Recognition and Enrichment of Archival Documents

D5.7.
Development and Implementation of mobile Crowd-Sourcing Tools

Rory McNicholl, ULCC
Maximillian Bryan, ASV
Matti Jokinen, NAF
Berthold Ulreich, UIBK

Distribution: Public

http://read.Transkribus.eu/

READ
H2020 Project 674943

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

Project ref no.	H2020 674943
Project acronym	**READ**
Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	01 January 2016 / 42 Months

Distribution	Public
Contractual date of delivery	31/12/2018
Actual date of delivery	28/12/2018
Date of last update	28/12/2018
Deliverable number	D5.7
Deliverable title	Development and Implementation of mobile Crowd-Sourcing Tools
Type	Report on demonstrator
Status & version	Final
Contributing WP(s)	WP5, WP4
Responsible beneficiary	ULCC
Other contributors	ASV, NAF
Internal reviewers	Günter Mühlberger (UIBK)
Author(s)	Rory McNicholl, Maximillian Bryan, Matti Jokinen, Berthold Ulreich
EC project officer	Christophe Doin
Keywords	Handwritten Text Recognition, Web interface, crowd sourcing, mobile

Table of Contents

Executive summary

This document gives an overview of the web user interface to the Transkribus suite of tools represented by the by the “TranskribusWeb” website. This report will set out the changes and progress made during 2018.

1. Introduction

TranskribusWeb is a general term used to describe the presentation of the Transkribus functionality to people via a web browser as distinct from the Expert client desktop application (Transkribus-X). A beta-test of TranskribusWeb as it was at the end of 2017 gave the working group the opportunity to reflect on the usability of the application and what was required to deliver an application of real value.

1.1. 2017 Beta-test response

After feedback from a selected cohort of beta-testers, the overriding message was that the application as it stood contained too many bugs to allow for a consistent user-experience. Problems included errors when retrieving certain documents, truncation of longer lines, problems with right-to-left transcription, certain images not being displayed. In general though much effort had been made to provide a user-friendly interface, what was actually presented could behave in unexpected and off-putting ways.

1.2. Review and refactor of code base

During 2017 features and amendments had been frequently requested by the project management and other partners. Many features were subsequently added to the code-base without due consideration of the overall integrity of the code. During the first months of 2018 analysis of the underlying code base highlighted a high level of interdependence within the client-side code. This had meant it was difficult to add new features without causing unexpected results in some other existing features. In turn leading to the introduction of bugs (including those reported during the testing).

The decision was made by the group to refactor the client side code so that a set of well-defined components could be developed that would continue to operate as expected as other components were added. During this time the group was joined by Berthold Ulreich (UIBK) who had worked on the Transkribus e-learning application in the previous year.

Some work was performed to simplify the server side code and introduce a more consistent approach to handling the data obtained from the Transkribus API.

In general the focus of the group in 2018 was to provide robust features and present them in a clean and intuitive manner. This meant in some cases removing features previously requested by the wider project.

2. Developments in 2018

2.1. Componentry and the sandbox

A number of UI components were developed and tested in a sandbox environment (https://Transkribus.eu/r/testing/sandbox/). These sandbox components were not intended as finished solutions but allowed the group to assess the use of different techniques and libraries. These sandbox components tended to concentrate on executing one of the many specific tasks involved in delivering the TranskribusWeb UI. Only once it was found that this task could be reliably performed the process of integrating the features could begin.

