READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D5.10
ScriptNet Large Scale Dataset P3

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	31.12.2018
Actual date of delivery	28.12.2018
Date of last update	13.12.2018
Deliverable number	D5.10
Deliverable title	ScriptNet Large Scale Dataset P3
Type	report
Status & version	Final report
Contributing WP(s)	WP5
Responsible beneficiary	CVL
Other contributors	UPVLC, DUTH, NCSR, ABP, CITLAB
Internal reviewers	UPVLC,NCSR
Author(s)	Florian Kleber, Markus Diem, Stefan Fiel
EC project officer	Christophe DOIN
Keywords	baseline, KWS, writer identification, HTR, table, page detection, page segmentation, 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.

Two competitions have been carried out with newly created datasets (H-DIBCO 2018 and HTR 2018). Additionally, two datasets are presented which form the basis for competitions that will be submitted to ICDAR 2019 (table recognition and page segmentation). One enriched dataset (cBad) and a Writer Identification dataset has also been published within READ. The datasets are freely available on Zenodo or on the competition website.

These datasets (see listing) are additionally made publicly available through Zenodo (see D5.8, D5.9 and D7.3 for a description of the datasets).

ICFHR-2014 Bentham Dataset, see `http://doi.org/10.5281/zenodo.44519`

ICDAR-2015 ICDAR 2015 Competition HTRtS: Handwritten Text Recognition on the tranScriptorium Dataset, see `http://doi.org/10.5281/zenodo.248733`

ICFHR-2016 HTR Dataset, see `https://doi.org/10.5281/zenodo.1164027`

ICDAR-2017 Dataset for ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset, see `https://doi.org/10.5281/zenodo.835488`

Dataset for HTR and Layout A dataset of Spanish notarial deeds (18th Century) for Handwritten Text Recognition and Layout Analysis of historical documents, Quirós, Lorenzo, Serrano, Lluís, Bosch, Vicente, Toselli, Alejandro H., Congost, Rosa, Saguer, Enric, and Vidal, Enrique, Oficio de Hipotecas de Girona, see `http://doi.org/10.5281/zenodo.1322666`.

2 Competition Datasets

In this section newly created competition datasets are presented in detail. The datasets comprise the topics binarization, handwritten text recognition, table detection and recognition, page segmentation and page detection. For a detailed description see the referenced Zenodo pages.

2.1 H-DIBCO 2018 Dataset (ICFHR2018)

The H-DBICO 2018 testing dataset comprises 10 handwritten images for which the associated ground truth was built for the evaluation. The document images of this dataset originate from the READ project in various collections such as:

- The protocols of the city or municipal council of Bozen, from the 15th century to the 19th century.
- Reconstructed Alexander von Humboldt's "Kosmos-Lectures"

Figure 1: Sample pages of H-DIBCO 2018 dataset.

- Archive Bistum Passau (ABP) collection that contains sacramental register and index pages like baptism, marriage and death entries.

The dataset is available on the competition homepage, see `https://vc.ee.duth.gr/h-dibco2018/benchmark/`. Two sample pages are shown in Figure 1.

2.2 Handwritten Text Recognition Dataset (ICFHR2018)

The presented dataset was used for the ICFHR2018 Competition on Automated Text Recognition. The key numbers of the dataset are:

- 22 different documents of (roughly) 25 pages each
- one writer per document
- 16984 lines, 98239 words and 601877 characters in total
- heterogeneous data set: different writers, time periods and languages
- containing line images, info file (containing e.g. a surrouning polygon) and ground truth (not for test set)

The data is available on SciptNet since the competition was planned as an ongoing competition. For a detailed description of the competition and the dataset see the ScriptNet Competition Site `https://scriptnet.iit.demokritos.gr/competitions/10/`.

2.3 cPAS

This competition will facilitate research in page segmentation. An annotated database is collected from 6 archives with 3000 page images. The objective is to automatically locate and correctly label regions in document page images. The GT annotation was prepared this year and it is planned to use the cPAS dataset for a page segmentation competition in conjunction with ICDAR 2019. Figure 2 shows example images of the dataset which consists of 2700 images in total.

