Country-scale infrastructure for creation of full text versions of historical documents from Polish Digital Libraries

Adam Dudczak, Miłosz Kmieciak, Marcin Werla
Poznań Supercomputing and Networking Center
{name.surname}@man.poznan.pl

Abstract.
With its beginning in late 1990s development of Polish digital libraries came to point were 79 such services give access to more than 850 000 objects, mostly digitised cultural heritage materials. Information about these objects is gathered by the Digital Libraries Federation (http://fbc.pionier.net.pl, DLF) portal. This portal gives convenient access to unified metadata search mechanisms which simplifies life of both users and librarians involved in digitisation. The DLF portal plays also important role as an intermediary, passing cleaned and normalized metadata from Polish digital libraries to international services like Europeana (http://europeana.eu).

Having one common and highly interoperable access point to metadata is the first and inevitable condition to increase the usage of digital resources available in digital repositories. But even the most detailed metadata cannot replace machine-readable textual content of described object, especially in the context of humanities research. The challenge to be able to provide such content in an automated way and on a country-scale is addressed by ongoing works which PSNC undertakes in the SYNAT[1] project, this paper outlines current state of these works.

To have a full view on state of practices related to Optical Character Recognition (OCR) in Polish digital libraries a survey was conducted. It consisted of questions about number of objects with machine-readable text, OCR tools used and their features, methods used for quality assurance etc. Received responses covered 26 institutions responsible for creation of more than 70% of resources available in Polish digital libraries at that time[2]. Only 40% of these documents were a subject of OCR, no one does perform manual corrections on results obtained from (mostly) ABBYY FineReader. This is caused by limited human resources and the lack of tools which can integrate correction in mass digitisation workflow. Only 3 institutions reported the usage of advanced features like custom recognition profiles, which led to improvement in results of text recognition (especially in case of historical documents), but because of software limitations could not be easily integrated into digitisation workflows.

Outcome of this survey allowed to create initial list of assumptions and requirements which should be considered to improve quality and quantity of available machine-readable text in Polish digital libraries. It was assumed that tools to be developed should support correction of existing machine-readable text and improve results of OCR by simplifying the process of creation of custom recognition profiles.

[1] Polish national research project SYNAT (http://www.synat.pl/) is financed by National Center for Research and Development, grant number: SP/I/1/77065/10.

[2] The survey was conducted on September 2010

In the work described in this paper we have used Tesseract OCR engine as a base for our text recognition service and developed custom applications and scripts to simplify its training process. The most important part of this workflow is based on webapp called “Cutouts”, which allows to crowdsource preparation of Tesseract training material by allowing volunteers to work on glyph extraction from scanned historical documents. Additionally every document has associated metadata record which is in this context the source of information about the origin of each glyph evaluated by volunteers. Having such information it is possible to create custom recognition profiles which can be used to OCR similar documents.

Integration of text correction into digitisation workflows can be achieved thanks to the second developed web application named “Virtual Transcription Laboratory” (VTL). Users of VTL can upload scanned images (in future also direct import from digital library will be possible) and create transcription project on this basis. This project can be a subject of several activities of the project owner:

- uploaded image can be OCR-ed using custom or standard recognition profile, alternatively existing transcription can be imported to the project,
- the project can be published for crowd-driven correction,
- final version of transcription can be export in hOCR format.

VTL has a simple and useful transcription editor which allows to interconnect machine-readable text with its position in original scanned image. History of all changes made in the transcription is stored in the system, so the project owner can remove unwanted modifications at any time.

Results of work in VTL can be exported and used for many purposes, including the update of a full-text search index in the digital library from which the processed object came. As created transcription has also information about position of text, it can be also used to enrich the user experience.

Future developments will include work on additional crowd-driven correction tools (e.g. games) and closer integration with the infrastructure of Polish digital libraries.