LIBER QUARTERLY 20 (1), September 2010 – ISSN: 1435-5205. P119–128
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

User Collaboration for Improving Access to Historical Texts

Clemens Neudecker

Koninklijke Bibliotheek, Prins Willem-Alexanderhof 5,
2595 BE, The Hague, The Netherlands,
clemens.neudecker@kb.nl

Asaf Tzadok

IBM Israel – Science and Technology Ltd,
Mount Carmel, Haifa 31905, Israel,
asaf@il.ibm.com

Abstract

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

The paper will present the work of the IMPACT (Improving Access to Text) project, a large-scale integrating project funded by the European Commission as part of the Seventh Framework Programme (FP7). One of the aims of the project is to develop tools that help improve OCR results for historical printed texts, specifically those works published before the industrial production of books from the middle of the 19th century.

Technological improvements to image processing and OCR engine technology are vital to improving access to historic text, but engaging the user community also has an important role to play. Utilising the intended user can help achieve the levels of accuracy currently found in born-digital materials. Improving OCR results will

allow for better resource discovery and enhance performance by text mining and accessibility tools. The IMPACT project will specifically develop a tool that supports collaborative correction and validation of OCR results and a tool to allow user involvement in building historical dictionaries which can be used to validate word recognition. The technologies use the characteristics of human perception as a basis for error detection.

Key Words: digitisation; OCR; user collaboration; correction; crowd sourcing; IMPACT

1. Introduction

In recent years, advanced libraries all over the world have been setting the pace for mass digitisation of entire collections. Huge quantities of scanned images from a diverse set of domains are being added to their institutional repositories on a daily basis. However, in the undertaking of mass digitisation projects, especially when dealing with historical printed material, many challenges are yet to be met. One of the most important challenges involves the accurate transformation of digital images into high-quality searchable text that researchers require to make proper use of these rich resources.

Although several commercial software tools are available for Optical Character Recognition (OCR) that already achieve very accurate results on modern prints, none of them perform very well when applied to historic source material. There are many reasons for this: besides the large variation of typefaces and fonts as well as complex layouts, another challenge is posed by the use of historical spelling variations. Also, distortions can be found in the text that are caused by the item's storage conditions (warping, discolouration, mould, shrinkage, fading etc.) and usage (folds, tears, annotations, stains, repairs, holes etc.).

The EU funded IMPACT project — Improving Access to Text[1] — aims at addressing these challenges by developing a variety of software tools to enhance the digital image, improve the state-of-the-art OCR software and enrich the results of text recognition by making use of lexical resources and language technology. But tools do not always suffice. Some historic documents are so complex or have been captured in such a way that efforts to improve them in an automated way are to no avail. That is when only correction and validation

by human beings can ensure that the final result will be close to 100 percent accurate as compared to the original content. Human beings have the cognitive abilities to recognise text that no computer software can read. But there efforts have to scale up to the millions of pages that are digitised every day.

2. A Growing Digital Collection

In the four next years the KB, National Library of the Netherlands, will concentrate its efforts on building a digital library aimed at providing every Dutch citizen access to all digital and printed publications that are published in and about the Netherlands. In its role as a national library, the KB will also foster the establishment of a digital information infrastructure for the scholarly record. In order to realise its objectives, the KB has established five strategic priorities:

1. Offering every citizen access to all resources published in and about the Netherlands.
2. Improving the national information infrastructure.
3. Guaranteeing long-term access to digital information.
4. Maintaining, presenting and strengthening its collections.
5. Developing the KB into a challenging organisation and an attractive employer.

To achieve these goals, a coherent programme for mass digitisation is one of the key factors. Already from the middle of the nineties the KB gained experience in digitisation by means of a large number of projects and programmes. In its Strategic Plan 2010–2013, the KB has made a clear choice to concentrate its efforts on the digital future (KB, 2010). The digital library offers great opportunities to improve and extend the service to customers — be they scholars or other customers. By large-scale digitisation of the paper collection and by offering digital access to the collections to everyone, everywhere, the barriers preventing customers from making use of the services of the KB is lowered considerably. It is regarded essential that information becomes easy to find — and text that is not digital is virtually invisible to the user on the internet. That is why the KB and a number of partners initiated the IMPACT project in 2008 — to significantly improve the processes that lead to highly accurate searchable text, accessible to everyone through the internet via the digital library.

