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	31.12.2017
Actual date of delivery	24.11.2017
Date of last update	20.11.2017
Deliverable number	D8.4
Deliverable title	Large Scale Demonstrators – Zurich
Type	report
Status & version	final
Contributing WP(s)	WP8
Responsible beneficiary	StAZH
Other contributors	URO
Internal reviewers	Maria Kallio (NAF), Eva Lang (ABP), 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. Five 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 handwritten text recognition (HTR) and alignment processes. Third, the delivery of documents as data for training and evaluation for READ partners (closely tied to number two). Four, the networking across archives and memory institutions with similar needs/work in similar areas. Five, the dissemination activities carried out in order to inform and enable interested projects to work with tools developed within READ and contribute themselves by providing documents for training. All five 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 (LSD), 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 expertise 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 into 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 D. 7.20.).

2 Main task: Strategies, Evaluation and Bootstrapping on Large Scales

The main 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 (GT) that can be used for training and evaluation of HTR models, as well as for purposes of benchmarking.

2.1 Bootstrapping of alignment tools

For **bootstrapping** and text-image alignment, the dialogue with URO yielded promising results and two volumes of aligned texts (about 1200 pages) from the StAZH corpus. After treatment with text-to-image (T2I) the line-finder was able to identify and correctly assign more than 90% of the lines. Subsequent training of the model led to a character error rate (CER) of slightly above 5% on a test set (the test set contains the same hands, but particular pages were not part of the training). This proved that the identified ceiling of the error rate can be reached with models spanning multiple hands (see also below 2.2.2).

Second, more than 100'000 pages from StAZH/TKR have been treated with a specialized T2I routine that builds on image files (JPEG) and transcriptions (in simple TEI XML).[1] A prerequisite was the identification of page numbers for image files, this has been outsourced (using money foreseen for Ground Truthing). The results are more than 100'000 pages of StAZH documents in PAGE XML on Transkribus that can be used for further training and testing. They will also be made available via ZENODO in year three.

Further, the T2I has been implemented in Transkribus X, where it has been tested by StAZH using different use cases from different fields:

- "Parzival" manuscripts (600 pages): Medieval hands, transcriptions provided by a digital edition project at the University of Berne (MoU partner). Matching of more than 95% of lines possible. Trained model resulted for one hand in CER below 4%.

- "Semper" manuscripts (500 pages): Semper was a famous architect of the 19th century, transcriptions provided by a digital edition project at the ETH Zurich (MoU partner). Matching of about 90% of lines possible (due to difficult layout slightly below other matching jobs).

- "Itinera Nova" (about 10'000 pages): Crowd sourcing project in the city of Leuven (BE) regarding medieval and early modern city books, data provided and supported by University of Cologne and Institut für Dokumentologie und Editorik (IDE, MoU partner). The task is due to the amount of data currently ongoing.

For year three, data sets of edition projects and earlier transcriptions will be ingested and made accessible to READ. With the stable and correct alignment, the training of HTR models for documents stemming from 13th to 18th century will broaden the scope of documents available for recognition. Since most of transcriptions and edition projects

[1] For the T2I see: D7.20.: Semi-supervised HTR Training. The TEI XML are provided as open data on ZENODO: http://dx.doi.org/10.5281/zenodo.803239

as of today are not aligned on line level but only on page level, the tool developed will make vast numbers of documents available for HTR training (and testing).

At the same time, this increases the appeal for using HTR enormously: For most writing styles transcriptions exist and can now be matched in order to train models to check their applicability. Still, digital or digitized texts are a prerequisite in order to follow the workflow described here and below (2.2.2).

2.2 Strategies for Transcription/Recognition in Large-Scales

Within the second year, three strategic partnerships were established with institutions intending to use or already using READ tools to recognize large amounts of documents.

1. Federal Archives of Switzerland (BAR): Recognition of the minutes of the Federal Council (1848-1903): About 150'000 pages, without correction (see also business implementation in D 3.2., MoU partner).

2. University Library of Basle (UBB): Recognition of oldest journal in Basle (AVIS Blatt) using recursive neural networks (rNN) for German gothic script: About 70'000 pages (MoU partner).

