READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D5.9
ScriptNet Large Scale Dataset P2

Florian Kleber, Markus Diem, Stefan Fiel
CVL

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	28.12.2017
Date of last update	11.12.2017
Deliverable number	D5.9
Deliverable title	ScriptNet Large Scale Dataset P2
Type	report
Status & version	in progress
Contributing WP(s)	WP5
Responsible beneficiary	CVL
Other contributors	UPVLC, DUTH, NCSR, ABP
Internal reviewers	UPVLC,NCSR
Author(s)	Florian Kleber, Markus Diem, Stefan Fiel
EC project officer	Martin MAJEK
Keywords	baseline, KWS, writer identification, HTR, table, Dataset

Contents

1 Executive Summary

This task comprises the selection of the document images, the definition of the Ground Truth (GT) for the corresponding task, the management of the data production, the distribution of data to training and evaluation sets and the description of the datasets. Based on the planned datasets in D5.8 five competitions have been carried out at ICDAR 2017. All datasets have been made publicly available together with their GT (downloadable at the competition website (see ScriptNet) or from Zenodo.org, e.g. cBad). The competitions are summarized in Section 2. Section 3 describes new datasets and also further planned datasets for D5.10. For the creation of the datasets the presented benchmarking tool[1] in D5.8 is partly used.

2 Competition Datasets

In D5.8 datasets have been prepared for the planned competitions at ICDAR 2017. Finally, the following competitions have been carried out at ICDAR:

- ICDAR2017 Competition on Baseline Detection (cBAD) [1]
- ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI) [2]
- ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset (ICDAR2017 HTR)
- ICDAR2017 Competition on Document Image Binarization (DIBCO 2017) [3]

The planned datasets for the four competitions (1-4) have been described in D5.8 and are now publicly available together with the annotated ground-truth. Additionally a dataset for keyword spotting has been prepared[2]. The datasets can be downloaded either from the scriptnet[3] site or from Zenodo[4]. The results of all competitions have been published at ICDAR 2017 [1, 2, 4, 3].

2.1 ScriptNet: Dataset for Document Image Binarisation. ICDAR 2017

The DIBCO 2017 testing dataset consists of 10 machine-printed and 10 handwritten document images for which the associated ground truth was built manually for the evaluation. The selection of the images in the dataset was made so that representative degradations appear. The machine-printed documents of the dataset originate from collections that belong to the IMPACT project, while the handwritten document images originate from collections that belong to READ project. The testing dataset along with

[1] `https://github.com/TUWien/Benchmarking`
[2] https://scriptnet.iit.demokritos.gr/competitions/7/
[3] `https://scriptnet.iit.demokritos.gr/competitions/`
[4] e.g. cBad, `https://zenodo.org/record/835441#.WhgH8OcxlhE`

the associated ground truth as well as the evaluation software are publicly available at: `http://vc.ee.duth.gr/dibco2017/benchmark`.

3 Datasets

For the Large Scale Demonstrator (LSD) the data of the Passau Diocesan Archive is the basis. The dataset is named *ABP_S_1847-1878* and is described in detail in the following section.

3.1 ABP_S_1847-1878

The *ABP_S_1847-1878* dataset contains information about the parishioners who died within the geographic boundaries of the various parishes of the Diocese of Passau between the years 1847 and 1878. The dataset holds a total of 26,579 scanned pages. The scans originate from 212 pastoral districts (mainly parishes) with their own record keeping in the time between the uproar of 1848 and the beginning of the German Empire in 1871. This period is marked by massive social, economic and technical transformations.