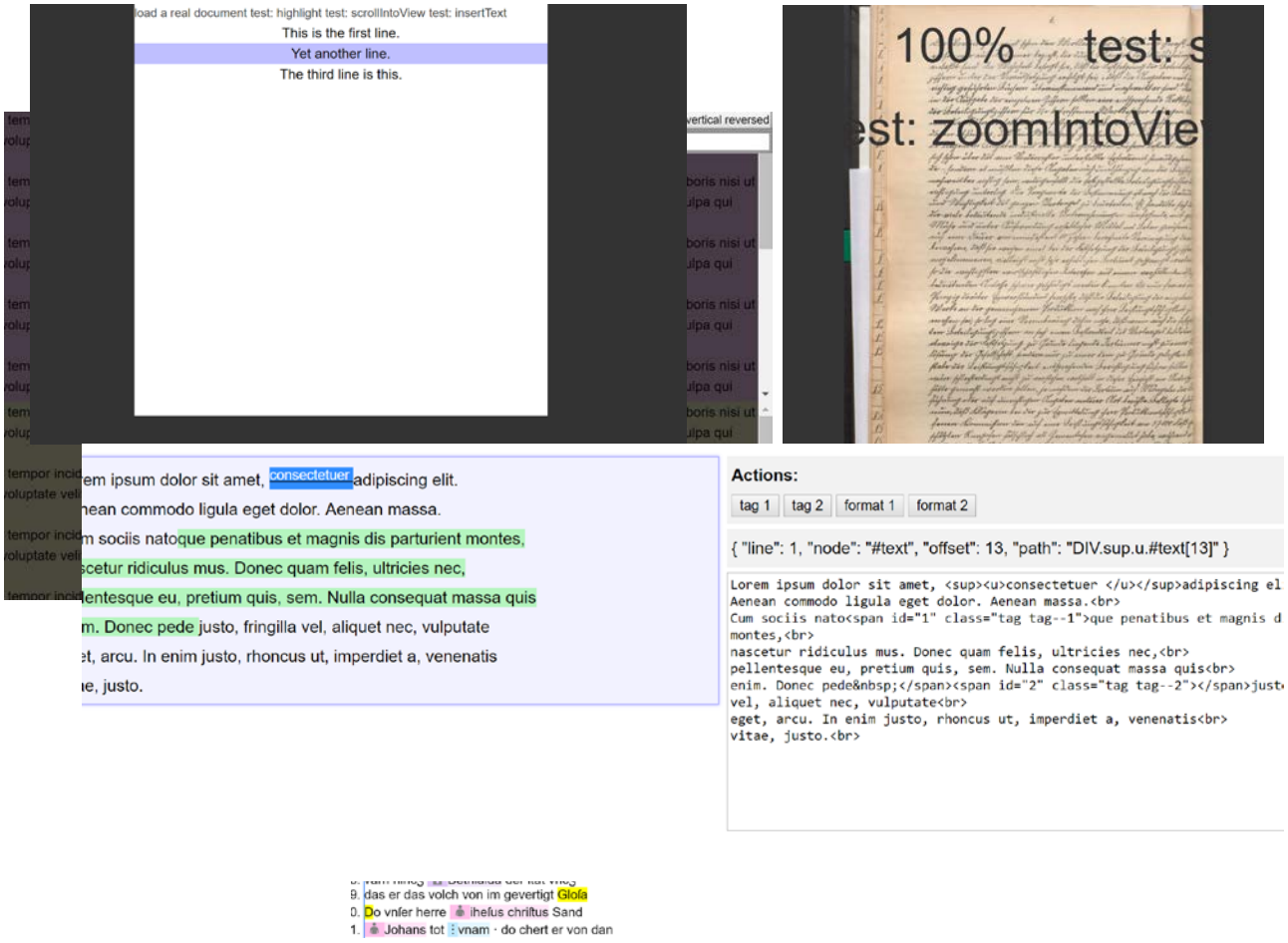

Figure 1 Examples of the sandbox components used to assess and develop features like simple text input, image zooming, page layout and markup manipulation.

2.2. Website structure

As mentioned above removal of a number of features was required to fulfil the purpose of providing an intuitive and uncluttered user-experience.

At the end of last year, after a user has logged in, there was a list of collections shown. These collections were shown with a thumbnail, title, description, number of documents, the user's role in that collection and a small table showing statistics.

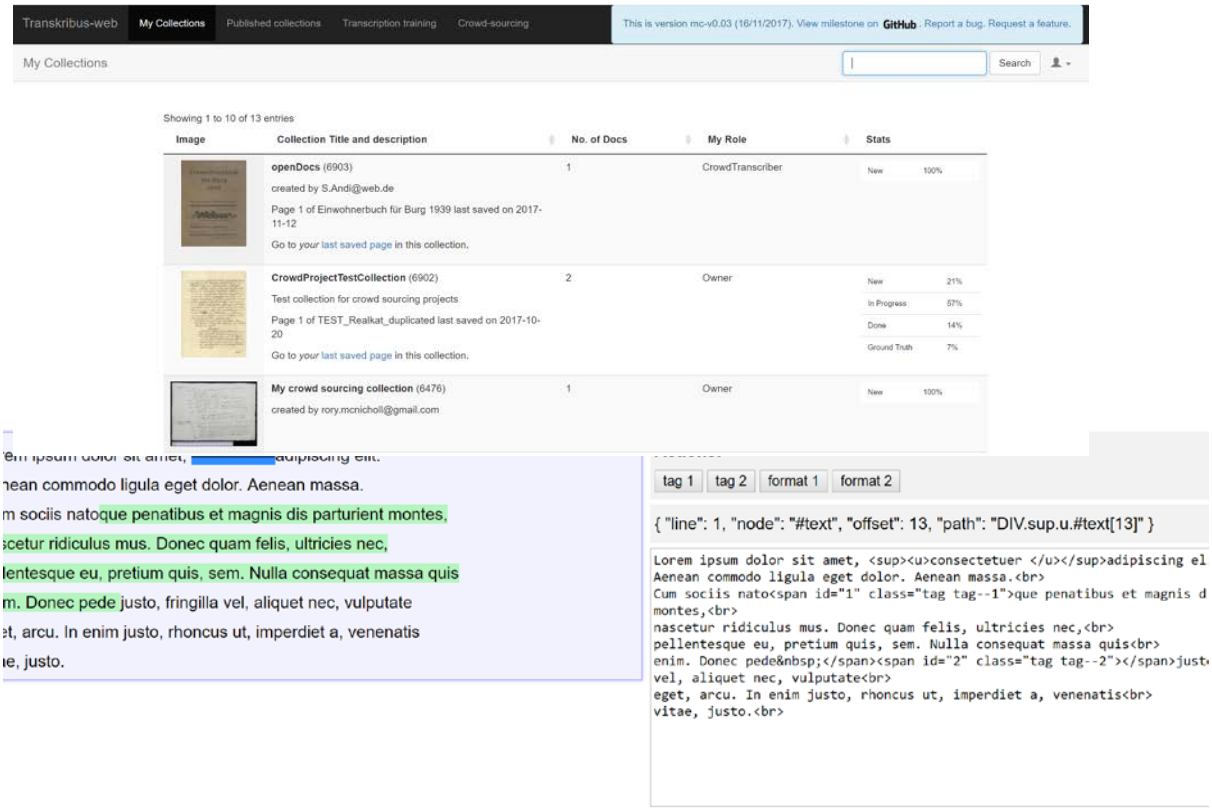

Figure 2 Collections list 2017

Apart from a color change, this year's version of the website is showing much less information. Due to no descriptions in differing lengths and images in various sizes and colors, the page appears much calmer.

Figure 3 A clear list of collections available to the public

When opening a collection, in last year's version, the user was shown the same kind of datatable that was also used for the collection table. It again contains a lot of information.

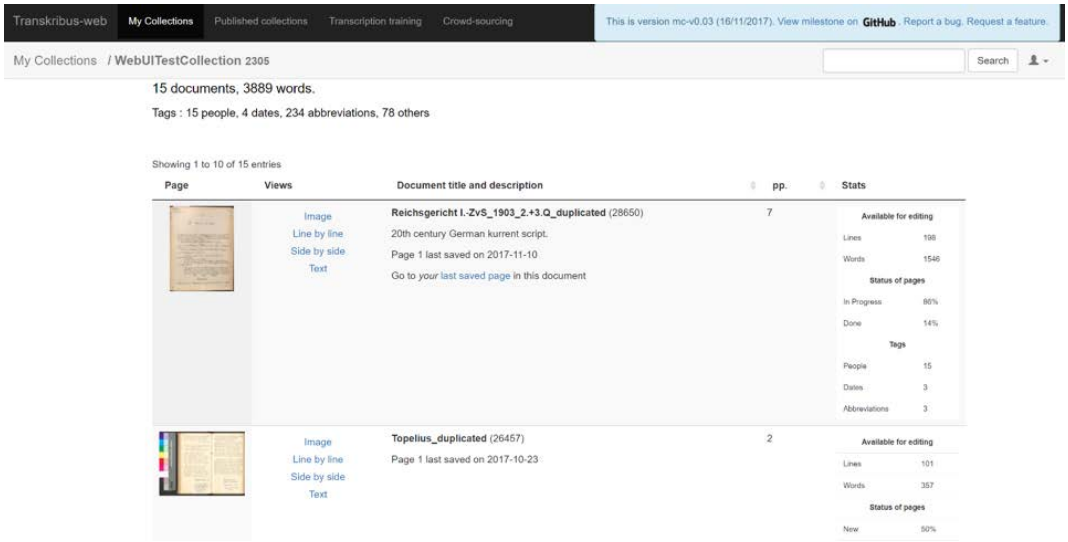

Figure 4 Collection page 2017

The overhauled version of the documents view contains not a table but so-called cards. Each card contains a thumbnail of the document and the document's status.