The first goal of the KB's Digital Library Programme is to make ten percent of all publications published in and about the Netherlands available in digital form by 2013. In the long term 100 percent of these documents should become available in digital form, including quality and availability control.

Important progress towards this goal was made when the Historical Newspapers Website recently became available, providing users with free access to (currently) more than one million Dutch newspaper pages from 1618 to 1995; at the end of the project in December 2011 some eight million pages will be available. Other important digitisation projects at the KB include 'Staten-Generaal Digitaal/Dutch Parliamentary Papers', covering a range of parliamentary documents from the period of 1814–1995, the 'Dutch Print Online' project, which gives free access to a total of two million book pages from 1781–1800 and the Periodicals Digitisation Project that makes available 1.7 million pages from a selection of Dutch 19th and 20th-century periodicals.

For all of these digital collections, full text has been produced, making use of state-of-the-art OCR software. However, because of the historic nature of much of the material, or because of a complicated layout or structure the accuracy of the recognised text is not as high as required by researchers for conducting their research. Also, the conversion of scanned images into searchable text is still a costly undertaking and manual re-keying of material by service providers can easily cost one euro per page. As we are dealing with millions of pages here, funding cannot be secured.

Therefore, the KB expects that IMPACT not only improves access to these resources by providing tools supporting the entire text recognition process and leading to better OCR accuracy of historical material, but that the project also delivers a coherent programme aimed at sharing expertise and best practice in digitisation in order to reduce costs and speed up the process of mass digitisation.

3. Crowd Sourcing — a Potential to Capitalise

The rise of Web 2.0 technologies that build upon users to create content has recently moved on from completely user-driven communities like Wikipedia, Youtube or SourceForge towards involving the users more and more in

initiatives in the public sector as well. For example, a wide range of so-called 'crowd-sourcing' projects have been established during the last years that leverage the input of volunteers in creating or improving cultural heritage resources on the web.

When one looks at the challenges presented by automated text recognition of historical printed material, it becomes apparent that the huge amounts of digital resources can only be transformed into highly accurate resources by using this potential and involving volunteers in the correction of bad OCR results.

One of the most prominent undertakings in this area is the Distributed Proofreaders Project, which currently offers access to more than 18,000 titles from the public domain which have been converted by volunteers into high quality e-books and added to the Gutenberg Project repository. More than 4,000 users regularly contribute to this project; on average around 1,000 users are actively participating in proofreading at least once a week.

Recently, the Bibiliothèque nationale de France has entered into a collaborative partnership with Wikimedia France in order to release 14,000 volumes already digitised from the BnF's collection into the public domain in order to allow the more than 13,000 users involved in the French Wikisource community to correct the OCR results of this material.[2]

Another example worth mentioning is the Australian Newspaper Digitisation Program of the National Library of Australia. As part of the programme, an online service has been created to facilitate public collaborative correction of OCR results from digitised Australian historic newspapers (Holley, 2009). Started in March 2007 after an initial project phase to set up architecture and workflows, a public beta version of the collaborative OCR correction interface went online in July 2008. Until today, already 12.5 million lines of newspaper text have been corrected by more than 9,000 volunteers.

It is important to note that moderation of the users was deliberately left to the community itself; the project's secondary aim was to investigate how people would engage with the service and interact with each other. The concept of self-regulation has worked, for no vandalism of texts has been detected by the library up to date. Indeed, users observe accidental mis-corrections of others within a short space of time and correct them, reporting the trust and

respect given to them by the library as one of the main motivating factors for engaging with the system.

The above examples showcase the potential that can be leveraged by giving volunteers the opportunity to contribute to the task of providing access to cultural heritage; such involvement results in highly accurate text resources from mediocre to bad OCR results. The participation of the public can help fill the gap in public institutions that is caused by insufficient resources to perform extensive quality checks and corrections on such large quantities of text. It can also provide libraries with an opportunity to establish direct interaction with large user groups.

