READ

Recognition and Enrichment of Archival Documents

D2.4. Dissemination and Awareness Plan

Günter Mühlberger (UIBK),

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	31.12.2016
Actual date of delivery	30.01.2017
Date of last update	30.01.2017
Deliverable number	D2.4
Deliverable title	Dissemination and Awareness Plan P1
Type	Report
Status & version	Final
Contributing WP(s)	All WPs
Responsible beneficiary	UIBK
Other contributors	All beneficiaries
Internal reviewers	Maria Kallio, Louise Seaward
Author(s)	Günter Mühlberger
EC project officer	Martin Majek
Keywords	Dissemination

Table of Contents

Executive Summary

We distinguish between general and specific dissemination actions. General actions comprise all kinds of papers, talks, workshops, mailings, blog posts and similar, whereas specific actions are defined to be unique for the service, tool or product developed in the READ project. The focus of this paper is laid on these specific dissemination actions.

1. Introduction

The overall objective of READ is to "revolutionize access to archival documents". In order to reach this ambitious goal it is not sufficient to provide "only" cutting edge technology in the fields of Machine Learning, Pattern Recognition, Layout Analysis and Natural Language Processing. It is necessary to translate it into services and tools which are of direct benefit to archives/libraries, scholars, computer scientists and the public (family historians). Simple services and tools are often more beneficial than the most sophisticated and elaborate technologies.

2. General dissemination and awareness activities

The dissemination and awareness strategy of the READ project reflects this general consideration and can therefore be separated in two strands: One strand comprises traditional dissemination and awareness activities in order to keep everyone interested in the project up-to-date. The main activities are:

- Keep the website up-to-date
  - This is done by maintaining the pages of the website and posting regular news posts about technological developments, user activity, new partnerships, short portraits of the "people behind READ" etc.
- Keep the Wiki site up-to-date
  - The Transkribus Wiki is an important means of supporting users of the Transkribus platform. The Transkribus "How to Guides" are made available on this website.
- Run the Twitter account
  - Regular posts on the Twitter account
- Apply for conference papers, workshops, etc.
  - The list of papers and workshops from Y1 of the project shows clearly that this was one of the most successful activities in the project and it will be continued in the same way in Y2.
- Organize specific workshops
  - We will continue to organise a number of independent Transkribus workshops over the next year, especially in cooperation with MoU partners in various European countries.

3. Specific activities

In addition to the activities outlined briefly above, we will implement a number of specific actions which can only be performed by the READ project on the basis of the technology we are developing. We believe that such actions are of eminent importance since they clearly showcase the benefit of the technologies and innovative services and tools for our specific target groups.

3.1. Transkribus 1st User Conference

Due to the overwhelming interest in the Transkribus platform we plan to organise a series of dedicated Transkribus User Conferences for 2017, 2018 and 2019.

The first user conference shall take place in 2017. Its main objective is to gather all users who are already working with Transkribus or are interested in Handwritten Text Recognition to get detailed information about existing and planned services. The conference will take place in a location in central European location and address all of our target groups. The idea of "information exchange, synergy and cooperation" will be the main focus. The conference will include the following items:

- Practical advice on how to use Transkribus for scholarly work and family history projects
- Updates on advances in core tools and services: Layout Analysis, Writer Identification, Information Extraction, Table Recognition
- Demonstrations of Innovative products: E-Learning app, ScanApp/ScanTent, FamousHands etc.
- Use cases and stories to show how Transkribus is already being used by scholars and archives
- Feedback and feature requests

An important aspect of the user conference will be that users of Transkribus get to know each other and learn from each other. For example, we know that there are several transcription projects which plan to involve volunteers. It is obvious that the experience gained in the course of such projects in Finland, United Kingdom, Denmark or Italy is in general comparable and will be of high interest to all project managers.

Due to the fact that all Transkribus users register using their email address, it will be very simple to disseminate the invitation for this conference to a large audience comprising all user groups.

Nevertheless special invitations will be provided to funding agencies and DARIAH representatives. The main idea behind the conference is to demonstrate the benefits of the READ/Transkribus Virtual Research Environment for our target groups.

3.2. Dissemination to Scholars

Humanities scholars are an important user group of READ. Most of our invitations for talks and workshops come from this group. They are not only interested in (Handwritten) Text

Recognition but in several other aspects of "Digital Humanities" such as digitisation or digital editions.