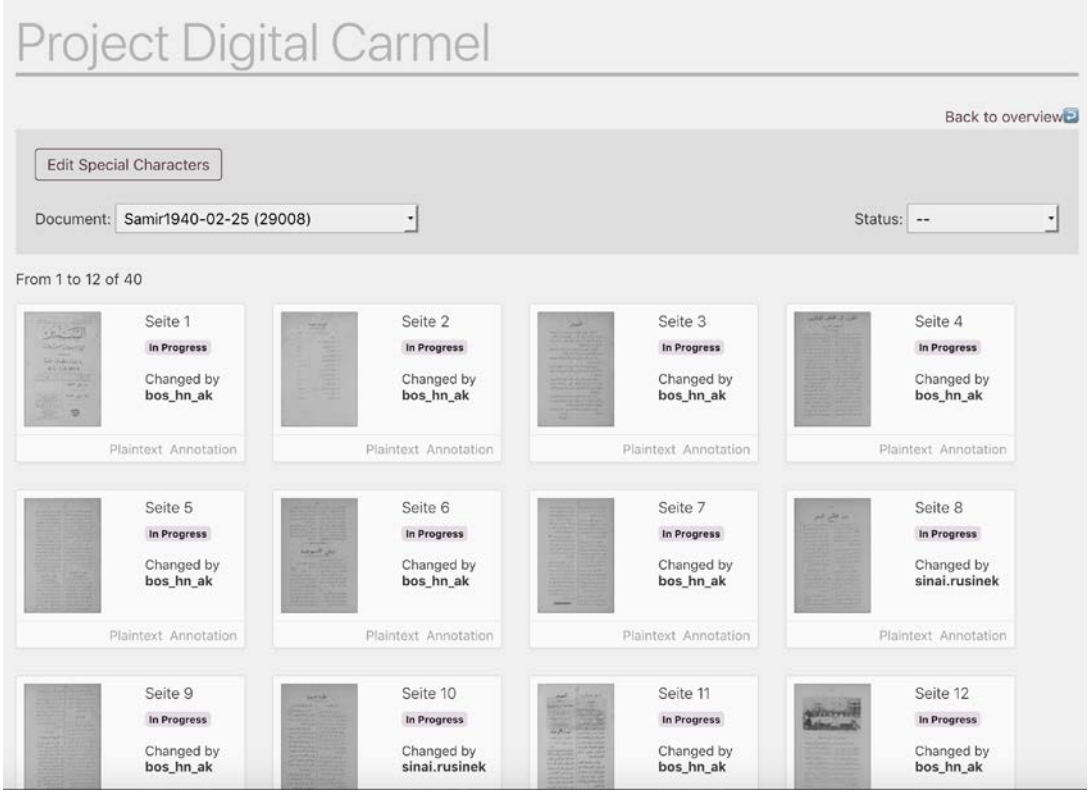

Figure 5 A collection's documents are presented on cards

2.3. Transcription

In an attempt to please as many different project partners as possible 2017's website contained four different viewing modes for the transcription pages. In this year's development, less demanded views were dropped in favour of refactoring and improving of existing ones. Instead of an image, line-by-line, side-by-side and sole text view, there now only exists a split view with image on top and lines on the bottom.

Components for display and zooming of the document image and rendering of transcript regions were combined with the text editing components to present a "plaintext" transcription interface.

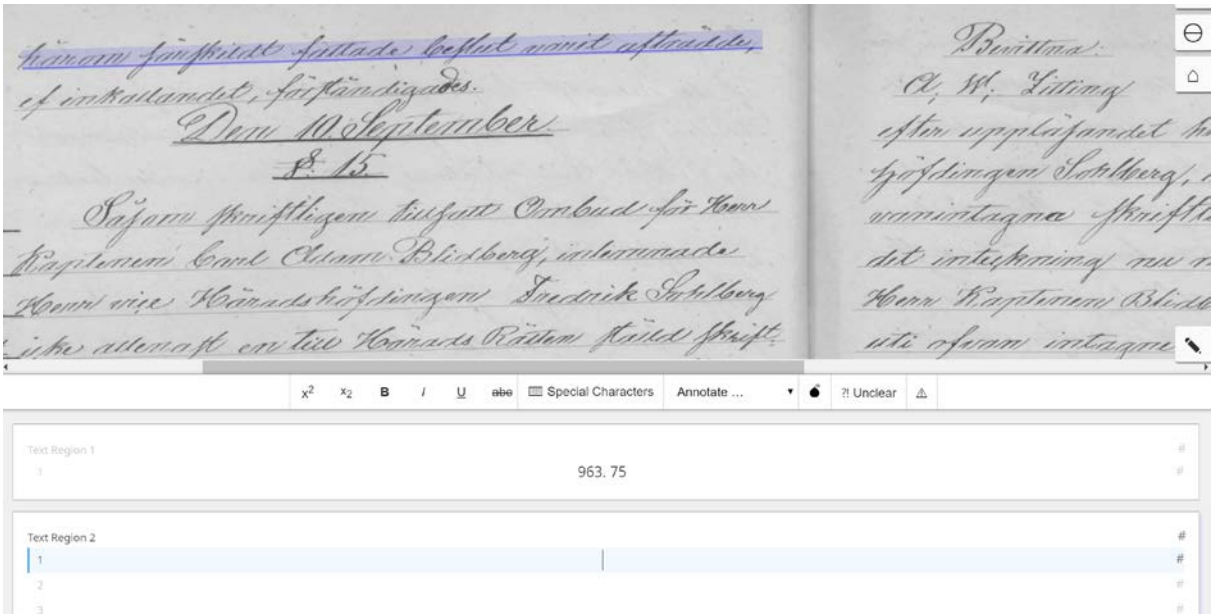

Figure 6 Plaintext transcription 2018

A tool bar between image and transcript allows the transcriber to record text styles and tag words phrases or any arbitrary section of the transcript.