4. The Concept of Collaborative Correction in IMPACT

Although the initiatives listed above have shown astonishing results, the very scale of the daily output of mass digitisation projects requires a more efficient approach to collaborative correction, if only to maintain user involvement and prevent participants from being discouraged by the gigantic challenge still ahead of them.

The focus of the IMPACT approach to collaborative correction was therefore to improve the overall efficiency of the correction process and re-purpose the contribution of the users for a number of tasks, thereby drastically improving overall productivity. The web-based CONCERT (COllaborative eNgine for Correction of ExtRacted Text) platform developed in the course of IMPACT intends to significantly improve overall throughput and encourage public participation by letting volunteers know that their efforts are highly valued and will not be wasted.

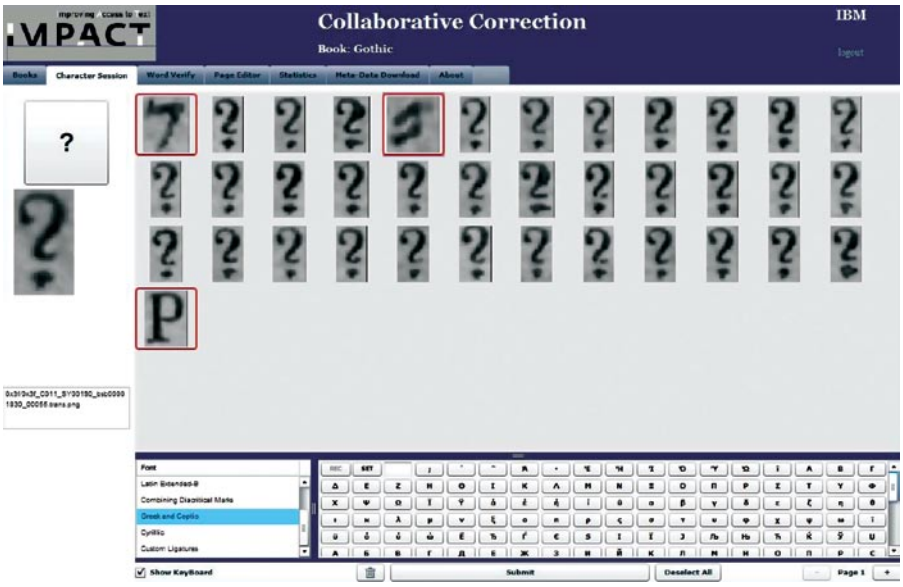
The system itself consists of an integrated adaptive OCR engine for initial text recognition, followed by three layered modes for performing the actual correction on the obtained results. After uploading the scanned images of a complete volume and processing of the pages by the OCR engine, the first stage that is presented to the user is the so-called 'carpets session' (Kluzner *et al.*, 2009).

A carpet consists of a screen displaying all occurrences of a single character found in a certain book that have been recognised by the OCR engine with a confidence level below a particular threshold (e.g. in the range of 90–99%),

usually referred to as 'suspicious characters' by commercial OCR engines. Instead of correcting every single one of these characters separately, the operator can easily flag just those characters that have not been recognised correctly by the engine and the remaining ones will automatically be approved. This means that in order to correct, say, one hundred suspicious question marks, the user only indicates the obvious errors, which usually results in just a few mouse clicks on a carpet and speeds up the correction process tremendously. User feedback obtained from such a session is also automatically used to further train the underlying OCR engine in the font type.

The notion of a carpet is illustrated in Figure 1.

Fig. 1: Screenshot of a character session: the 'carpet' contains '?'; the image shows three OCR errors flagged by the operator.



It may appear that in the character session even a human being is not capable of correctly indicating a character as being either a question mark or a capital letter P, for example due to noise introduced in the capture. In this case, the character can be rejected by the operator who then moves on to the next level in the application which is the word session. The word session presents the operator with the contextual information on the word level that enables

him to correctly identify correct words. A clipping from the original image is shown with one or several interpretations, most often already containing the correct interpretation, which is then simply selected by the user with a single mouse click. The correct word is then also automatically added to a dictionary in order to enhance the engine's vocabulary and language analysis features.

