Project Number: 674943

Project Acronym: READ

Project title: Recognition and Enrichment of Archival Documents

Periodic Technical Report for 2017 – Public version

(Y2 of the project)

Part B

Period covered by the report: from 1.1.2017 to 31.12.2017

Periodic report: 2nd

Table of Contents

Executive Summary

Year 2 of the READ project brought major achievements: In some fields, such as Layout Analysis, HTR/Keyword Spotting or the ScanTent/DocScan app real breakthroughs can be reported.

Moreover the READ platform, Transkribus, is now for the first time able to cover the complete workflow comprising all major steps: (1) document digitisation, (2) training of specific text recognition models and (3) indexing and keyword spotting of historical documents of any language or writing style.

About 3500 users registered in 2017, so that e2017 altogether 8400 users were registered in Transkribus and 30 cooperation agreements (MoUs) with institutions all over the world were concluded.

Another highlight in Y2 was the arrangement of the first Transkribus User Conference (TUC) in Vienna in November 2017. More than 90 participants from 18 countries took part in this two-days conference where the READ team introduced new developments. Several "power" users (scholars, libraries, archives) reported about their usage of the Transkribus tools. The conference is also available on the YouTube channel of Transkribus.[1]

As in Y1 also in 2017 several dozens of workshops, presentations and webinars were given by the READ team to strengthen involvement of our target groups. A highlight in this respect was a Transkribus workshop at the Digital Humanities 2017 conference in Montreal.

Public deliverables for Y2 are online on the READ website. Software developed in the READ project is available via GitHub, Research Data for scientific competitions can be accessed via ZENODO.

Following some of the main recommendations of the first review meeting in 03/2017 we were working on a detailed sustainability plan. This plan is now set out in D3.2. Business plan implementation. Currently we envisage the foundation of a cooperative (working title: READ-COOP) as the best model to express that READ and Transkribus are built strongly on the idea of a fundamental cooperation among archives, libraries, scholars, universities and the public.

The Transkribus platform together with the ScanTent/DocScan will be the main products of the envisaged cooperative.

1. Overview of the progress for Y2

1.1. Achievements and highlights

If we briefly want to characterize the first two years in the project we can say, that Y1 set the scene, but Y2 brought several breakthroughs in research, technology and innovation. Main highlights are:

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

(1) Significantly improved layout analysis

One of the most urgent research challenges in READ was the development of a layout analysis tool which is capable to cope with handwritten documents and their non-standardized and complex layout. With the introduction of machine learning approaches it can today be regarded as resolved. One of the key factors for this breakthrough was the availability of a large dataset and the organisation of a scientific competition at the ICDAR 2017 conference by READ members.

(2) Handwritten Text Recognition as a standard technology

The breakthrough in HTR was achieved on two levels: Firstly in late 2016 we implemented the HTR engine from the CITlab team, but in 2017 it become really productive. The Transkribus client was extended with a GUI (Graphical User Interface) which enables non-technical users to select documents for training and testing and to start the learning process for the neural networks themselves. Since the beginning of 2017 more than 300 models were trained covering not only many languages in Europe, but also Arabic and Hebrew. Secondly the UPVLC team implemented a new HTR engine based on neural networks which shows a dramatic improvement, from e.g. 26% Character Error Rate for HMM based approaches to 9% CER with the new system.[2]

(3) Keyword Spotting and (probabilistic) indexing

In the second half of 2017 also one of the several Keyword Spotting (KWS) engines developed in READ was integrated into the Transkribus platform. For the first time users are now able to perform KWS immediately after having applied their HTR model on their documents. In contrast to full-text search where noisy text data are a fundamental problem KWS brings excellent results even if the recognition results are not usable or readable for human beings or machines. Moreover significant progress was made by several groups (UPVLC, NCSR, DUTH) on building indexes for large document collections to enable fast image retrieval. To further exploit the full potential of KWS more experiments are foreseen for 2018.

(4) ScanTent prototype and DocScan released

In addition to achievements in research also in the fields of innovation we can report a major progress. In 2017 it became very clear that users are highly interested in the ScanTent device and its connected DocScan app. Therefore a series of 15 prototypes for the ScanTent was produced and distributed among several test users. Moreover the connected Android app was further developed. The user is now able to take images automatically, as well as to upload them directly to the Transkribus server. Reactions of users are very positive or even enthusiastic so that in 2018 we hope to produce a first set of 1000 devices.

[2] Cf. Deliverables D7.1. and D7.2: D7.2 Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal: HTR Engine Based on HMMs P2. UPVLC. 2017. Online: https://read.transkribus.eu/wp-content/uploads/2017/12/D7.2.pdf

With these breakthroughs we are now able to cover the complete workflow as it is necessary to carry out digitisation projects in the Digital Humanities as well as in the archives and libraries domain. Users are now enabled to prepare specific text recognition models for their documents, to run them on large document collections and also to search them via keyword spotting.

In detail:

Task in the workflow	Implementation
Digitising documents with the smartphone in a fast and reliable manner	Fully implemented with DocScan. The ScanTent is still under development, respectively a first series of 1000 devices needs to be produced.
Uploading documents directly to the Transkribus server either from existing resources (repositories) or from the DocScan app.	Fully implemented service, both GUI and API.
Sharing digitised documents with archives and libraries	QR Code detection in order to connect scans with archival records is implemented in DocScan. Full implementation of the service is foreseen for 2018.
Detecting lines and regions in complex documents	Fully implemented. Improvements such as the ability to train a Layout Model for specific documents is foreseen to be added in 2018 to Transkribus.
Generating training data either manually or automatically	Fully implemented. Especially the automated matching of text with images is a major step forward (t2i).
Training of specialised and/or global models for text recognition	Fully implemented. Currently only selected power users have access to the training interface in Transkribus. The interface will be opened up to every Transkribus user in 2018 leveraging GPU servers.
Involving users via the web-interface for transcription and correction	A first beta version is implemented. More interfaces as well as a mechanism to expose collections as crowd-sourcing collection will be added in 2018.
Processing large document collections with trained models.	Fully implemented. Full scalability (i.e. to process millions of images) will be a focus in 2018.
Measuring the results of the recognition process and KWS	Fully implemented. For KWS a Precision/Recall curve is foreseen and will be implemented in 2018.
Using Keyword Spotting to search documents	Fully implemented. KWS using indexes is foreseen for 2018.

Exporting results of the recognition process and the KWS	Partly implemented, KWS export will come in 2018.
Using the API for all services from above	Fully implemented, ongoing process.

Table 1 Overview of the complete workflow covered by the Transkribus platform

It has to be emphasized that the Transkribus platform is the only research infrastructure worldwide offering such a comprehensive digitisation cycle including all relevant steps in enriching historical documents.

The impact of these achievements for our user groups is significant:

- **Humanities scholars** are enabled to digitise thousands of pages quickly and independently of any digitisation efforts from archives and libraries or, if digitised documents are already available they can upload them to the platform. They can train their own recognition models and directly apply them to their documents which become immediately accessible via Keyword Spotting. They can invite colleagues or involve students and share their work with archives and repositories. They can measure results, deal with the uncertainty of automated processes and carry out deep investigations in large collections of documents independently of any infrastructure provided by a specific archive or library. In this way Transkribus becomes a real research infrastructure and a personal research site for humanities researchers.

- **Archives and libraries** are enabled to offer keyword spotting to their users also for handwritten historical collections with a minimum of effort. Based on general models (e.g. English writing M1) large scale recognition can be carried out in a fully automated mode. In this way their collections become much better accessible to users of any kind, may they be researchers or family historians. Especially the text2image matching tool will play a role in this respect: Based on existing transcriptions (in whatever format they are available) large text recognition models can be trained quickly. Archives may also benefit from digitisation efforts carried out by users with the ScanTent and DocScan app in their facilities, since Transkribus will enable them to store the images taken by users also in their own repositories.

- For **public users (family historians, volunteers)** we expect that the DocScan app will be one of the most important means to satisfy their needs. These users will be able to take images from their personal documents (owned by them or by an archive) and to request a transcription directly via the DocScan app from the Transkribus team. This is an innovative service which was not foreseen in the GA.

  But also in their role as volunteers and crowd-users Transkribus will offer new and more effective ways to involve the public in digital projects. Instead of keying or correcting automated text the recently introduced keyword spotting service will offer them the chance to validate and correct "important" words of a document, e.g. proper names or specific keywords. Last but not least they will be able to contribute in the form of "Scanathons" – using the ScanTent and their mobile phone to digitise significant amounts of archival documents within a group of volunteers.

- For **researchers** in the domain of **Pattern Recognition** and Computer Vision a number of competitions were organized in 2017. The most important one was the cBAD competition as well as the HTR competition (both at the ICDAR 2017 conference) In this way READ/Transkribus becomes better known to the research community and this will also be continued in 2018.

Apart from these highlights of Y2 we also prepared the ground for one of the most challenging tasks, namely the scalability of the Transkribus platform. First tests were performed with all Large Scale Demonstrators foreseen in the project: the National Archive Finland, the Diozesan Archive Passau, the State Archive Zurich and the Venice Time Machine. In Y3 we assume that we will be able to offer services for processing not only thousands or tens-of-thousands of pages, but to run Layout Analysis and Text Recognition on hundreds-of-thousands or even millions of page images. Together with the launch of the governance and business models this will be another milestone in establishing Transkribus as "the" research infrastructure for digitisation, transcription, recognition and searching of historical documents.

1.2. Objectives[3]

The overall objective of READ is to implement a Virtual Research Environment where archivists, humanities scholars, computer scientists and volunteers are collaborating with the ultimate goal of boosting research, innovation, development and usage of cutting edge technology for the automated recognition, transcription, indexing and enrichment of handwritten archival documents.

Until the end of 2015 just a few experts had noticed that some ground breaking work with respect to Handwritten Text Recognition (HTR) was going on in the Pattern Recognition, Machine Learning and Computer Vision domain. The tranScriptorium project (FP7 2013-2015) provided a first prototype and at workshops in London, The Hague and Vienna archivists and humanities scholars became aware of these developments (http://transcriptorium.eu/).

Now, at the beginning of 2018 the picture has changed significantly: thousands of archivists, humanities scholars and public users have not only heard about the READ project and its service platform "Transkribus" but also demonstrated their actual interest by registering at the Transkribus platform, downloading the expert tool, trying it out for their own purposes and contacting project members for further support. A number of research projects already uses Transkribus for their daily work or has included it into their grant applications.

READ had a significant impact how professionals are dealing with historical documents and how they are planning and shaping future projects for digitisation, digital scholarship and research. Many of these users are recognizing READ and Transkribus as a "game changer" in their professional domain. The claim which we use on the web site of the READ project is therefore ambitious but more valid than ever: "**READ revolutionizes access to handwritten documents**".

The main objective of the READ project as it was set out in the GA was therefore fully achieved and Y2 of the project contributed with significant progress. The success of READ and Transkribus is based on five main concepts which we defined as the basis for our work plan.

[3] Citations from the Grant Agreement are set in italic.

In the following we re-examine these five concepts in light of the progress we made in Y1 and Y2 of the project.

Key Concept 1: Culture of cooperation

Establish a "culture of collaboration" among archivists, computer scientists, humanities scholars, and volunteers as the main prerequisite to enabling scientific progress and innovation in the field of Handwritten Text Recognition and its related fields.

In general presentations given by the READ project members, one slide is always shown to the audience which illustrates this "culture of collaboration" as one of the core aspects of the project. The service platform Transkribus is the focal point for our user groups and it is built to satisfy their needs by making the newest technology and growing amounts of data available.

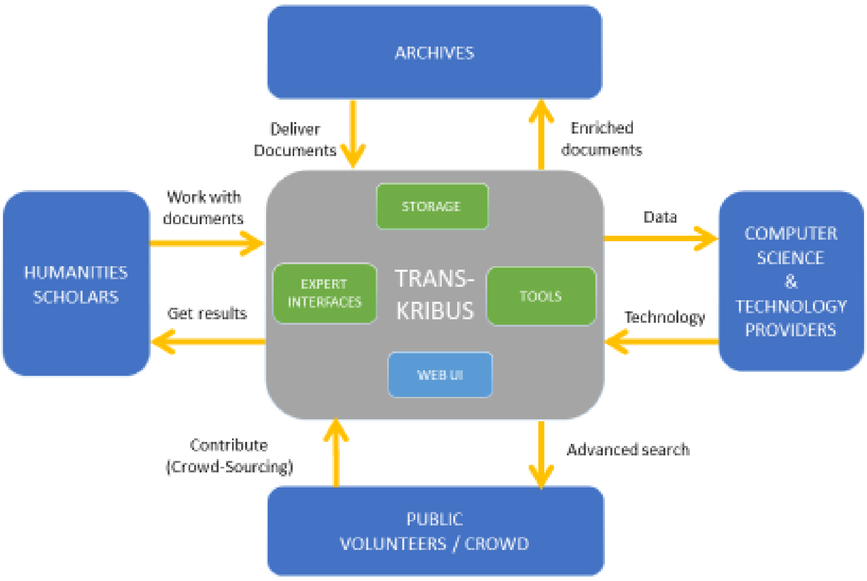


Figure 1 User groups of the READ platform /Transkribus

It may sound a common place but none of the four user groups is able to "revolutionize access to historical documents" on its own. Of course technology will come from computer scientists and technology providers and it plays the leading role in this revolution, but all machine learning approaches are still highly dependent on the availability of large quantities of training data. Training data on the other hand can only be generated if archives and collection holders are not only willing to share their collections, but also to digitize large amounts of their holdings – where the trained models can be applied and real benefit can be gained.

Humanities scholars on the other hand are the experts when it comes to generating training data. Their capability to read and correctly transcribe historical scripts is one of the main requirements to be able to "feed" neural networks. Without the involvement of these experts from the humanities the lack of training material will be a huge obstacle for any progress in applying the technology to large amounts of documents. This is even more true since – in contrast to modern documents – the market for historical documents is small and has very specific requirements. And finally the public will also be able to contribute documents and transcriptions and benefit from the availability of other documents, technology and free services, such as the searching of large amounts of data.

