Recognition and Enrichment of Archival Documents

D8.8
Layout analysis and crowdsourcing

Maria Kallio

Distribution:

http://read.transkribus.eu/

READ
H2020 Project 674943

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

Project ref no.	H2020 674943
Project acronym	**READ**

Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	1 January 2016 / 42 Months
Distribution	Public
Contractual date of delivery	
Actual date of delivery	
Date of last update	19 December 2017
Deliverable number	D8.8
Deliverable title	Layout analysis and crowdsourcing
Type	Report
Status & version	
Contributing WP(s)	WP8 Large Scale Demonstrators
Responsible beneficiary	NAF
Other contributors	
Internal reviewers	Günter Mühlberger (UIBK), Tobias Hodel (StAZH), Stefan Fiel (CVL), Ioannis Pratikakis (DUTH), Eva Maria Lang (ABP)
Author(s)	Maria Kallio
EC project officer	Martin Majek
Keywords	Reference data, ground truth, handwritten text recognition

Table of Contents

Executive Summary

The main objective of this task is to process large amounts of data from the digitised collections of the National Archives of Finland and involve a great number of users who are willing to contribute to the enhancement of the digitised documents. This serves as a growing basis for building a large-scale demonstrator with the collections of the National Archives Finland.

1. Introduction

Based on the experience gained in the first year, it was known that one of the biggest challenges for the National Archives was efficient production of the ground truth. The main reason for this is that there are hardly any existing transcriptions made from the collections of the National Archives. That is why the plan was to focus only on a limited amount of material and to get one applicable HTR model. To achieve this goal, the aim was use the Transkribus web interface, which would have made the production of ground truth with the help of volunteers much easier. As the development of the web interface was still delayed, the National Archives decided to launch small-scale test projects with targeted user groups using the Transkribus expert tool.

2. Ground truth production

2.1 Renovated court books – notification records

One of the largest collections in the National Archives of Finland is the series of renovated court books. They are transcribed court records from different jurisdictions in Finland that were provided for the Court of Appeal by the lower courts. The series of renovated court records start from 1623, when the Court of Appeal was established, and is continued until the 1970s. Due to the enormous amount of material, it was decided that, for the test project, it had to be limited somehow. Since the notification records from the 19th century were digitised only recently from the original manuscripts, it was quite easy to select them as test material. Their advantage is also the clarity of handwriting and the fact that usually there is only one scribe per manuscript. The notification records alone contain more than 600,000 images, so the benefits of handwritten text recognition are quite obvious.

It was decided to produce the transcripts in cooperation with the Genealogical Society of Finland. The promotion of the test project was carried out by the society, and around 30 volunteers took part in the Transkribus workshop, which was organised by the National Archives. The actual crowdsourcing started in mid-June, and by early August, volunteer genealogists had produced

more than 200 pages of ground truth. The first HTR model was trained already in August, based on 75,000 words of training data. The result was very good, with an average character error rate (CER) of only 12 per cent.
The genealogists have continued to work on the material throughout the autumn, and already around 400 pages have been transcribed. The goal is to collect as much training data as possible in order to reach a CER of 5–10 per cent.

2.2 War diaries

The second group of material selected for ground truth production was Finnish war diaries. The whole collection at the National Archives includes over 26,000 diaries from the time of the Second World War. The collection is very popular among public users, due to the enormous interest in Finnish military history. Despite being digitised, the use of the collection is cumbersome, as a limited amount of metadata is available. In practice, researchers are forced to browse the diaries one by one, since the only information available at the moment is the name of a division or a troop.

Again, the idea was to produce transcriptions with the help of volunteers, in co-operation with the Finnish Military History Society and the Finnish Military History Forum on Facebook. The Society and the forum have over 9,000 members altogether. However, the lack of the Transkribus web interface led to the decision to postpone the launch of the collaborative project.

In any case, it was necessary to test the suitability of the material for handwritten text recognition, especially because the material has been digitised from microfilms. During the spring of 2017, ten Finnish war diaries were segmented and transcribed by employees of the National Archives and a couple of volunteers. The amount of ground truth was around 70,000 words, and the CER of the first HTR model was 22 per cent. More training data was provided with the help of university students, and another model based on 144,000 words was trained in the autumn. The CER dropped to 17 per cent, but more ground truth is still needed in order to train an applicable HTR model with CER between 5-12%.