For this group we will offer three new specific services and tools which will all be introduced with targeted dissemination activities.

MyModel

The new user interface within Transkribus 1.0 as well as the rights management on the platform enables us to allocate Text Recognition models to individual collections. This means that e.g. a scholar who is working on the remains of a famous writer will be able to train a model with his specific material following exactly those rules (e.g. for special characters) which shall be applied in his edition. The more data which is produced, the more accurate the model becomes (after retraining).

The user can view the training curve of the model which indicates also the performance on the test set. When new material is available the model can be retrained and updated.

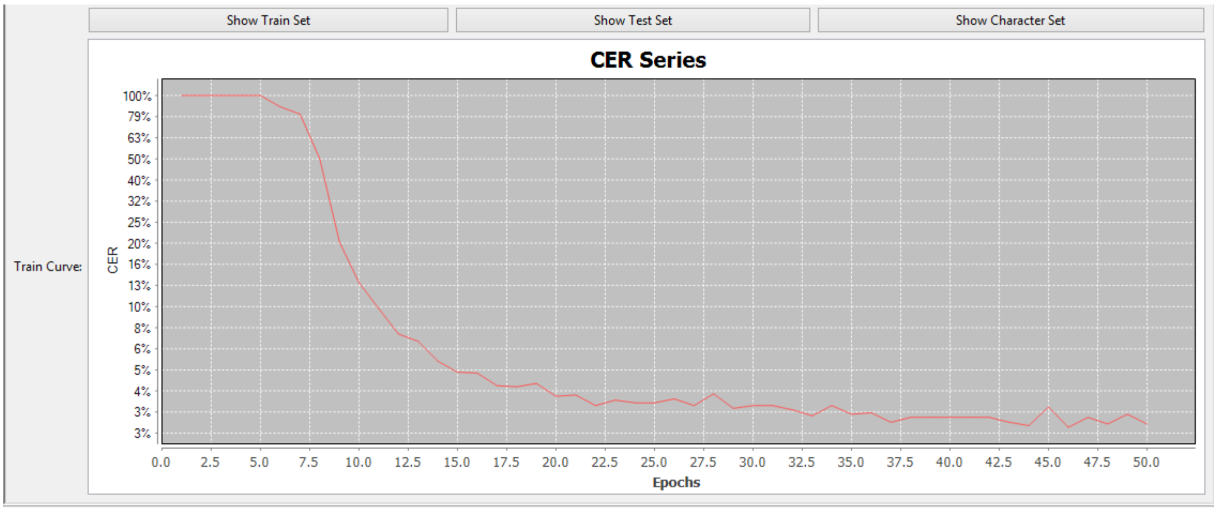
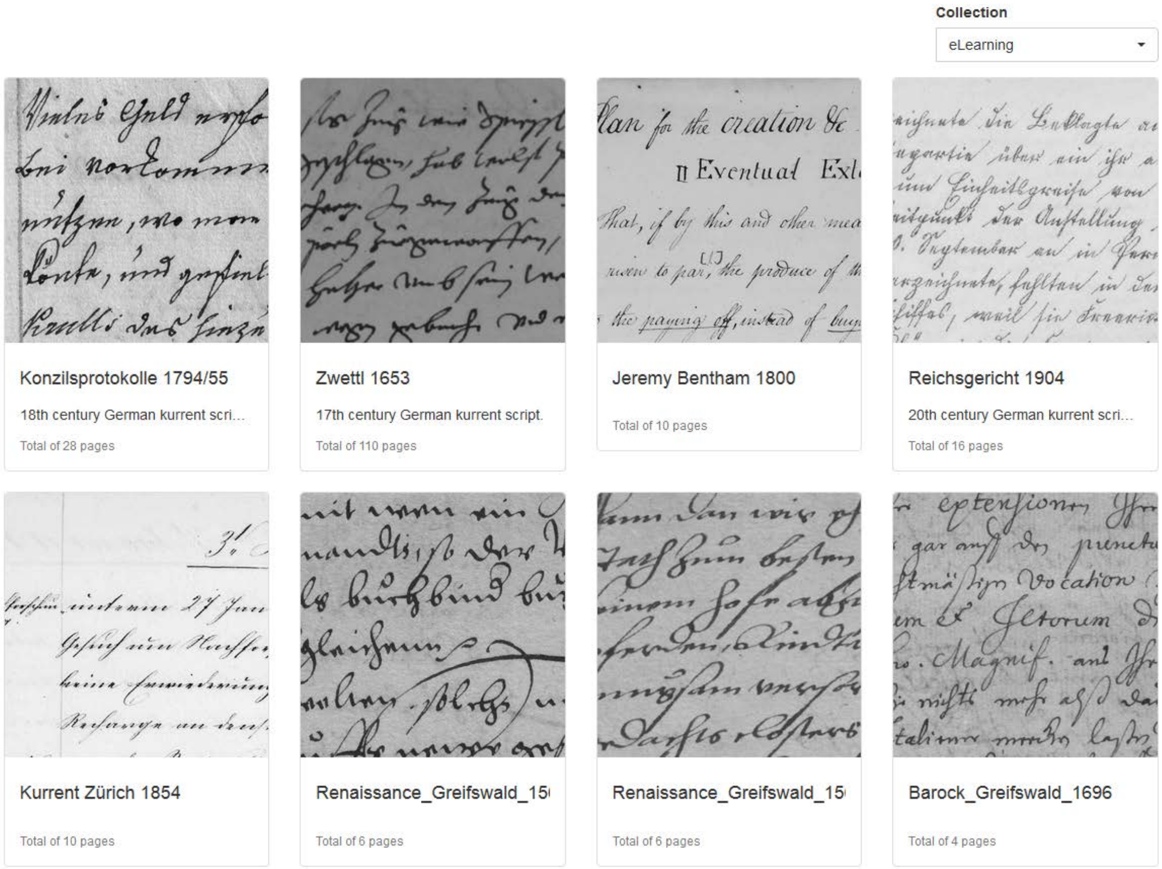