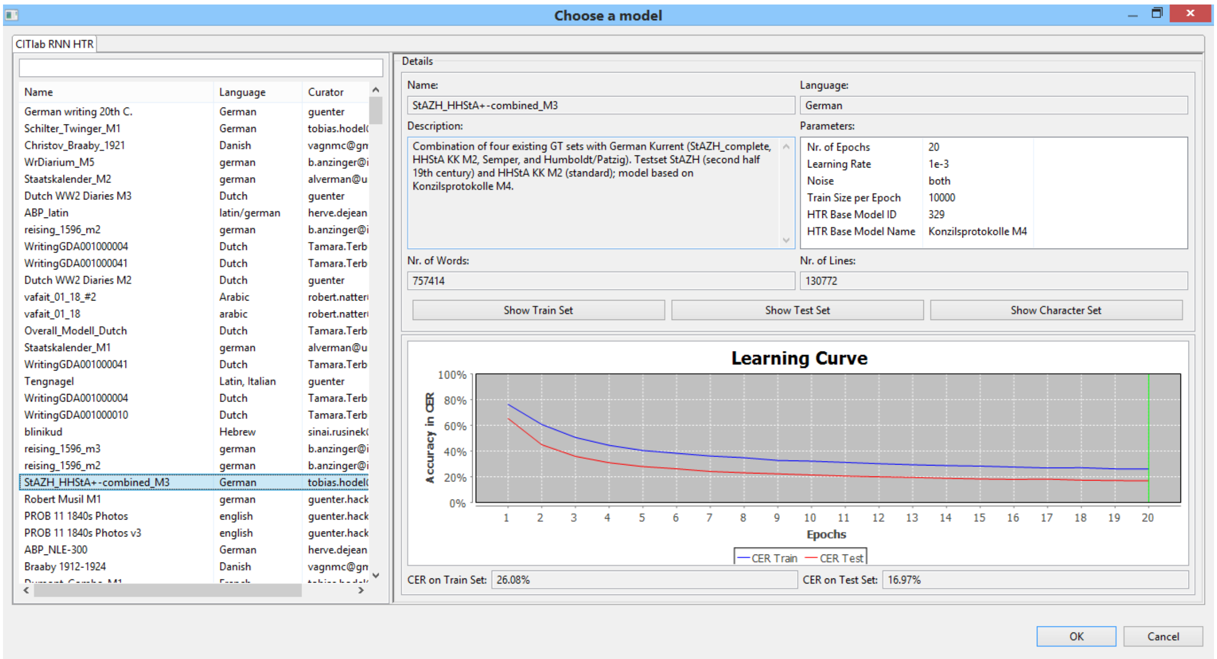
The other main message from this “culture of cooperation” is that the platform is able to generate network effects and synergies which cannot be achieved by one single group on its own. The most striking example is that training data for HTR models can easily be reused by other users without infringing any copyright or personal rights.

The governance and business model which we drafted in Y2 tries to transform this “culture of collaboration” into a viable model for sustaining the Transkribus platform. It suggests to set up a cooperative as the legal entity and a freemium model for all member organisations. The main purpose of the cooperative will be to enable shareholders (owners) to fulfil their missions with the best quality and the lowest costs. The common denominator which unifies all groups is to “revolutionize " access to historical documents. As with all cooperatives owners and customers are identical which means that the cooperative itself is a for-profit organisation, but not for the sake of pure profit, but for the sake of realizing the common goal of its shareholders. Therefore business will only take place between the cooperative and its member organisations. We believe that in this way the “culture of cooperation” is best realized and will create the chance to work together without neglecting the positive aspects coming from a free market, such as competition and dynamic development.

Archives, Memory Organisations and Collection Holders

In the first year 25 institutions signed a Memorandum of Understanding with the READ project, now in 2017 this number even increased: 30 institutions signed a MoU and are now connected with READ. They are located in several different countries (Australia, Austria, Canada, Finland, France, Germany, Greece, Italy, Luxembourg, Netherlands, Norway, Serbia, Spain, Switzerland, United Kingdom, United States). Two universities from the United States of America also recently signed a MoU. All of these institutions are testing the READ tools and platform in order to explore the options for their own institutions.

In Y2 we were not only able to produce more than 300 HTR models for all interested institutions but also underpin these test projects with scientifically proofed data on their performance. The learning curve, important parameters, and especially the Character Error Rate achieved against a test set are highly valuable data for every project manager or decision maker in the involved institutions. A good portion of the training data provided by MoU partners will be publicly available so that other users will also be able to evaluate the performance of the HTR engines against different datasets.



Name	Language	Curator
German writing 20th C.	German	guenter
Schilter_Twinger_M1	German	tobias.hodel(
Christov_Braaby_1921	Danish	vagnmc@gn
WrDiarium_M5	german	b.anzinger@i
Staatskalender_M2	german	alverman@u
Dutch WW2 Diaries M3	Dutch	guenter
ABP_latin	latin/german	herve.dejean
reising_1596_m2	german	b.anzinger@i
WritingGDA001000004	Dutch	Tamara.Terb
WritingGDA001000041	Dutch	Tamara.Terb
Dutch WW2 Diaries M2	Dutch	guenter
vafait_01_18_#2	Arabic	robert.natter
vafait_01_18	arabic	robert.natter
Overall_Modell_Dutch	Dutch	Tamara.Terb
Staatskalender_M1	german	alverman@u
WritingGDA001000041	Dutch	Tamara.Terb
Tengnagel	Latin, Italian	guenter
WritingGDA001000004	Dutch	Tamara.Terb
WritingGDA001000010	Dutch	Tamara.Terb
blinikud	Hebrew	sinai.rusinek(
reising_1596_m3	german	b.anzinger@i
reising_1596_m2	german	b.anzinger@i
StAZH_HHStA+-combined_M3	German	tobias.hodel(
Robert Musil M1	german	guenter.hack
PROB 11 1840s Photos	english	guenter.hack
PROB 11 1840s Photos v3	english	guenter.hack
ABP_NLE-300	German	herve.dejean
Braaby 1912-1924	Danish	vagnmc@gn

Name: StAZH_HHStA+-combined_M3

Language: German

Description: Combination of four existing GT sets with German Kurrent (StAZH_complete, HHStA KK M2, Semper, and Humboldt/Patzig). Testset StAZH (second half 19th century) and HHStA KK M2 (standard); model based on Konzilsprotokolle M4.

Parameters:

Nr. of Epochs	20
Learning Rate	1e-3
Noise	both
Train Size per Epoch	10000
HTR Base Model ID	329
HTR Base Model Name	Konzilsprotokolle M4

Nr. of Words: 757414

Nr. of Lines: 130772

CER on Train Set: 26.08%

CER on Test Set: 16.97%

Figure 2 Transkribus GUI for trained HTR models

In Y3 we will focus our work in two ways: Firstly a complete relaunch of the HTR engine is planned and already prepared. The CITLab team was working in 2017 on a Tensorflow implementation which not only brings a significant improvement of the recognition rates (up to 30 or even 50% reduction of the Character Error Rate) but also enables us to run training and recognition on GPU servers.[4] This will reduce computing time by a factor of 5-10. Typically a specialized model will then be trained in 2-3 hours, instead of a full day. Based on these improvements we will be able to open up our training interface to all registered users in Transkribus. Secondly we will start to train general models capable to recognise very different writing as it is the case in large document collections hosted by archives and libraries. This will be one of the main requirements to offer Keyword Spotting "out-of-the-box".

Computer Scientists

The involvement of computer scientists is mainly undertaken via the launch of competitions as well as the availability of datasets. In Y1 the main prerequisites were prepared such as the development of the ScriptNet website which supports teams to organize their own competition, as well as the collection of datasets.

Y2 actually fulfilled our expectations. With the launch of the largest research dataset for line detection the significant improvement of layout analysis and line detection tools became reality. The dataset, known as cBAD, is available via ZENODO and formed the basis for the cBAD competition at The International Conference on Document Analysis and Recognition (ICDAR) 2017 in Kyoto.

[4] Cf. Deliverable D7.8 Max Weidemann, Johannes Michael, Tobias Grüning, Roger Labahn: HTR Engine Based on NNs P2. Building deep architectures with TensorFlow (2017) Online: https://read.transkribus.eu/wp-content/uploads/2017/12/Del_D7_8.pdf. We will also investigate the options to integrate the CNN implementation of UPVLC in Transkribus taking into account that running two systems in parallel may be a challenge in terms of maintenance.

The University of Fribourg / DIVA (Document Image and Voice Analysis) group who are one of the leading groups for digital palaeography and historical documents organised a workshop on Open Source tools at the ICDAR 2017. The Transkribus team introduced the platform to the community and got highly valuable input from several colleagues, among them members of the Google research team from Tesseract/Ocropus.

The READ team was also approached by several other research groups to support them or to get involved in grant applications. E.g. Transkribus is part in the application sent to the Canadian Science Funds from the University of Toronto in the area of medieval Latin manuscripts collected by the DEEDS (Documents of Early England Data Set) project group.

Already successful was Dirk van Hulle from the University of Antwerp who got a grant for his CATCH 2020 (Computer-Assisted Transcription of Complex Handwriting) project. The group from Antwerp will use approaches and tools from the computational linguistics field in order to improve text recognition and correction in Transkribus. The software will be available under an Open Source license and will directly be integrated in Transkribus.

Humanities Scholars

Lots of useful feedback also comes from the humanities domain. And indeed, scholars who are actually transcribing handwritten documents are among the most interested users.

One example among many: Already in Y1 of the project we set up a collaboration with Thomas Wallnig from the University of Vienna. A first HTR model was trained for a Latin writer of the 17th century using the HMM solution. It brought a CER of about 16%. In Y2 we retrained the same dataset with the new HTR engine and improved the recognition rate to 8%. In parallel our collaborator applied at the Austrian Science Funds for a digital edition project where the Transkribus platform was included as the main tool to create a scholarly transcription. The application was successful and in 2018 now a three-years project started where Transkribus will be used as research infrastructure for this editing project.

Similar efforts are going on in several countries, e.g. Transkribus is part of an application drafted by scholars from the United States, by Canadian researchers and also by Swedish and German scholars.

Success stories which we reported from Y1, such as with the University of Greifswald, are continued and transformed into larger projects. Apart from the transcription of thousands of pages with the support of automated processing done in Y2 this collaboration will be extended in several ways. The University of Greifswald and the Transkribus team will apply for a digitisation project where HTR and KWS will be key technologies. The whole project shall be managed by the Transkribus platform.

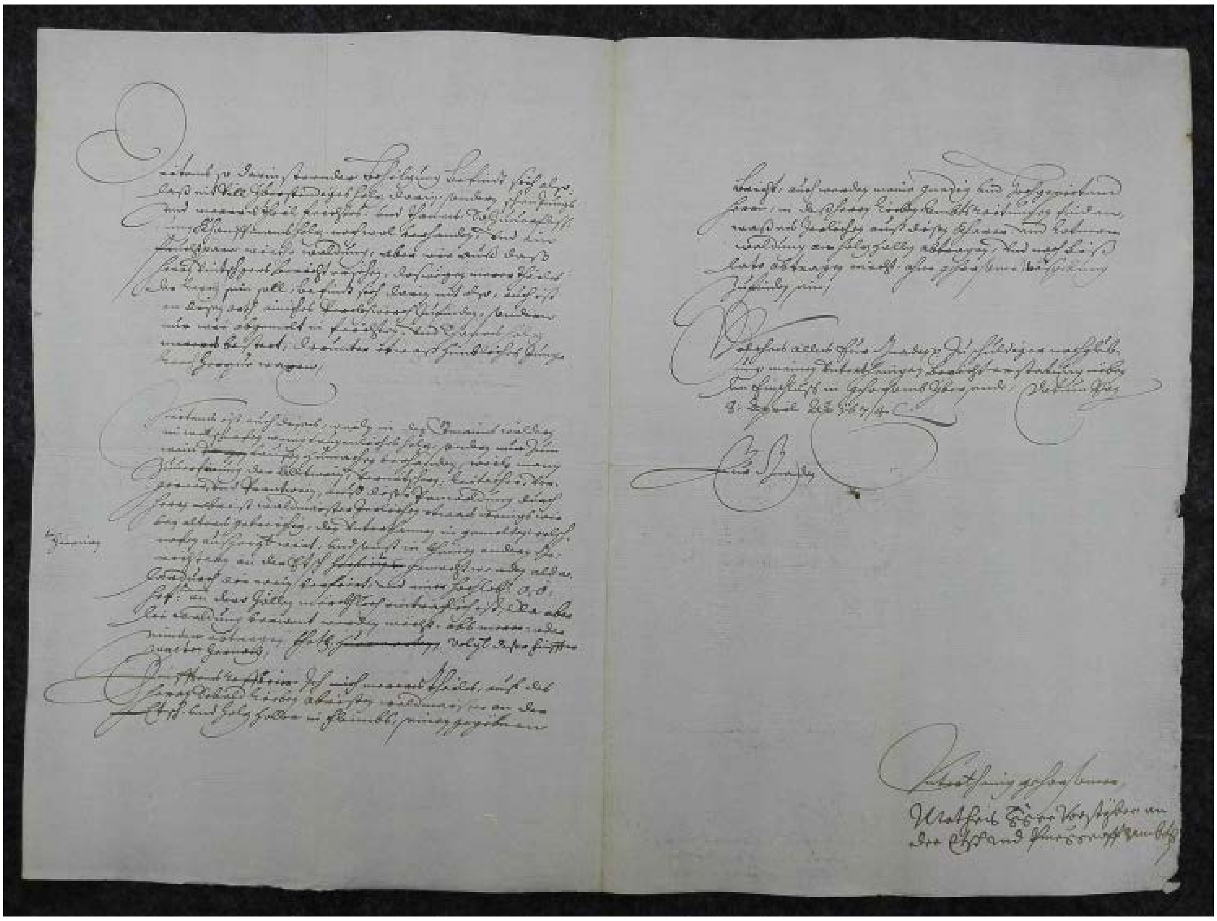


Figure 3 Images taken with the ScanTent and DocScan in the Tyrolian State Archive

Greifswald is one of the pilot users for the ScanTent and DocScan application as well where the images taken by users are stored and forwarded to the digital library (Goobi installation). The digital repository was already prepared to output QR codes for their records so that users are able to scan the QR code directly from the screen with DocScan and have not only the metadata available for their digital images, but their images can later on be directly sent to the archive as well.

There are similar success stories which can be mentioned here just briefly such as the collaboration with the Haifa University and the Digital Humanities group in Israel, the collaboration with the National Archives London, the Archivio Ricordi (owned by Bertelsmann media house) and many, many others.

