

READ

Recognition and Enrichment of Archival Documents

D8.11

Large Scale Demonstrators

Keyword Spotting in Registry Books P2

ABP

Herbert W. Wurster, Hannelore Putz, Eva M. Lang, Wolfgang Fronhöfer,
Andrea Fronhöfer, Elena Mühlbauer

Distribution: Public

<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
Contractual date of delivery	31.12.2017
Actual date of delivery	22.12.2017
Date of last update	18.12.2017
Deliverable number	D8.11
Deliverable title	Passau – Keyword Spotting in Registry Books
Type	Report
Status & version	In progress
Contributing WP(s)	WP5, WP6, WP7, WP8
Responsible beneficiary	ABP
Other contributors	NLE, CVL, UPVLC
Internal reviewers	NLE; CVL; NAF; StAZH
Author(s)	Eva M. Lang
EC project officer	Martin Majek
Keywords	Large Scale Demonstrator, Ground Truth, Archives, Reference Data, Handwritten Text Recognition, Key Word Spotting, Document Understanding, Table Matching, Table Processing, Information Extraction

Table of Contents

Executive Summary	4
1. Clarification of ABP goals within WP 8	4
2. Data Processing and Ground Truth Production	5
2.1. Data Selection	5
2.2. Segmentation and Layout	5
2.3. Transcription	6
2.4. Quality Assurance	6
3. First results from automatic processing	6
3.1. Table Processing	6
3.2. Key Word Spotting	8
3.3. Information Extraction	9
4. Activities supporting activities in WP 8	9
4.1. Software Development focusing on the user perspective	9
4.2. Dissemination Activities	10
4.3. Inter-Archival Working Group	10
5. Papers and Publications	10
5.1. Papers and presentations	10
5.2. Publications	10

Executive Summary

The Archives of the Catholic Diocese of Passau (ABP) represents one of the Large Scale Demonstrators within the READ project. During the 2017 reporting period, the main efforts of the ABP have been the sharpening of research scope and project goal as suggested by the project reviewers during the 2017 review meeting and work towards meeting this goal (see Section 1). Therefore, the selection of data and the production of Ground Truth (GT) transcriptions (Section 2), the implementation of the technical results for the Passau demonstrator, (Section 3) development, testing and enrichment of software modules in Transkribus (Section 4) and several publications and talks (Section 5) were the main tasks carried out.

1. Clarification of ABP goals within WP 8

The comments of the reviewers at the first Brussels meeting asking for a streamlined application of forthcoming technology to institutional demands, the current status of available technical methods and tools as well as ABP experiences with GT production in 2016 called for a more focused approach to tackle the large ABP collection within the READ project.

ABP has pledged to direct all efforts towards creating the highest possible number of Ground Truth transcribed pages for death records of the Passau Diocese between 1847 and 1878. ABP produces GT pages using the “expert transcription” mode, including character-by-character transcription and tagging abbreviations, text style and mark-ups as well as providing expansions for abbreviations. Therefore, the precision of the transcription, as available only through experts, has shown to be very important.

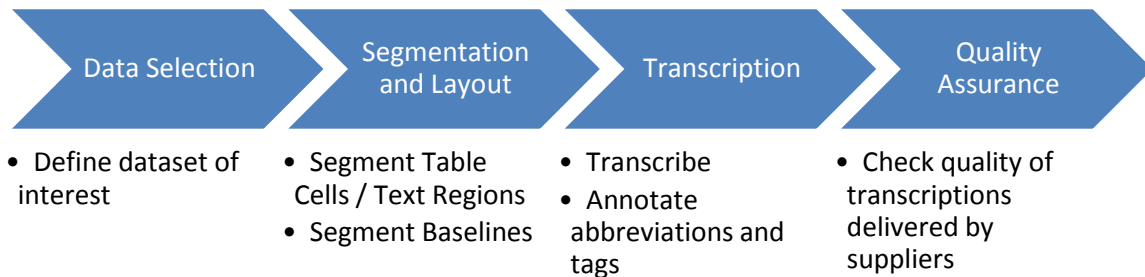
A total number of 600 pages, corresponding to 20 selected scribes was produced for the ABP_S_1847-1878 dataset in 2017. In addition, 40 pages of GT transcription produced in early 2017 were passed on to READ partners at URO and UPVLC for setting up a general KWS demonstrator and showing the possibilities of indexing and searching the full collection of 800,000 images.