2.3 Estate inventory deeds of Finnish nobility

The third crowdsourcing test project was carried out as part of teaching co-operation with the History department at the University of Turku. The National archives organised a course called “Handwritten Text Recognition and Historical Research”, where students were introduced to using Transkribus in their own research. The goal was not only to present the opportunities offered by Transkribus for historical research, but also to produce training data for an HTR model. The material was 19th-century estate inventory deeds of Finnish nobility, and based on earlier tests, the material was thought to be sufficiently consistent with the foundation of a good HTR model. The course had a total of 20 students, and together they transcribed about 99,000 words of GT. The resulting HTR model had a precision of approximately 24 per cent CER. The reason for this rather poor result is likely to be due to the diversity of the material and, secondly, to the inconsistent quality in GT. This will be taken into account when the course will be held again in the

following academic year at the Department of Philosophy, History, Culture and Art Studies at University of Helsinki.

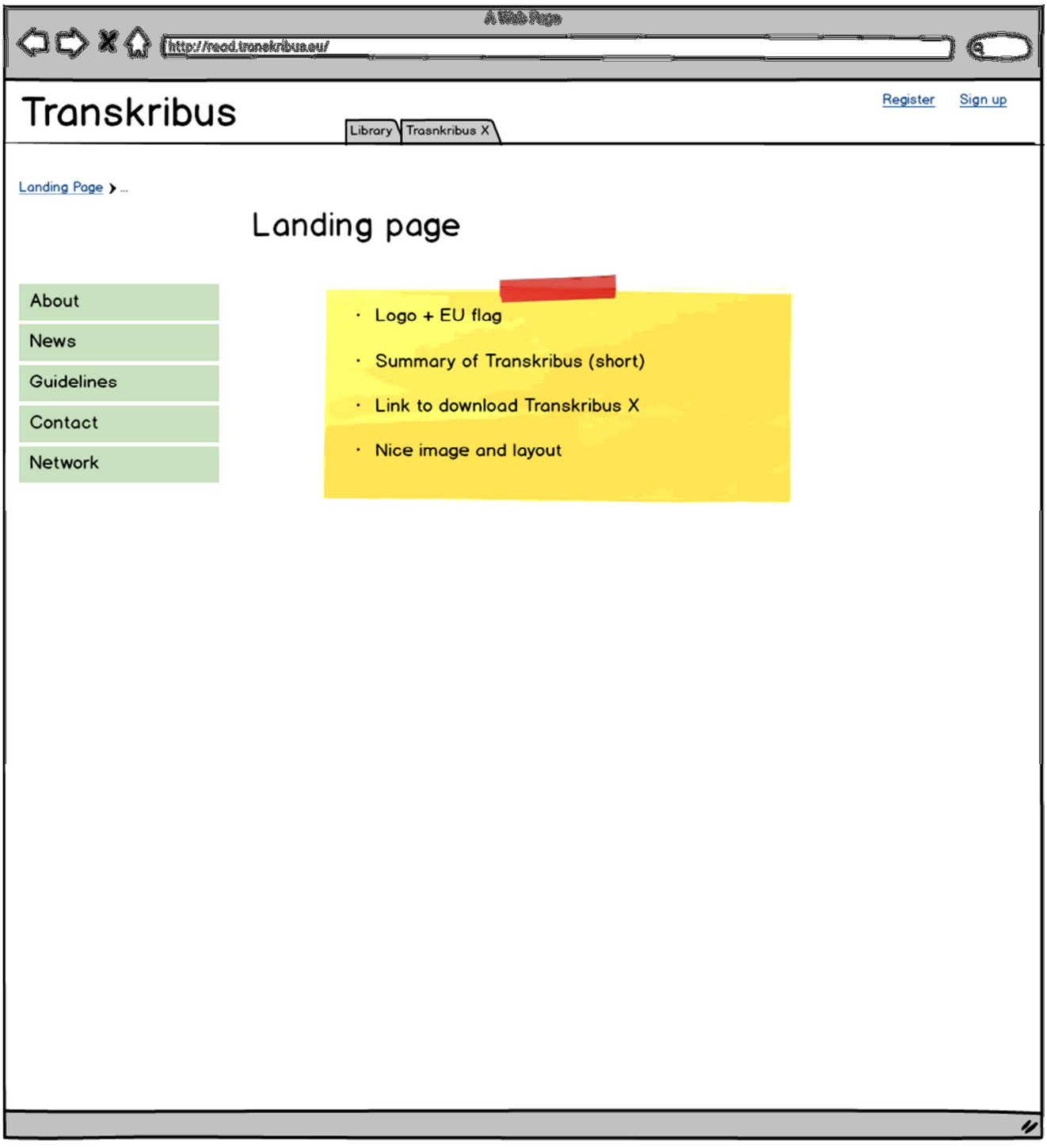
3. Transkribus web interface

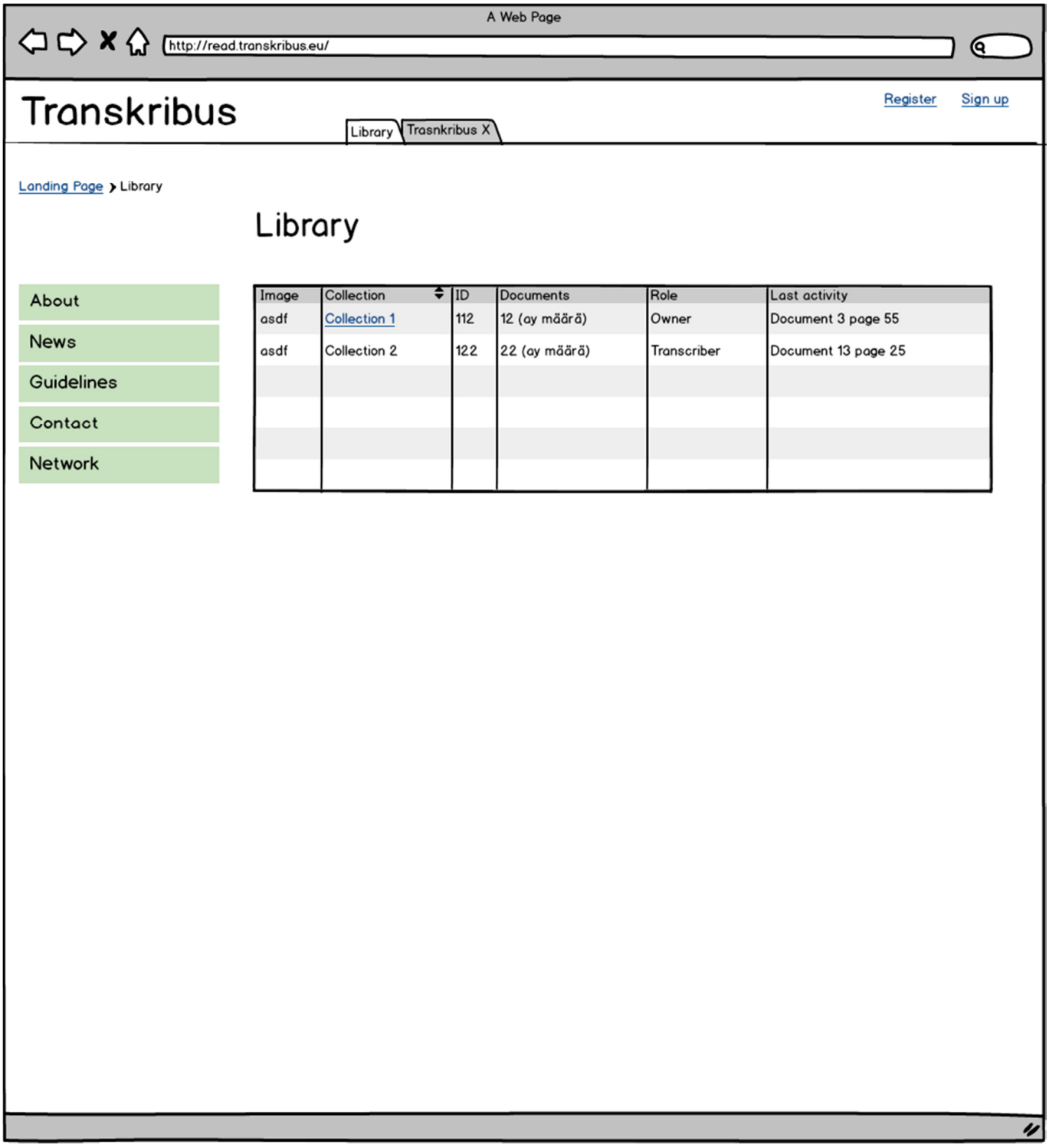
Since the main task of the National Archives within the READ project is to involve a large amount of users, the completion of the Transkribus web interface is an important priority. That is why the National Archives has been actively involved in its development (see deliverable 5.6.). In addition to the technical development, we have been trying to provide knowledge and know-how about the users and data management. One tool for this is the mock-ups and wireframes of the platform, which have been designed in co-operation with other LSD partners. The goal of the wireframes is to demonstrate the different functionalities and desired features for the developers. The designed frames are attached to this report as an appendix.