Figure 1 MyModel Training curve

The dissemination activities will focus on users who are already working with Transkribus. Therefore direct mailing, pop-up messages within Transkribus and information on the Transkribus Wiki and READ website will be preferred means to reach the target group.

Script Practising Tool (eLearning)

A large number of scholars in the humanities and especially those who are interested in (digital) editing of historical documents are employed at universities and are teaching students in palaeography. Though there are some digital tools available for digital palaeography the application developed within READ is the first one which enables teachers to prepare training and test documents directly for their students. The Transkribus Script Practising Tool does not replace or substitute theoretical instructions on the history of handwriting, but it offers a tool for students to practice and get familiar with historical handwriting. As outlined in deliverable *D5.3. E-Learning Application* every document within the Transkribus platform can be made available as a "practice document" as well. This gives university teachers the freedom to individually choose and prepare the training material for their students. On the other hand there is of course the chance to share the training material and to benefit from documents prepared by other teachers.



Collection

eLearning

Konzilsprotokolle 1794/55

18th century German kurrent scri...

Total of 28 pages

Zwettl 1653

17th century German kurrent script.

Total of 110 pages

Jeremy Bentham 1800

Total of 10 pages

Reichsgericht 1904

20th century German kurrent scri...

Total of 16 pages

Kurrent Zürich 1854

Total of 10 pages

Renaissance_Greifswald_15

Total of 6 pages

Renaissance_Greifswald_15

Total of 6 pages

Barock_Greifswald_1696

Total of 4 pages

Figure 2 E-Learning - Practising Tool

The promotion of application will rely on a direct mailing action where especially history and philology departments will receive information about the tool and its benefits for teaching palaeography. For this purpose we will collect addresses and names of these departments and prepare mass mailings and information brochures. This will of course be accompanied by the usual channels such as talks, videos, demos, blog posts and Twitter. We are convinced that with this dissemination activity we will reach many scholars who are still working in a more traditional way and may have heard about Transkribus but are still reluctant to use digital tools for their daily work of transcribing or editing documents. Via the “detour” of supporting this group in their task to prepare a course and teach palaeography, we hope to demonstrate a clear benefit for them and to stimulate their interest in other aspects of the Transkribus platform as well.

