READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D8.4
Large Scale Demonstrators – Zurich

Evaluation and Bootstrapping

Tobias Hodel
StAZH

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	31.12.2016
Actual date of delivery	28.12.2016
Date of last update	25.11.2016
Deliverable number	D8.4
Deliverable title	Large Scale Demonstrators – Zurich
Type	report
Status & version	in process
Contributing WP(s)	WP8
Responsible beneficiary	StAZH
Other contributors	URO
Internal reviewers	Maria Kallio (NAF), URO
Author(s)	Tobias Hodel
EC project officer	Martin Majek
Keywords	Evaluation, Large Scale Demonstrator, Archive, Transcription, Bootstrapping, Alignment, Digitizing

Contents

Executive Summary

This document gives an overview of the foundation of the involvement of StAZH in READ. Three main topics cover the work provided for READ by StAZH: First, the development of strategies for the execution of mass-transcriptions. Second, the evaluation of HTR- and alignment processes. Third, the delivery of documents as data for training and evaluation for READ partners. All three parts are reflected and described in this paper as they are currently carried out.

1 Introduction

The state archive of Zurich (StAZH) is one of the four large scale demonstrators, testing, implementing, and using the technologies developed in READ in a typical environment (archive, libraries, etc.). StAZH has been digitizing documents for more than ten years in order to make documents accessible for users or for long-term preservation. For six years, important series of documents of the archive are made accessible by adding full text to the digitized materials. Furthermore selected documents are being prepared and published for scholarly editions. Hence, the archive has gathered expertises for the manual extraction and description of text in digital environments.

For READ the transcriptions as well as the digitized images are prepared for training, evaluation, and benchmarking of the software as well as the developed algorithms. By assessing the algorithms on different, esp. larger scales, it will be possible to estimate costs for implementation and execution of the technology in typical institutional environments. Scholars as well as computer-scientists gain insight in the consequences, the benefits, as well as the risks of the application on larger scales. A subordinate part of the task is the enlargement of the available material (esp. for training and evaluation), by the implementation of tools connecting images with texts on line basis (developed mainly in WP 7.2.).

2 Strategies, Evaluation and Bootstrapping

The task is carried out aiming at three trajectories:

First The identification of strategies to execute (mass-)transcriptions in institutional settings.

Second The evaluation of processes developed in READ for roll-out on larger scales, focusing on the identification of resources needed and resources saved for long-term mass-transcription projects.

Third The bootstrapping of text-image alignment tools helping to generate ground-truth that can be used for training and evaluation of HTR-models, as well as for purposes of benchmarking.

2.1 Strategies for Transcription in Large-Scales

For the next year the set-up of a larger transcription project is in planning. Currently we are identifying documents and series suitable for such tests as well as methods of implementation into the structures of the archival information and dissemination systems. For that matter graphical user interfaces of READ (developed in WP 4.3.) will be used and tested. The **strategic view** on transcription projects will benefit from this approach as well as approaches to the business plan (see also deliverable 3.1.).

For a large transcription project,[1] currently in its final phase (end mid-2017), all transcriptions were produced by experienced transcribers (recruited students). For the 198'709 pages the equivalent of 30 years of work had to be invested. The cost of the endeavour is roughly at 3 million Swiss Francs. The project will be the basis for future comparisons of costs involving HTR versus traditional methods. Such tests will be possible in year 2 and 3 of the project as soon as more data regarding the quality of the HTR-processes becomes available and more experience in dealing with different sets of documents is gained. The same goes for the gained experience for a larger set of transcriptions treated with the methods developed and corrected in a later step.

In year 3 the ingested sets of documents will be made searchable using Key Word Spotting (KWS) tools in order to test the accuracy and the possibility to find words and/or strings in larger scales. KWS is promising for archival institutions in order to make large collections searchable and therefore usable for their target groups. Beside technical aspects of the compatibility of the used algorithms, it needs to be evaluated what target groups are expecting as results of KWS and how they experience problematic results (such as false-positive).

2.2 Evaluation of HTR-processes

For the first phase of the **evaluation** process ground-truth was produced manually in order to train first models of neural networks. From the first models as well as tests of other models, first insights of the ceiling of the methods applied could be gained. In several scenarios Character Error Rates of less than 5% could be achieved and first patterns determining the quality of the outcome identified. The results are comparable to the experience of other involved institutions (project partner UCL, Transcribe Bentham and MoU partner University of Greifswald) dealing with manuscripts. The discussion about achieved results was opened towards the end of the year by contacting and gathering information from institutions working with algorithms developed in READ. A presentation with comparison of the achieved results including the evaluation of all involved partners is planned for the "Digital Humanities 2018" (Mexico City).

The application of the first model for StAZH resulted in acceptable results (5-20% Character Error Rate) for texts written in similar but not the same writings/hands. Of less importance for the moment was the evaluation of the recognition by involving spe-

[1] See "Transkription und Digitalisierung der Kantonsratsprotokolle und Regierungsratsbeschlüsse des Kantons Zürich seit 1803", URL: http://www.staatsarchiv.zh.ch/internet/justiz_inneres/sta/de/ueber_uns/organisation/editionsprojekte/tkr.html.

cialized dictionaries. This issue will be addressed in the next phase. Identified problems concerning the results of the recognition process are to be found in the quality of the provided scans as well as in the tasks of layout detection: Short, intertwining, or wrongly assumed baselines are posing problems. As for the pure handwritten text recognition, the quantity of training material for the recurrent neural networks increases the results and furthermore makes recognition more robust. Enough data for ground-truth was prepared to produce another two to three models, which will broaden the approach of the evaluation and help getting more insight into the stability of the technologies applied in the next two years.

Another way to get evaluation data for HTR processes is the dissemination of the software Transkribus (specialized GUI). In order to collect feedback from memory-institutions as well as interested scholars, several workshops and talks (part of WP 2.5.) were held. As a result, new MoU partners were acquired and documents are currently prepared or ingested. The contact with partners and institutions will be broadened and a flow of information invoked to get to know more about specific needs and problems.

2.3 Bootstrapping of alignment tools

For **bootstrapping** and text-image alignment the dialogue with URO has been opened and suitable documents have been shared by uploading roughly 200'000 pages with alignment of text and images on page basis. First tests show that the alignment is possible but the error-rates need to be evaluated in depth. The already produced ground-truth will support this task in the next phase. As soon as stable processes are being achieved, ground-truth will be produced semi-automatically and the process will be presented to the public in order to be able to ingest more data for Ground-Truthing and to train more models.

For year 2 and 3 data sets of edition projects (mainly carried out at StAZH) will be made accessible to READ. As soon as stable and correct alignment is possible, training of HTR-models for documents stemming from 16th to 18th century will be supported. Since most of transcriptions and edition project as of today are not aligned to line level but only page level, the tools developed will make vast numbers of documents available for HTR-training (and testing).

2.4 Further Involvement in READ

The produced ground-truth of as well as the documents provided by StAZH will be used for competitions (part of WP 3.7.) and tasks in document understanding (part of WP 6.5.). Both tasks are supported with domain knowledge of StAZH. Especially document understanding could be exploited in archival contexts, since the understanding of documents leads to the (semi-)automatic production of metadata for archival information systems. Those databases are currently filled by specialized archivists: A task that could be supported and partly automatized by the developed approaches, independent of the nature of the documents (print or handwritten).