Recognition and Enrichment of Archival Documents

D2.5. Dissemination and Awareness Plan

P2

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	31.12.2017
Actual date of delivery	21.02.2018
Date of last update	21.02.2018
Deliverable number	D2.5
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

For Y3 we plan several specific actions to channel the overwhelming interest in the project and the Transkribus platform. These actions comprise on the one hand the progress in Handwritten Text Recognition, the involvement of more archives, libraries and scholars in the production of training data, a higher coverage in terms of European countries contributing to the establishment of modern technologies in the archives domain. Specific actions are also planned for volunteers and the public as well as for computer scientists.

1. Specific dissemination activities

1.1. Introduction

The READ project comprises a large number of different strands of activities which all can be summarized in one sentence:

READ revolutionizes access to archival documents

After two years of work we are more convinced than ever that this ambitious claim holds true.

READ/Transkribus is the only platform for historical documents where users are enabled to transcribe, train, recognize and search handwritten documents with cutting edge technology. READ has clearly demonstrated the progress made in this respect with impressive figures concerning the ability of machine learning methods to recognize historical documents and make them searchable. Of course our work in READ strongly benefits from the progress made in general in the artificial intelligence and machine learning domain, however there is no project or platform worldwide which has put such a strong emphasis on the recognition of historical documents with such a comprehensive approach.

It is the task of the "Dissemination and Awareness Plan" to spread this message to all target groups, such as archives, libraries, humanities scholars, computer scientists and the broad public.

1.2. Messages and claims for Y3

If we go into more detail we can specify our general message from above into several ones which are clearly dedicated to user groups and use cases. These messages will form the basis for all of our dissemination activities in Y3 and above.

The key messages for our target groups in respect to the READ/Transkribus platform are:

1. Enjoy the privileges of a **private environment**. The usage of documents in the platform does not include any making-available nor the distribution of the documents to the public therefore copyright regulations are much less important than for public

repositories or digital libraries.[1] Also benefit from the **General Data Protection** laws in the European Community which guarantee a high level of security and trust with respect to the content of the documents uploaded to the platform.

2. Work with **any kind of historical document** and upload thousands or even tens-of-thousands of scanned pages easily to your private collection in the platform. The technology is language and script agnostic and can be used for modern documents as well as for medieval ones.

3. Get independent of the digitisation efforts of archives and libraries by using your own **smartphone** and the **ScanTent** for document digitisation. Upload your files from your smartphone directly to Transkribus and share them with archives and libraries.

4. Benefit from **HTR models which were already trained** on material from other Transkribus users, e.g. run the English Writing Model on 19th and 20th C. documents and receive results on English documents without any further training.

5. **Train the HTR engine** according to your requirements – e.g. to cover a specific hand or a specific way to transcribe (i.e. abbreviations or special characters). Use **existing transcriptions** to create an HTR model in a fully automated way and also retrain the HTR models as often as new material or new technology become available.

6. **Share HTR models** with other users – without infringing copyright or personal rights.

7. Search any recognized document with **keyword spotting technology** – a much more powerful way to explore historical documents than with conventional full-text search.

8. **Involve users and volunteers** with a web-interface for viewing, validating and editing documents in a simplified browser environment.

9. **Encourage students and volunteers** to train themselves in reading historical handwriting with the Transkribus eLearning component.

10. **Contribute to the evolvement of research and development** in the domain of historical documents by making data available to computer scientists, namely via open scientific competitions and challenges.

These 10 important messages need to be completed by another – even more important – message which has to make a clear statement on the future of the Transkribus platform.

The two key messages about the future of the Transkribus platform after the end of the READ project (06/2019) are:

11. In 2018 READ members are aiming to set up a **legal entity** which will run and **maintain the Transkribus platform** after the end of the project. This legal entity shall be based on a membership model where platform operators and platform customers will collaborate to achieve the highest **benefit for all**.

[1] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society

12. Services in Transkribus will remain **free for everyone** up to a certain amount of processed pages per year. After the end of the READ project a **subscription fee** will be charged above this limit. **Large scale projects** will be calculated on the basis of image based fees.

These 12 messages will be the basis for all our dissemination and awareness activities in Y3 of the project. Of course they need to be detailed and adapted to specific target groups and services but we believe that they are highly important for all users who want to make their plans and to further benefit from Transkribus services.

1.3. Specific actions for Y3

In Y3 we plan four to five specific actions towards disseminating the most important results of the READ project respectively for the Transkribus platform.

These actions are:

- 2nd Transkribus User Conference
- HTR+
- Training data from and for everyone
- learn.transkribus.eu (eLearning)
- ScanTent and DocScan
- The digital historian

1.3.1. 2nd Transkribus User Conferences

One of the highlights of Y2 was the arrangement of the 1st Transkribus User Conferences in Vienna (2.-3. November 2017).

Its main objective was to gather all users who are already working with Transkribus or are interested in Handwritten Text Recognition and to offer them a forum for information exchange but also an update on existing and planned services. The conference took place at the Technical University Vienna (CVL). The idea of "information exchange, synergy and cooperation" was the main focus and acknowledged by the participants.

The interest in the conference was overwhelming. Though we were very cautious with announcing the event already several weeks before its start we had to close registration. More than 90 people took part, coming from 18 different countries including the U.S., Russia and Turkey.