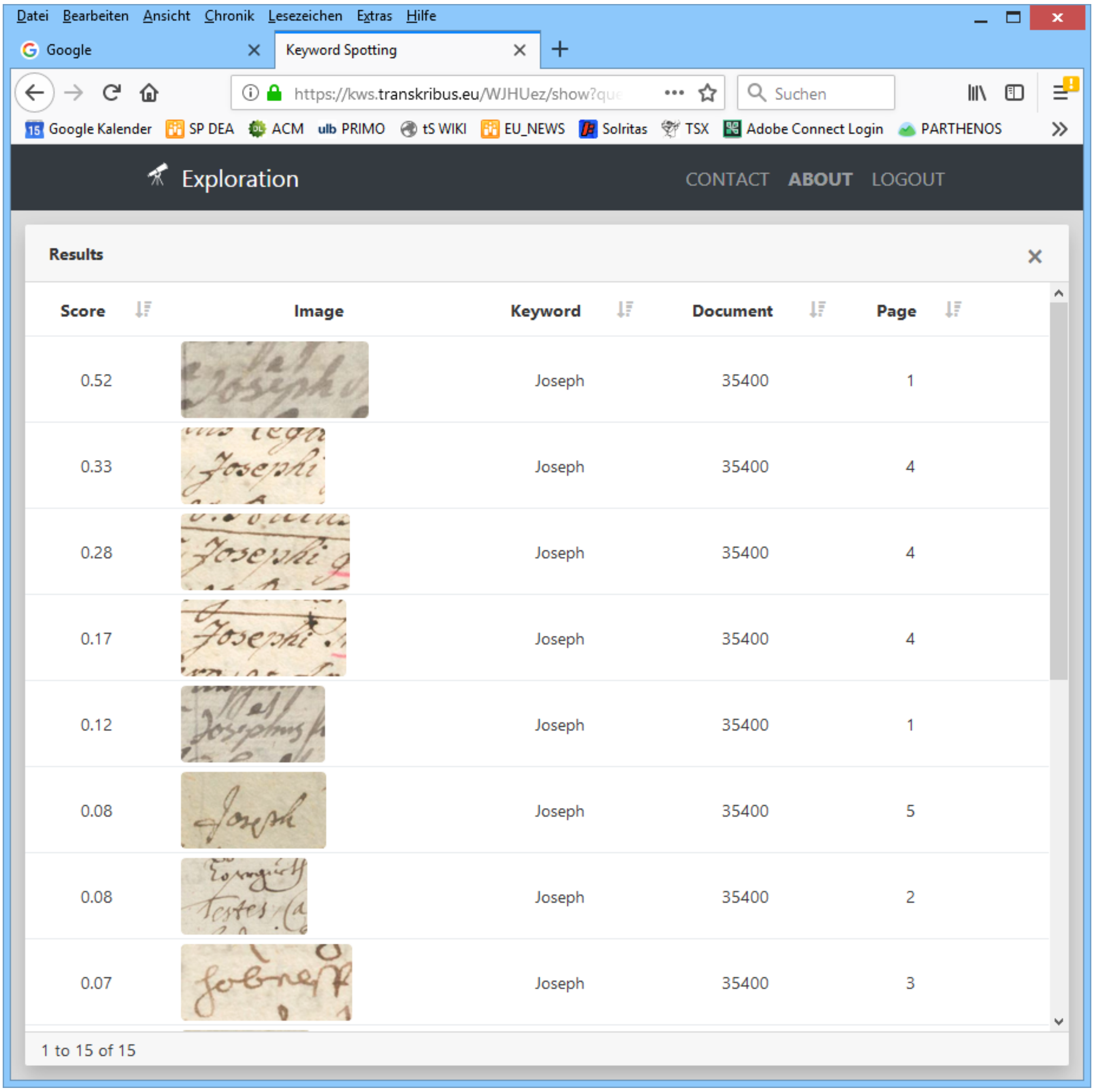
Crowd-users and Volunteers

In Y1 of the project we were approached by many users who could be summarized under the term “family historians”. In Y2 we continued a volunteering project at the City Archive Bozen/Bolzano in Italy.[5] Here about 60 people are working together on the transcription of council meetings from the late 18th century. One of the main findings is that careful and reliable transcription of historical texts is a hard task and requires a lot of engagement but also continued support. The expert interface is definitely a challenge for volunteers who are often people with less affinity to computers. But this is not necessarily the main reason why

[5] Our distinction between crowd- and volunteering projects is that volunteers are expected to work with the Transkribus expert interface, whereas the crowd will be working with simplified tools.

about 80-90% dropout quote can be observed. And finally also this volunteering project made the experience that all transcriptions coming from volunteers need to be checked carefully since a number of errors will be included. This confirms the experiences of the Transcribe Bentham crowdsourcing project at UCL.

In Y3 we will explore new ways for crowd-sourcing projects. The ScanTent and the DocScan app are much easier to learn and much more motivating for crowd-users than the transcription task itself. In March 2018 we will therefore organize a first Scanathon at the World Archives Day in order to spread the idea that users can contribute directly to the digitisation of archival material.



Score	Image	Keyword	Document	Page
0.52		Joseph	35400	1
0.33		Joseph	35400	4
0.28		Joseph	35400	4
0.17		Joseph	35400	4
0.12		Joseph	35400	1
0.08		Joseph	35400	5
0.08		Joseph	35400	2
0.07		Joseph	35400	3

1 to 15 of 15

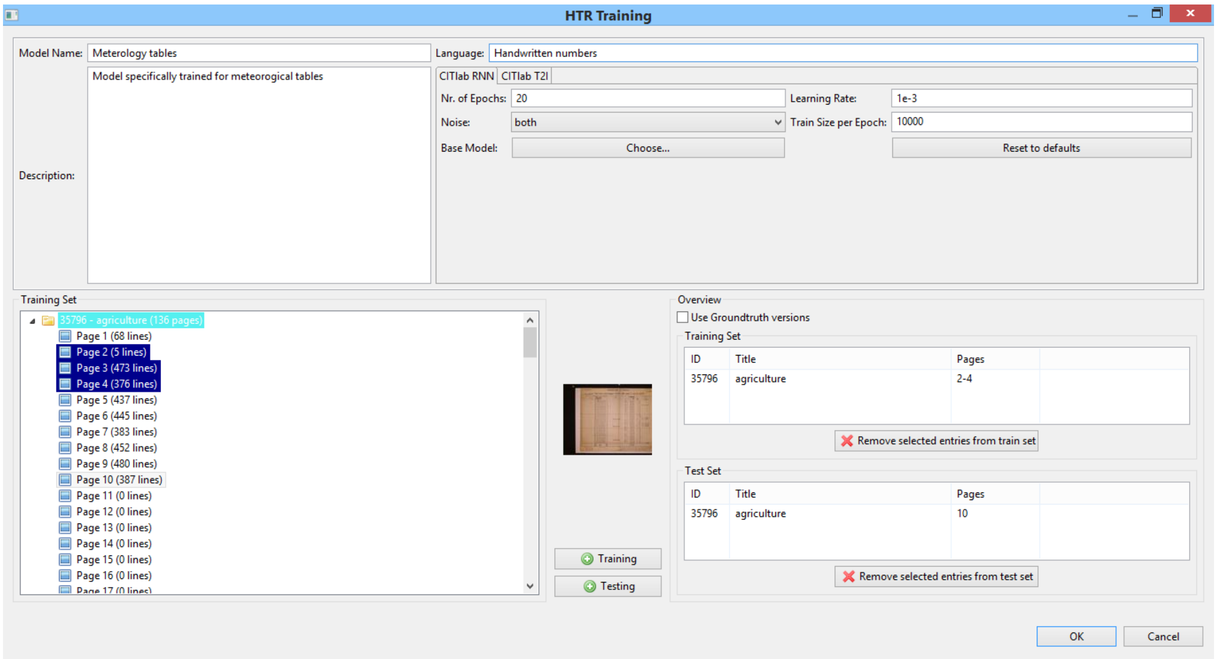
Figure 4 Result hits for KWS. Search was "Joseph"

Another direction will be to involve volunteers and the crowd in validating the results of keyword spotting (cf. figure above). We imagine that project managers can send several query words against a collection and the crowd will be able to view and validate the accurateness of the hit list. KWS is for such a task very well suited since all result hits come with a confidence value. The lower the confidence the higher the amount of “false alarms” which can be observed among the hit list. To reduce these false alarms seems to be a very valuable task for crowd-users.

Key concept 2: Breakthrough in HTR

Enable a “breakthrough” in Handwritten Text Recognition, so that HTR will be transformed from a cutting-edge research field to a mature and well-understood technique which can be exploited and used in real world applications.

In Y1 we were able to launch Transkribus as the world’s first publicly available implementation of a state-of-the-art Handwritten Text Recognition engine based on neural networks. In Y2 the main progress was that Transkribus is now equipped with a Graphical User Interface enabling users to carry out all steps of the recognition process by themselves. Users can create training data, start the training process of the neural networks, receive an HTR model and apply it to their documents. Processing time for applying the HTR model has been reduced to below one minute per page (it was about 20-30 minutes for the first HMM implementation), therefore it is in the range of a commercial OCR engine. We are especially proud that we were able to reach a TRP Level of 8-9 in the field of training and applying text recognition. Downtime and failed trainings are extremely seldom and are in the range of industrial applications. This work will be continued in Y3 when we will implement a new version which significantly reduces processing time, both for training as well as recognition.



HTR Training

Model Name: Meterology tables

Language: Handwritten numbers

Description: Model specifically trained for meteorogical tables

Nr. of Epochs: 20

Learning Rate: 1e-3

Noise: both

Train Size per Epoch: 10000

Training Set

- 35796 - agriculture (136 pages)
  - Page 1 (68 lines)
  - Page 2 (5 lines)
  - Page 3 (473 lines)
  - Page 4 (376 lines)
  - Page 5 (437 lines)
  - Page 6 (445 lines)
  - Page 7 (383 lines)
  - Page 8 (452 lines)
  - Page 9 (480 lines)
  - Page 10 (387 lines)
  - Page 11 (0 lines)
  - Page 12 (0 lines)
  - Page 13 (0 lines)
  - Page 14 (0 lines)
  - Page 15 (0 lines)
  - Page 16 (0 lines)
  - Page 17 (0 lines)

Overview

□ Use Groundtruth versions

Training Set

ID	Title	Pages
35796	agriculture	2-4

Test Set

ID	Title	Pages
35796	agriculture	10

Figure 5 Training interface for Text Recognition models

But the real breakthrough in this domain comes from the new Layout Analysis algorithms developed in the project. These algorithms are now based on machine learning and adapt in a much better way to the complex layout of handwritten documents than any other “hand-crafted” algorithm. The difference is dramatic and can be experienced by every user immediately. In the figure below red lines indicate automatically recognized baselines. Also the regions are detected automatically.

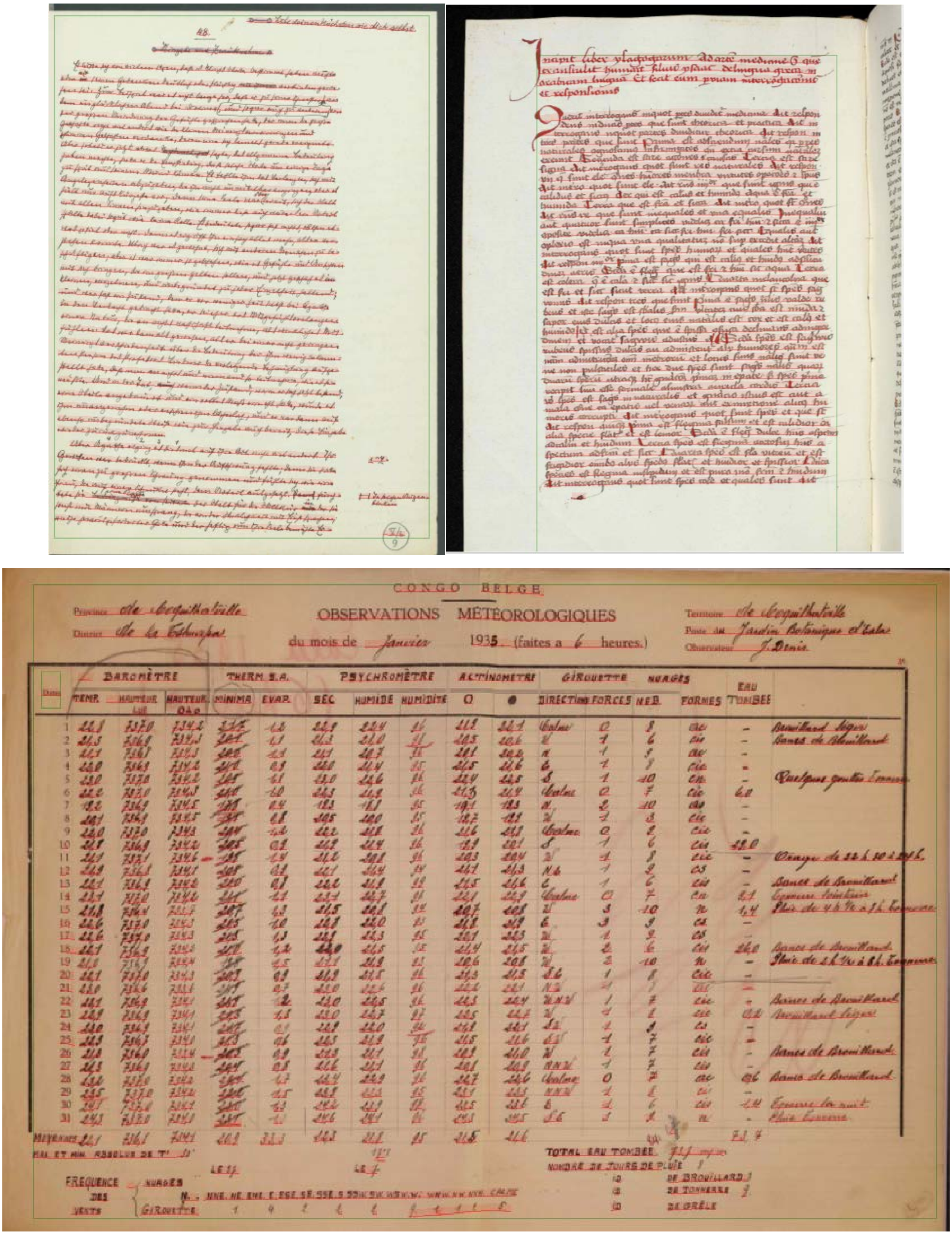


Figure 6 Example pages for complex layout recognized with the new layout analysis tool

Today we can say that nearly all documents – even with complex layout – can be processed automatically without any tuning of the layout analysis tools. This makes Keyword Spotting a much more attractive option, since users can be very sure that (nearly) no text in a document will be missed by the text recognition process. It will also be the most important requirement to carry out Large Scale Demonstrator projects with the archives and memory organisations of the READ project.