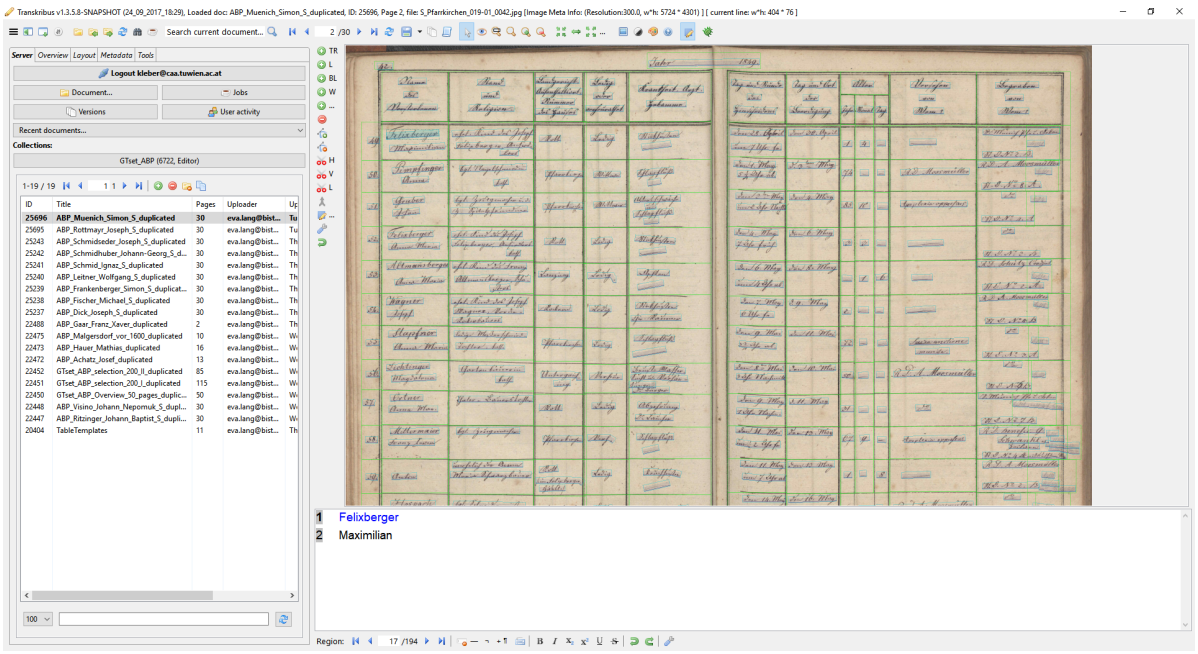

Figure 1: Example image of the GTset_ABP with annotatedt GT (tables, baselines, transcription).

According to the official order of the Catholic Church, the parish scribes had to record name, profession, religion, court, address, marital status, reason of death, dates of death and burial, age, names of doctor and priest as well as additional information in written form. The images display the records mainly in tabular format referring to one person per row. Stemming from than 590 individual hands, the data set, recoding an estimated number of 295,000 casualties, is also highly diverse with regards to writers.

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 eleven template categories. Most of these printed layout categories comply with the given normative for content imposed by the Church. The vast majority of scans (15,147 images) even fall into one single template category. On 4,578 pages, the requested information was recorded in manually drawn tables or manually extended table prints. The images are openly available through the matricula online platform [5], records can be queried using a search engine supplied by the Diocese of Passau [6]. The data are used by family historians as well as by historian scholars interested in age of those who died, the development or the spread of deadly diseases, etc.

The dataset is availabe in Transkribus (Collection ID 6285). Based on the *ABP_S_1847-1878* dataset a subset with annotated tables, baselines and tables is provided. The subset contains about 370 pages and is on Transkribus in Collection *GTset_ABP* (Collection ID 6722).

Figure 1 shows an annotated image of the *GTset_ABP* (Collection ID 6722) dataset. The table and the baselines are annotated, as well as the transcribed information.

3.2 Girona Dataset

The Girona Collection, has been provided by the Centre de Recerca d'Historia Rural (CRHR) from the Universitat de Girona (MoU partner of the READ project). The collection, called "Oficio de Hipotecas de Girona", is composed of a large number of notarial documents from the 17th century. It contains much more than morgage deeds, including such notarial acts as sales and purchase agreementes, leasing and Emphyteusis contracts, cretid operations, marriage contracts, manufacturing and commercial compaines, etc. The CRHR is interested in the perfect transcription of this collection.

So far, around 400 pages has been processed and the ground truth at two levels has been generated. First, a layout analysis of each page was manually done to indicate text blocks and lines, resulting in a dataset of around 16000 lines. Second, the pages were completely transcribed line by line by an expert paleographer using the CATTI technology (see Deliverable D7.5). Additionally to the diplomatic transcription of the document, some tags are added to the corresponding words in order to provide a more rich information about the content of the document: toponyms, antrhoponyms, trades, registry tupology, abbreviations and hyphenated words.