In Y2 we will repeat the conference. This time the focus will be put on the following items:

- Launch of READ-coop and the business model
- New services for users, such as model sharing, extended web-interface, transcription-on-demand service, keyword spotting with validation service, ground truth for everybody,…
- Updates on advances in core tools: Writer Identification, Information Extraction, Table Recognition
- Demonstrations of products: learn.transkribus, ScanTent/DocScan
- Use cases and stories to demonstrate 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 but also understand that they contribute to the sustainability of the Transkribus platform if they go home and convince their archives, libraries or universities to become member in READ-coop.

1.3.2. HTR+

As we have indicated in our yearly report we can expect that in 2018 we will achieve a significant improvement of the HTR accuracy rate. A reduction of 30-50% or even more is realistic and can be guaranteed already today. This "good news" is an excellent starting point to (re-)contact users from Y1 and Y2 where we got training material and already trained first models.

We will therefore set up an information campaign and a standard workflow where we

- retrain a HTR model with the new engine
- contact the owner of the training data
- arrange a Skype or Hangout meeting with screen sharing
- demonstrate the new results
- take this as an opportunity to report about other achievements in the platform and
- try to intensify or re-establish the existing contact with the clear offer to join the Transkribus platform respectively READ-coop as a member

Such sessions may be organised as a single meeting or via a series of webinars with several users at once.

A rough calculation shows that about 100-150 institutions need to be contacted for this dissemination activity and about 350 new models need to be trained.

1.3.3. Training data from and for everyone

As a matter of fact the success of HTR, layout analysis and document understanding depends strongly on the availability of meaningful training data (ground truth). Since the READ project has a dedicated budget for the generation of such training data we plan to make even more institutions and potential users aware of READ and the Transkribus platform by organising a campaign which offers a "deal" to archives and libraries, as well as humanities scholars, researchers but also family historians and genealogists.

This deal can be formulated in the following way:

- Send us all kinds of documents for which you would be interested in text recognition and information extraction.
- READ/Transkribus will create ground truth data, e.g. for Layout Analysis, Table Recognition, HTR or Document Understanding – using the remaining ground truth budget from the project
- This training data will be the basis for automated processing, but they will also be shared between the content providers and the Transkribus platform members

Since Transkribus is already well-known in countries such as Austria, Belgium, Finland, France, Germany, Ireland, Luxembourg, Netherlands, Norway, Sweden and UK we will first and foremost approach the following countries with this campaign:

- Bulgaria, Czech Republic, Croatia, Estonia, Greece, Hungary, Italy, Lithuania, Latvia, Malta, Poland, Portugal, Rumania, and Spain.

For this purpose we will contact the archives associations in these countries as well as use our address list for history departments in Europe to directly approach the two most important target groups.

1.3.4. lern.transkribus.eu

In Y1 and Y2 we prepared the eLearning application which has now reached a professional level and can be launched to a large number of users. In Y2 we gathered nearly 2000 addresses of history departments in Europe but also from the US and Canada.

These data will be used to start a specific campaign for making the site known to scholars from the history domain as well as their students. The campaign will start during spring 2018 once we can offer a good selection of different documents in several languages and covering different scripts. We will strongly work with instruction videos, both for students or volunteers as well as for scholars who want to include specific documents in the Transkribus learn collection.

1.3.5. Market ScanTent and DocScan

It is too early for a solid timetable but we expect that during 2018 we will be able to produce a first set of 1000 pieces from the ScanTent. Once a date is foreseeable we will prepare an information campaign to market this device. In contrast to other cases where we preferred a soft-launch this will be different with the ScanTent. Here it will be important to reach the highest possible number of potential customers in a short time period.

The two target groups are humanities scholars who are working professionally with archival material and have access to resources, such as project grants or students working in their team.

The other target group are archives and libraries which want to provide a cost-effective alternative for users who want to take images from their archival holdings.

The more than 10.000 registered users (we will reach this amount in April or May 2018) will be our first addressees. They will receive not only information on the ScanTent via a personal email, but also get links to videos showing how the ScanTent can be used at home for the digitisation of personal documents, but also in archives and libraries. We will also explain the idea that the ScanTent enables volunteers to easily contribute to the digitisation of archival collections via Scanathons and similar events.

Due to the fact that the DocScan app also tracks the location of each image taken we can automatically provide anonymous information for users, where the ScanTent has been used. E.g. in March 2018 a researcher from the University of Würzburg will go to Egypt and take images in archives from there. Also users from the U.S. already approached us and once can imagine that the idea to be part of a great “scanning endeavour” will be motivating to many users.

1.3.6. The digital historian

Since HTR, Layout Analysis, Keyword Spotting and similar tasks are – more or less – solved problems from a research point of view we should think on the next generation of challenges. And in this respect it makes sense to have a look to the big players in the domain, such as IBM or Google. Their way to make a “grand challenge” (IBM) tangible for a wide audience is to construct a simple looking challenge which internally requires to solve fundamental technical questions.