Transkribus ScanApp and ScanTent

As outlined in *D8.1. Open Innovation Forum* the ScanTent is a completely new product arising from work carried out in *Task 5.6. Crowd-Scanning*. It directly addresses one of the main drawbacks which humanities scholars often face in their daily work: the material that they find most interesting has not yet been digitised by an archive or library and Digitisation-on-Demand or reproduction services are often prohibitively expensive (several EUR for one image).

The combined product ScanApp and ScanTent will enable scholars to use their mobile phone to take high quality pictures of archival documents. The main benefit is that they have both hands free for manipulating the document (opening it, keeping it still) and that images are directly sent to the Transkribus platform for further image processing. These images can be equipped with a geo-tag so that the images can be automatically assigned to an archive. This would be very helpful for researchers who need to know which archive a document comes from. . In a more advanced stage, the archive could also attach a QR code to the document, which provides details of the shelf-number of the document. This will allow the researcher to identify the document and also facilitate the simple easy transfer of images to the archive's digital repository.

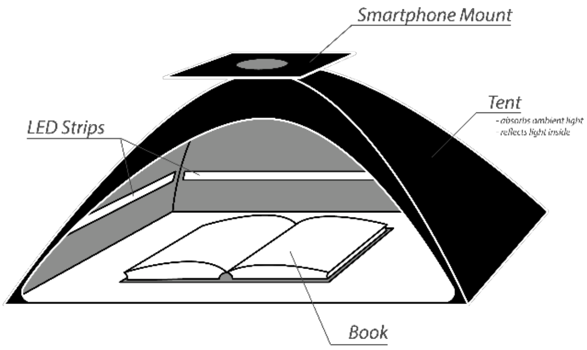


Figure 3 ScanTent

We have shown some early previews of the prototypes to archivists from Austria and scholars from the École nationale de chartes in Paris – and their reaction was enthusiastic. They see the clear benefit to get good, reliable images from documents in quick and easy.

The dissemination campaign for the Transkribus ScanApp and ScanTent will depend on the further marketing of the ScanTent as a product which can be bought in online shops such as Amazon or Conrad. This will take some time and may not take place during 2017. But in order to make the product known to scholars we will start to build up a group of early adopters who will receive prototypes of the product and test it in real-world conditions. For this purpose we are also planning a "scanathon" where a group of volunteers is equipped with ScanTents and asked to take pictures of historical documents for 2-3 hours. If we assume that one user can produce some hundreds of images within an hour, such a scanathon with 10 users would produce some 10,000 images or 20,000 pages (since most documents can be scanned double sided). A typical yearbook running for some 50 years in two volumes with 500 pages per volume could be completely scanned within such a scanathon.

Again such dissemination activities will not only showcase services and tools with a clear benefit for users, but also raise attention and interest for the more elaborate services in the Transkribus platform.

3.3. Dissemination to Archives/libraries

ScanTent accounting system

One of the main dissemination activities for archives is built on the ScanApp and ScanTent application. The main idea is that archives and libraries may use the product for two purposes. Firstly, the product allows them to get hold of the images of documents which are taken by users in their reading rooms. If archives are able to pass over a QR code containing the shelf-mark of the document, it will be possible to build up a digital repository of user-generated image files which fit into the hierarchical structure of the archive. Secondly, the product offers archives the chance to earn some money from the self-service scanning performed by the users. The accounting system for running this service could also be provided by the Transkribus platform, meaning that archives and libraries would not have to use their resources to maintain this service themselves.

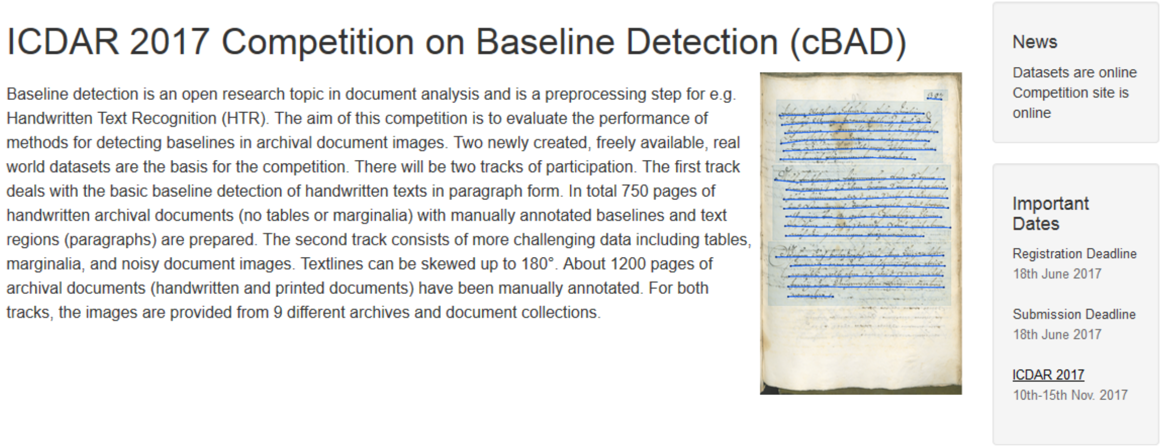
Similar to the early adopters group of scholars, we plan to provide a targeted group of archives with the ScanApp and ScanTent options (first of all without accounting) so they can get their users to test them out. First talks with MoU partners are already taking place.