These pages are available in the READ platform in the RH_Girona collection. Image 2 shows an image of the Girona dataset.

3.3 Planned Datasets

For D5.10 the LSD dataset will be extended. Additionally, the *GTset_ABP* will be the basis for further competitions (e.g. table recognition) at ICFHR 2018 (proposal planned). Also the cBad dataset will be extended with new document types (tables,

[5] `http://data.matricula-online.eu/de/deutschland/passau/`
[6] `http://gendb.bistum-passau.de/`

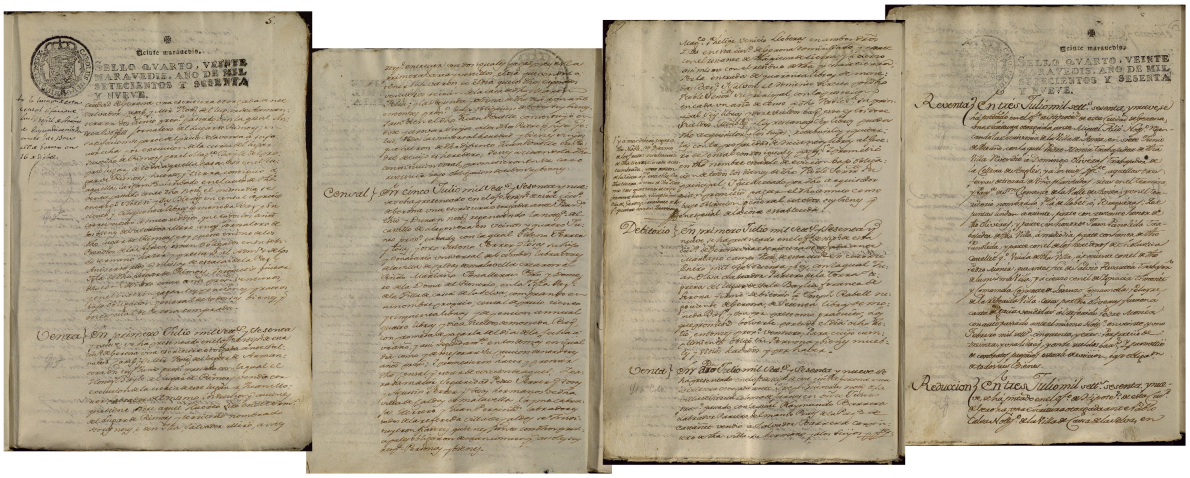

Figure 2: Example image of Girona dataset.

postcards) to create a new layout analysis dataset. Also a follow up competition of cBad is planned.

Currently a new dataset for writer identification is in progress. This dataset is created using the document images of the Passau Diocesan Archive, will be the first dataset which covers the evolving handwriting style over years. Currently, 28 writers have been selected with pages written within a time period of up to 20 years. Resulting in an uneven distributed dataset with about 1.800 document images. The pages will be analyzed further if preprocessing steps are needed and then the new dataset will be created. Furthermore, new tasks on this dataset for researchers have to be described which reveals the invariance of the methods to handwritings over decades.

References

[1] M. Diem, F. Kleber, S. Fiel, T. Grüning, and B. Gatos, "Icdar2017 competition on baseline detection (cbad)," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR17)*. IEEE Computer Society Press, 2017, pp. 1355–1360.

[2] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, N. Stamatopoulos, and B. Gatos, "Icdar 2017 competition on historical document writer identification (historical-wi)," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR17)*. IEEE Computer Society Press, 2017, pp. 1377–1382.

[3] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "Icdar 2017 competition on document image binarization (dibco 2017)," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR17)*. IEEE Computer Society Press, 2017, pp. 1395–1403.

[4] J. A. Sanchez, V. Romero, A. H. Toselli, M. Villegasy, and E. Vidal, "Icdar2017 competition on handwritten text recognition on the read dataset," in *14th IAPR*

International Conference on Document Analysis and Recognition (ICDAR17). IEEE Computer Society Press, 2017, pp. 1383–1388.