The third breakthrough which needs to be reported here is the implementation of the CITlab KWS engine in Transkribus. After recognising a document KWS can be performed on a document by every user on his own documents. Since KWS will be "the" most important feature for archives and libraries we expect high interest in this technology in 2018. Here a simple example how it works:

The Australian National Library which is one of the MoU partners in the project has published a letter of James Cook on their Trove digital library.[6] If we download the image, integrate it into Transkribus, run the Layout Analysis and use a general model for English handwriting (based on writings of Jeremy Bentham and secretaries) we receive results which may not look very convincing at first glance.

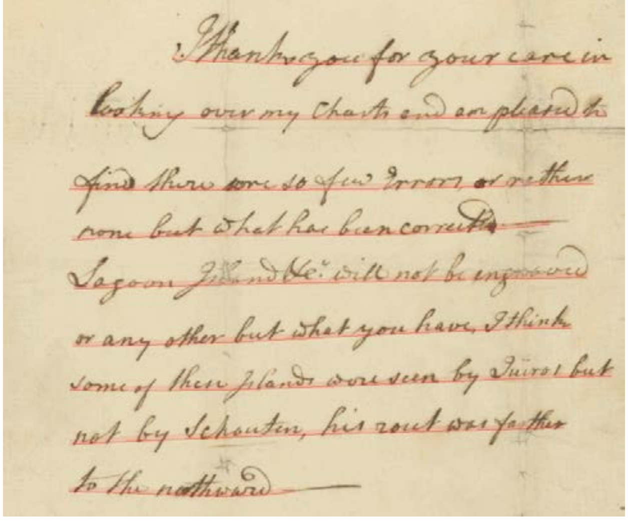
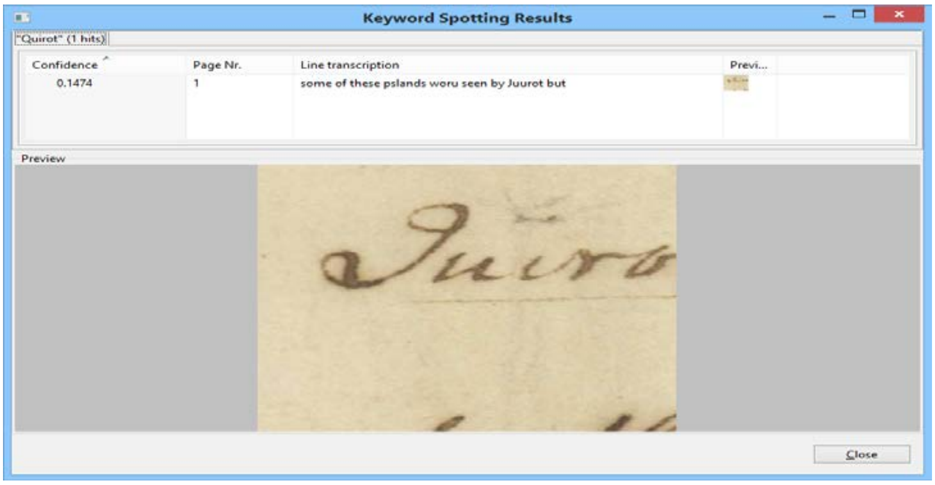
 I thank you for your care in looking over my Charts and am pleased to find there are so few Errors or rather none but what has been corrected Sagoon Island &c. will not be engraved or any other but what you have, I think some of these Islands were seen by Quiros but not by Schouten, his rout was farther to the northward	Ithankrgou for gourcare in Ibokiny over my charts end ar polased the find thie ore so fw Errors, or rether none but what har bien corrects Sagoon Idand bte: witt not be engnsbd or any other but what you have, I thint some of these pslands woru seen by Juurot but not by Scouten, his rout wes father to the naothoare

Figure 7 Automated transcription of the writing of James Cook without any training

We find here a Character Error Rate of 19,3 and a Word Error Rate of 52,5 %. Indeed to use this transcription as a basis for correction is not recommended since it will require more time to correct than to key it from scratch. But the result must be completely differently rated if we look at the KWS facility.

[6] MS 4307 - Letter by James Cook, 1772 July 11 [manuscript] : Plymouth Sound, to George Perry, Victualling Office, London Item nla.obj-547501553: Letter by James Cook, 1772 July 11. (http://nla.gov.au/nla.obj-547501553/view) one item, 5 images - comprising a letter (4 images but only two pages with text) and transcript (the 5th image in the sequence)



Keyword Spotting Results

"Quirot" (1 hits)

Confidence	Page Nr.	Line transcription	Previ...
0.1474	1	some of these pslands woru seen by Juurot but	

Figure 8 KWs in Transkribus: search for "Quirot"

With the error rate from above all interesting words can be found with rather high confidence. E.g. the search for Quirot retrieves actually the image with this word though the automated transcription is spelled as: *Juurot.* It is obvious that this is exactly what archives, libraries and their users are looking for – to have the chance to search in handwritten or printed historical documents. For a large scale project this would mean that e.g. a representative set of images is selected by the library or archive, several tests are done with the Transkribus GUI to find out the quality of the available HTR models, and – if the quality is sufficient to e.g. retrieve 95% of all words of a collection – it is then just a matter of transferring images into the Transkribus platform and processing them. Once this is done the documents are immediately available for searching.

HTR – a "European" research field

All HTR competitions of the past few years were organised and won by research teams from Europe. E.g. the International Conference on Frontiers in Handwriting Recognition (ICFHR) 2014, the International Conference on Document Analysis and Recognition 2015 (both by the CITLab team from Rostock) and the ICFHR 2016 (by the RWTH Aachen team). In 2017 this changed, now a team from the US (Young Bingham University) won the ICDAR 2017 HTR competition organised by UPVLC.

This is an indication that nowadays more groups became aware of this research fields. For ICDAR 2017 the HTR competition was organized by the UPVLC team based on the largest dataset ever used in this domain coming from the Alfred Escher Foundation (Switzerland). The challenges in this competition were not only to create a good HTR model from a given training data, but also to find a way to match existing transcriptions with the corresponding images. The competition was organized via the ScriptNet platform which was set up by NCSR. It was won by the Young Bingham University from California with excellent figures in text recognition.

Comprehensive approach

A Virtual Research Environment dealing with historical documents and their recognition, indexing, searching and enrichment needs to cover many aspects and be able to fulfil very different requirements. From our point of view this approach turned out to be extremely feasible and well-received by users.

As already outlined above Y2 brought an important milestone with respect to this "comprehensive" approach. Until Y2, we were dependent on archives and libraries and their digitisation efforts, but this has now changed fundamentally. With the ScanTent and the DocScan app three target groups will be able to directly benefit from this "Crowd-scanning" approach:

- Humanities scholars can digitise significant amounts of historical documents on their own. First tests have shown that about 200-300 images (or 400-600 pages) can be taken per hour with a ScanTent. If we assume that remains of famous persons often do not exceed 20-30.000 pages it is obvious that some students can do the digitisation in a few days or weeks. Since the ScanTent will come for some hundreds of EURs it is easily scalable – several students may work in parallel. The documents can be directly uploaded to the Transkribus platform, segmentation and HTR can be performed which means that also KWS can be applied immediately after the digitisation process. This means that scholars can search in significant amounts of documents in principle some hours after they have scanned the documents in the archive.
- Archives and libraries on the other hand can directly incorporate the scanned images via the Transkribus platform. They just need to use the QR facility of Transkribus so that every document which is scanned by a user contains the identifier for the Finding Aid. The scanned document is then not only available for the user himself, but also for the archive and of course for all other users.
- Public users and volunteers can contribute to the digitisation of archives in a simple way by scanning documents with their smart phone and making them available to the archive.

All of these users not only get in contact with Transkribus via the ScanTent and DocScan but will also be able to use all other features of the platform and in this way get familiar with the services offered by Transkribus. This service is to our best knowledge unique and from our point of view a major innovation in the domain.

Data driven research

It was one of the remarks of the final review meeting for the tranScriptorium project (held in 03/2016) that the project always "returns to the Bentham dataset". This remark is no longer valid for READ. The datasets processed in READ are not only heterogeneous, they are also coming directly from users who are interested to see how the READ tools and here namely Layout Analysis and HTR are performing on their datasets. As already mentioned above, more than 300 HTR models were trained covering all types of languages and scripts.

The success of our data driven research is also very well demonstrated with the breakthrough in the Layout Analysis domain. The cBAD dataset contains nearly 2000 images with around 80.000 lines from several different archives, with different languages and different challenges. It was this data set which made the success possible and which proofs now to be representative for many, many documents since the selection process was really based on a large amount of data. E.g. we received a sample of one millions of pages from the National Archive of Finland, representing their complete collection.

What has to be mentioned here is that it was extremely helpful to have a dedicated sub-contracting budget available for supporting ground truth production and test projects.

Key concept 3: Service Platform

Operate, from the first day of the project onwards, a comprehensive service platform offering the automated and cost-effective recognition and transcription of archival documents.

From the very beginning we understood the requirements set up by the e-Infrastructure call as a chance to provide a service platform directly dedicated to the needs of our user groups. The teams from UIBK, URO, ULCC and partly ASV, NAF and ABP are mainly involved in the service components of READ. The other teams are supporting them by adapting their research tools according to the specifications set up for the platform.

In Y1 we were able to professionalize the platform and its interfaces (expert interface, API). In Y2 we enlarged our service portfolio with the Layout Analysis tools, the Table Recognition, and the Language Tools. In Y3 a focus will be laid on enabling the exchange of data among the users of the platform ("sharing models") as well as to improve the scalability of the processes.

Open Platform

The Transkribus platform is "open" in many ways, for human beings as well as machines:

- It can be used by everyone who creates an account on the platform.
- The expert client can be downloaded for free to access the platform.
- All services in the platform can be accessed for free.
- A RESTful API is available to connect machines to the platform.
- Most of the software tools are Open Source and can be accessed or downloaded via GitHub.

Content uploaded to the platform is private by default but this does not necessarily contradict the concept of "openness". Two arguments have to be mentioned here.

First of all, scholars, researchers, and members of the public want to work with their own documents in a private sphere. The idea that all research activities will take place in public is naïve and does not take into account that research and scholarship are of course not only based on cooperation and collaboration but also on competition and exclusiveness. Nevertheless, we support of course that results achieved in research should be made publicly available (Open Access but also Open Research Data).

The second important argument is that private collections and documents help to avoid urgent issues of copyright and personal data which will appear obviously in research dealing with documents from the 20th century.

The user agreement of the Transkribus platform reflects exactly these considerations and limits therefore the usage of the documents within the platform to research and improvement of services.

From a strategic point of view openness also includes connectivity. A future research landscape will consist of several service platforms similar to Transkribus but with their own, specific portfolio. We are already in discussion with such e-Infrastructure providers to strengthen these components. Just to list a few:

- THOR provides a DataCite system (ORCID) where "relations between contributors, research artefacts (including data) and organizations" are established. It would be a great benefit for users if their activities in Transkribus (e.g. transcription, proof reading, extending abbreviations, tagging named entities, training networks, etc.) could be directly submitted to their research record built on the ORCID service.
- ZENODO and OpenAire are providing repository services to guarantee citation and long-term preservation of research data (articles, data sets). One of the most important advantages is again the quotability of the data, e.g. via a Document Object

Identifier (DOI). Connecting ZENODO with Transkribus would enable users to preserve their data in a long-term preservation infrastructure.

- CLARIN centres in Europe are gathering language data, and also especially data from historical documents. Users may want to make the automated transcription of their documents – which were processed by the HTR engine – available to CLARIN, enabling new research on aspects which they will not focus on.
- The DIVA group of the University of Fribourg (Switzerland) are providing very specific services for palaeographers, a field which READ touches indirectly but will not focus on. For users it would be great if they could send a document from Transkribus directly to the DIVA services, and receive in return an updated or enriched annotation of their document.

The list could go on but these examples clearly demonstrate that the ability of the Transkribus platform to connect with other platforms in the field is an important cornerstone for research infrastructures and virtual research environments. The "unique selling point" of Transkribus is that it is the only comprehensive platform dealing with the recognition, indexing and enrichment of historical documents with no limitations in time and alphabet.

Y2 brought in this respect a stronger collaboration especially with DARIAH and CLARIN based services, such as a collaboration with the PARTHENOS project.

HPC Integration

The task of running machine learning processes on a HPC cluster is a perfect example how fast technology changes. In Y1 we worked hard on the integration of HTR into the HPC server of the Central Computing Centre of the University of Innsbruck but had to accept that due to technical reasons this was not possible or doable with a reasonable effort. Main reasons are described in detail in Deliverable D4.16 HPC Integration and Maintenance).

In Y2 the decision was taken to go for a completely new option and to implement a new version of the CITlab HTR engine on the basis of Tensorflow. This machine learning platform was developed by Google and has become increasingly popular in 2017. Apart from enabling researchers to develop their algorithms in a convenient way Tensorflow also supports the use of GPUs. As already outlined above the performance gain is significant and will reduce recognition and training by a factor 5-10. In Y3 the final implementation will take place and will enable us to provide HTR services with a drastically reduced processing time. This will enable us to open up the training module for every registered user in Transkribus.

Graphical User Interfaces: Expert and Crowd-sourcing Tools

In Y1 we released the expert interface as version 1.0. In Y2 the expert interface was regularly improved, mainly with the Train Interface, the Text2Image matching interface[7] and several other updates and improvements. At the time of writing version 1.3.7. of Transkribus is the stable release. The expert interface represents a TRP level of 8-9 used by hundreds of researchers and volunteer for their daily work.