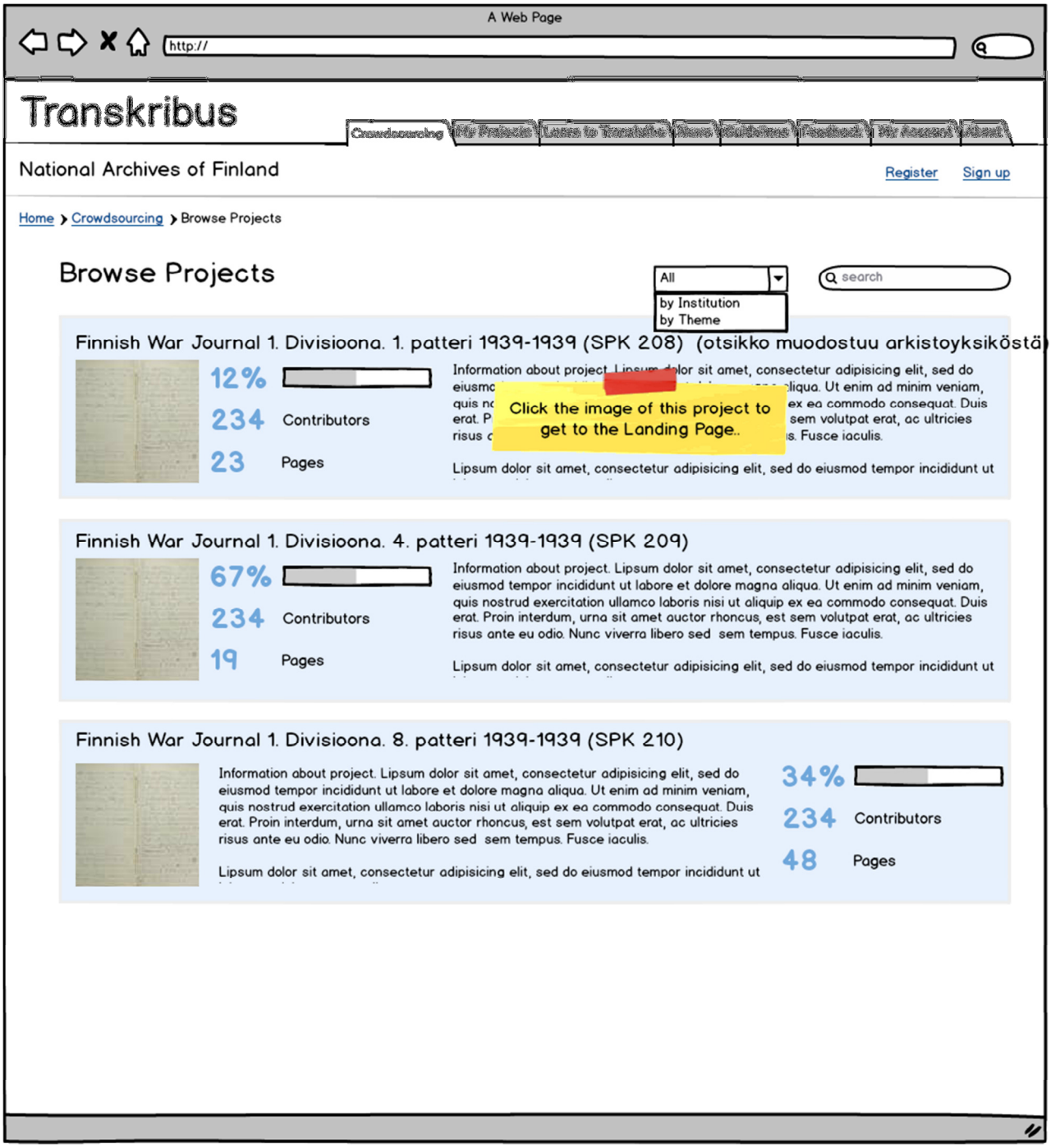
4. Summary

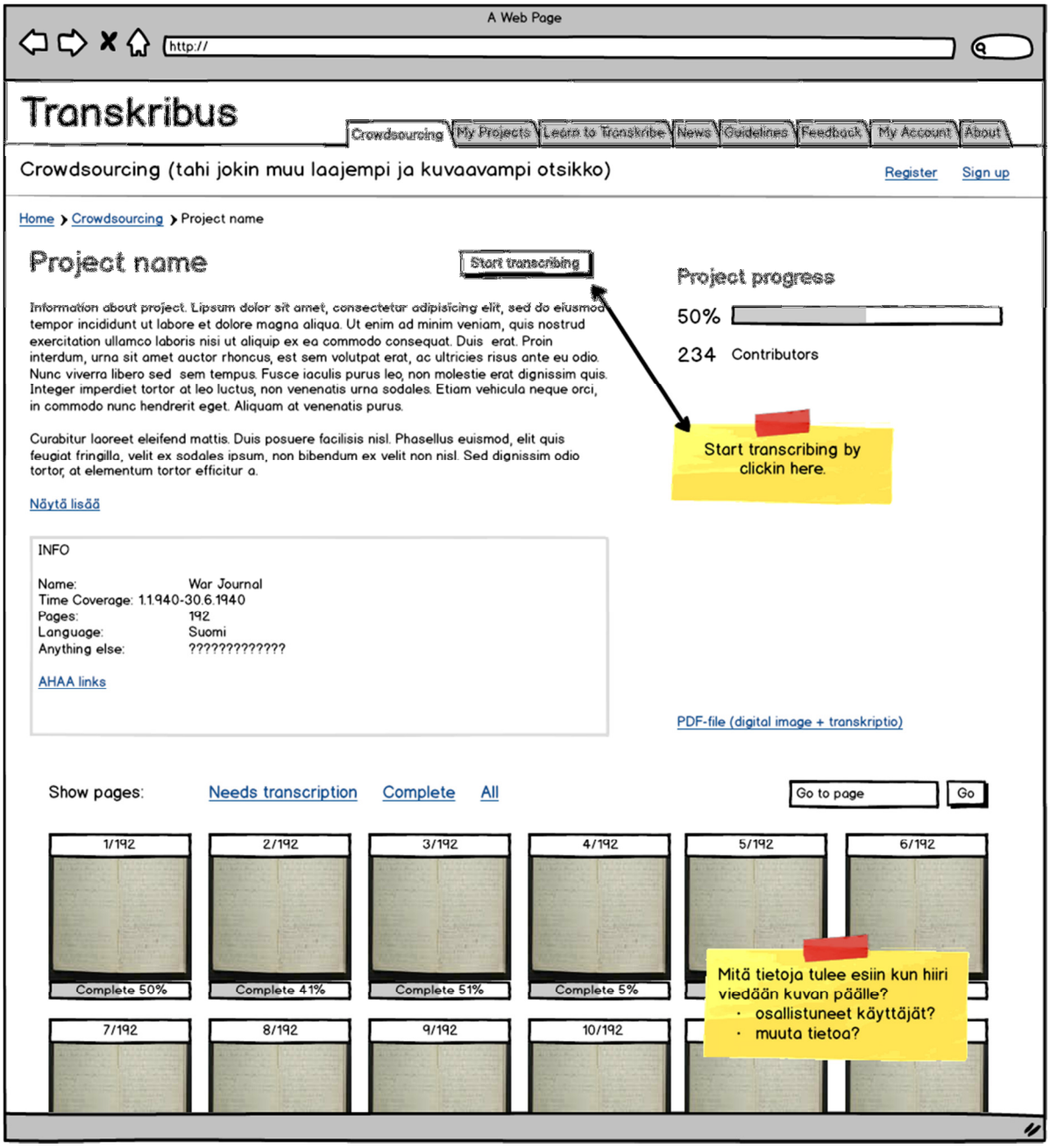
As part of the role of large-scale demonstrator, the National Archives has organised several workshops and talks about Transkribus (see deliverable 2.12.) and published videos and guides for users in Finnish. The archive has taken an active part of the Dissemination working group as well as the Archives working group of READ sharing best practices in cooperation with other LSD partners. One of the main goals at NAF was also to train our own staff to use Transkribus, and perhaps to re-think the expertise needed in the archives. The workshops that we organised for the staff were very popular and they raised important questions about the future. Over the course of two years, the National Archives has collaborated with various volunteer groups to produce 1,500 pages of training data for the handwritten text recognition, and also achieved the first promising results. During the third year of the project, the aim is to provide text-recognised collections for the users of the National Archives, launch crowdsourcing projects, and continue co-operation with Finnish universities.