From a historical point of view, the chosen timeframe covers the era between the uproar of 1848 and the foundation of the German Empire in 1871. This period between the eve of the revolution and the big economic boom of the establishment of the German Empire in the 1870s marks the silent finale of the so-called “Sattelzeit” and the entry into the industrial age. The death registry books, which cover the rural as well as the urban population within the Diocese of Passau, serve as a highly valuable source and present a vast perspective on the history of society, administration and social conditions.

We would envision the selected data set to serve as a good basis for producing reliable confidence values and promising results in key word spotting applications.

2. Data Processing and Ground Truth Production

Ground Truth production follows a workflow similar to automatic processing. The standard workflow in our case is as follows:



2.1. Data Selection

The total number of scanned ABP death records in the 1847-1878 timeframe consists of 26,579 pages. In this data set we estimate to find 295,000 entries of dead inhabitants. The focus on the selected time ensures a large enough test set of feasible size. Given the more than 590 individual hands, the data set is highly diverse with regards to writers. Therefore, ABP sharpens the selection to represent only those scribes which cover more than two thirds of the research timeframe. Thus, the number of individual scribes in the data selection phase decreases to 348.

This data set is manageable in such a way that the chosen GT pages will represent the total scope of images and the appearing hands therein in sufficient quality. The GT set already contains 30 images per selected hand, which reflect the complete working period of the respective scribe.

As the selection of data per scribe also requires research of the writing period of the individual clergy filling the registry book pages and his stations of work, each record keeper can be tracked over his professional life. The recorded data such as details of the scribe, stations of work, period of writing of the individual were handed over to READ partner CVL to support the task of Writer Identification (see WP 7, D7.17).

2.2. Segmentation and Layout

ABP has done the full cycle of selecting, segmenting and transcribing the chosen samples of hands. Our experience has demonstrated that data selection, segmentation, transcription, tagging and mark-up of abbreviations as well as quality assessment of the final transcriptions of one page varies heavily depending on the readability of the hand. While easy layouts and clear handwriting take approximately 30 minutes per page, folios with many lines and hardly legible writing style may take up to four hours for manual preparation.

Based on these findings, ABP outsourced the segmentation process to subcontractors. A total number of 640 pages were produced in 2017.

ABP experiences show that the subsequent transcription process is more user-friendly if the segmentation is based on text regions instead of table cells. With the valuable and great support from READ partners NLE and CVL, a tool to convert from the machine-friendly table area and cell format to the user-friendly text regions was implemented, tested and has been used up to now.

Recent results show that automatic layout analysis and segmentation of table structures and baselines has improved majorly, especially due to contributions of a new line finding tool provided by URO. Evaluations of how this is integrated into the production of GT pages are currently carried out.

2.3. Transcription

ABP decided on providing exact, character-by-character transcriptions which can only be provided by expert transcribers. We contracted a professional historian for this task. The tightening of scope further allowed us to test and outsource the complete transcription process. Thus, in the last quarter of 2017, the focus of the contracted transcriber shifted from transcribing towards proof-reading, correcting and tagging the transcriptions.

2.4. Quality Assurance

Quality assessment and control of manually and automatically produced transcriptions is of utmost importance to get the best possible results for machine learning and to ensure the quality of the input for automatic processing.

This includes the very first steps of preparing the data set, researching the biographical data of the scribes as well as the last steps of checking the produced quality of the transcription. For this task, senior archival knowledge is a prerequisite and was applied by ABP.