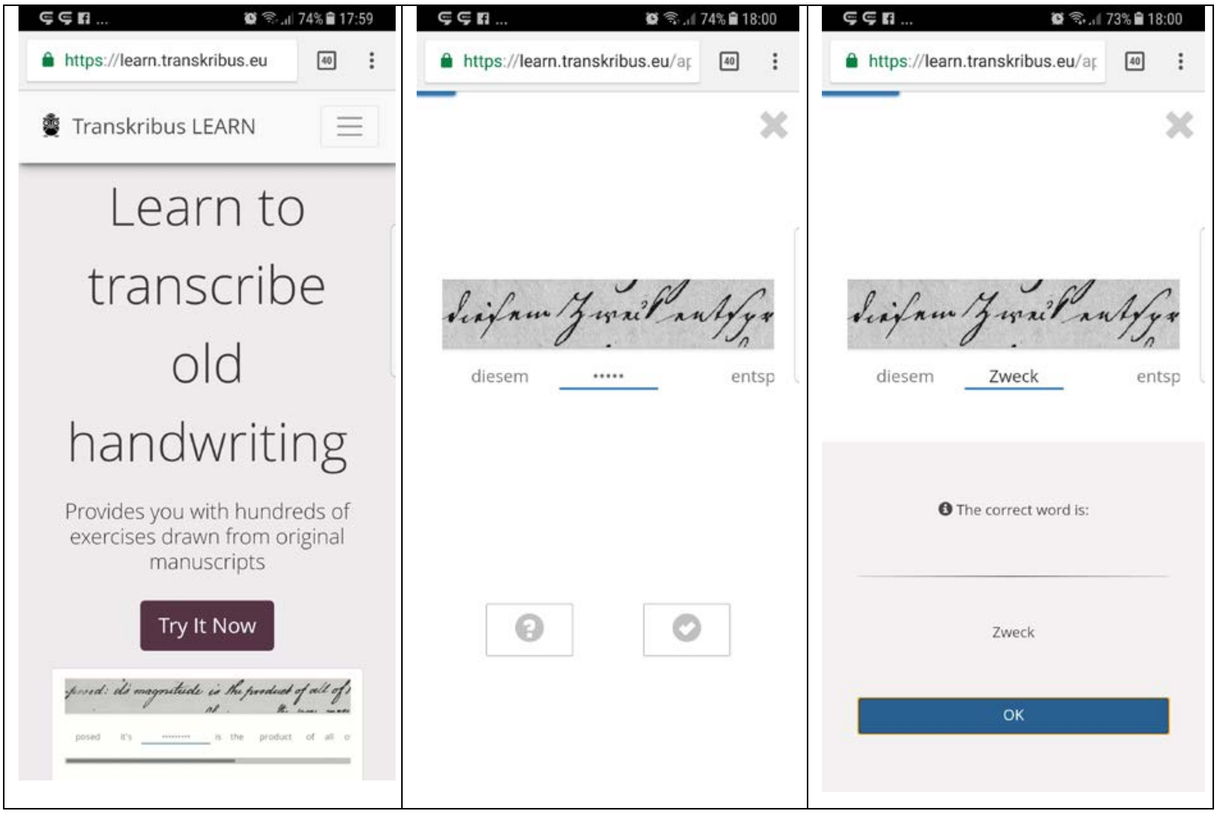
The crowd-sourcing tools were also prepared in the expert client as well as in the web-interface but the focus was set on improving the editing interface for text correction. This is the main prerequisite and it turned out that the complexity of such an interface was clearly

[7] Cf. Deliverable D7.20. Gundram Leifert, Tobias Strauß, Roger Labahn: Model for Semi- and Unsupervised HTR Training P2. How to get a good HTR without expensive ground truth production (2017) Online: https://read.transkribus.eu/wp-content/uploads/2017/12/Del_7.20.pdf

underestimated by the project team. Nevertheless at the end of Y2 we were able to release a first beta version.

eLearning application

In Y1 a prototype was developed for the e-Learning application. In Y2 the prototype was significantly improved and released to test users. Also a dedicated landing page was launched. The eLearning application can be used on a computer screen, but it is especially designed to work with a mobile phone. When doing their exercises users need not to key text on the mobile phone (which is not convenient) but two buttons (No idea – I know) are sufficient to interact with the application. The following screenshots are taken from a mobile phone:


Learn to transcribe old handwriting

Provides you with hundreds of exercises drawn from original manuscripts

The correct word is:

Zweck
17:59
https://learn.transkribus.eu
Transkribus LEARN
Try It Now
posed: it's •••••••• is the product of all o
18:00
https://learn.transkribus.eu/ap
diesem
•••••
entsp
18:00
https://learn.transkribus.eu/ap
diesem
Zweck
entsp
OK

Figure 9 Transkribus learn (eLearning interface) on a mobile phone

In addition we did a thorough investigation on all history and philology departments of all European and US universities in order to collect important members of our target group which are mainly those university lecturers who are teaching students in historical handwriting. More than 1500 such persons where identified and recorded and will be the target group for the application.

We had originally planned to make a public launch already in Y2 but we postponed this mainly due to the fact that we did not want to overstress ourselves. It is clear that the interest from the community will be very high and that we need to have a good mechanism in place to answer all questions and requests. In Y3 we will open the interface to every registered user in Transkribus and have a soft launch during spring 2018. This will go together with a targeted mailing campaign to the most important university history departments in Europe and abroad. We also have a clear business model in mind how the eLearning component is able to contribute to the overall success of the Transkribus platform. The current thinking is that of course access to all documents will be free, but that the incorporation of specific documents will be charged with a specific fee. This fee will not be necessary if a basic subscription is concluded with READcoop.

The potential impact of the eLearning application is high, thousands of historians and tens of thousands of students are supposed to respectively teach and learn historical handwriting.

ScanRead

The original plan was to create a document scanning app on basis of an existing app from ABBYY.[8] But it turned out to be more feasible to use technology available at the Computer Vision Lab of the Technical University Vienna (CVL). When developing the app, the idea of a ScanTent was born. The ScanTent is a simple device which enables users to digitize historical documents in an accurate and fast manner. With the ScanTent in the background, it turned out to be a great advantage to have the source code of the scanning app under full control of the project partners. Important features, such as the automated scanning mode which relieves users to press the release button on their mobile phone could be implemented easy and quickly.

We changed the name of the app from ScanREAD to DocScan. A very first prototype of the app as well as of the device were developed in Y1. In Y2 we launched the scanning app on Google Play, but we decided to wait for the availability of the ScanTent to launch a marketing campaign.

When presenting the ScanTent and DocScan at various occasions it turned out that users reacted enthusiastically on the device and the app. The prospect of becoming somehow independent of the scanning facilities and digitisation projects of libraries and archives and to be able to produce thousands of document images within a relatively short time period, is fascinating for researchers, volunteers and family historians. In the same way the prospect that users are sharing their digital images with the archive is fascinating for archive holders and an easy way to extend their digital collections. So what we learned in Y2 is that the ScanTent and DocScan have a great potential and are ideal completions of our overall objectives.

A first set of 15 prototypes was produced in Y2 and distributed to several users. Important feedback was gathered and first negotiations with suppliers took place. Unfortunately it turned out that the production of such a device is not trivial and requires a lot more work and competences than we had originally envisaged.

In Y3 we will produce another set of 30 prototypes in order to gain more user feedback and we will finalize a model which will serve as the benchmark for suppliers. We expect therefore to be able to place an order for a first series of 1000 devices during 2018.

Key concept 4: Cycle of Growth

Initiate a "cycle of growth" so that the READ Virtual Research Environment is constantly growing in terms of user involvement and collections made available. This will be a key to transform READ into a self-sustained platform offering its service portfolio in a business environment.

Transkribus as the READ Virtual Research Environment is based on the assumption that even if all images of historical documents of all archives in Europe would be collected in one single place this would – from a technical point of view – be a minor challenge for a supercomputing

[8] Cf. Deliverable D5.14: Günter Mühlberger (UIBK), Markus Diem (CVL), Stefan Fiel (CVL), Florian Kleber (CVL): D5.14. ScanREAD. (2016) Online: https://read.transkribus.eu/wp-content/uploads/2017/01/READ_D5.14_ScanREAD.pdf

centre such as Leibniz Rechenzentrum in Garching or similar institutions in Europe. The average size of images in the Transkribus platform is 5 MB which means that about 200.000 images can be stored at 1 TB. In other words: 200 million images will require just one Petabyte of storage.

In order to support the scalability of the Transkribus platform the University of Innsbruck supports the Transkribus platform with 100 TB dedicated storage, summing up to 20 mill. page images.

The possibility of storing millions of documents in one place offers completely new opportunities to process historical documents in a comparable way to Google. But the main obstacle for such a solution is not cost but rather the fact that neither archives/libraries nor their stakeholders (ministries, councils, communes) are yet prepared to understand that a truly digital science/cultural heritage requires new solutions which go far beyond local interests and insular approaches. The mind-set that cultural heritage is directly relevant to the “local”, “regional”, or “national” level is one of the main obstacles. An important aspect is here also that archives and libraries are often understanding their holdings mainly as “cultural heritage”. However for digital humanities scholars these holdings are “research data” which implies a completely different approach to their availability and usability.

Fortunately the launch of the European Open Science Cloud by the EU commission is a clear statement against this narrow-minded thinking and it will help us in the mid- and long-term to convince archives and memory organisations to make their holdings available also via central platforms such as READ/Transkribus.

In Y3 we will address the Cycle of Growth in three ways:

(1) Scalability of Transkribus

Currently a user is able to process thousands and even tens-of-thousands of pages in Transkribus with a few mouse clicks. But if we assume that several users are doing this per day we would run into performance issues. Both, the hardware as well as the job management need therefore be adapted to the requirement of being able to handle thousands of jobs at once. The integration of GPU servers, the usage of new software implementations (Tensorflow) will be important steps in this direction.

(2) Data sharing

One of the strongest feedbacks from the 1st Transkribus User Conference was that users are not only willing to share their data but are requesting such options to be able to do data sharing. And indeed, with several hundreds of models trained for different languages, time periods and scripts the chance that several users are working independently on similar data is getting higher and higher. This is also supported by the integration of Keyword Spotting into the Transkribus platform since KWS does not require high-quality transcriptions. A robust model which has “seen” a lot of different scripts will be in many cases fully sufficient that users find (nearly) every word in an arbitrary document.

(3) Cooperative

The idea of a truly collaborative setup of the Transkribus platform as a “cooperative” will be an important addition to the overall corporate identity of Transkribus. The cooperative will give all memory organisations the chance to deliver their documents for the sake of handwritten text recognition and keyword spotting, but also to keep control over their holdings also with respect to a legal background. The second main argument which may change the attitude will be that a centralized access to large collections will support research and scholarship in a much more effective way than setting up dozens of local installations.

Incorporation of Archives and Volunteers

The interest of archives and memory organisations was high in Y1 and we concluded about 25 Memorandum of Understandings. In Y2 – and especially after the Transkribus User Conference – the interest was even higher, about 25 MoUs were concluded, so that today we can count altogether 65 signed cooperation agreements.

In Y1 we have set up an innovative service which enables archives to connect directly with the Transkribus platform. This service was developed within a cooperation with the German software company Intranda (http://intranda.com/) which provides a well-known repository system ("Goobi").

The service enables Intranda users to select a document from a public repository and to transfer it directly to the Transkribus platform. The main benefit for users is that they need not to work with different repository systems, but that they can collect their documents from various sources in one place. The benefit for archives on the other hand is that e.g. transcribed text or enriched documents can be transformed back to the repository since main standards in the field (METS/ALTO/TEI) are supported by the Transkribus platform. The service was introduced to the Intranda users in autumn 2016.

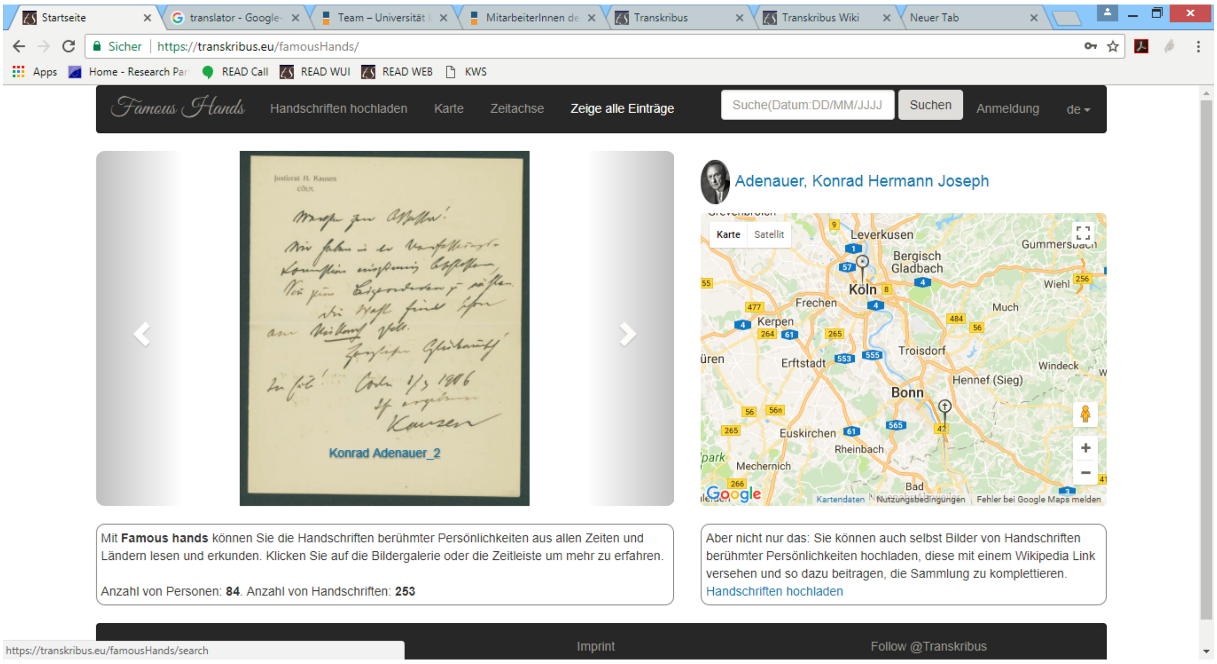
In Y2 the Transkribus team was approached by the University of Gent with a request for cooperation where exactly this service will play an important role to establish Transkribus as a tool for the Digital Humanities community in Belgium. In Y3 an application shall be drafted for implementing this service on an operative level. We believe that more archives will follow this example especially once the KWS service becomes available also via the web-interface.