This was the case with Deep Blue playing chess against Garry Kasparow (IBM), Watson taking part in the Jeopardy quiz (IBM) and the DeepMind team playing Go against the best players of the world (2016) and now against their best computer programmes from 2014 with a new method.[2]

If we transfer this idea to our domain we believe that we are able to set up a challenge which will be of interest to archives, libraries as well as humanities scholars and the public. We call this challenge "The Digital Historian" and the main idea behind is that – similar to the Jeopardy quiz – now serious historical questions and problems will need to be answered on the basis of a large historical collection of primary papers.

The challenge could be drafted in true continuation of the famous Turing test in the following way:

- A team of academics consisting of "the" experts in the field will compete against a team of computer programmes (including software for HTR, text mining, natural language processing, etc..).
- The basis of this competition are e.g. the remaining papers of a famous person or any other large collection of unpublished papers.

The challenge will be:

- Both teams will be asked some meaningful questions on the content of the Bentham collection (about 100.000 pages).
- The questions will come from the jury consisting of experts in the field but also from the public.
- These questions may include trivial ones, but also very abstract ones:
    - A trivial one could be to ask if Jeremy Bentham ever said something about Norway, and if yes, where and what was the content.
    - An abstract one could be to ask for the general attitude of Jeremy Bentham towards the French Revolution, including the development of his thinking over time and his main arguments and a statement on the relationship to the English political system.
- Both teams would provide their answer within a given time frame and afterwards the jury would publish the results which need to come in natural language.
- The evaluation could be done by the jury, but of course also by the public and provide a good impression how far computer technology already would be in terms of historical research.

Such a competition could either be organised as an open competition or as part of an Artificial Intelligence or Machine Learning conference. In Y3 of the project we will discuss this idea with leading researchers in the domain and prepare the ground for it in terms of selecting appropriate datasets, contacting teams of historians and archivists and computer scientists.

[2] IBM today goes even a step further and has launched the AI xPrice competition where AI teams all over the world were asked to take part in a multi-year competition. The winner will take 5 mill. $ at home.

2. General dissemination and awareness channels

2.1. READ website

http://read.transkribus.eu

In Y1 this website was set up and in Y2 the website was maintained and updated according to the progress in the project. All technical deliverables are available as well as all new MoU partners were included.

Most importantly regular posts where published about progress in the project, about conferences or about new partnerships and short portraits of the "people behind READ".

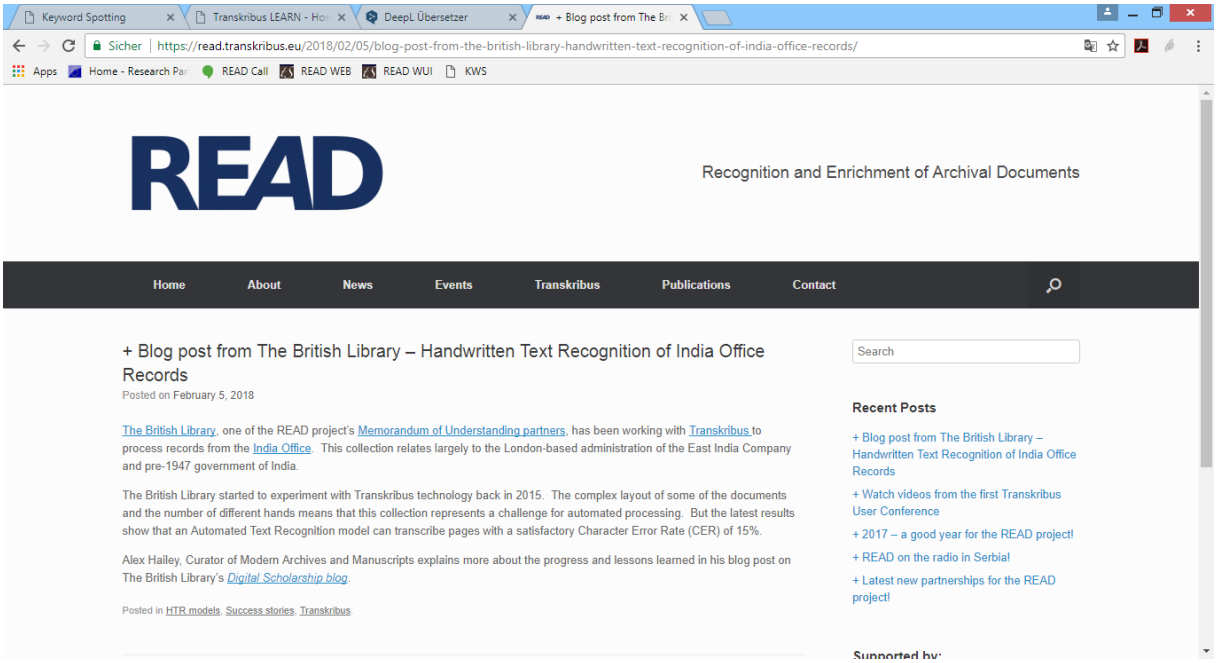


Figure 1 Blog post about a blog post by the British Library

In Y3 this website will be continued but the focus will shift to the new Transkribus website (see below).

2.2. Transkribus website

http://transkribus.eu/

This website was set up already before the start of READ. In Y3 we will create a new version of the Transkribus website. The current site is clearly outdated and needs to be adapted to the current status of the Transkribus platform. Very likely we will use a bootstrap template to construct the site in a modern and fresh way.