2.4. Annotation

A version of the editor with more focus on annotation was made available by combining the features of the plaintext interface with some additional features to allow more information regarding tagged text to be recorded.

Figure 7 Transcription with enhanced annotation

Layout editing

Previously only transcript editing and the addition of metadata tags was allowed via the TranskribusWed UI. This meant that the layout of the document needed to be perfect before release to the crowd user. This meant either perfect automated layout analysis, or intervention b the collection owner using the Transkribus-X desktop client. Adding layout editing to

TranskribusWeb means that “near” perfect automated layout analysis can be amended by the general public.

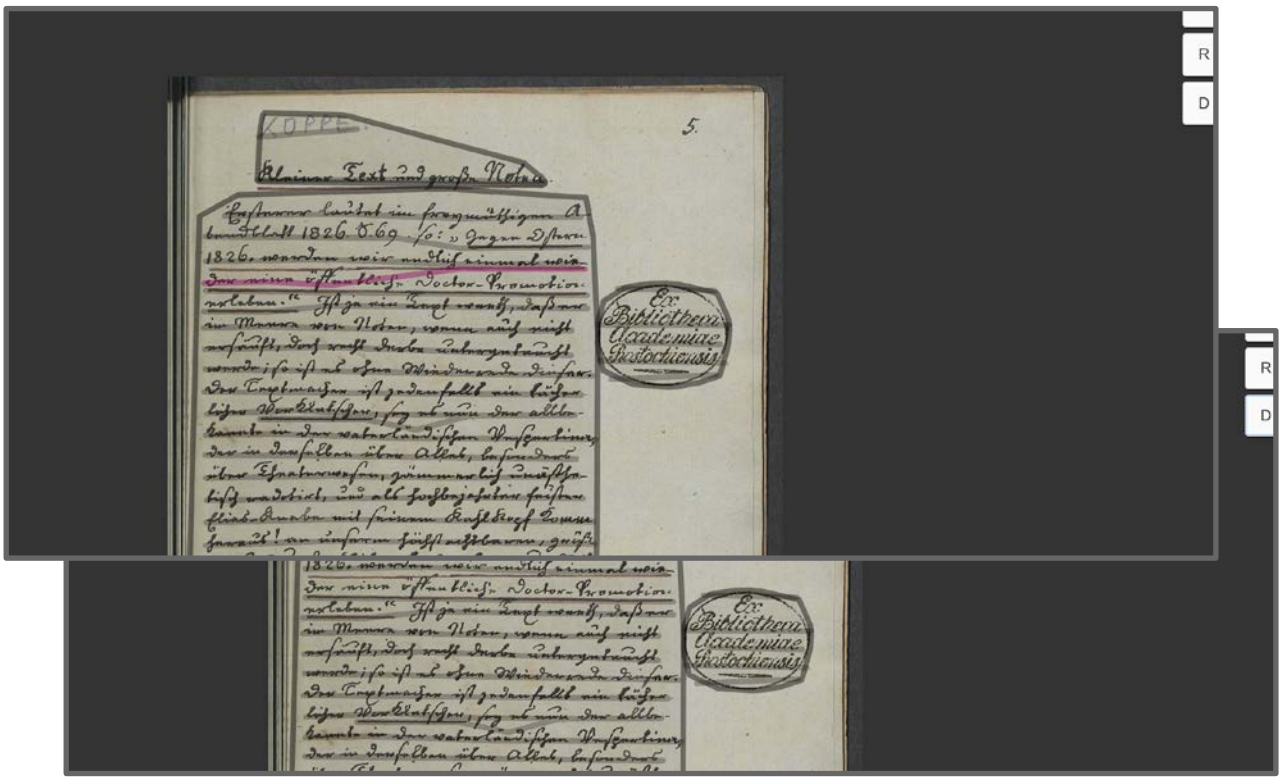

Figure 8 Before and after layout editing with TranskribusWeb

The example above shows how the layout editor can be used by the crowd to “tidy up” automatic document layout. A renegade base line is removed and a region added for the page number.

2.5. Crowdsourcing

The Transkribus service has been modified to allow collection owners to designate a collection open for public transcription. Collections with this designation will made available for transcription using TranskribusWed to any Transkribus registered users. A low-friction subscription method allows Transkribus users to join a project that has been flagged as open for crowdsourcing.

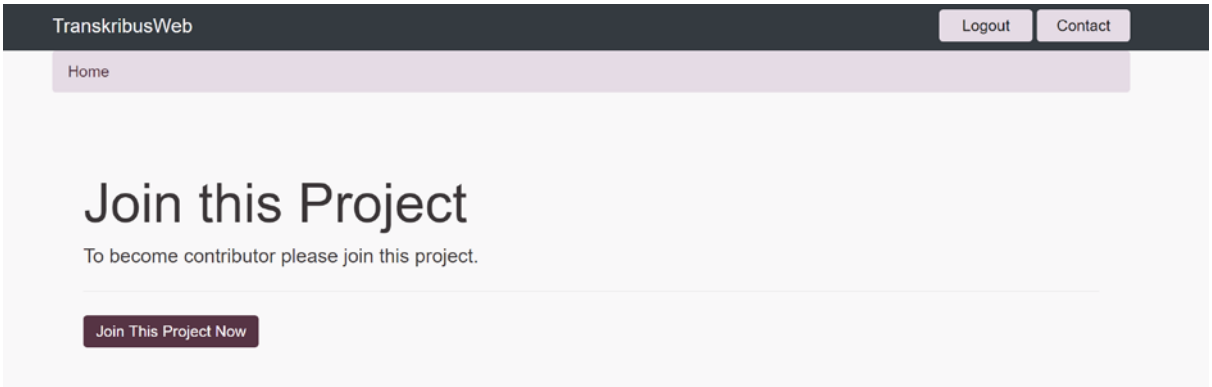

Figure 9 One click subscription for users to join a crowdsourcing project as a contributor

The contributor user type allows users to transcribe, save and submit for review their contributions. Other actions including access the collection with the Transkribus-X desktop client are restricted.