3.4. Dissemination to Computer scientists

ScriptNet

During Y1 we have developed the ScriptNet application which enables not only research groups from READ, but is also open to other groups to organize scientific competitions in a simple way. ScriptNet is equipped with a scoreboard and upload mechanism for result files so that participants of scientific competitions can take part and submit their methods and results on a specific challenge.

The main specific dissemination activity is that in 2017 several competitions will be organised via ScriptNet and therefore the Image and Document processing community will become aware of this tool.



ICDAR 2017 Competition on Baseline Detection (cBAD)

Baseline detection is an open research topic in document analysis and is a preprocessing step for e.g. Handwritten Text Recognition (HTR). The aim of this competition is to evaluate the performance of methods for detecting baselines in archival document images. Two newly created, freely available, real world datasets are the basis for the competition. There will be two tracks of participation. The first track deals with the basic baseline detection of handwritten texts in paragraph form. In total 750 pages of handwritten archival documents (no tables or marginalia) with manually annotated baselines and text regions (paragraphs) are prepared. The second track consists of more challenging data including tables, marginalia, and noisy document images. Textlines can be skewed up to 180°. About 1200 pages of archival documents (handwritten and printed documents) have been manually annotated. For both tracks, the images are provided from 9 different archives and document collections.

News

Datasets are online

Competition site is online

Important Dates

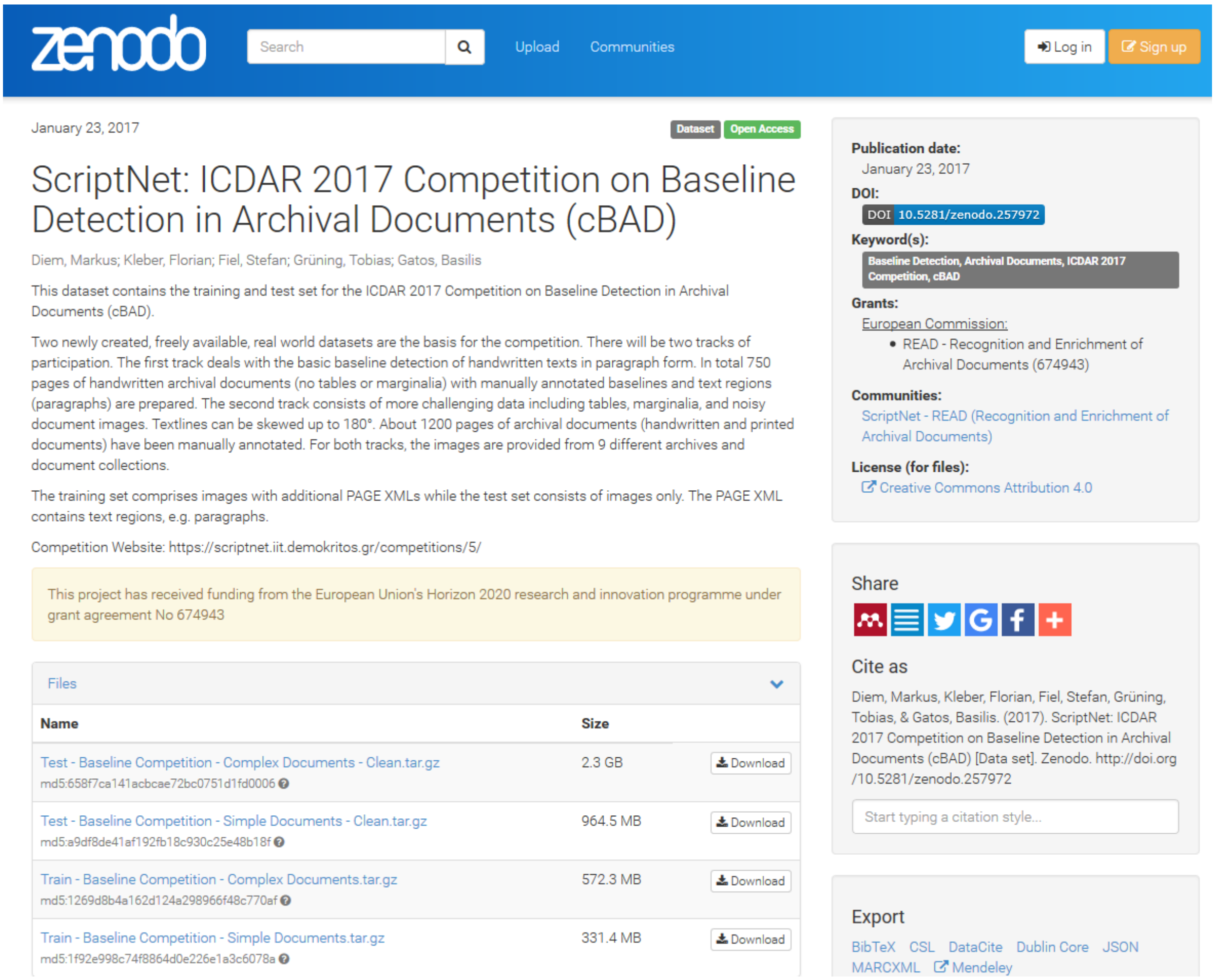
Registration Deadline
18th June 2017

Submission Deadline
18th June 2017

ICDAR 2017
10th-15th Nov. 2017

Figure 4 ScripNet – an example for an active competition on baseline detection

This activity is complemented by the open and long-term availability of the datasets which form the basis of these competitions. This data has been uploaded to Zenodo, where it is available as Open Research Data and citeable with a Digital Object Identifier (DOI).



January 23, 2017

ScriptNet: ICDAR 2017 Competition on Baseline Detection in Archival Documents (cBAD)