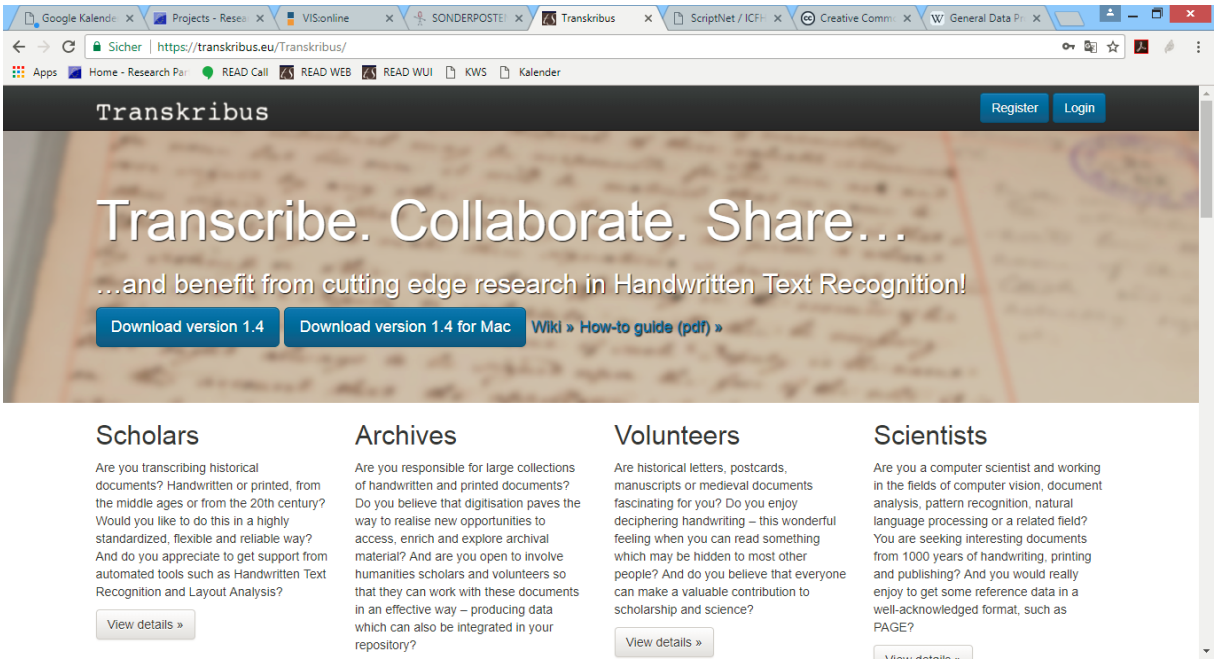


Figure 2 Transkribus website

Especially the corporate structure of the Transkribus platform shall be emphasized on the new website. Here we have to find a way to provide a good overview on the many activities going on in the READ project but also in the Transkribus platform as there are:

- Transkribus Expert client
- Transkribus document library
- ScanTent/DocScan
- learn.transrkribus.eu
- ScriptNet
- FamousHands

The following new features which are dedicated to the idea of sharing and collaborating will be implemented in this website as well:

- An overview of HTR models which are available to the public. The HTR model will be described by the users who created it but also contain some example pages of the dataset so that everyone interested can easily see on which scripts the model was trained.
- An overview of research projects working with Transkribus. With the growing number of users more and more overlapping of research fields can be observed. E.g. there are several groups working on Latin documents from the middle ages, or several groups working on WW2 documents. In many cases an exchange of information or data may be of benefit for each of the groups. Of course the MoU partners will be the first addressees for this list.

The site will of course also contain usual elements such as blog posts, news, HowToPapers and links to the YouTube channel.

Last but not least we are considering to implement a single sign-on system will enable users to switch directly between several Transkribus applications.

2.3. Transkribus web-interface – library

http://transkribus.eu/read/library/

In Y1 and Y2 we were working on the Transkribus web-interface / document library in WP4. But this site will also be an important communication channel since all users who are working in the platform also have access to their private collection via the web-interface. We believe that the chance to view and edit documents directly in the browser will make Transkribus even more attractive to many users, but also open up new use cases, such as the involvement of volunteers and the crowd.