Figure 2: Sample pages of the cPas dataset.

2.4 ScriptNet Table Dataset

The table dataset is the basis for the a new table detection and recognition competition which is planned in conjunction with ICDAR 2019. The dataset will contain the following annotation marks for each document:

- Table region (multiple table regions possible)
- Table col/rows/cells
- Visibility of border region
- Table header information
- Table caption
- Spine region
- Baselines (also table-running-text)
- Text regions outside of the table

The dataset contains contributions from 25 institutions around the world. The images show a great variety of tables from hand-drawn accounting books to stock exchange lists and train timetables, from record books to prisoner lists, simple tabular prints in books, production census and many, many more. The entire dataset consists of 1000 images. Figure 3 shows a document with a printed table, the annotated table caption/cells and the corresponding baselines.

A handwritten table is shown in Figure 4, again with annotated table caption/cells, corresponding baselines and additional text regions present in the image. The dataset will be published after ICDAR 2019.

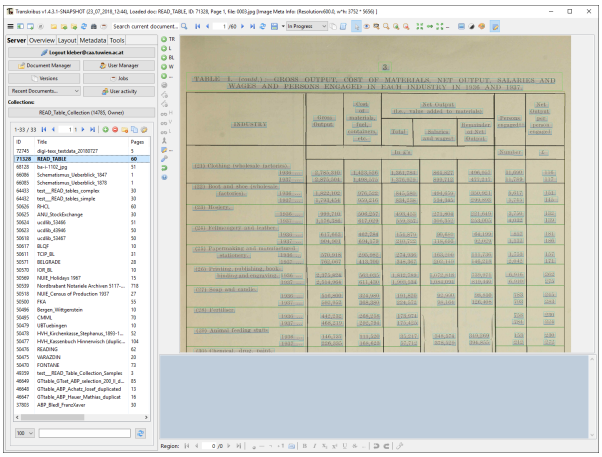

Figure 3: Sample page of the table dataset showing annotations of table caption/table headers/cells and baselines.

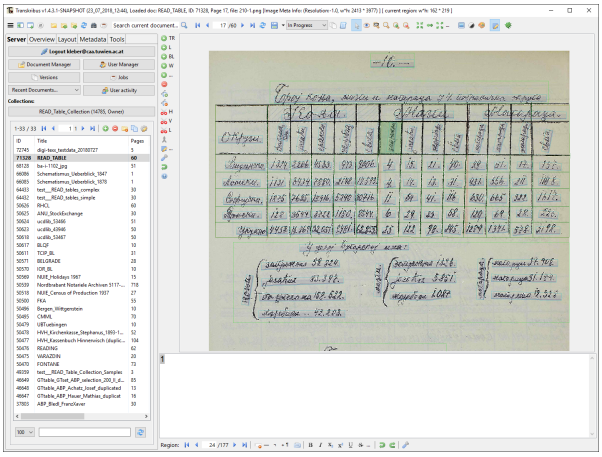

Figure 4: Sample page showing annotations of a handwritten table and additional text regions.

3 Published Datasets

Datasets that are summarized in this section were created within the READ project and made publicly available through Zenodo. They mainly serve as supplementary material of published papers that allow to reproduce the results published.

3.1 Enriched cBad Dataset

The cBad Dataset was created for the ScriptNet - Dataset for Baseline Detection in Historical Documents, ICDAR 2017 Competition (see D5.8 for a detailed description). The dataset contains the annotated baselines and for Track A also the text regions. The annotation of the dataset was enriched with the page area, and in case of double pages the page split (spine region). The dataset can be used to train and evaluate a page detection. A sample page with the annotated page border region and spine region is

shown in image 5.

