective. Figure 1 provides an example of a carpet session where the user is asked to identify all items which are not the letter e.

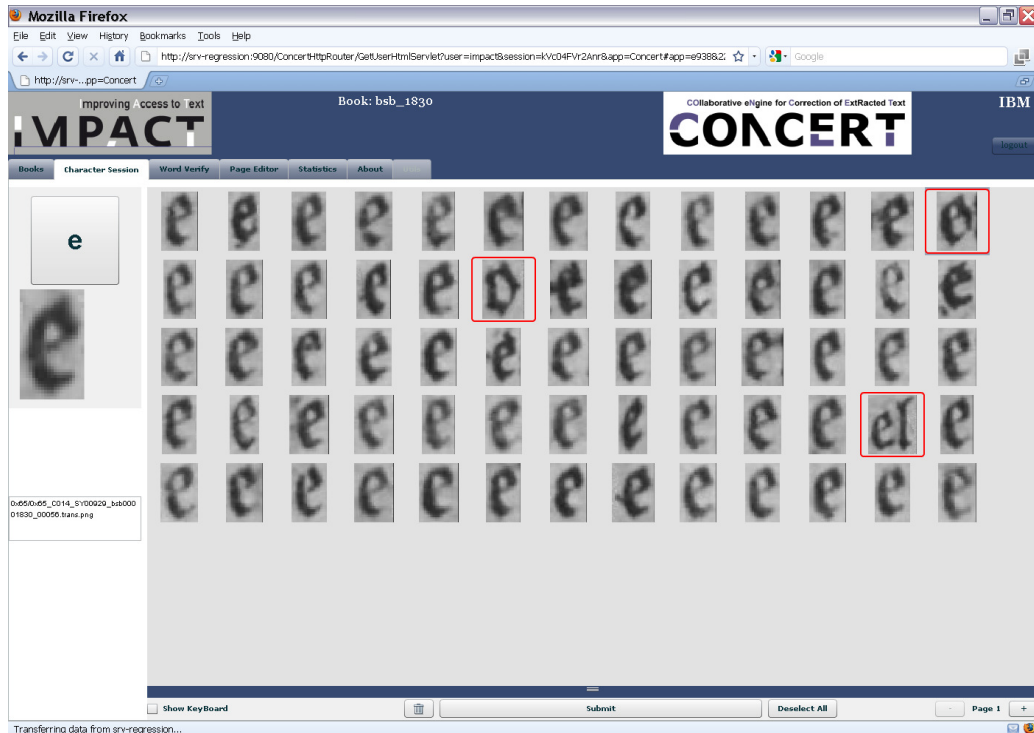


Figure 1 A View of a carpet session

At the end of this process, some words may remain unrecognised. Indeed, if image quality is very poor, it maybe impossible to recognise a given character without the context of the entire word or sentence. For these infrequent cases, word-based data entry will be introduced and context information made available as necessary. In this way the user can also add words to the dictionary supporting the OCR process by identifying words which are not currently included.

This dictionary is part of an Adaptive OCR Engine and user collaboration will therefore not only correct words but train and enhance the engine’s vocabulary and language analysis features. The system will dynamically decide how much manual intervention is needed to achieve a certain level of accuracy.

Quality monitoring will be enabled by feeding known errors into user “carpets” and seeing whether the user detects those errors. If the user doesn’t, their results will be weighted as less accurate than those of someone who does. In extreme cases, the user may be defined as a malefactor and his/her contribution discarded altogether. Quality monitoring will be done online to facilitate the adaptive utilisation of the results.

To summarise, the IMPACT approach to collaborative correction involves:

- creation of a data validation/correction application that is simple and intuitive enough to be attractive to untrained users and yet effective enough to ensure high productivity
- developing a powerful control system capable of analysing and segmenting books and other documents into individual small jobs, ensuring successful job completion, and reassembling the final result
- creating a web-based application suitable for mass volunteer participation.

## 4. Summary

Large-scale digitisation of historical printed text resources is a reality and is transforming the research landscape for this type of material. With the research benefits there are weaknesses, a key one being the quality of OCR text.

Advances have been made and will continue to be made in advancing the state-of-the-art in the automated translation of printed text into machine-readable text, and IMPACT is active in this area.

Significant benefit can be derived by using humans to bridge the gap between what can be done in an automated way and desired accuracy levels. Examples of this have been deployed with great success and will significantly improve digitised resources. IMPACT is introducing a new paradigm in this area, where not only is the text corrected but the user input is used to improve future OCR and language processing in a seamless manner.

Material that is born digital can have extremely high levels of accuracy, user collaboration will allow us to approach those levels of accuracy for historical texts in a cost effective manner.

## 5. References

[1] Hans Petschar, et al. “EDL Report on Digitisation in European National Libraries 2006-2012” February 2008. Available at < [http://www.cenl.org/docs/report\\_digitisation\\_nls.pdf](http://www.cenl.org/docs/report_digitisation_nls.pdf) >

[2] Simon Tanner, Trevor Muñoz, Pich Hemy Ros. “Measuring Mass Text Digitization Quality and Usefulness” D-Lib Magazine. July/August 2009, vol. 15 no 7/8 < <http://www.dlib.org/dlib/july09/munoz/07munoz.html> >

[3] < <http://blog.flickr.net/2009/10/12/4000000000/> >

[4] < <http://www.pgdp.net/c/> >

[5] < <http://recaptcha.net/learnmore.html> >

[6] < <http://markmail.org/message/orklsq2e4csqr2s4> >



MINISTERO  
PER I BENI E  
LE ATTIVITÀ  
CULTURALI



[7] Holley, Rose (2009) “Many Hands Make Light Work: Public Collaborative Text Correction in Australian Historic Newspapers.” ISBN 978-0-642-27694-0. Available at <[http://www.nla.gov.au/ndp/project\\_details/documents/ANDP\\_ManyHands.pdf](http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf)>