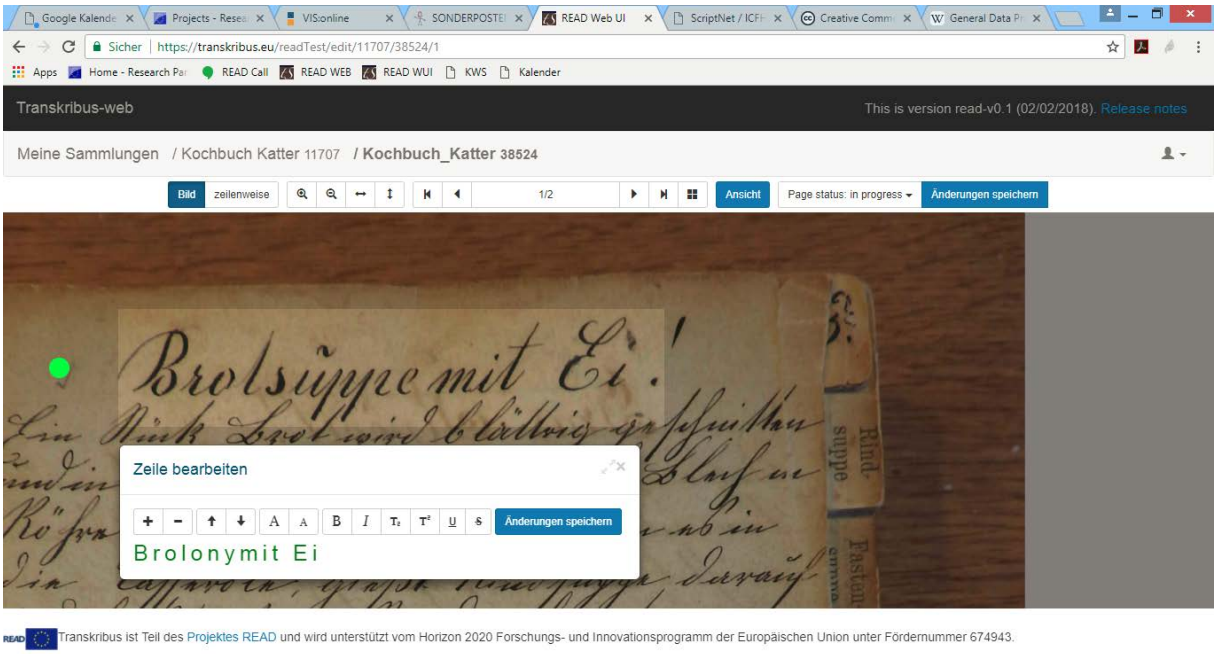


Figure 3 Transkribus web-interface / document library

The Transkribus web-interface / document library will be an important means to involve new users in the platform. In Y3 we will put a strong emphasis on this site and further develop it in strong connection also with the DocScan app.

2.4. Transkribus wiki

http://transkribus.eu/wiki/

The Transkribus Wiki was set up in Y1 and also maintained in Y2. Nevertheless in Y3 we will streamline the complete support service for Transkribus users and go more in the direction of the Transkribus “How to Guides” as well as video instructions (see below).