Finally, if the information given to the operator on the word level is still insufficient (e.g. due to hyphenation), the final step is to move on to the page level where the complete page image is displayed with the recognised text merged between the lines. In this session, the operator can conveniently correct words, split merges or merge erroneously separated strings or paragraphs. Additionally, the operator can identify and define new text areas which were not detected by the OCR Engine.

The system will have online operator monitoring to detect potential pranksters and also motivate the most prolific volunteers. User monitoring will happen by from time to time introducing known errors into user carpets and checking whether the user correctly identifies these as errors. This allows for the spotting of users who do sloppy work but also of those high-quality contributors qualifying for potential awards.

Altogether, the Concert system significantly advances the state of the art in collaborative correction by applying a highly efficient concept that scales up with the requirements of mass digitisation, as well as by utilising human work for a number of tasks: not only the correction of OCR results, but also training of the adaptive OCR engine and enriching the dictionary with additional entries.

5. Summary

In view of the enormous amounts of text that are to become available in digital format during the next years and the difficulties even elaborate OCR software has in dealing with historical material, ways must be found to distribute the work on many shoulders. There is a huge potential to be leveraged from crowd sourcing to improve the accuracy of digitised texts. The only remaining question is: how to utilise this potential in an effective and rewarding way.

A key factor in this undertaking lies in providing concepts and tools that can scale up to the challenge and use the resources in a way that provides the

users with an instant feeling of success rather than of being inefficient and the work being tedious.

One of the major aims of the European Commission's i2010 vision of the European Digital Library, 'to bring cultural heritage on par with born digital material' will become much more attainable when users all over the world can contribute to this vision themselves. In the CONCERT tool for collaborative correction, IMPACT makes available a powerful means and incentive to do so.

References

Abdulkader, A. and M.R. Casey (2009): 'Low cost correction of OCR errors using learning in a multi-engine environment', ICDAR 2009 Conference (26–29 July, Barcelona, Spain), pp. 576–580.

Balk, H. and L. Ploeger (2009): 'IMPACT: Working Together to Address the Challenges Involving Mass Digitization of Historical Printed Text', *OCLC Systems & Services*, 25(4), 233–248.

Conteh, A. and A. Tzadok (2009): 'User collaboration in the mass digitisation of textual materials', Cultural Heritage Online Conference (15–16 December, Florence, Italy), http://www.rinascimento-digitale.it/eventi/conference2009/proceedings-2009/conteh.pdf.

Holley, R. (2009): ´Many Hand Make Light Work: Public Collaborative Text Correction in Australian Historic Newspapers', http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.

KB (2010): KB - National Library of the Netherlands, Strategic Plan 2010–2013, http://www.kb.nl/bst/beleid/bp/2010/index-en.html.

Kluzner, V., A. Tzadok, Y. Shimony, A. Antonacopoulos and E. Walach (2009): 'Word-Based Adaptive OCR for Historical Books', ICDAR 2009 Conference, (26–29 July, Barcelona, Spain), pp. 501–505.

Websites Referred to in the Text

Australian Newspaper Digitisation Program, http://www.nla.gov.au/ndp/

Digitalisering tijdschriften (Periodicals digitisation), http://www.kb.nl/hrd/digitalisering/tijdschriften/index.html

Distributed Proofreaders Project, http://www.pgdp.net/c/

Dutch Print Online, http://www.dutchprintonline.nl/

Gutenberg Project, http://www.gutenberg.org/

Historische kranten online (Historic newspapers online), http://kranten.kb.nl

Staten-Generaal Digitaal/Dutch Parliamentary Papers, http://www.statengeneraaldigitaal.nl/

Notes

[1] IMPACT project (2008–2011): Improving Access to Text. IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National library of the Netherlands. http://www.impact-project.eu.

[2] ´Wikimédia France signe un partenariat avec la BnF´, http://www.wikimedia.fr/wikim%C3%A9dia-france-signe-un-partenariat-avec-la-bnf.