As already explained above the incorporation of members of the public (crowd, volunteers) was not our focus in Y1 and also in Y2 we were cautious to send out any information to genealogists and family historians – despite the fact that we are completely aware that this group is highly interested in Transkribus. The reason for our cautiousness is simply that we do know that we are currently not in the situation to "simply" satisfy the needs of this user group and that the communication effort is high. In Y3 this will change significantly. With applications such as the ScanTent and Doc/Scan as well as with KWS we are able to offer specific services for this target group.

Marketing Campaigns: EuropeanHands and e-learning and mobile applications

The "cycle of growth" shall also be fed by applications such as the eLearning application (now: transkribus.learn) or FamousHands where we hope to encourage a large number of users to take part in the project in various ways.

In Y1 we decided to radically simplify the concept for FamousHands (renamed from EuropeanHands). Instead of putting Handwritten Text Recognition into the focus of the campaign we have built it around the idea of collecting the handwriting of "famous persons" – wherever they may come from. Transkribus will play only an indirect role.



Mit **Famous hands** können Sie die Handschriften berühmter Persönlichkeiten aus allen Zeiten und Ländern lesen und erkunden. Klicken Sie auf die Bildergalerie oder die Zeitleiste um mehr zu erfahren.

Anzahl von Personen: **84**. Anzahl von Handschriften: **253**

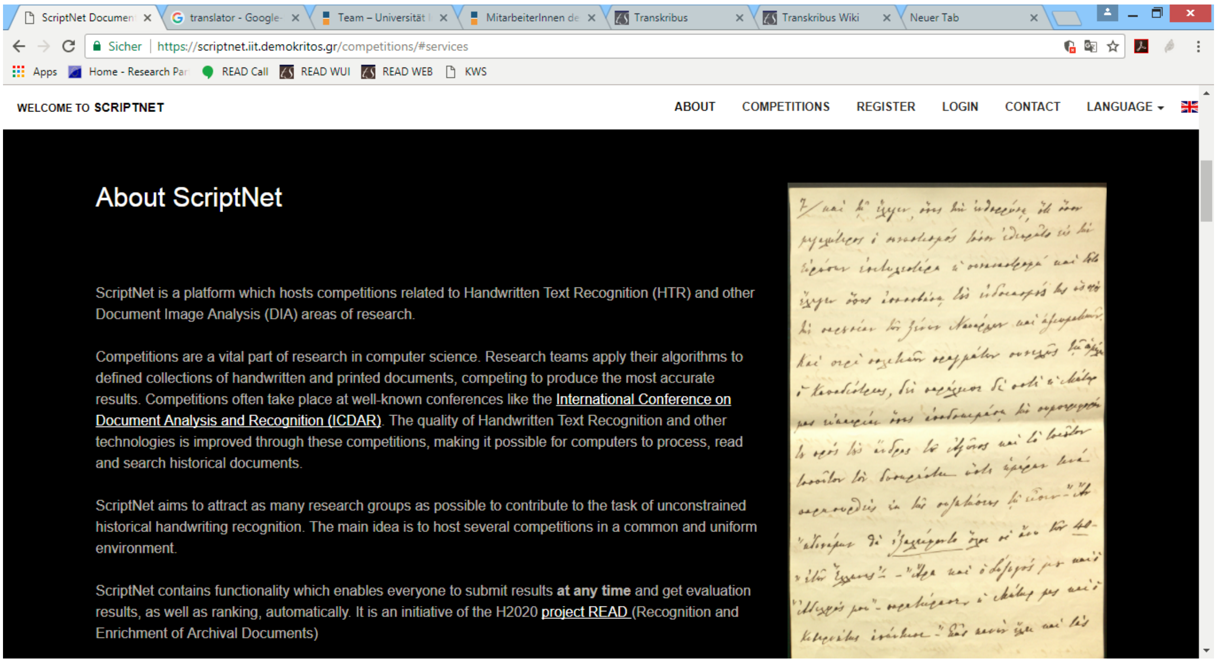
Aber nicht nur das: Sie können auch selbst Bilder von Handschriften berühmter Persönlichkeiten hochladen, diese mit einem Wikipedia Link versehen und so dazu beitragen, die Sammlung zu komplettieren.
Handschriften hochladen

Figure 10 Famous Hands website (http://transkribus.eu/famousHands)

The eLearning app and the DocScan app will follow a similar approach and address the needs of specific user groups (students, scholars and members of the public). In this way the user basis of the READ platform shall be significantly extended.

ScriptNet competition

In Y1 another innovative service was developed for organizing research competitions in the Document Image Analysis domain. Although competitions were organized for many years as part of the two top conferences in the field, the International Conference on Document Analysis and Recognition (ICDAR) and the International Conference on Frontiers in Handwriting Recognition (ICFHR) so far no research team has provided a dedicated and open site for managing research competitions. In Y1 the site was implemented, in Y2 several competitions were managed with ScriptNet.



About ScriptNet

ScriptNet is a platform which hosts competitions related to Handwritten Text Recognition (HTR) and other Document Image Analysis (DIA) areas of research.

Competitions are a vital part of research in computer science. Research teams apply their algorithms to defined collections of handwritten and printed documents, competing to produce the most accurate results. Competitions often take place at well-known conferences like the International Conference on Document Analysis and Recognition (ICDAR). The quality of Handwritten Text Recognition and other technologies is improved through these competitions, making it possible for computers to process, read and search historical documents.

ScriptNet aims to attract as many research groups as possible to contribute to the task of unconstrained historical handwriting recognition. The main idea is to host several competitions in a common and uniform environment.

ScriptNet contains functionality which enables everyone to submit results **at any time** and get evaluation results, as well as ranking, automatically. It is an initiative of the H2020 project READ (Recognition and Enrichment of Archival Documents)

Figure 11 ScriptNet website

In Y3 we will proceed this approach and are planning to launch an Open Competition (not connected to any conference) for document understanding.

User-driven Approach

One of the main advantages of a central Virtual Research Environment is that it enables us to understand much better the way users are interacting with their historical documents compared to anonymous environments such as digital libraries or repositories. In our case we have data which documents a user uploads to the system, which actions are performed, how often users return or leave the interface. But more importantly we receive daily feedback from a large number of users working with the platform. This feedback is an extremely valuable input for the further improvement of the platform.

Key concept 5: Make Innovation Happen

We need to understand that innovation requires different strategies and methods compared to research. Innovation is always "disruptive" and provides a new view, a new concept, a new idea which may lead to a new product or service. In order to address innovation adequately we have "reserved" special fields within the project where we want to encourage the creation of new ideas.

We believe that we clearly exceeded the goals for this key concept. On the one hand we got early feedback from users concerning our services summarized under this heading, such as the eLearning application, the DocScan app or ScriptNet, and on the other hand completely new ideas and services came up.

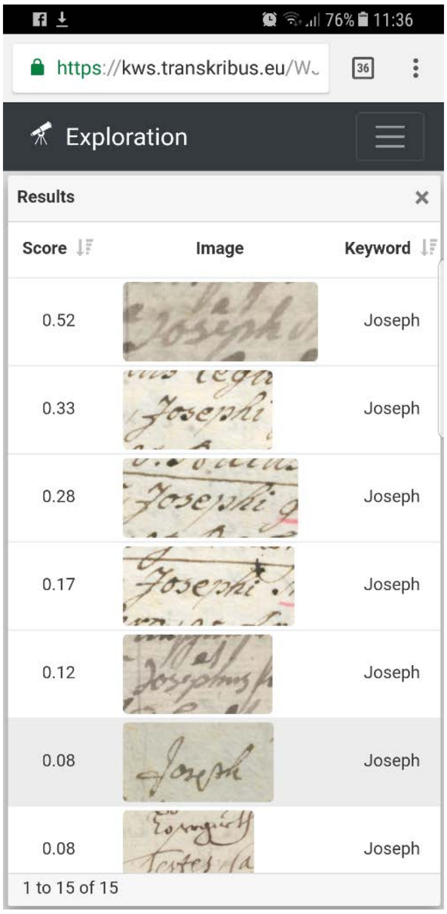
It is a characteristic of innovations that they "appear" and cannot be planned in the same way as research or services. The only way to encourage innovation is therefore to put a "place holder" into the work plan where innovation may happen. Task 8.2. Open Innovation Forum is such a "place holder".

Service Innovation - Open Innovation Forum and Large Scale Demonstrators

Services which go beyond our original work plan are all directly connected with our core objectives but extend them significantly. We already described the ScanTent and the DocScan app but we also want to mention here two other tools/services which came up during Y1 and Y2. E.g. some attempts to deal with abbreviations in historical documents, or with cadastre maps.

Other innovative services – which were not foreseen in the GA – are connected with Keyword Spotting. KWS is by sure one of the cornerstones of the "revolution" we would like to put into motion in the archives and humanities field. Especially the precision-recall trade-off is one of the most useful features for anyone interested in searching and investigating historical documents. Precision-recall trade-off means that the user is able to decide whether he wants to get very precise results (but will miss some hits hidden in the document) or if he accepts to go through rather inaccurate results – but will miss only a very few hits from the document. Inaccurate results will mean that e.g. up to 50% of all hits displayed are "false alarms".

Therefore it is obvious that a further step of "validating" the search results will be useful to deal with the results in a convenient way. And exactly for this step in the workflow we will come up with first solutions in Y3 of the project. The following screenshot shows an early concept for displaying results and involving users in this process:



76% 11:36

https://kws.transkribus.eu/W...

Exploration

Results

Score	Image	Keyword
0.52		Joseph
0.33		Joseph
0.28		Joseph
0.17		Joseph
0.12		Joseph
0.08		Joseph
0.08		Joseph

1 to 15 of 15

Figure 12 Mobile interface for validating KWS results

The search term was “Joseph” and starting with high confidence hits the list returns finally also a false alarm. The task of the volunteers would be similar to the selection process in well-known applications like “Tinder”: image to the left would mean that it is incorrect (false alarm), image to the right would mean that it is correct.

If we take the example from above we can also see that even if someone is not very familiar with this German script from the 18th century he will be able to compare the hits and starting from the hits with a high confidence (e.g. 0,52 for the first hit) he will be able to decipher hit no. 6 correctly as “Joseph” – even if he would not be able to do that without having the context of the other hits.

Another advantage of KWS can be illustrated in this example very well: As we can see the engine finds not only “Joseph” but also “Josephi” and “Josephus”. This is especially useful if we deal with historical spelling variants and KWS solves this problem en passant.

Nevertheless it becomes also very clear that such an interface will enable many more people to take part in crowd-sourcing projects than nowadays. Especially the fact that a mobile phone can be used for this work will increase the chance that many people are contributing without any need to sit in front of a computer screen and do “real” work.

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.

2. Dissemination activities

2.1. 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.

2.2. 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.[9] 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.

[9] 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

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**.

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.

2.3. 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

2.3.1. 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.

2.3.2. 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.

2.3.3. 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.

2.3.4. 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.

2.3.5. 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.

2.3.6. 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.[10]

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

[10] 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.

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.

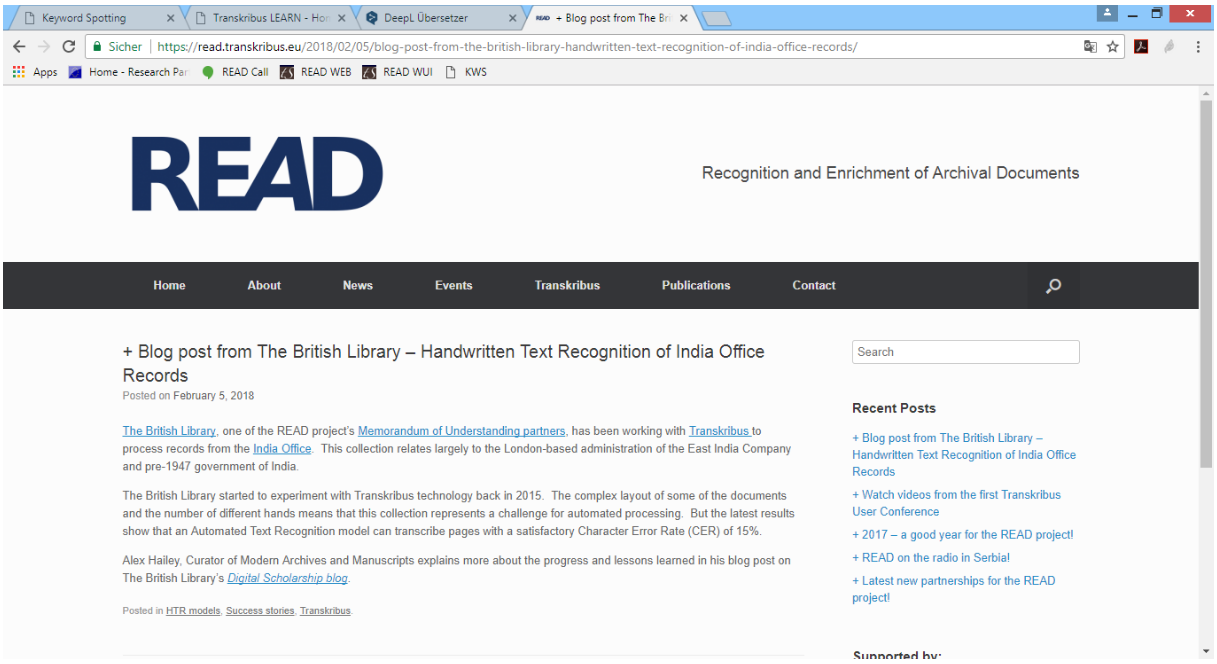
3. Dissemination channels

3.1. 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”.



+ Blog post from The British Library – Handwritten Text Recognition of India Office Records

Posted on February 5, 2018

The British Library, one of the READ project's Memorandum of Understanding partners, has been working with Transkribus to process records from the India Office. This collection relates largely to the London-based administration of the East India Company and pre-1947 government of India.

The British Library started to experiment with Transkribus technology back in 2015. The complex layout of some of the documents and the number of different hands means that this collection represents a challenge for automated processing. But the latest results show that an Automated Text Recognition model can transcribe pages with a satisfactory Character Error Rate (CER) of 15%.

Alex Hailey, Curator of Modern Archives and Manuscripts explains more about the progress and lessons learned in his blog post on The British Library's *Digital Scholarship blog*.

Posted in HTR models, Success stories, Transkribus.

Recent Posts

- + Blog post from The British Library – Handwritten Text Recognition of India Office Records
- + Watch videos from the first Transkribus User Conference
- + 2017 – a good year for the READ project!
- + READ on the radio in Serbia!
- + Latest new partnerships for the READ project!

