READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D5.8
ScriptNet Large Scale Dataset P1

Markus Diem, Stefan Fiel, Florian Kleber
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	31.12.2016
Actual date of delivery	28.11.2016
Date of last update	21.12.2016
Deliverable number	D5.8
Deliverable title	ScriptNet Large Scale Dataset P1
Type	report
Status & version	in progress
Contributing WP(s)	WP5
Responsible beneficiary	ULCC
Other contributors	UPVLC, DUTH, NCSR
Internal reviewers	UPVLC,NCSR
Author(s)	Markus Diem, Stefan Fiel, Florian Kleber
EC project officer	Martin MAJEK
Keywords	baseline, KWS, writer identification, HTR, Dataset

Contents

1 Executive Summary

This task comprises the selection of the page 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 database. For the selection of document pages from large digital collection to create Document Image Analysis (DIA) datasets an open source tool, *CVL Benchmarking Module*, has been developed within this task. A detailed description of the tool is given in Section 2. Also the requirements for the GT and the evaluation metrics regarding the following tasks have been defined: Baseline Detection, Handwritten Text Recognition (HTR), Writer Identification (WI) and Keyword Spotting (KWS). For each tasks datasets have been created and the GT was manually labeled. For a detailed description see Section 3. The databases are the basis for the *ICFHR 2016* and the planned *ICDAR 2017* competitions and will be made public via the ScriptNet site: `https://scriptnet.iit.demokritos.gr/competitions/`.
The definition of the GT for the different tasks as well as the CVL benchmarking tool are the basis for the *Large Scale Data and Reference Set* which will be defined in D5.9 and D5.10.

2 CVL Benchmarking Module

A Python script has been developed to sample a defined amount of images from large scale databases of different collections. An equidistant sampling is chosen to get representative samples from all available collections. The file hierarchy of the sampled images can either be flat or hierarchically. Additionally, statistics are created. The tool is Open-Source under LGPLv3 and available at: `https://github.com/TUWien/Benchmarking`. The tool has been used to create the baseline competition dataset.

3 Competition Datasets

A short description of prepared datasets including GT is given in the following Sections. The datasets are also the basis for the ICFHR 2016 and the planned ICDAR 2017 competitions and are published at the ScriptNet Site `https://scriptnet.iit.demokritos.gr/competitions/`

3.1 ScriptNet: Dataset for Baseline Detection in Historical Documents. ICDAR 2017

A dataset for the detection of baselines in handwritten and printed documents has been prepared. In contrast to previous competitions, we propose a new evaluation which is based on baselines rather than utilizing pixel- or area based error metrics. Two different tracks are proposed: Track A evaluates baseline methods on handwritten document images (no tables or marginalia) where text regions (paragraphs) are labeled in the input image. Track B comprises a full text detection and localization system. Hence, only

the page area in the image is labeled as input (no text regions). Images of Track B have a complex layout, e.g. tables, empty pages, marginalia and skewed text-lines up to 180°. The planned competition will be successor of the *ICDAR-2015 ANDAR Text Lines competition*. The dataset (training and test set) and the competition will be published on ScriptNet: `https://scriptnet.iit.demokritos.gr/competitions/5/`

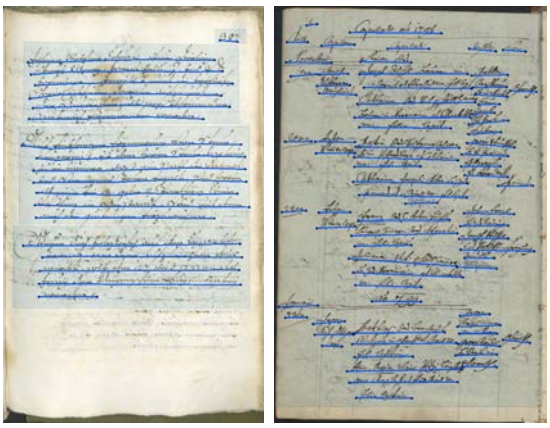

Figure 1: Two examples of document images of Track A (left) and Track B (right) with annotated baselines and text regions.

Track A consist of 755 images, Track B of 1284 images, randomly extracted from 9 different collections of different archives. The images were randomly chosen using a specifically designed database crawler[1]. For each image in the dataset, a PAGE XML file is provided containing the GT. Figure 1 shows two examples of the proposed dataset. A more detailed listing can be found in Table 1. One third of the images is used for training.

Collection	Track A	Track B
ABP_FirstTestCollection	88	138
BHIC_Akten		241
Bohisto_Bozen_SetP	91	146
EPFL_VTM_FirstTestCollection	6	207
HUB_Berlin_Humboldt	178	70
NAF_FirstTestCollection	85	144
StAM_Marburg_Grimm_SetP	158	79
UCL_Bentham_SetP	53	160
unibas_e-Manuscripta	96	96
Total	755	1284

Table 1: Collection names and number of pages used per track. Intially 250 pages were randomly selected from each collection.