In the course of the project it became apparent that it would be difficult to produce a wide range of user-reward features in a “one-size-fits-all” crowdsourcing platform, especially one that could hope to displace more well established existing crowdsourcing frameworks. Concentrating on producing a suite of re-usable web tools based on the Transkribus services has allowed for greater the potential penetration of the Transkribus services within the existing landscape crowdsourcing platforms in use by organisations around Europe.

2.6. Embedding and integration

The transcription app as shown above can be be embedded into existing websites with easily inserted html and javascript code. This in turn allows for the development of plugins for specific well-used platforms.

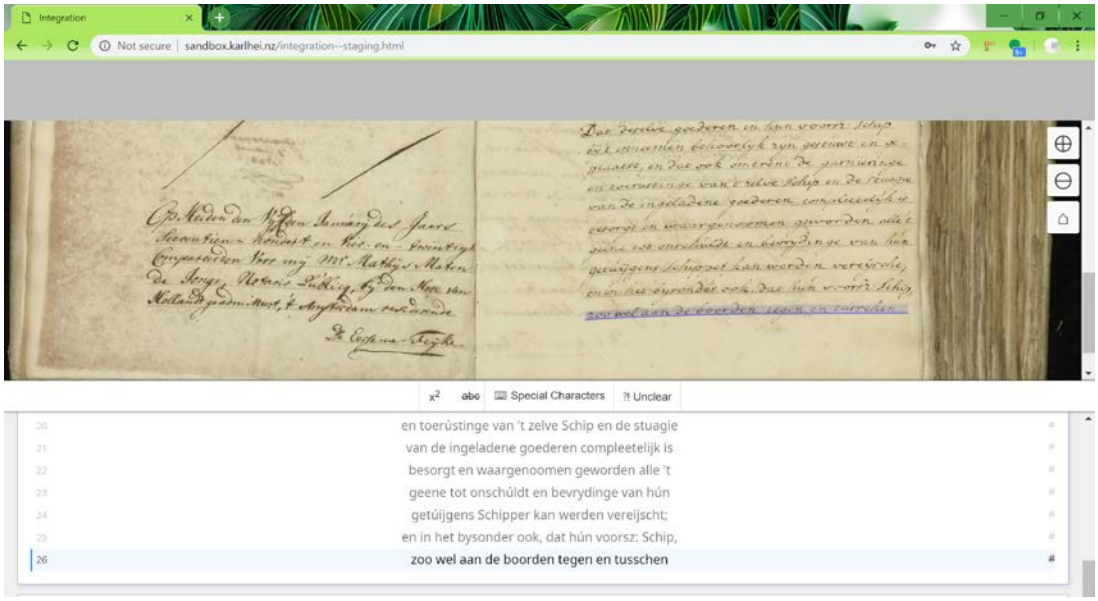

Figure 11 Example of transcription plaintext transcription interface embedded in a third party website

Any of the transcription interfaces (plaintext, annotation or layout editing) can be embedded in this way and transcripts saved to and retrieved from the Transkribus servers.

Figure 10

An example of such a plugin is a MediaWiki plugin. This has been demonstrated as a proof of concept within the Transcribe-Bentham website. This is a long running project based on the commonly used open-source MediaWiki platform.

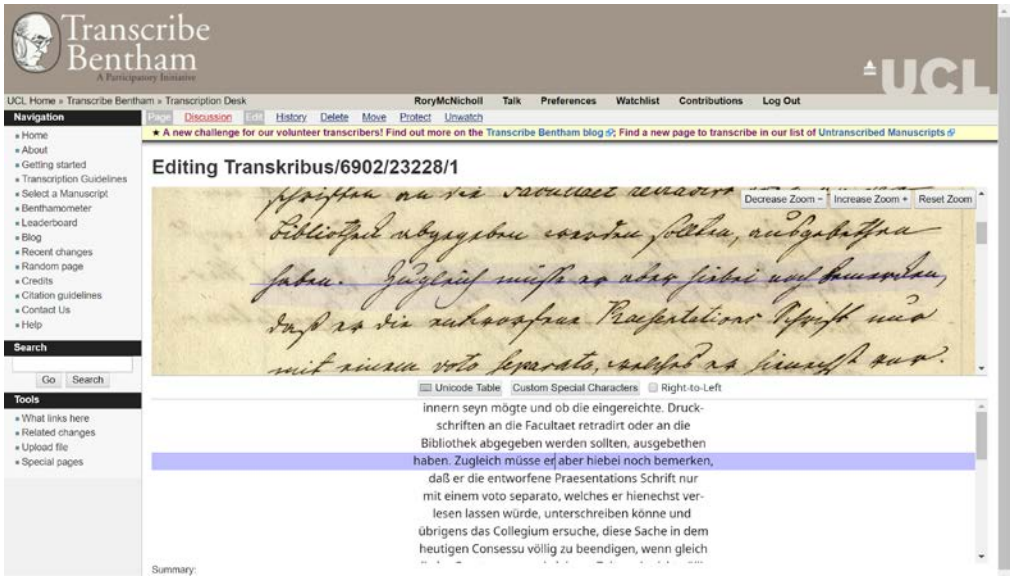

Figure 12 Transcribe Bentham with Transkribus MediaWiki plugin

Over a number of years a dedicated user community has grown up around this platform and much of the existing project and user data are stored and managed from within the platform. The integration plugin can allow the project to choose to retain their existing infrastructure whilst taking advantage of the enhanced transcription experience offered by Transkribus.

2.7. External adoption

The provision of easily integrated transcription tools has been one of the factors that has aided an uptake of the TranskribusWeb tools by organisations external to the READ project. Below are some examples of organisations that have integrated or will be integrating the transcription interface.

2.7.1. Picturae and the Amsterdam city archives

This company from the Netherlands have integrated TranskribusWeb into their existing crowdsourcing platform. The Transkribus based service went live in October 2018.