Supported by:

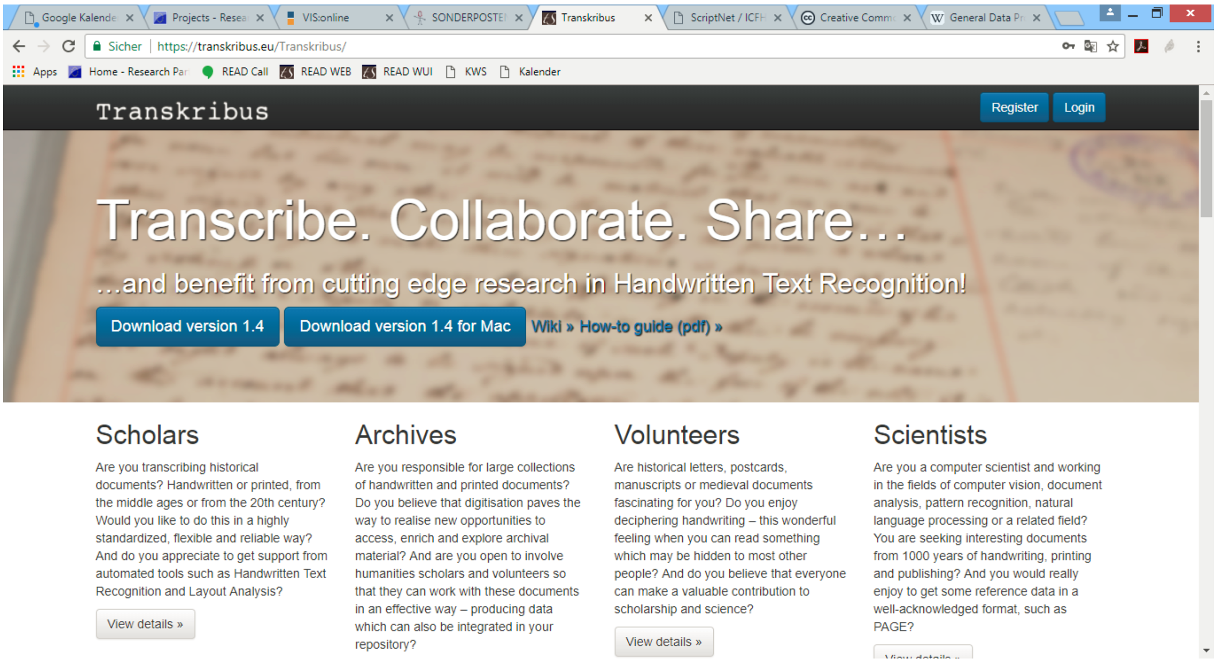
Figure 13 Blog post about a blog post by the British Library

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

3.2. 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.



Transcribe. Collaborate. Share…

…and benefit from cutting edge research in Handwritten Text Recognition!

Scholars

Are you transcribing historical documents? Handwritten or printed, from the middle ages or from the 20th century? Would you like to do this in a highly standardized, flexible and reliable way? And do you appreciate to get support from automated tools such as Handwritten Text Recognition and Layout Analysis?

Archives

Are you responsible for large collections of handwritten and printed documents? Do you believe that digitisation paves the way to realise new opportunities to access, enrich and explore archival material? And are you open to involve humanities scholars and volunteers so that they can work with these documents in an effective way – producing data which can also be integrated in your repository?

Volunteers

Are historical letters, postcards, manuscripts or medieval documents fascinating for you? Do you enjoy deciphering handwriting – this wonderful feeling when you can read something which may be hidden to most other people? And do you believe that everyone can make a valuable contribution to scholarship and science?

Scientists

Are you a computer scientist and working in the fields of computer vision, document analysis, pattern recognition, natural language processing or a related field? You are seeking interesting documents from 1000 years of handwriting, printing and publishing? And you would really enjoy to get some reference data in a well-acknowledged format, such as PAGE?

Figure 14 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.

3.3. 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.

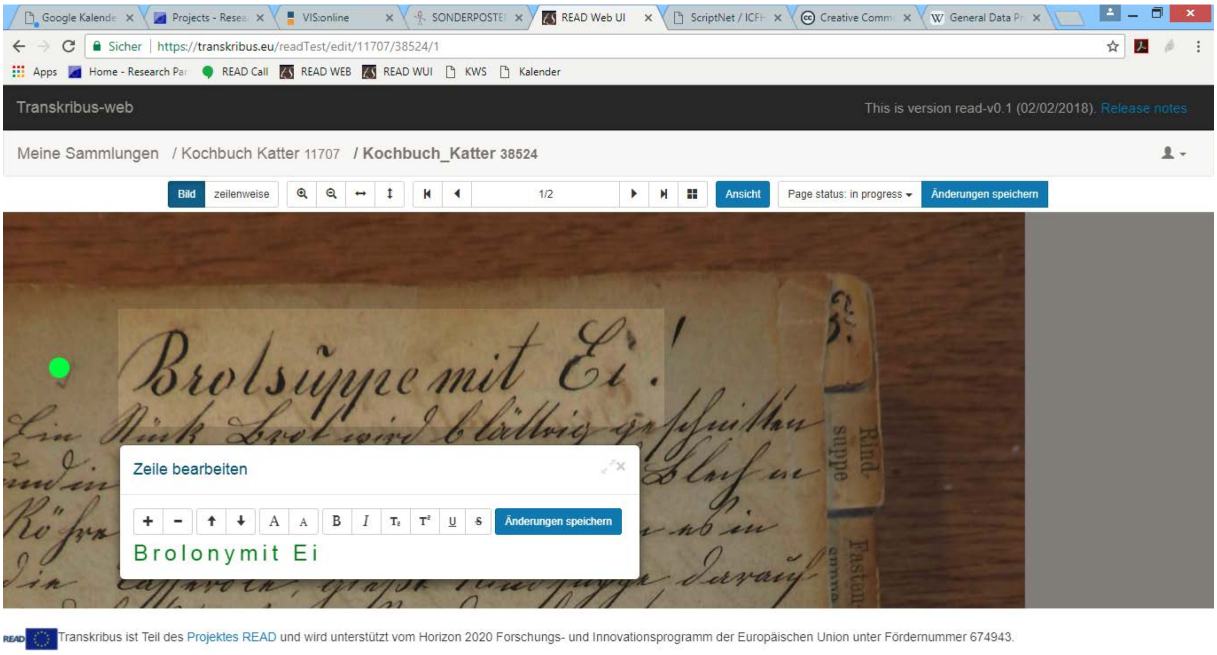

https://transkribus.eu/readTest/edit/11707/38524/1
Transkribus-web
This is version read-v0.1 (02/02/2018). Release notes
Meine Sammlungen / Kochbuch Katter 11707 / Kochbuch_Katter 38524
Bild
zeilenweise
1/2
Ansicht
Page status: in progress
Änderungen speichern
Zeile bearbeiten
Änderungen speichern
Brolonymit Ei
Transkribus ist Teil des Projektes READ und wird unterstützt vom Horizon 2020 Forschungs- und Innovationsprogramm der Europäischen Union unter Fördernummer 674943.

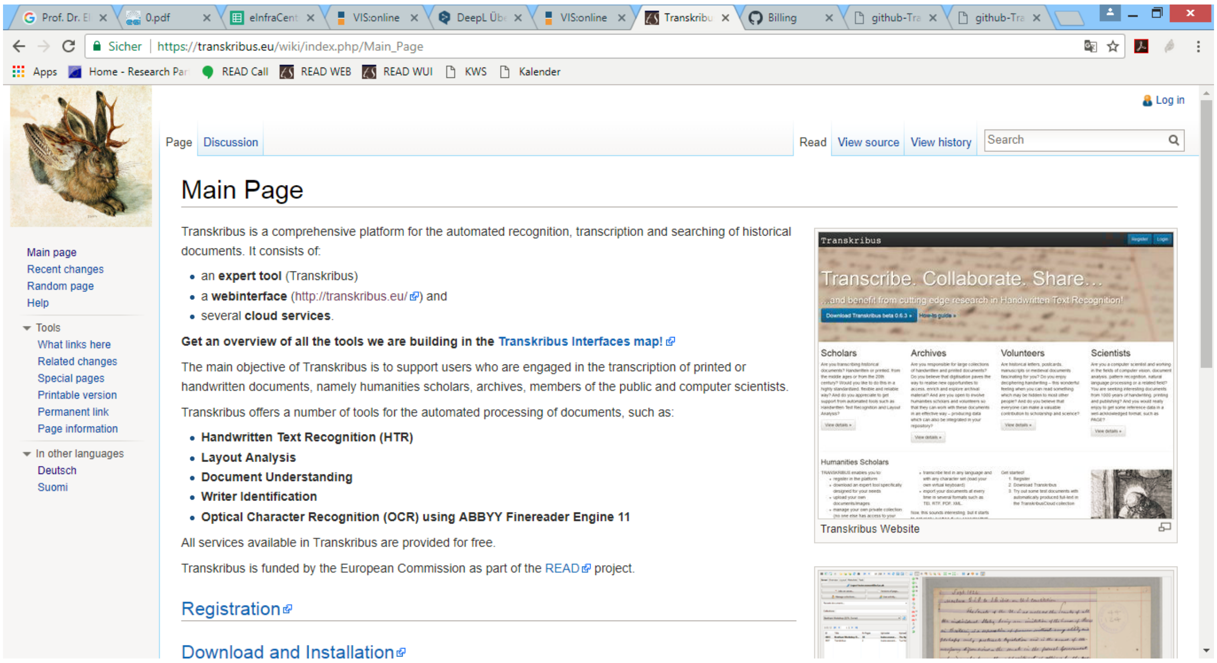
Figure 15 Transkribus web-interface / document library

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.

3.4. 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).



Main Page

Transkribus is a comprehensive platform for the automated recognition, transcription and searching of historical documents. It consists of:

- an **expert tool** (Transkribus)
- a **webinterface** (http://transkribus.eu/) and
- several **cloud services**.

Get an overview of all the tools we are building in the Transkribus Interfaces map!

The main objective of Transkribus is to support users who are engaged in the transcription of printed or handwritten documents, namely humanities scholars, archives, members of the public and computer scientists.

Transkribus offers a number of tools for the automated processing of documents, such as:

- **Handwritten Text Recognition (HTR)**
- **Layout Analysis**
- **Document Understanding**
- **Writer Identification**
- **Optical Character Recognition (OCR) using ABBYY Finereader Engine 11**

All services available in Transkribus are provided for free.

Transkribus is funded by the European Commission as part of the READ project.

Registration

Download and Installation

Figure 16 Transkribus Wiki

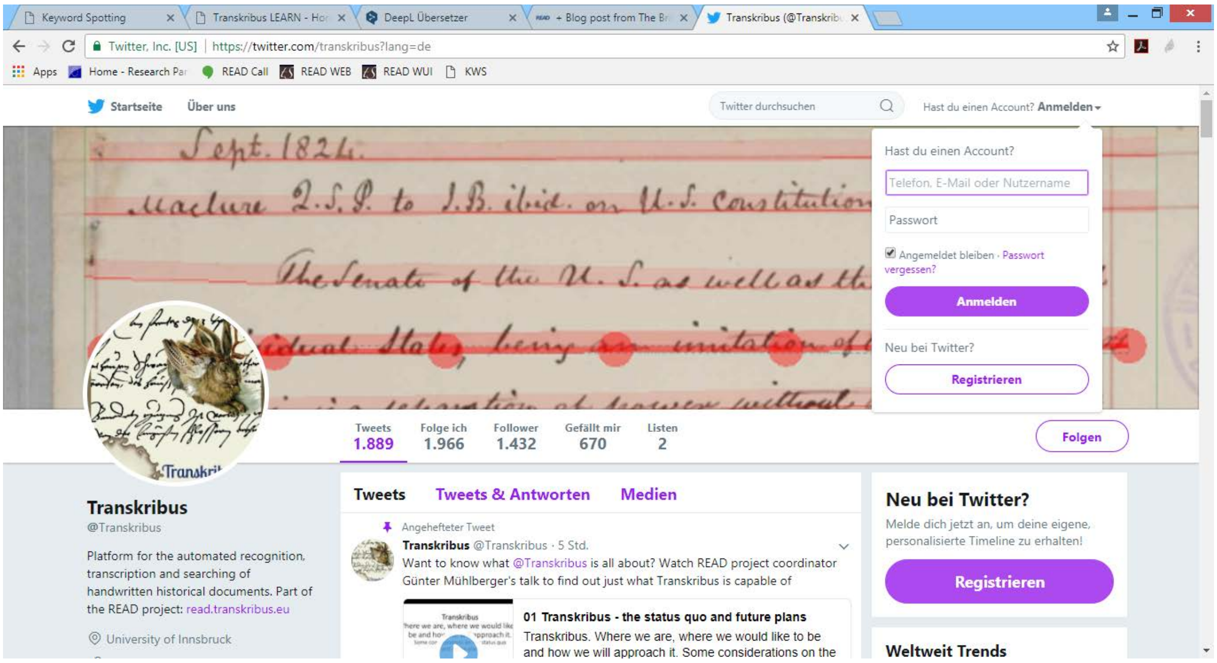
In order to reduce the maintenance effort for the Wiki we will drastically reduce the amount of content made available via this site.

3.5. 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.