[1] `https://github.com/TUWien/Benchmarking`

3.2 ScriptNet: Dataset for Writer Identification in Historical Documents. ICDAR 2017

A dataset for writer identification and writer retrieval has been prepared. The documents originate from the Universitätsbilbiothek Basel. In total 140.418 images were analyzed. The first pre-processing step is searching for an author in the metadata file. If no author is found this page is skipped. If an author appeared twice and a different year of birth or death is noted in the metadata, then the author with more pages is taken. This results in 51.543 pages from 3403 different writers. The next step is the elimination of non-text pages like drawings and music scores. This is achieved by analyzing the histogram of the gray values. Pages with handwriting on it have a characteristic signature in the histogram. After this step 39.000 written by 3133 remain in the dataset. The last step of the automatic pre-processing is the estimation of the writing area, since the dataset contains pages with only a few lines of text. This is done by calculating SIFT features on the document image and by analyzing their distribution over the document image. Again, roughly 10.000 pages are removed with this step. One goal of this dataset is that the distribution of pages a writer contributes is equal. At this stage the database consists of 1275 writers which have contributed more than 5 pages. 5 pages were selected by equidistant sampling and these 5 pages were controlled manually if they are suitable for writer identification. Since all thresholds of the automated steps are set not too strict still some of these pages contain drawing or music scores, or the quality of the text was too low for example due to bleed trough. These pages are, if possible, replaced. This results in a dataset of 3610 pages of 722 writers. The text area is then manually cropped so that writer identification methods only calculate their features on the handwriting and not on the background. Also a binarized version of the dataset is created. Figure 2 shows two sample pages of the dataset written by two different writers.

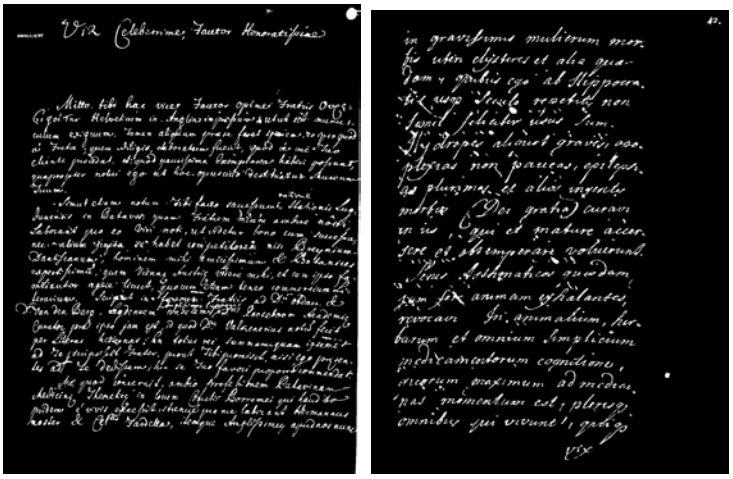

Figure 2: Two binarized and cropped pages of the dataset.

3.3 ScriptNet: Dataset for Handwritten Text Recognition. ICFHR 2016

A dataset for the ICFHR 2016 for HTR was prepared [1]. The dataset for this competition was composed of 450 page images, each encompassing of a single text block in most cases, but also with many marginal notes and added interlines. These pages entailed several line detection and transcription difficulties and the corresponding ground truth (GT) was produced semi-automatically and manually reviewed. The GT information was registered in PAGE.

These 450 pages contained 10,550 lines with nearly 43,500 running words and a vocabulary of more than 8,000 different words. The last column in Table 2 summarizes the basic statistics of these pages.

Table 2: The Ratsprotokolle dataset used in the HTR contest.

Number of:	Train	Validation	Test	Total
Pages	350	50	50	450
Lines	8,367	1,043	1,140	10,550
Running words	35,169	3,994	4,297	43,460
Lexicon	6,985	1,526	1,656	8,120
Character set size	92	80	83	92
Running Characters	208,595	26,654	25,179	260,428

The dataset was divided into three subsets for training, validation and testing, respectively encompassing 350, 50 and 50 page images. Since it was not possible to accurately identify the writers in all cases, this characteristic was not taken into account for distributing them over these two subsets. This means that some writers could appear in the three sets.

The GT in both training and validation sets was in PAGE format and it was provided annotated at line level in the PAGE files. The transcriptions at line level were also included in the PAGE files. On the other hand, the PAGE files of the test set contained the line regions, but the transcripts were removed.