Diem, Markus; Kleber, Florian; Fiel, Stefan; Grüning, Tobias; Gatos, Basilis

This dataset contains the training and test set for the ICDAR 2017 Competition on Baseline Detection in Archival Documents (cBAD).

Two newly created, freely available, real world datasets are the basis for the competition. There will be two tracks of participation. The first track deals with the basic baseline detection of handwritten texts in paragraph form. In total 750 pages of handwritten archival documents (no tables or marginalia) with manually annotated baselines and text regions (paragraphs) are prepared. The second track consists of more challenging data including tables, marginalia, and noisy document images. Textlines can be skewed up to 180°. About 1200 pages of archival documents (handwritten and printed documents) have been manually annotated. For both tracks, the images are provided from 9 different archives and document collections.

The training set comprises images with additional PAGE XMLs while the test set consists of images only. The PAGE XML contains text regions, e.g. paragraphs.

Competition Website: https://scriptnet.iit.demokritos.gr/competitions/5/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Files

Name	Size	
Test - Baseline Competition - Complex Documents - Clean.tar.gz md5:658f7ca141acbcae72bc0751d1fd0006	2.3 GB	Download
Test - Baseline Competition - Simple Documents - Clean.tar.gz md5:a9df8de41af192fb18c930c25e48b18f	964.5 MB	Download
Train - Baseline Competition - Complex Documents.tar.gz md5:1269d8b4a162d124a298966f48c770af	572.3 MB	Download
Train - Baseline Competition - Simple Documents.tar.gz md5:1f92e998c74f8864d0e226e1a3c6078a	331.4 MB	Download

Publication date:
January 23, 2017

DOI:
DOI 10.5281/zenodo.257972

Keyword(s):
Baseline Detection, Archival Documents, ICDAR 2017 Competition, cBAD

Grants:
European Commission:
- READ - Recognition and Enrichment of Archival Documents (674943)

Communities:
ScriptNet - READ (Recognition and Enrichment of Archival Documents)

License (for files):
Creative Commons Attribution 4.0

Share

Cite as

Diem, Markus, Kleber, Florian, Fiel, Stefan, Grüning, Tobias, & Gatos, Basilis. (2017). ScriptNet: ICDAR 2017 Competition on Baseline Detection in Archival Documents (cBAD) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.257972

Start typing a citation style...