3. First results from automatic processing

3.1. Table Processing

The preparatory steps with regards to the selection of 1847-1878 death records in the Passau Diocese also lead to the evaluation of the printed table forms in the images. A thorough analysis of the dataset shows that for 22,001 images 88 different table prints were used. These unique layouts were further categorized into 11 template categories, four of which are depicted in Figure 1. Most of these printed layout categories comply with the given normative for content imposed by the state government.



Figure 1 Examples of table mark-ups for four different categories

While the presented information in this era had been standardized and the prints mainly strive to fulfill the legal requirements, some layouts diverge from the norm by either providing more specific details, e.g. three columns for day, month and year of the burial date, or leaving out some details.

Our findings also show that the vast majority of prints in our selected 26,579 images fall into mainly two table categories: 15,147 images were identified in category ABP_S_1847-1878-01, 4,203 images fall into category ABP_S_1847-1878-07. A total of 4,578 images were fully drawn or amended by hand. For each of the 11 category representatives, table templates were created and headers transcribed using the table editor and the transcription tools in TranskribusX.

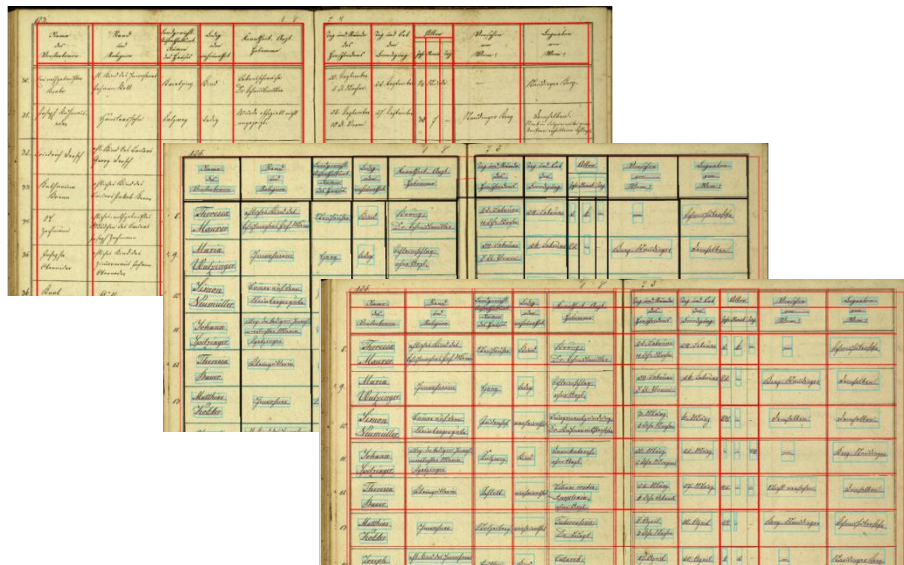


Figure 2 Different steps of table segmentation show the mark-up of graphical lines, text baselines as well as the final layout ready for transcription purposes (left to right).

As technical tools and methods progressed both with regards to layout (see Figure 2 for graphical table mark-up) and HTR quality in the course of the year due to the ongoing development and great support by READ partners URO, CVL and NLE, first HTR tests were carried out in September 2017. Several HTR models were trained and tested by ABP based on