Table 2 contains basic statistics of these partitions. The rows "Running words" and "Running OOV" show the total number of words and Out-Of-Vocabulary (OOV) words, respectively. The OOV words in the Validation column are words that do not appear in the training set. The OOV words in the Test column are words that do not appear neither in the training set nor in the validation set. The row "OOV Lexicon" shows the number of *different* running OOV words.

3.4 ScriptNet: Dataset for Keyword Spotting in Historical Documents. ICFHR 2016

The H-KWS 2016 dataset [2] comprises a series of documents from two different collections prepared in the READ: the Alvermann Konzilsprotokolle and the Botany in British India collections. The former, in good preservation state, belongs to the University Archives Greifswald and involves around 18 000 pages. This collection contains fair

copies of the minutes, written during the formal meetings held by the central administration between the years 1794-1797. The documents belong to the University Archives and were digitized and provided by the University Library in Greifswald. Transcripts were provided by the University Archives (Dirk Alvermann). On the other hand, the Botany in British India[2] is from the India Office Records and provided by the British Library. This collection covers the following topics: botanical gardens; botanical collecting; useful plants (economic and medicinal). Figure 3 shows an example page from each dataset.

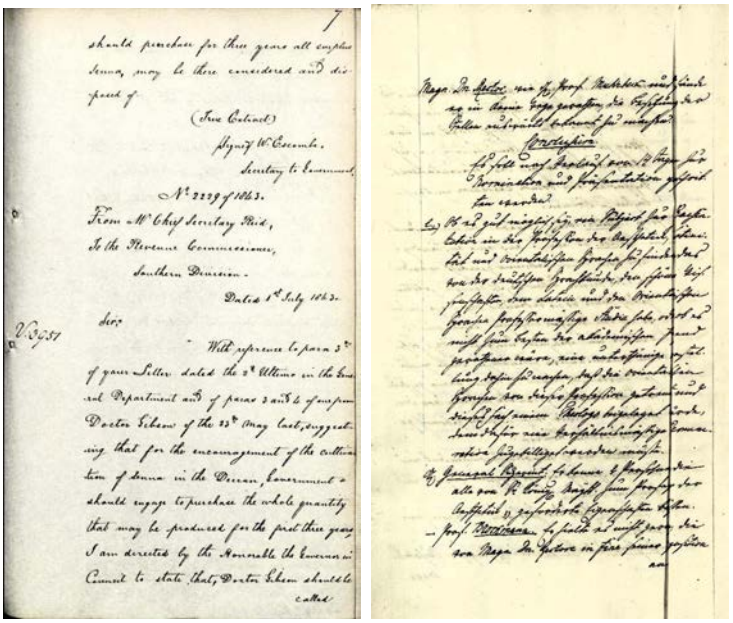

Figure 3: Examples document page images from the Botany (left) and Konzilsprotokolle (right) collections

For each collection, several training set partitions were released sequentially in order to evaluate the competing systems under different amounts of available training data. Details on the number of pages, lines and words on each partition of the training data and the test data are given in Figure 4. For each partition, the set of page images and two XML files, containing the word-level and line-level transcription and segmentation, were given. However, only three pages from the first training partition of each dataset were manually segmented at a word-level. The word-level bounding boxes of the remaining training pages were obtained by means of Viterbi forced alignment using the line-level segmentation, which was performed manually by human operators.

Each test dataset comprises 20 pages wherein the bounding boxes of all words were manually obtained. The query set of each dataset is provided in UTF-8 plain text format (QbS) and word image queries (QbE) of various length and frequency. 150 and 200 different words were manually selected for the Botany and the Konzilsprotokolle datasets, respectively. Figure 4 shows the frequency and the query length distribution for

[2] http://www.bl.uk/reshelp/findhelpregion/asia/india/indiaofficerecords/botanymat.html

			Botany	Konzilsprotokolle
Train	I	Pages	10	10
		Lines	275	263
		Words	1684	1849
	II	Pages	30	30
		Lines	676	824
		Words	3611	5968
	III	Pages	114	45
		Lines	2961	1235
		Words	16686	9102
Test		Pages	20	20
		Lines	607	524
		Words	3318	3891

Figure 4: Statistics on the content of the training and testing dataset

each query set. All data used in the competition, including transcriptions and evaluation ground-truth for KWS, was released after the competition and it is available through the competition's webpage[3].

3.5 ScriptNet: Dataset for Document Image Binarisation. ICFHR 2016

The H-DIBCO 2016 testing dataset [3] consists of 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 document images of this dataset originate from collections that belong to READ project contributed by the Archive Bistum Passau (ABP) and by Staatsarchiv Marburg (StAM) which concerns the Grimm Collection. The ABP collection contains sacramental register and index pages like baptism, marriage and death entries containing around 18000 document images. The StAM – Grimm collection contains around 36000 document images from the Grimm brothers comprising mainly letters, postcards, greeting cards, etc. The original images along with the corresponding ground truth are shown in Figure 5, respectively.