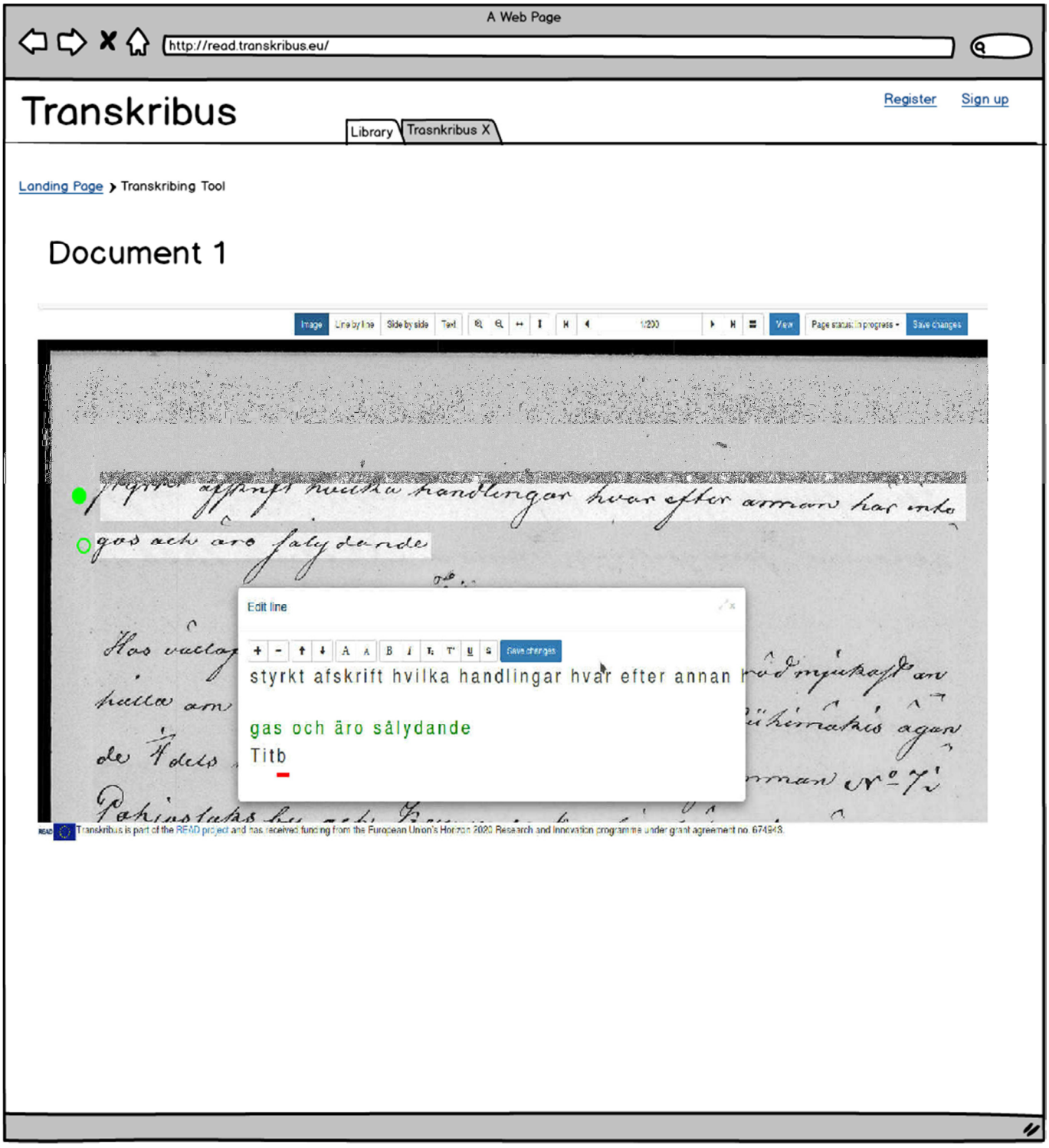
Appendices

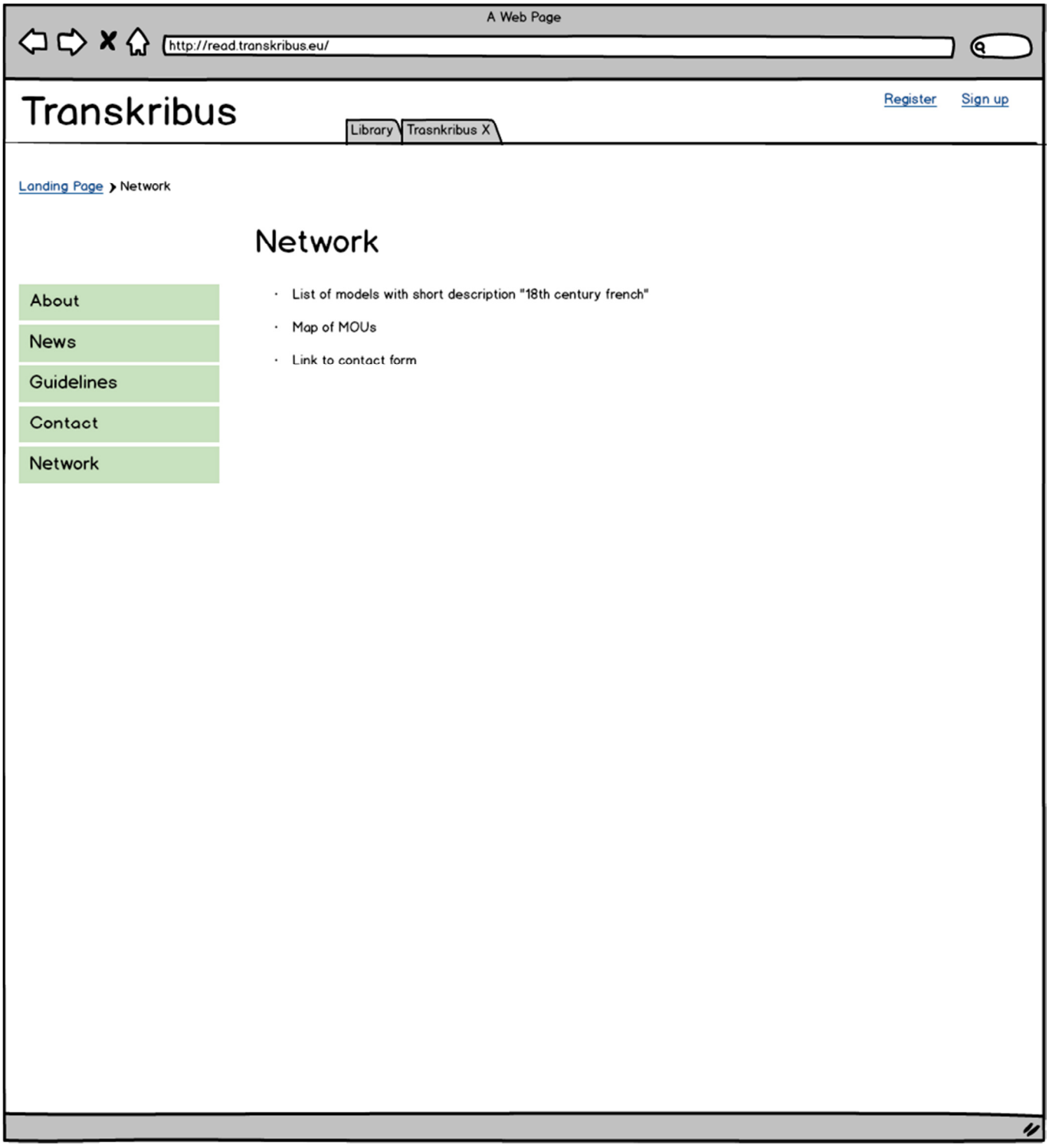

A Web Page
http://read.transkribus.eu/
Transkribus
Register
Sign up
Library
Trasnkribus X
Landing Page ›
Landing page
About
News
Guidelines
Contact
Network
Logo + EU flag
Summary of Transkribus (short)
Link to download Transkribus X
Nice image and layout