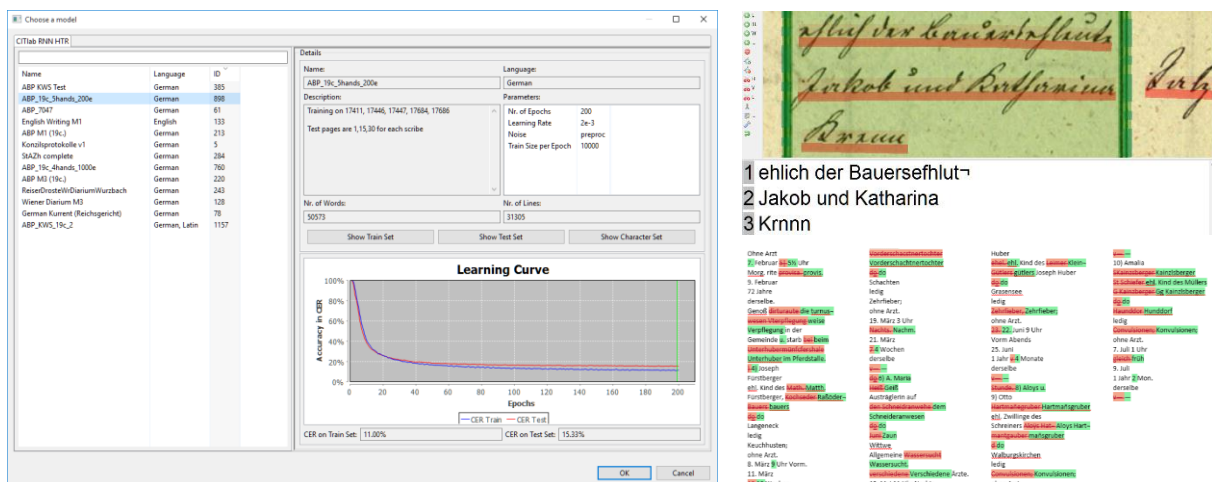


Figure 3 HTR model parameters (left) and example results of HTR application (HTR results for one entry; top right) as well as comparison between GT and HTR transcriptions. Highlighted words (bottom right) refer to differences in GT (green) and HTR (red).

the selected hands, resulting in a CER of 13-17% (see Figure 3). While these numbers seem fairly low already, the visual control of the automatic transcription reveals challenges in providing the correct notation of last names, places and specific characteristics in our records (for further details see WP 6, D 6.8).

3.2. Key Word Spotting

READ partner UPVLC set up a Key Word Spotting demonstrator based on available GT segmentation and transcription provided by ABP. The demonstrator can be found at <http://transcriptorium.eu/demots/kws-Passau/> (see also Figure 4). Thanks to the efforts of the UPVLC team, this first test setup clearly shows the opportunities which the KWS technology provides for accessing the presented information. The challenges caused by low recall/precision error rates are the next steps to be tackled (see also WP 5, D 7.14).

READ **BETA**


Johann|Johan|Joh Search Confidence: 50 Max. results:

Need help?

You are here: [home](#) » [passau](#) » [passau.chap1](#) » [page 32](#)

7 matches found for "Johann|Johan|Joh" with an average confidence of 64.1!
 Hits: 6, Misses: 0, False positives: 1

← Previous | Next →



© 2017 READ & PRHLT, UPV
 XHTML CSS AA Browse Happy

Figure 4 A search for the name Johann produced six high qualitative matches and one false positive hit on the given test page.

In November and December 2017, an HTR model was trained in order to test Keyword Spotting within TranskribusX using the KWS functionality provided by READ partner URO. This supplements the KWS demonstrator set up by READ partner UPVLC.

Results and evaluations of a comparison between both approaches are forthcoming (see also in the technical report WP 5, D 7.14).

3.3. Information Extraction

ABP worked closely with READ partner NLE to assess the quality of the automatic information extraction process. The task of ABP was testing the methods for functionality and usability on our given dataset.

```
<PAGE number="1" years="NA" nrecords="0">
<RECORD lastname="Behan" firstname="Marin" occupation="Austräger" location="Pocking" deathreason="Alterschwäche ohne Paul" burialDate="Febr"/>
<RECORD lastname="Madl" firstname="Theresia" occupation="Bäuerin," location="Riggerding" deathreason="Wassersucht" burialDate="t"/>
<RECORD lastname="Riesinger" firstname="Korona" occupation="Wirth" location="Neuhofen" deathreason="Bösartige Geschire, de Reiner" burialDate="April"/>
<RECORD lastname="Deixlberger" firstname="Joseph" occupation="Inwohner," location="Geishofin" deathreason="Bösartige Geschwüre"/>
<RECORD lastname="Ratzinger" firstname="Mathias" occupation="Bauer" deathreason="Schlagfluß der Arzt" burialDate="April"/>
<RECORD lastname="Weber" firstname="Theresia" occupation="Inwohnerstochter" deathreason="Fraisen ohne Arzt"/>
<RECORD lastname="Veit" firstname="Johan" deathreason="Abzehrung"/>
<RECORD lastname="Wiser" location="Häuzing" burialLocation="im"/>
<RECORD occupation="Bauer-" location="Berg" deathreason="Fraisen ohne Arzt" burialDate="April"/>
<RECORD firstname="Joseph"/>
<RECORD deathreason="Schwäche, Landarzt"/>
```