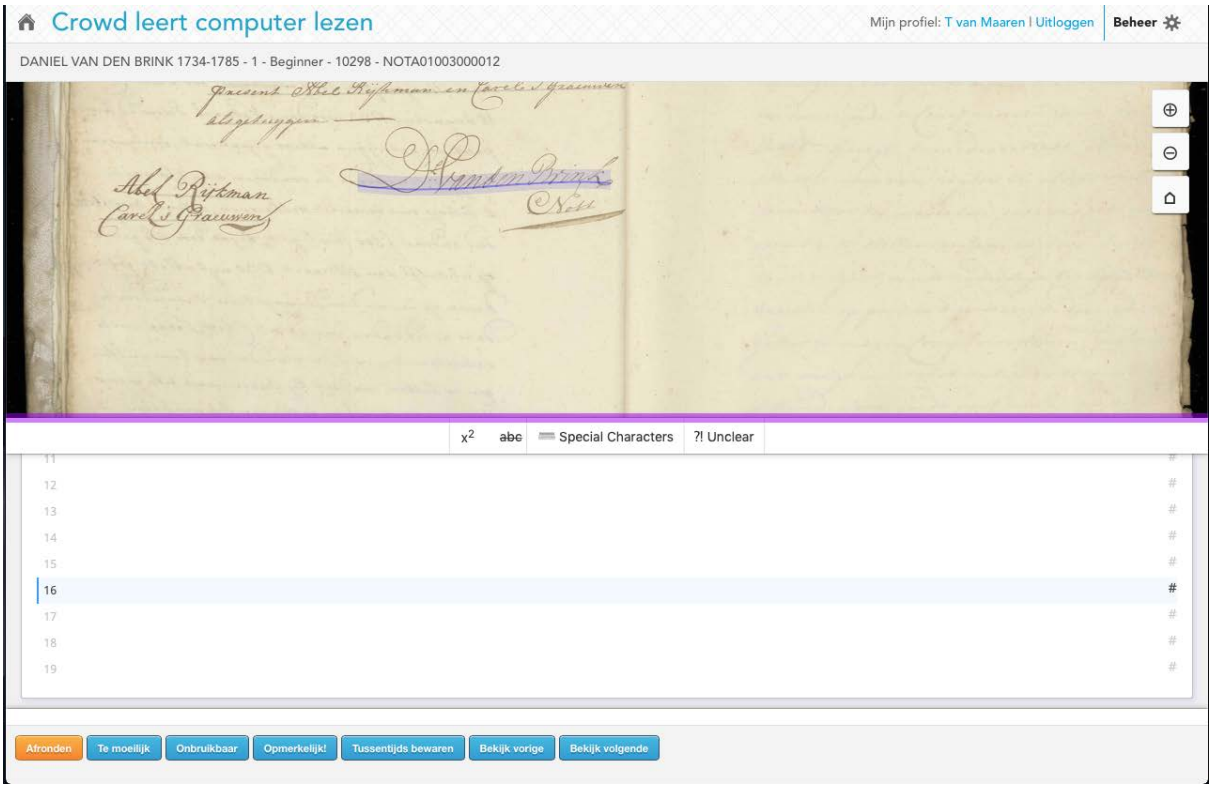

Figure 13 Screenshot from https://heritagehelpers.co.uk/ (English) https://velehanden.nl/ (Dutch) by Picturae, NL

2.7.2. NZAC Alpine journal

A digitisation project in association with the Innsbruck linguistics department with regard to material from the NZAC Alpine journal.

https://alpineclub.org.nz/nzac-alpine-journal-digitisation-project/

The Transkribus services have been used to digitise and perform content recognition of 17,500 pages. A crowd-editing project is now underway inviting volunteers to use the the TranskribusWeb interface to correct errors. Now open for contributions at https://Transkribus.eu/r/read/projects/13547/

2.7.3. Trug und Schein: A Correspondence

A project from the University Missouri, Kansas City with an associated event. This is planned to transcribe the handwritten letters from the second world war

https://info.umkc.edu/dfam/en/

2.8. Usage statistics

The impact of these external projects is already apparent from the recent data on international visitor by country.

Figure 14 Visitor sessions on Transkribus.eu/r/read/ by country, 7 days to 6/12/2018

2.9. Conclusion and next steps

During 2018 the web-based transcription tools have been made production ready as evidenced by their ongoing use in a number of "live" projects. During the last phase of the project broader dissemination and use of these tools can be sought. A good place to start this may be the 17,000 current Transkribus subscribers who until this point have been concentrating on the desktop application, but may in some cases benefit from a more portable version of the Transkribus transcription functions.

An upcoming enhancement to the integration-by-embedding of the TranskribusWeb tools is a configurable workspace. This will allow third parties that use the tools to configure the features they require for their particular crowdsourcing project. Different feature configurations allow project owners to tailor the presentation of the tools for the needs of their project, The process is currently developer led but very simple, suggesting that this may be possible to add as an administrative function to add to the TranskribusWeb site in the future.

Much of the feedback during 2018 has suggested there is an appetite for more administrative functions within TranskribusWeb. The group have been cautious not to overload the interface with features, as its main purpose is to encourage the non-experts to try their hand at editing the Transkribus generated transcripts. It is also the case that administrative features are already available via the desktop application and the number of users (collection owners, etc) requiring such features is relatively small in comparison to the potential "crowd". It may be wise to add some administrative features to TranskribusWeb in the future, especially if there is a well-defined benefit

in doing so in addition to the desktop client. However keeping in touch with the core purpose of this tranche of the Transkribus services means that priorities should lie with presenting a clean and intuitive interface that can be easily incorporated into third party platforms.

Many existing transcription frameworks (including Zooniverse and Picturae) tend to deliver material to the crowd transcriber in a random manner. The draw being the act of transcription itself, rather than what the material is about. A distinction with potential to exploit is that Transkribus presents the material as a whole, allowing the transcriber to understand the context of the page, document and collection. This technique of access to the collection as a whole has been used to great effect by projects like transcribe-Bentham that use a wiki to structure the documents as one might find them in the archive. Such projects can cultivate a loyal core of transcribers with an in depth knowledge of the material that can improve their performance as transcribers.

3. learn.transkribus.eu as an example of a mobile crowd-sourcing tool

3.1. General

Based on the technology developed in 2018 a new version of the eLearning application – *learn.transkribus.eu* has been developed. Due to the fact that there were several generic tools available which were developed in the course of the TranskribusWeb application this development required much less work than originally expected. The site is already online and will be announced to the broader public in early 2019.

3.2. learn.transkribus.eu

The concept for the eLearning application *learn.transkribus.eu* was kept but adapted according to user needs and user requests gathered during 2018.

Main issues raised by the users were to have a better overview on available documents, to be able to filter them, to get a more detailed description of the documents but also to have a more comprehensive feedback of the system (cf. fig. 15) . The layout as well as the look-and-feel of the site was changed significantly resulting in a much better convenience than before. In order to increase the convenience a user is now not only able to view a single line but the whole image can be zoomed and moved easily.