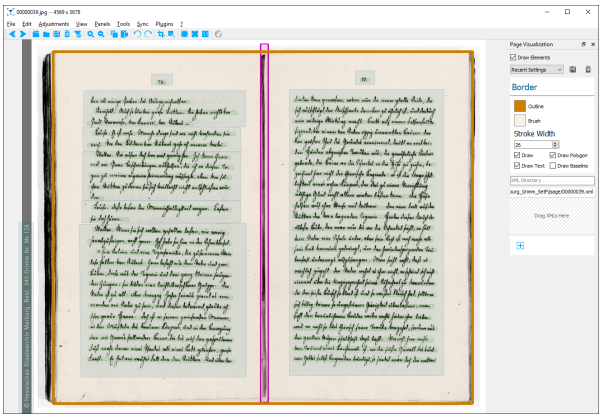

Figure 5: Sample page of cBad with annotated page border region and spine region.

The dataset with the enriched annotation is freely available on Zenodo `https://doi.org/10.5281/zenodo.746925`.

3.2 READ ABP WI Dataset - Writer Identification over decades

A hand is usually considered as a unique characteristic of a person. However, it may slightly change over their whole lifespan. This change might be due to some physical or mental issues. To the best of our knowledge, there is no dataset available, which covers this aspect of evolvement of handwriting of a single person.

When dealing with archival documents, it is important to show that methods are invariant against these changes or investigate how much of these changes are covered. Thus, the READ ABP WI Dataset was created with data of the Passau Diocesan Archives (ABP, https://www.bistum-passau.de/bistum/archiv), one sample page is presented in Figure 6.

The documents originate from death records of different villages or towns in the Diocese of Passau. Usually the writer of these records (mostly the priest) remains the same over several years.

Figure 7 shows how many page Writer 3 contributed to the dataset in each year. In total this writer has 59 pages in the dataset, which have been written in a time period of 12 years.

The READ ABP WI dataset consists of 1766 pages, which originate from 28 different writers. The number of pages per writer varies from 7 up to 311. For some writers, we only have data from 3 different years, whereas the largest time span between two documents of the same writer is 31 years. The dataset has been published on Zenodo[1].

3.3 READ ABP Table Dataset

The READ ABP Table 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 records show for each sacramental event the record name,

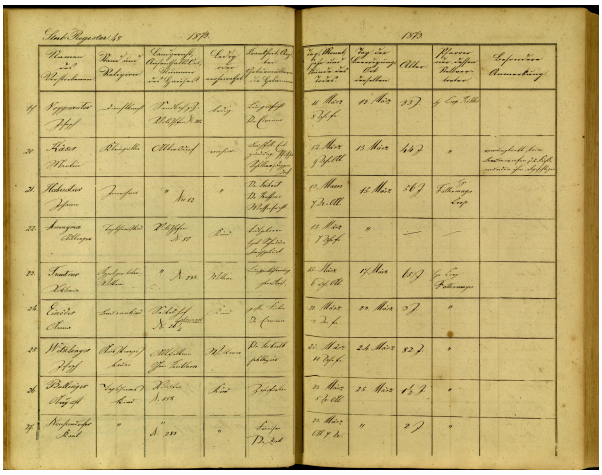

Figure 6: Sample page of the writer identification dataset.

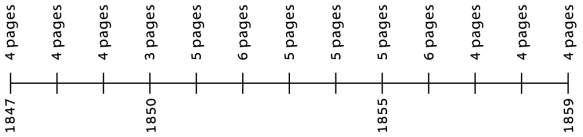

Figure 7: The figure shows how many pages one writer (Id 3) contributed to the dataset in which year.

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 mainly in tabular format referring to one person per row.

Two sample datasets of 150 and 100 images each were compiled and are openly available through Zenodo: `https://doi.org/10.5281/zenodo.1226878` and used by Clinchant et al. [2], *"Comparing Machine Learning Approaches for Table Recognition in Historical Register Books"*.

References

[1] S. Fiel, F. Kleber, E.-M. Lang, and W. Fronhöfer, "READ ABP WI Dataset - Writer Identification over decades," Sep 2018. [Online]. Available: https://zenodo.org/record/1421600

[2] S. Clinchant, H. Déjean, J. Meunier, E. M. Lang, and F. Kleber, "Comparing machine learning approaches for table recognition in historical register books," in *13th IAPR International Workshop on Document Analysis Systems (DAS)*, April 2018, pp. 133–138.