A Web Page
http://read.transkribus.eu/
Transkribus
Register
Sign up
Library
Trasnkribus X
Landing Page › Library
Library
About
News
Guidelines
Contact
Network
Image
Collection
ID
Documents
Role
Last activity
asdf
Collection 1
112
12 (ay määrä)
Owner
Document 3 page 55
asdf
Collection 2
122
22 (ay määrä)
Transcriber
Document 13 page 25


A Web Page
http://
Transkribus
Crowdsourcing
National Archives of Finland
Register
Sign up
Home › Crowdsourcing › Browse Projects
Browse Projects
All
by Institution
by Theme
search
Finnish War Journal 1. Divisioona. 1. patteri 1939-1939 (SPK 208) (otsikko muodostuu arkistoyksiköstä)
12%
234 Contributors
23 Pages
Information about project. Lipsum dolor sit amet, consectetur adipisicing elit, sed do
Click the image of this project to get to the Landing Page..
Lipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut
Finnish War Journal 1. Divisioona. 4. patteri 1939-1939 (SPK 209)
67%
234 Contributors
19 Pages
Information about project. Lipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis erat. Proin interdum, urna sit amet auctor rhoncus, est sem volutpat erat, ac ultricies risus ante eu odio. Nunc viverra libero sed sem tempus. Fusce iaculis.
Lipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut
Finnish War Journal 1. Divisioona. 8. patteri 1939-1939 (SPK 210)
Information about project. Lipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis erat. Proin interdum, urna sit amet auctor rhoncus, est sem volutpat erat, ac ultricies risus ante eu odio. Nunc viverra libero sed sem tempus. Fusce iaculis.
Lipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut
34%
234 Contributors
48 Pages


A Web Page
http://
Transkribus
Crowdsourcing
My Projects
Learn to Transkribe
News
Guidelines
Feedback
My Account
About
Crowdsourcing (tahi jokin muu laajempi ja kuvaavampi otsikko)
Register
Sign up
Home › Crowdsourcing › Project name
Project name
Start transcribing
Information about project. Lipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis erat. Proin interdum, urna sit amet auctor rhoncus, est sem volutpat erat, ac ultricies risus ante eu odio. Nunc viverra libero sed sem tempus. Fusce iaculis purus leo, non molestie erat dignissim quis. Integer imperdiet tortor at leo luctus, non venenatis urna sodales. Etiam vehicula neque orci, in commodo nunc hendrerit eget. Aliquam at venenatis purus.
Curabitur laoreet eleifend mattis. Duis posuere facilisis nisl. Phasellus euismod, elit quis feugiat fringilla, velit ex sodales ipsum, non bibendum ex velit non nisl. Sed dignissim odio tortor, at elementum tortor efficitur a.
Näytä lisää
Project progress
50%
234 Contributors
Start transcribing by clickin here.
INFO
Name: War Journal
Time Coverage: 1.1.940-30.6.1940
Pages: 192
Language: Suomi
Anything else: ?????????????
AHAA links
PDF-file (digital image + transkriptio)
Show pages:
Needs transcription
Complete
All
Go to page
Go
1/192
2/192
3/192
4/192
5/192
6/192
Complete 50%
Complete 41%
Complete 51%
Complete 5%
7/192
8/192
9/192
10/192
Mitä tietoja tulee esiin kun hiiri viedään kuvan päälle?
· osallistuneet käyttäjät?
· muuta tietoa?


A Web Page
http://read.transkribus.eu/
Transkribus
Register
Sign up
Library
Trasnkribus X
Landing Page › Transkribing Tool
Document 1
Image
Line by line
Side by side
Text
1/200
View
Page status: In progress
Save changes
Edit line
styrkt afskrift hvilka handlingar hvar efter annan
gas och äro sålydande
Titb
Transkribus is part of the READ project and has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 674943.


A Web Page
http://read.transkribus.eu/
Transkribus
Register
Sign up
Library
Trasnkribus X
Landing Page › Network
Network
About
News
Guidelines
Contact
Network
List of models with short description "18th century french"
Map of MOUs
Link to contact form