But the main difference to the former version is a simplification of the main features, as there are the “Study” and the “Test” mode.

In the “Study” mode users see just one button and pressing this button will reveal the correct transcription of one word in the line. For evaluation purposes just the number of words “studied” are shown to the user. This is a good measure of his general diligence, but does not make any prediction on the ability of the user to actually provide the correct transcription. This was actually the issue with the “Got it” functionality in the former version which left it open to the user to “cheat” or to provide correct input.

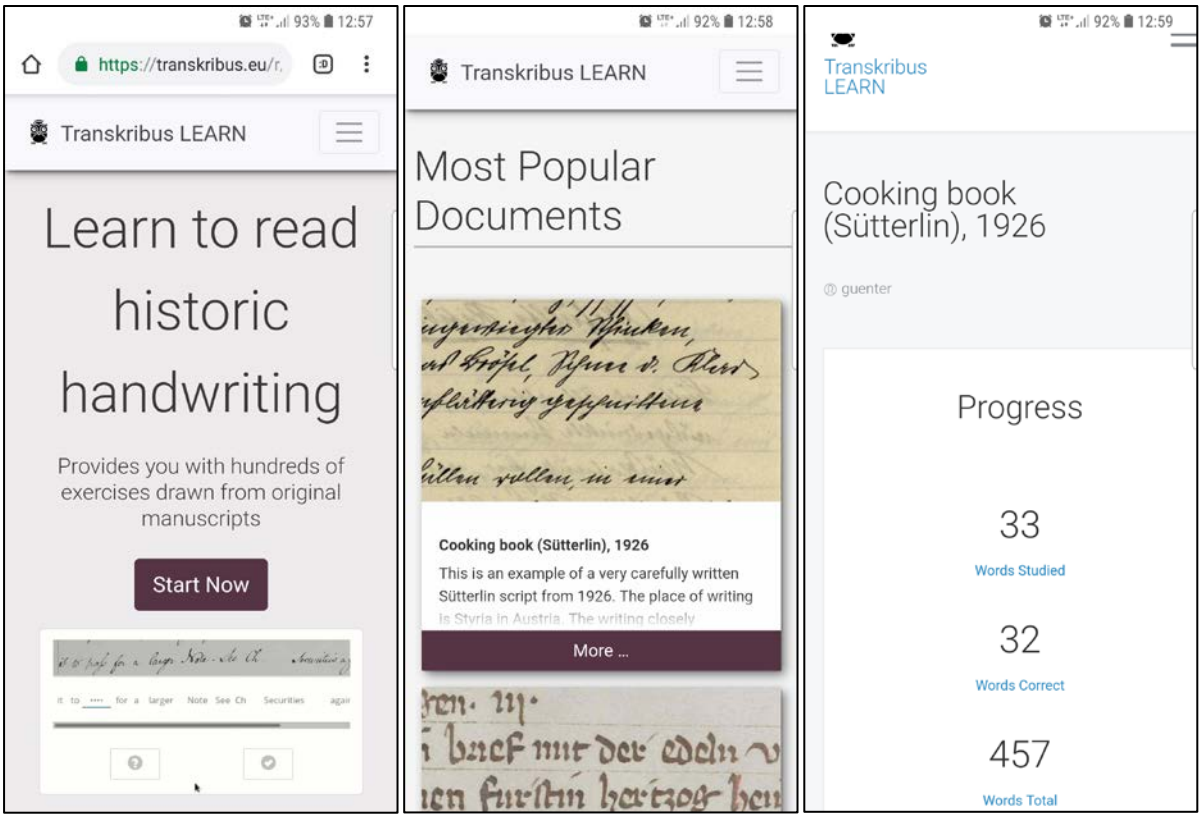

Figure 15 Extended ways to filter and view documents

In the "Test" mode the user is forced to enter the correct transcription. However and in contrast to the previous version he gets immediately a general feedback on the correctness of his input: either the next button is displayed in green or in red – together with the correct word.