Tweets	Folge ich	Follower	Gefällt mir	Listen
1.889	1.966	1.432	670	2

Transkribus
@Transkribus

Platform for the automated recognition, transcription and searching of handwritten historical documents. Part of the READ project: read.transkribus.eu

University of Innsbruck

Angehefteter Tweet
Transkribus @Transkribus · 5 Std.
Want to know what @Transkribus is all about? Watch READ project coordinator Günter Mühlberger's talk to find out just what Transkribus is capable of

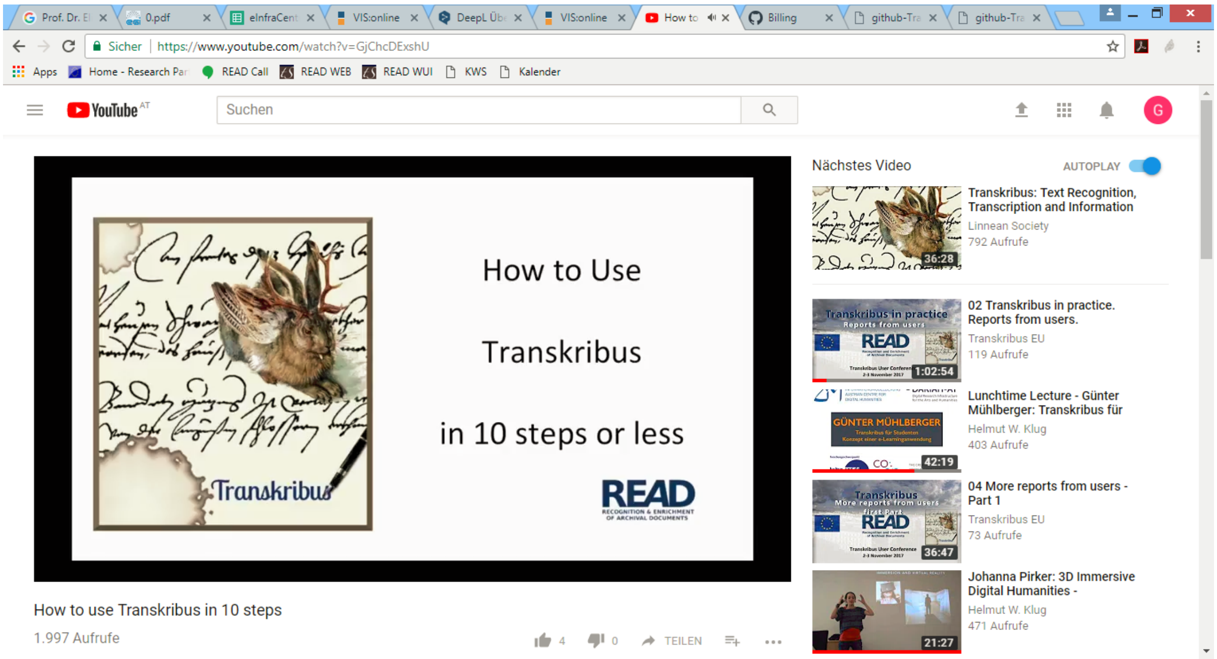
01 Transkribus - the status quo and future plans
Transkribus. Where we are, where we would like to be and how we will approach it. Some considerations on the

Melde dich jetzt an, um deine eigene, personalisierte Timeline zu erhalten!

3.6. 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.



https://www.youtube.com/watch?v=GjChcDExshU
YouTube
Suchen
How to Use
Transkribus
in 10 steps or less
READ
How to use Transkribus in 10 steps
1.997 Aufrufe
TEILEN
Nächstes Video
AUTOPLAY
Transkribus: Text Recognition, Transcription and Information
Linnean Society
792 Aufrufe
02 Transkribus in practice. Reports from users.
Transkribus EU
119 Aufrufe
Lunchtime Lecture - Günter Mühlberger: Transkribus für
Helmut W. Klug
403 Aufrufe
04 More reports from users - Part 1
Transkribus EU
73 Aufrufe
Johanna Pirker: 3D Immersive Digital Humanities -
Helmut W. Klug
471 Aufrufe

Figure 17 YouTube channel for Transkribus

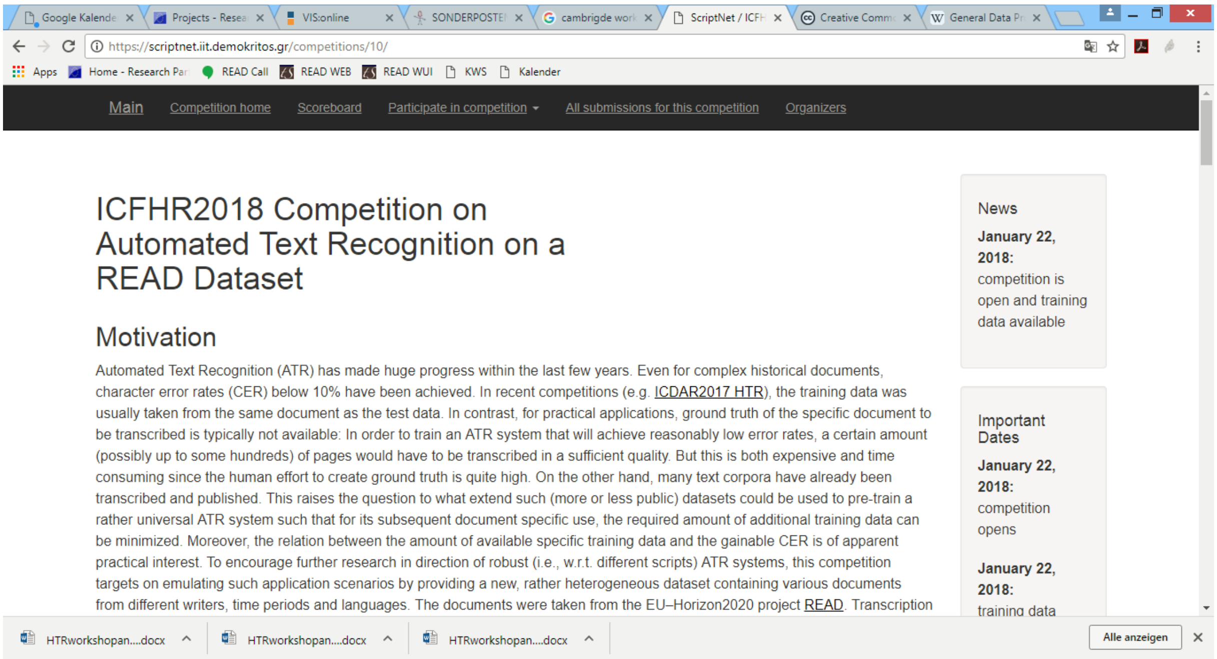
The 10 steps guide was viewed nearly 2000 times at the YouTube site of Transkribus.

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

3.7. 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.



https://scriptnet.iit.demokritos.gr/competitions/10/

Main | Competition home | Scoreboard | Participate in competition | All submissions for this competition | Organizers

ICFHR2018 Competition on Automated Text Recognition on a READ Dataset

Motivation

Automated Text Recognition (ATR) has made huge progress within the last few years. Even for complex historical documents, character error rates (CER) below 10% have been achieved. In recent competitions (e.g. ICDAR2017 HTR), the training data was usually taken from the same document as the test data. In contrast, for practical applications, ground truth of the specific document to be transcribed is typically not available: In order to train an ATR system that will achieve reasonably low error rates, a certain amount (possibly up to some hundreds) of pages would have to be transcribed in a sufficient quality. But this is both expensive and time consuming since the human effort to create ground truth is quite high. On the other hand, many text corpora have already been transcribed and published. This raises the question to what extend such (more or less public) datasets could be used to pre-train a rather universal ATR system such that for its subsequent document specific use, the required amount of additional training data can be minimized. Moreover, the relation between the amount of available specific training data and the gainable CER is of apparent practical interest. To encourage further research in direction of robust (i.e., w.r.t. different scripts) ATR systems, this competition targets on emulating such application scenarios by providing a new, rather heterogeneous dataset containing various documents from different writers, time periods and languages. The documents were taken from the EU–Horizon2020 project READ. Transcription

News

January 22, 2018: competition is open and training data available

Important Dates

January 22, 2018: competition opens

January 22, 2018: training data

Figure 18 ScriptNet site with ICFHR 2018 HTR competition site

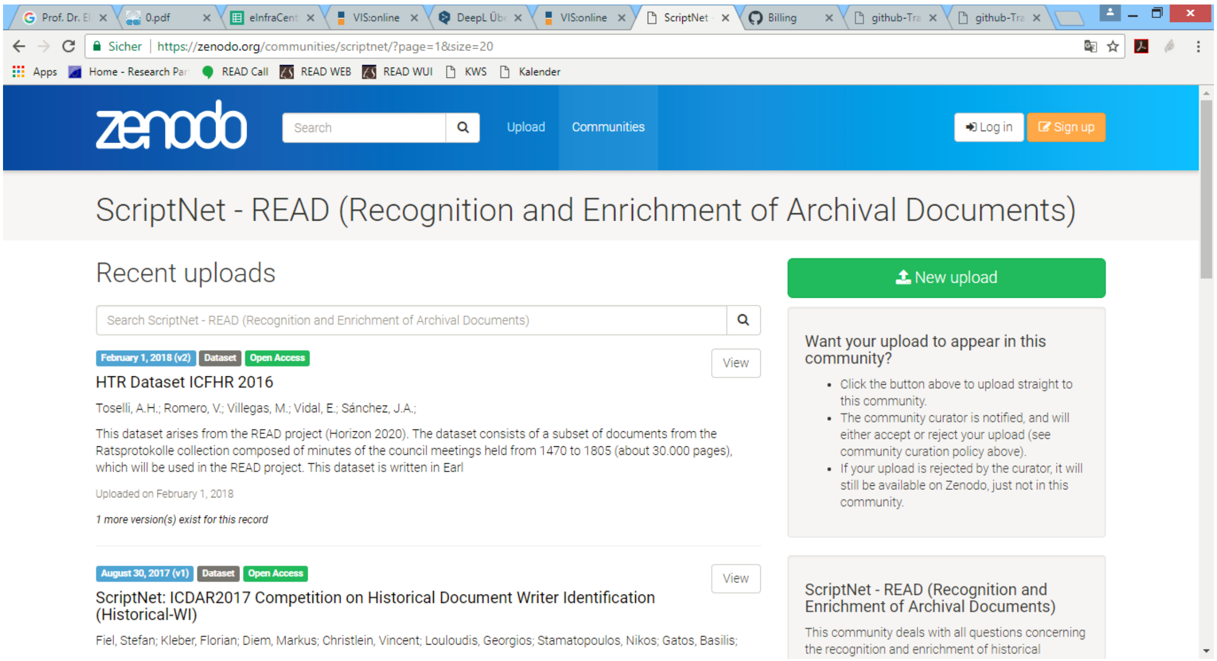
Five competitions were organised in 2017, and for 2018 already an HTR competition was accepted by the International Conference on Frontiers in Handwriting Recognition (ICFHR).

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.

3.8. 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.



ScriptNet - READ (Recognition and Enrichment of Archival Documents)

Recent uploads

Search ScriptNet - READ (Recognition and Enrichment of Archival Documents)

February 1, 2018 (v2) Dataset Open Access

View

HTR Dataset ICFHR 2016

Toselli, A.H.; Romero, V.; Villegas, M.; Vidal, E.; Sánchez, J.A.;

This dataset arises from the READ project (Horizon 2020). The dataset consists of a subset of documents from the Ratsprotokolle collection composed of minutes of the council meetings held from 1470 to 1805 (about 30.000 pages), which will be used in the READ project. This dataset is written in Earl

Uploaded on February 1, 2018

1 more version(s) exist for this record

August 30, 2017 (v1) Dataset Open Access

View

ScriptNet: ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)

Fiel, Stefan; Kleber, Florian; Diem, Markus; Christlein, Vincent; Louloudis, Georgios; Stamatopoulos, Nikos; Gatos, Basilis;

New upload

Want your upload to appear in this community?

- Click the button above to upload straight to this community.
- The community curator is notified, and will either accept or reject your upload (see community curation policy above).
- If your upload is rejected by the curator, it will still be available on Zenodo, just not in this community.

ScriptNet - READ (Recognition and Enrichment of Archival Documents)

This community deals with all questions concerning the recognition and enrichment of historical

Figure 19 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.

3.9. 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.

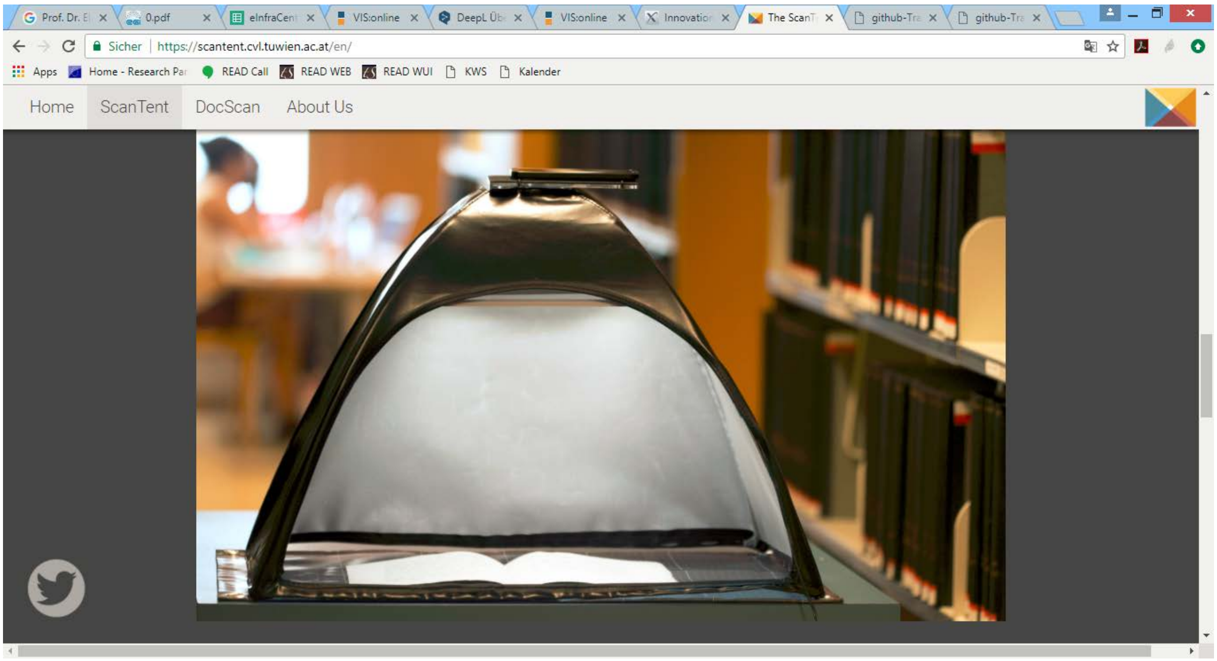


Figure 20 ScanTent website

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

6. References

READ Website

- http://read.transkribus.eu/

ScriptNet

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

ZENODO: ScriptNet/READ datasets

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

Transkribus

- http://transkribus.eu/

learn.transkribus.eu

- http://learn.transkribus.eu/

Transkribus YouTube Channel

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

Transkribus Twitter

- https://twitter.com/transkribus

FamousHands

- https://transkribus.eu/famousHands/