3. University of Fribourg/e-Codices (e-Cod): Preparation of HTR model to provide keyword spotting (KWS) for selected document collections (starting with the world renowned Carolingian codices of Saint Gall monastery: About 1'000 pages). The e-codices corpus as a whole is more than 300'000 pages (future MoU partner).

These three projects are used in order to describe needs and wants of large-scaled projects. The first project will be treated as business case, whilst the other two are using the open tools developed in READ under guidance of StAZH but without specialized technical support.

All three projects can be used as comparison to the large transcription project,[2] that was finished in September 2017 at StAZH (TKR). 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.

2.2.1 Prerequisite: Layout Detection

Thanks to a new **layout analysis** (LA),provided by URO, called "CITlab Advanced", most of the problems regarding Layout Analysis have been erased (missing lines, incorrect text regions) with regard to the recognition on the large scale:

- Regular book layouts (one or two columns) are no longer problematic;

[2] 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.

- Minutes with marginalia, indention, etc. are correctly identified;
- Text and images are correctly separated.

Furthermore, almost no lines are missed and lines do not get incorrectly split in halves. As a result the Layout Analysis (LA) is comparable to results from established (commercial) OCR engines.

2.2.2 Evaluation of HTR processes

For the first phase of the **evaluation** process GT was produced manually in order to train first models of HTR. 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 CER 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 (READ partner UCL, Transcribe Bentham and MoU partner University of Greifswald) dealing with manuscripts.

In a second phase, in 2017, it could be shown that even for more than one hand the same ceiling could be reached, by providing more GT (produced with T2I, see above).

The training of 760 pages of GT (different hands, spanning 80 years) at the end of 2016 led to an average of 18,6% of errors on character level. The model can be used not only for the documents written in Zürich but most hands in German kurrent. Although from a humanities standpoint the error-rate is underwhelming, the model demonstrates that it is possible to train HTR models that are suitable not only for the recognition of one but multiple diverging hands.

The question remains, how many writers can be fitted into one model with CER rates below >10% and whether it is possible to build models broad enough to recognize writers/hands unknown. These questions will be answered within the third year of the project.

For this reason, different document sets from different times have been gathered in order to build general HTR models with focus on 19th century German Kurrent and Medieval scripts (gothic book scripts and administrative hands of the 14th and 15th century).

In the second year, the broadening of quantity in terms of HTR models but especially in terms of GT prepared using semi-spervised training methods could be achieved. In order to be able to use semi-automatic training, the models built in year one were used and eventually re-trained with more GT.

The outcome of the evaluation let us state that given enough amount of GT an acceptable CER can be reached. The goal within READ and also regarding the input of interested institutions and projects, is therefore to allocate enough GT (via T2I or by human input) to train general HTR models.

Thanks to the T2I tool by URO (see above 2.1) the production of GT has been severely sped-up and eventually led to enorm amounts of training data. Consequently, a broad and dependable HTR for similar documents could be build.

2.2.3 Evaluation of KWS processes

Due to KWS based on confidence mHTRices (see below and D 7.14.), also with imperfect HTR recognition (i.e. the German kurrent model from StAZH with 18,6% CER) a very high recall (around 99%) can be achieved. A feature that will be helpful for archives and other memory institutions that want to make large amounts of images searchable without having to go through correction processes.

For year three, three big data sets are planned to be tested regarding the application of Key Word Spotting:

- The documents of the TKR project made available by StAZH, about 150'000 pages (with the possibility to compare results with human made transcriptions);
- The minutes of the Federal Council of Switzerland by BAR;
- Medieval Documents with no available transcriptions on e-codices.ch.

Besides 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 long loading times (KWS is time consuming) as well as problematic results (such as false-positive). Theses tests will be carried out with ongoing discussions with NAF and ABP who both aim in similar directions.

2.3 Networking and further involvement in READ

In order to get to know needs as well as ideas from stakeholders (with focus on scholars and archives), StAZH is part of several dissemination activities with one of the goals being to build a network of interested stakeholders.

Within Switzerland three institutions with matching interests have been identified (see above): All of them are using the tools provided within READ to make their documents better accessible, still they are separated as different use cases.