Export

BibTeX CSL DataCite Dublin Core JSON MARCXML Mendeley

Figure 5 The cBAD dataset in Zenodo

With regards to specific dissemination activities, we plan to organise workshops and tutorials for researchers from the computer science field to demonstrate how they can use the ScriptNet infrastructure for their purposes.

3.5. Dissemination to the general public

The main activity for this target group will be the launch of the *FamousHands* campaign. This activity is explained in more depth in *D3.4 European Hands*. It will enable users to contribute to an open database of the handwriting of famous persons. This dataset can be used to find writing of such persons within large amounts of digitized images. It will be relatively simple for users to upload images of the handwriting of famous people, along with some additional information. We therefore expect that a large number of users may be interested in getting involved with this initiative.

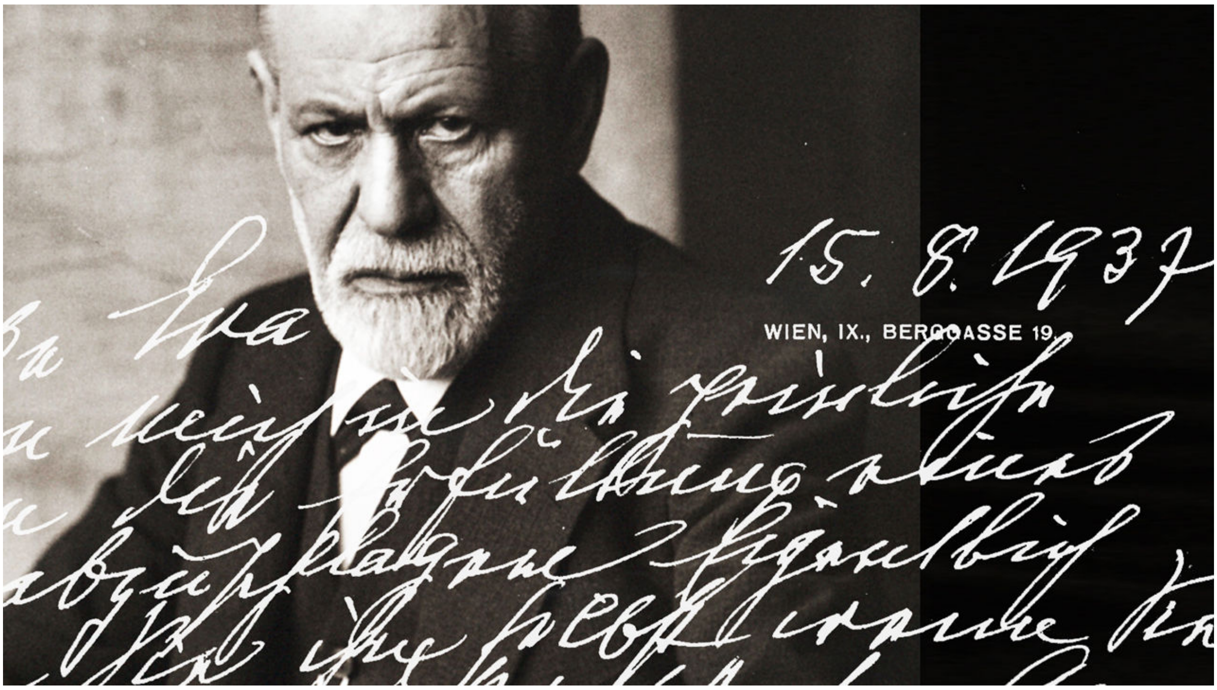


Figure 6 Sigmund Freud's handwriting

The main dissemination activity in this respect is to prepare a convincing press release for news media where the task of “building a public database of the handwriting of famous persons” is weaved into a convincing story. This campaign may also be used to set up a Transkribus Facebook account, where we can encourage members of the public to take part in the campaign.