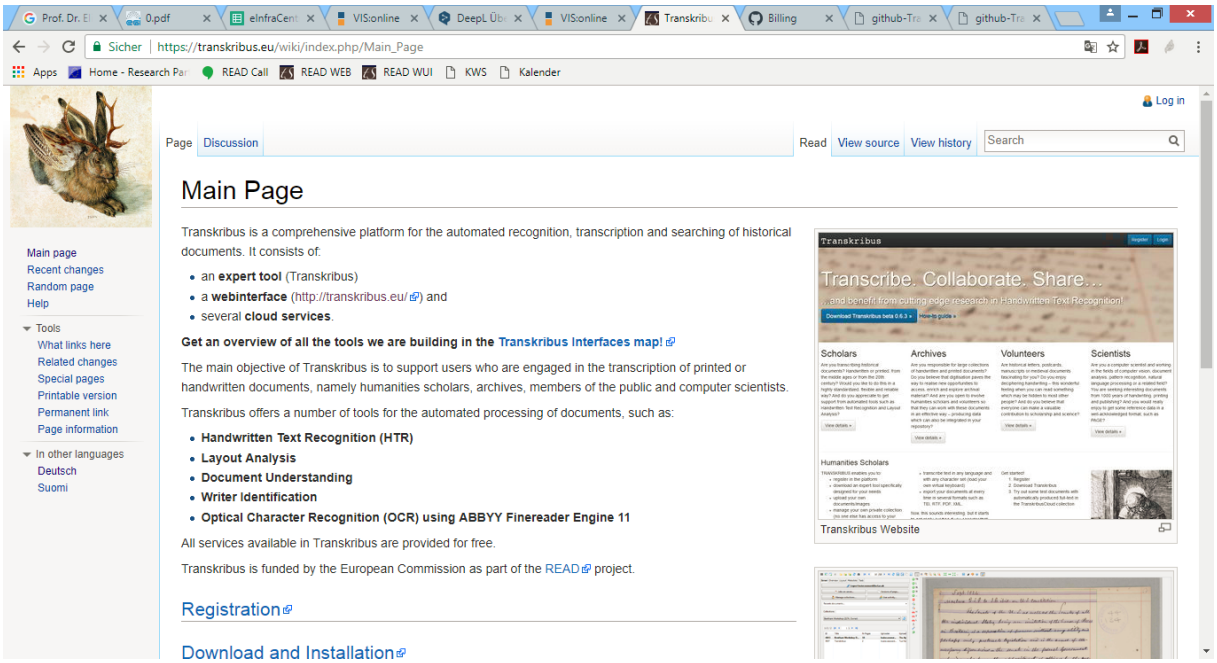


Figure 4 Transkribus Wiki

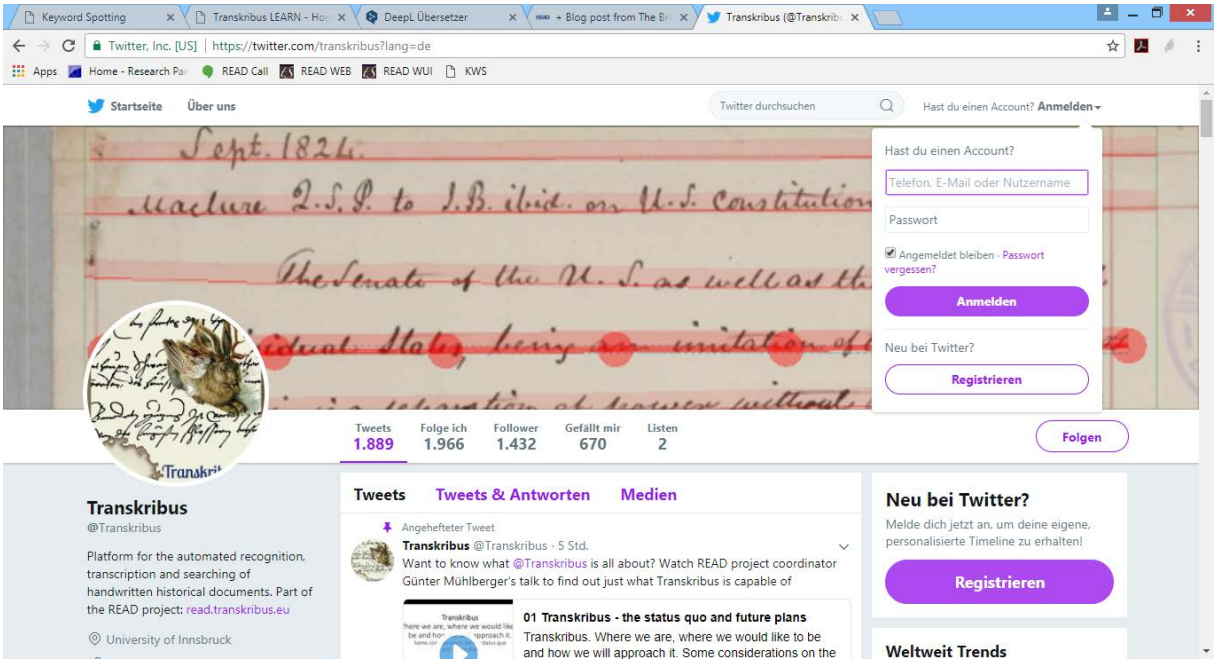
In order to reduce the maintenance effort for the Wiki we will drastically reduce the amount of content made available via this site.

2.5. Twitter

https://twitter.com/transkribus/

In the digital humanties domain the use of Twitter is – in contrast to the computer science domain – rather popular and a good means to reach this target group.

UCL, StAZh and CVL were very active on the Transkribus user account. An impressive number can be reported (as of 5th February): 1.889 tweets, 1432 follower and 670 likes. This will be continued in Y3.



2.6. YouTube video channel

https://www.youtube.com/channel/UC-txVgM31rDTGlBnH-zpPjA

Already in Y1 all talks from the Kick-off Meeting in Marburg where published as videos on YouTube. Later on a specific YouTube channel was created for Transkribus. Main content are HowTo videos for beginners. This aspect shall be emphasized in Y3: All HowToPapers will be accompanied by short screencast videos.

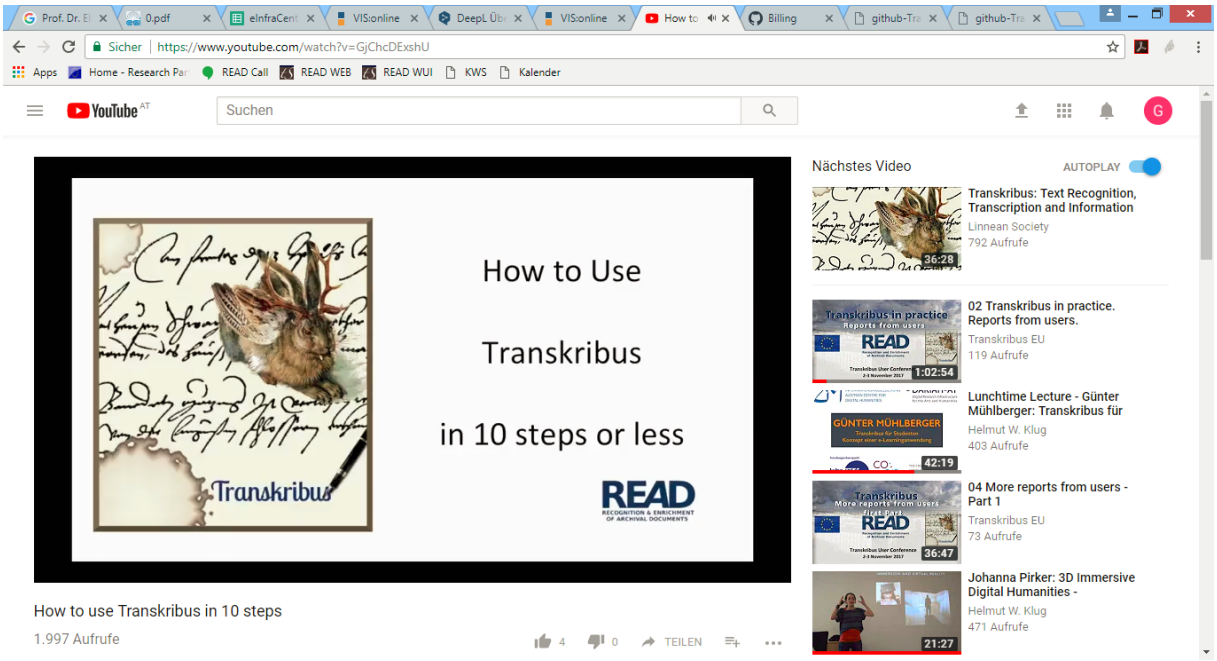


Figure 5 YouTube channel for Transkribus

The 10 steps guide was viewed nearly 2000 times at the YouTube site of Transkribus.