1. BAR is interested in transcriptions of their minutes without dealing themselves with the technology. This is a typical case of outsourcing (from the archival point of view) on a contractual basis. Therefore this type of deal could be demonstrating how a future business case might look like. Also, this case was chosen as demonstrator for **Document Understanding**, in order to test how much information about specific parts (header, heading, page number, etc.) could be provided (see also D 6.14.).

2. UBB and the University of Basle are interested in a high quality recognition, to use the text with Document Understanding tools to provide the different snippets as basis for historical research. The recognition projects stands at the beginning and the text needs to be exported into other formats and systems. GT and corrections are provided by student assistants in Transkribus X.

3. The cooperation with e-Codices has just started, they are interested in providing scholars access to their vast amounts of texts using KWS. With a scholarly audience

in mind, providing false-positives (erroneously found words) is not as big of an issue as long as high recall can be ensured. Providing GT for different models (starting with Carolingian minuscules as a pilot), along with discussions about the implementation of the search are currently underway.

The three trajectories demonstrate different needs of institutions and projects, giving an idea of the variety of ways dealing with archival (and other) documents. As a lesson learned, it can be said that the tools developed are of use to all of the institutions. But each case needs to be handled separately in order to determine what outcomes are expected and what goals should be deemed realistic. Also, financially the projects differ widely: From several tenth-of-thousands of Euros spent by the institutions to do the whole recognition process, to some hours invested by students.

For the third year, the mentioned projects will be followed up and presented as use cases in order to be able to draw "best practice" reports (together with UCL, NAF, and ABP). This is one of the main goals of regular calls with other archives acting as LSD. The group also exchanges opinions and solutions for implementations of READ products in the archival workflow and the connection of archival holdings.

The produced GT of, as well as the documents provided by StAZH were used for competitions (part of WP 3.7.) as well as tasks in document understanding (part of WP 6.5. and experiments in D 7.8, D 7.11, D 7.20).

2.4 Dissemination and Allocating GT

Another way to get evaluation data for HTR processes is the dissemination of the software Transkribus X. In order to collect feedback from memory-institutions as well as interested scholars, several workshops and talks (part of WP 2.6.) were held. As a result, new MoU partners were acquired and document collections have been tested out in Transkribus. For interested groups (especially the institutions mentioned above), the flow of information has beend invoked to get to know more about specific needs and problems.

Most of the dissemination of StAZH has been carried out in form of workshops in different places from Czestochowa (PL) to Leeds (UK).[3] In order to engage scholarly discussions, talks were given on the subject of machine learning in the historical sciences (Berlin and Basle) as well as problems of theorizing machine red text.

Within a specialized dissemination group, regular calls led to a mutual understanding of who was serving what user groups in order to spread the knowledge about READ but also to get meaningful feedback.

An often encountered problem are institutions foreseeing documents only for closed access. The topic was raised in a hands-on way by organizing a hackday at the State Archives of Zurich, part of the festivities of the international archives day, inviting interested people to play around with archival data sets.[4]

[3] See the Dissemination Report D2.2. for details.

[4] See reports: http://www.netzwoche.ch/news/2017-06-12/wenn-hacker-auf-archivare-treffen; https://www.societybyte.swiss/2017/06/16/hacken-erwuenscht/.

Interested parties and projects have been approached to share their documents in order to prepare general HTR models for different time frames. The search for suitable and accessible GT is still ongoing and for year three substantial results are expected.

Due to the heavy involvement in dissemination activities. One person month has been relocated from the pool for Large Scale Demonstrators (WP8) to Dissemination (WP2).

2.5 Sub contract

According to the Grant Agreement (GA) of the project sub contracts are mainly foreseen for generating GT and for involving institutions via a Memorandum of Understanding. Details are described in the GA, p. 89f.

According to the GA StAZH has therefore involved a subcontractor for producing GT, the amount will be billed in year three.

2.6 Publications

- Abstract: Piotrowski, Michael, et al.: Virtuelle Forschungsplattformen im Vergleich: MONK, Textgrid, Transcribo und Transkribus, in: DHd 2017, Abstractband, S. 66-69.
- In print: Hodel, Tobias: READing Handwritten Documents: Projekt READ und das Staatsarchiv Zürich auf dem Weg zur automatischen Erkennung von handschriftlichen Dokumenten. In: Geschichte & Informatik, 2018.