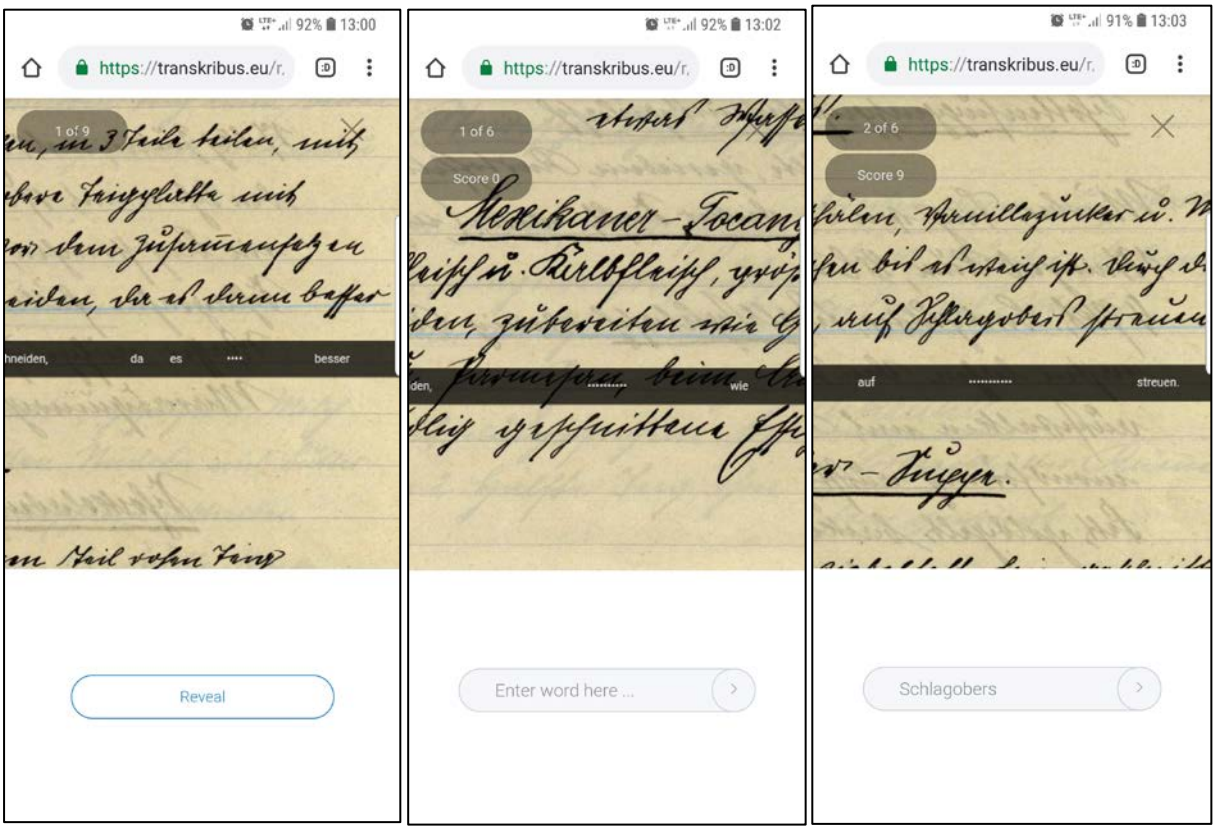

Figure 16 Study mode: Reveal – Test mode: Enter word here

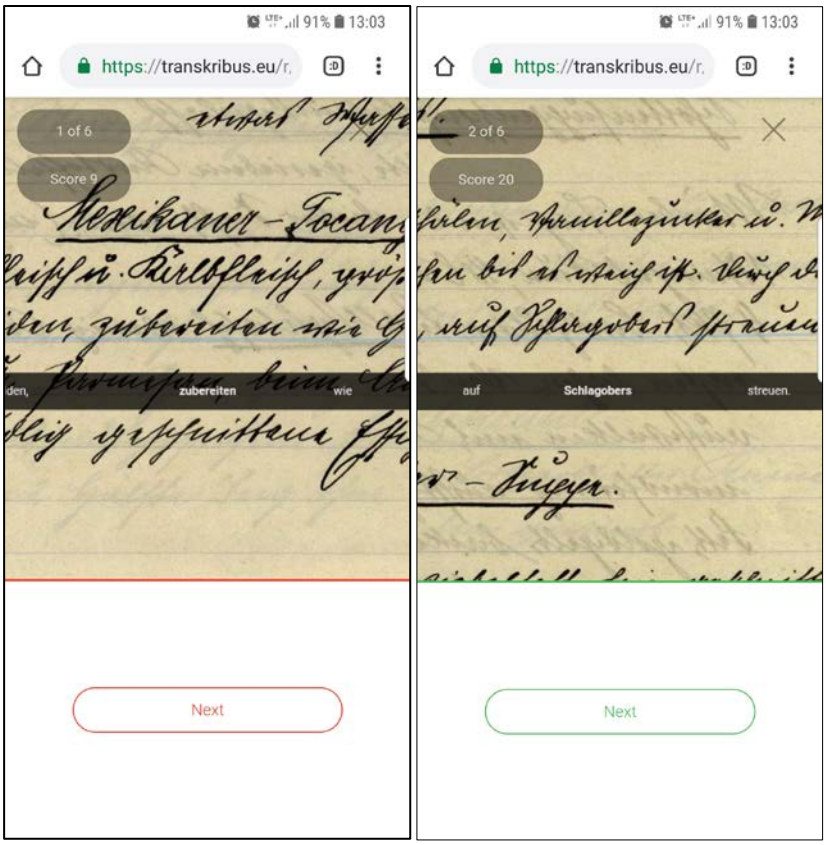

Figure 17 Test mode: immediate feedback and summary

After a “Test” session a user gets also feedback on the general score of his input. Moreover he also is able to review his input and to compare it with the correct transcription.

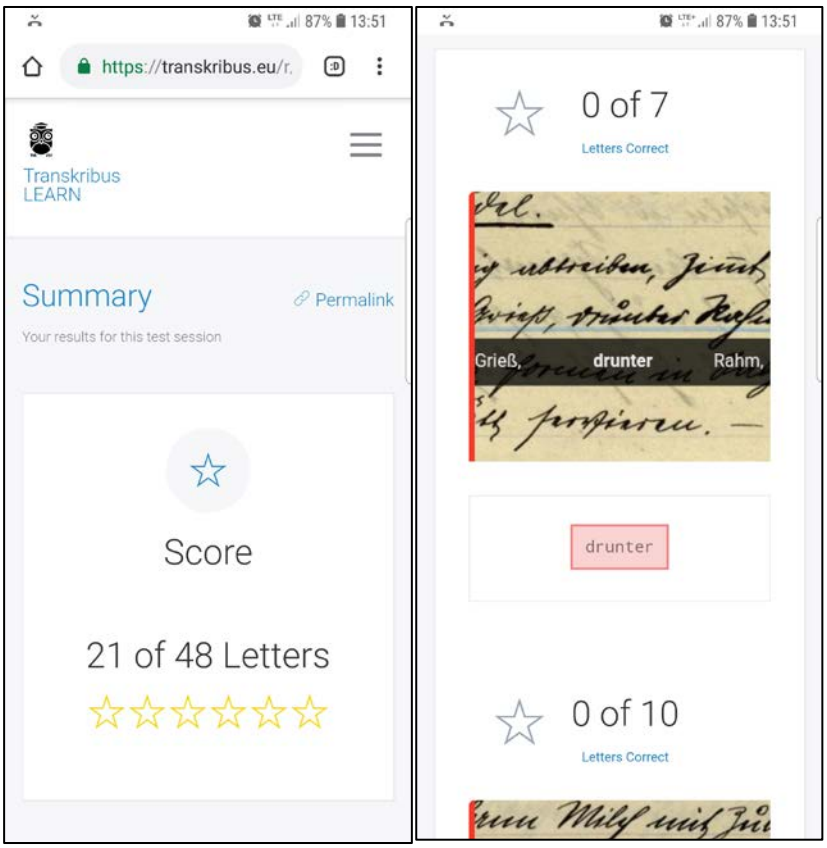

Figure 18 Feedback after test session

3.3. Outlook

With learn.transkribus.eu and the TranskribusWeb interfaces a number of powerful modules are available which can be used in many different ways. We are convinced that more and more people will explore archival collections not only with a desktop applications, but with their smartphone – responsive design is therefore a “must” for future web applications.

In our case we will use the technology developed in 2018 also for setting up a web-interface for the Keyword Spotting (KWS) application which will be developed in 2019 for the National Archives Finland.