Figure 5 Example result of extracted records based on automatic HTR runs and information extraction for one test page. Quality improvement is to be expected with a larger number of GT pages.

The information extraction process compiles the relevant record data for the given input image following the table recognition process described above. A first test run was carried out in the fourth quarter of 2017 leading to promising results as depicted in Figure 5, where the found and extracted records for one test page are displayed (see also WP 6, D 6.14). For 2018, we will work towards increasing the quality of the extracted records with the number of provided GT pages.

4. Activities supporting activities in WP 8

As stated in the WP description (see GA), ABP is focusing on the experience of the archival users. Due to regular feedback from the subcontractor and from our transcription experts, several requests for module development were passed on to the Transkribus developers, some of which were also taken over and developed by ABP.

4.1. Software Development focusing on the user perspective

ABP involvement in Software development and functionality enhancement of TranskribusX modules concentrated on providing development and input to the following tasks

- Synchronizing local Transkribus documents (PAGE XML) with existing remote Transkribus documents
- Improving the display of the search results for tag searches
- Improving the normalisation of tag attributes
- Fixing several bugs related to the transcription process and the keying experience, such as the deletion of empty characters
- Testing and debugging of web interface functionalities representing the end-user, archival role
- Testing and debugging the python framework supporting the automatic processing of images to information workflow by NLE
- Providing feedback for and implementing sub tasks related to table processing
- Participating in the regular architecture, table processing and web interface working group calls

4.2. Dissemination Activities

ABP participated in the regular conferences of the dissemination working group. For written publications by ABP, see Section 5 below. General dissemination activities within the READ project are recorded in WP 2.

4.3. Inter-Archival Working Group

During the 2016 all-staff meeting in Passau, a working group for inter-archival information exchange was started to ensure the communication flow and share best practices between the core READ archival partners ABP, NAF, StAZH. Five meetings were carried out in 2017 focusing on

- Discussion on usage and integration of metadata information within archival structures
- Informational exchange on current technical advances between the archival partners
- Usability and improvement of software interfaces available to the institutional users

5. Papers and Publications

5.1. Papers and presentations

- 22.-23.06.2017, Südwestdeutscher Archivtag, Bretten: „Männlich, alt, skurril“ – Vom Lesesaalkunden zum Onlinenutzer; Auswertung einer Benutzerumfrage im Archiv des Bistums Passau (Wolfgang Fronhöfer)
- 03.07.2017, Leeds, IMC 2017: From Tables to Transkribus (Elena Mühlbauer)
- 02.-03.11.2017, Wien, Universität Wien, Transkribus User Day 2017: Transkribus in Practice - Table Processing in Transkribus, https://read.transkribus.eu/wp-content/uploads/2017/07/Lang_Table_Processing.pdf (Eva Lang)

5.2. Publications

- Fronhöfer Andrea / Mühlbauer Elena: Archivnutzung ohne Limit. Digitalisierung, Onlinestellung und das Projekt READ für barrierefreies Forschen, in: Der Archivar, Zeitschrift für Archivwesen 70 (2017) 422-427
- Mühlbauer Elena: How to use Transkribus in 10 steps, <https://www.youtube.com/watch?v=GjChcDEXshU> (tutorial video by ABP)
- Lang Eva: How To Process Tables in Transkribus (written tutorial by ABP)
- Fronhöfer Wolfgang: <https://www.youtube.com/watch?v=gB-vHnbOfew> (video footprint at Südwestdeutscher Archivtag by Robert Reiter)