Making the most of digital resources with IIIF and OCR enabled transcription toolset

Tomasz Parkoła*†[1], Michał Kozak*[2], and Tomasz Kalota[3]

[1]Poznan Supercomputing and Networking Center – Poland
[2]Poznań Supercomputing and Networking Center – Poland
[3]Wrocław University Library – Poland

Abstract

Huge amounts of digital resources are available online - datasets or services are there to be reached/processed. However, one of the challenges in utilizing these resources, is to appropriately address researcher's needs. Although many repositories provide digital assets, their metadata or content is not always relevant for research practices. In the case of scanned/digitised resources, the main challenge pertains to availability of full-text historical documents with uncommon scripts/languages. There have been several approaches to tackle this issue, including manual transcription and automated OCR/HWR techniques (e.g. FromThePage and Transcribus projects). However, there is no approach developed yet to support digital repositories in enhancing their content via transcription tool with built in character recognition, IIIF-based streams and automated training routines. In 2019 Wroclaw University Library (WUL) in cooperation with Poznań Supercomputing and Networking Center (PSNC) launched a project to extend their digital infrastructure with such an approach. Since 2011 PSNC has been developing Virtual Transcription Laboratory (VTL) - an open environment for handling scanned textual resources, executing OCR and its post correction, conducting transcription as well as training OCR engine. This new project will add features to VTL or update existing ones, so that the overall solution is able to better respond to the needs of the end users.
The main idea behind the new approach is to fully integrate digital repository with the transcription toolset, so that it is possible to enhance digital assets with full-text. This approach will stimulate cooperation between researchers and content providers - WUL will provide digital content that can be used in the research practices, while the researchers will work with the content and create/enhance its full-text. The important part is that the full-text will be fed back to the original repository as an alternative representation of the digital resource.

VTL is composed of three modules: Import and export, Text recognition as well as Transcription and annotation. Import and export interacts with external systems and deals with various formats. It can import data from IIIF manifest, TEI P5 and METS. As a result all references to content files are either to the original source (IIIF stream) or to the internal IIIF streams which original files were converted to. Text recognition module can then be used to automatically recognize textual representation of the imported images. It is based on Tesseract 4.0 that utilizes neural net (LSTM) algorithm. Transcription and annotation

*Speaker
†Corresponding author: tparkola@man.poznan.pl

module is the place where users can transcribe documents manually or correct Tesseract's recognition results. By default transcription is done on a line level, but this can be changed by the user if necessary, e.g. by adding new regions or removing/replacing them. There are multiple ways to annotate the text itself, e.g. with headers, page numbers, font style or comments. Once the transcription is ready (even partially) the reviewers can verify if the transcription is correct. They can correct errors and provide feedback to the transcriber. Finally, once the transcription is verified the system can export full-text to plain text, hOCR, TEI or eBook (PDF, MOBI, ePub).

Keywords: transcription, annotation, IIIF, OCR