Due to some great support from volunteers we were also able to publish all talks (more than 10h) from the Transkribus User Conference.

2.7. ScriptNet

https://scriptnet.iit.demokritos.gr/

The ScriptNet website is dedicated to organise scientific competitions in the READ project (and above). It was developed in Y1 and further improved in Y2.

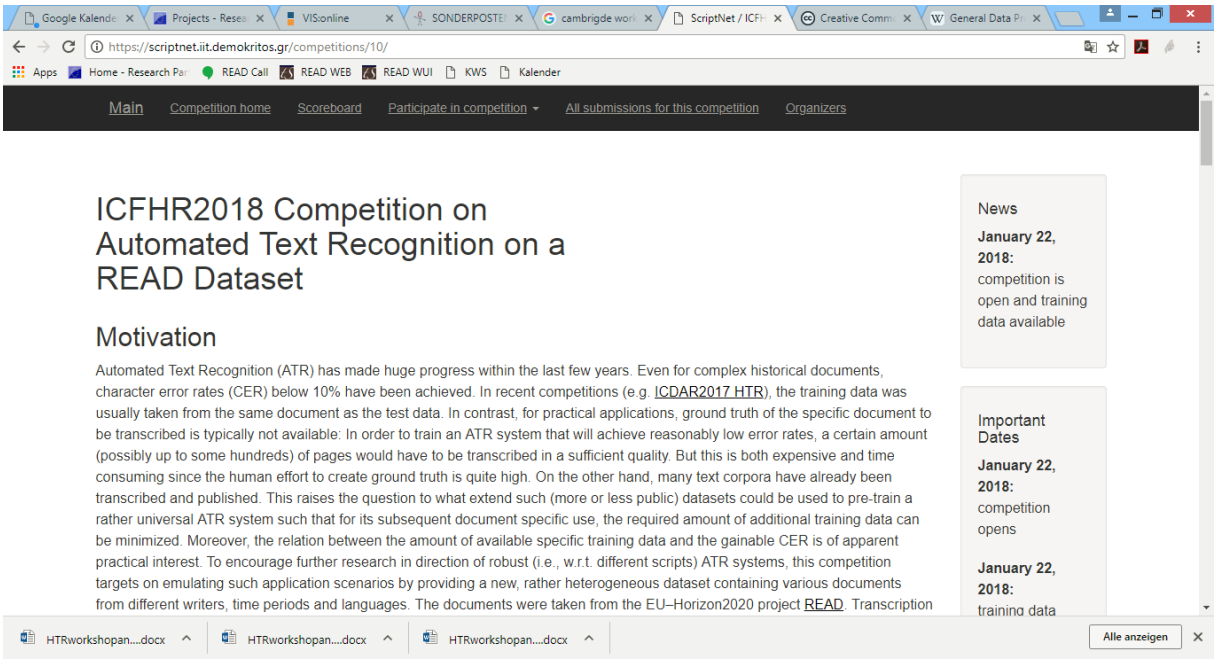


Figure 6 ScriptNet site with ICFHR 2018 HTR competition site

Five competitions were organised in 2017, and for 2018 already an HTR competition was accepted by the International Conference on Frontiers in Handwriting Recognition (ICFHR).

The target group for this site are computer scientists but also archivists and libraries who want to get an impression how their documents are processed as part of a scientific competition. This cross domain collaboration was also emphasized by respective blog posts on the READ website. In Y3 this aspect shall become even more important.

2.8. ZENODO

https://zenodo.org/communities/scriptnet/

In strong connection with ScriptNet and the concept of Open Research Data the datasets created and used in READ are – as far as owners agree – made available publicly via the ZENODO repository. This process was initiated in Y1 and became “routine” in Y2. Also a specific ScripNet – READ community was created.