3.6 ScriptNet: Alfred Escher Dataset for Handwritten Text Recognition

In cooperation with the Alfred Escher Foundation in Switzerland, we are working on a data set based on the collected letters of Alfred Escher. It ist the largest data set for historical handwritten text recognition world-wide. Here a couple of facts about the data set: Alfred Escher is known as one of the co-founders of modern Switzerland. He lived between 1819 and 1882. He was a Swiss statesman, politician, and industrialist. He is remembered mainly for his involvement in passing the 1849 railway law, which put railway construction and operation in private hands. He played a decisive role in the development of the modern Swiss railway system. Recognizing the need for engineers and specialists, Escher founded the ETH Zurich, a university for science and technology,

[3] `https://www.prhlt.upv.es/contests/icfhr2016-kws/data.htm`

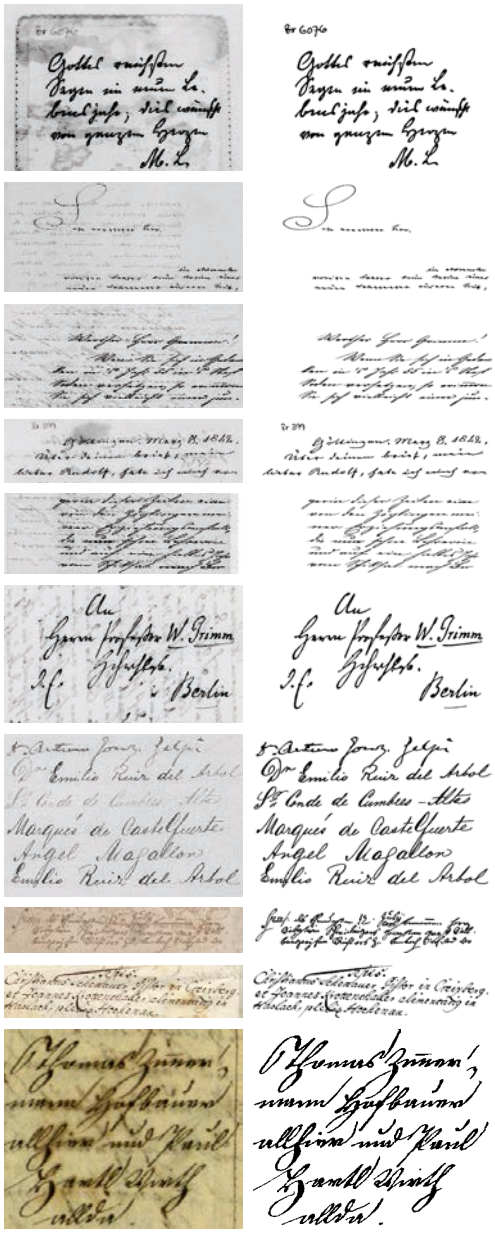

Figure 5: The h-DIBCO 2016 dataset.

in 1854. In order to secure financing for railway companies, he established a new bank, today known as Credit Suisse. On top of that, Alfred Escher was actively involved in the Gotthard project.

The Alfred Escher data set comprises all surviving correspondence between the Swiss pioneer and his associates. The body of 5,018 letters and 12,000 pages spans more than half a century. All documents have been carefully transcribed and edited by hand members of the Alfred Escher Foundation and Credit Suisse. They have identified individual lines, highlighted persons and places, and extracted detailed information for comparison with available data bases. About 33.000 working hours were invested into the project for the transcription only. We are very thankful that it was possible to

convince the Alfred Escher Foundation to make these transcripts and images available as dataset via the READ project. It will form a basis for upcoming HTR competitions.

3.7 Further Planned Datasets

Based on the GT definition and the developed CVL benchmarking module large scale data and reference set of at least three million images will be produced for the following deliverables. The planned datasets will also comprise the Alfred Escher Collection.

References

[1] J. Sánchez, V. Romero, A. Toselli, and E. Vidal, "ICFHR2016 competition on handwritten text recognition on the READ dataset," in *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, 2016, pp. 630–635.

[2] I. Pratikakis, K. Zagoris, B. Gatos, J. Puigcerver, A. Toselli, and E. Vidal, "Icfhr 2016 handwritten keyword spotting competition (h-kws2016)," in *15th International Conference on Frontiers in Handwriting Recognition (ICFHR16)*, 2016, pp. 613–618.

[3] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "Icfhr 2016 handwritten document image binarization contest (h-dibco 2016)," in *15th International Conference on Frontiers in Handwriting Recognition (ICFHR16)*, 2016, pp. 619–623.