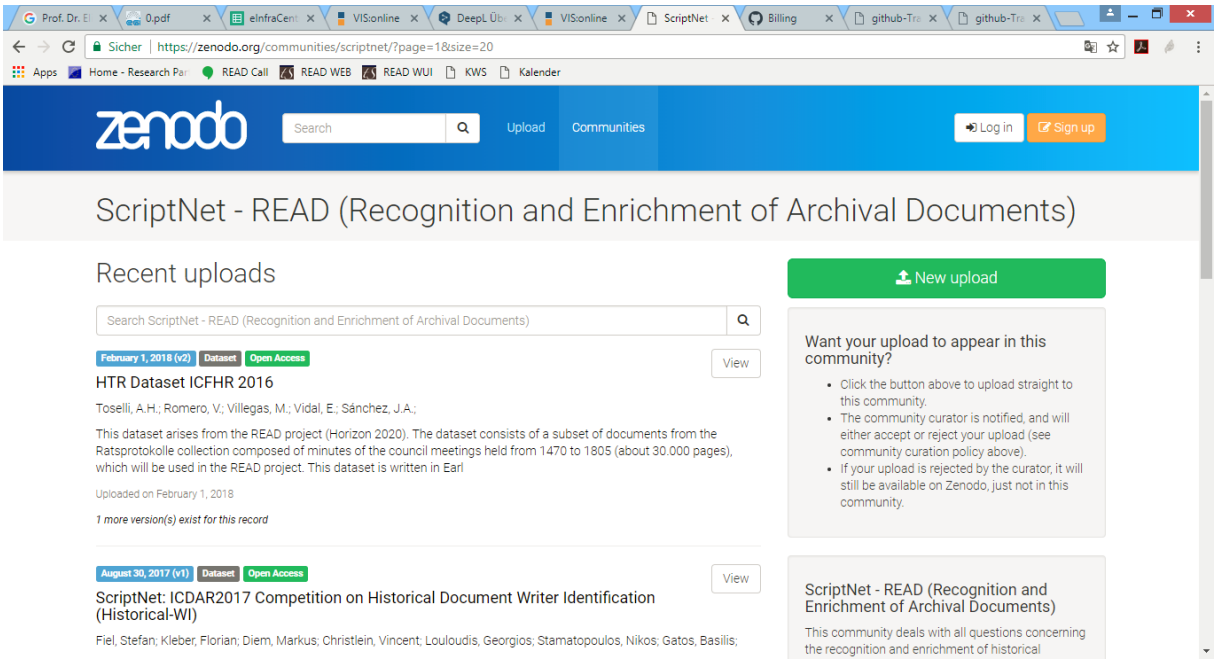


Figure 7 ScriptNet - READ datasets at ZENODO

In Y3 we will continue with this work but are also considering to make it even more convenient for Transkribus users to include their data as datasets in ZENODO. From a technical point of view it would be a doable task to use the ZENODO API directly for this purpose and to include a "ZENODO Export" directly in the Transkribus platform. The complete package including metadata could be directly sent to ZENODO and be made available via ScriptNet-READ. The implementation of such a service will depend on the availability of resources.

2.9. ScanTent/DocScan

https://scantent.cvl.tuwien.ac.at/

In Y2 CVL and UIBK set up a dedicated website for the ScanTent and DocScan. We did again a soft launch in order to avoid requests and raise expectations too high, but the site serves as a good starting point for those users who became already aware of this development or are part of the testing of the software and the device.

The site contains also two instructional videos how to set up the ScanTent.

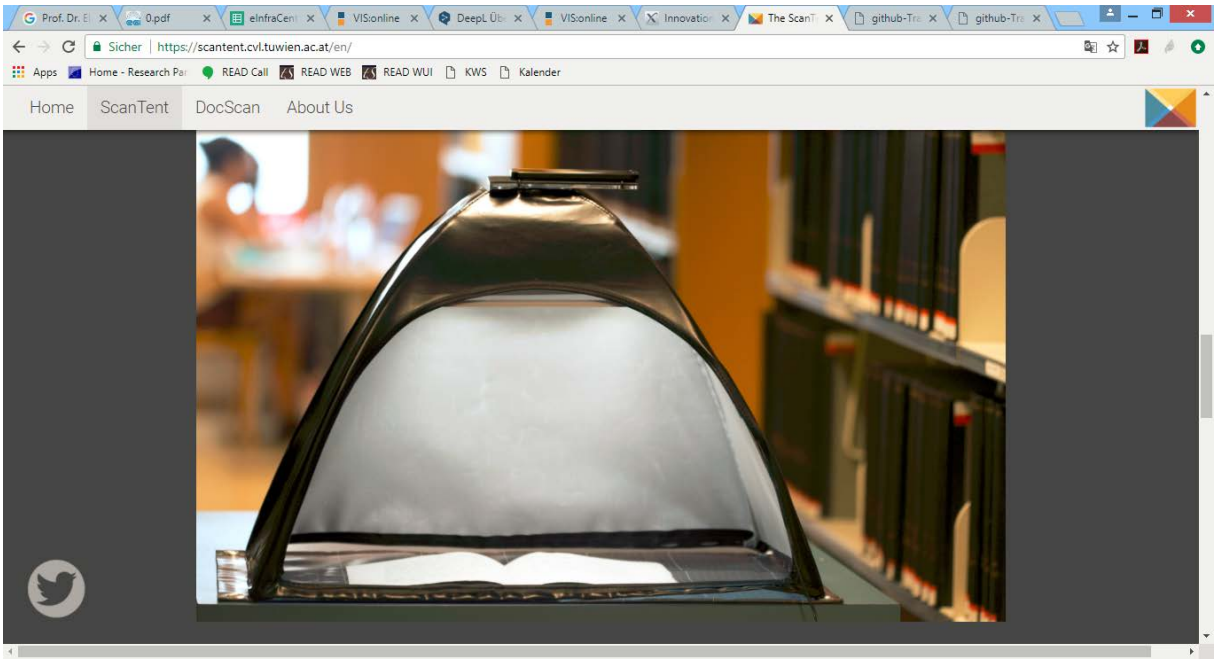


Figure 8 ScanTent website

The site will be adapted according to the progress made in Y3.