Recognition and Enrichment of Archival Documents

D8.7.
Layout analysis and crowd-sourcing

Maria Kallio, Matti Jokinen NAF

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	01 January 2016 / 42 Months
Distribution	Public
Contractual date of delivery	
Actual date of delivery	
Date of last update	15.12.2016
Deliverable number	D8.7
Deliverable title	Layout analysis and crowd-sourcing
Type	Report
Status & version	
Contributing WP(s)	WP8 Large Scale Demonstrators
Responsible beneficiary	NAF
Other contributors	INL, UCL, UPVLC
Internal reviewers	Günter Mühlberger (UIBK), Tobias Hodel (StAZH), Stefan Fiel (CVL), Ioannis Pratikakis (DUTH), Basilis Gatos (NCSR)
Author(s)	Maria Kallio, Matti Jokinen
EC project officer	Martin Majek
Keywords	Reference data, Ground Truth, Handwritten Text Recognition

Table of Contents

Executive Summary

The main objective of this task in the first year of the project was to process large amounts of data from the digitized collections of The National Archives of Finland and involve a great number of users who are willing to contribute to the enhancement of the digitized documents. This will serve as a basis for building a Large Scale Demonstrator with the collections of the National Archives Finland.

1. Introduction

The collections of The National Archives of Finland include more than 200 shelf kilometers of archival material from the 14th to the 20th century. At the moment, the digital collection contains about 45 million images and the number of digitized files is increasing by 4.3 million images per year. The online collection consists of, inter alia, court records, account books, church records, letters, maps and various kinds of registers. Despite the large number of digitized documents, the usability of the online collection is not very good. That is partly due to the fact that there is only a limited amount of metadata available and therefore it is challenging to identify suitable material. With the help of Handwritten Text Recognition, the accessibility of the online collection will improve significantly. For this reason, the main goal in this task is to process large amounts of data from the digital collections and provide full-text search of these materials to the customers of the archives.

2. Data processing

2.1 Selecting Data Sets

Due to the wide range of different types of documents in the online collection, the first data sets were formed by choosing a versatile collection of documents from different centuries. After careful selection, two data sets were assembled based on the format of the digitization. Approximately 80 % of the online collections of the National Archives of Finland are digitized from microfilm, while the number of original scans is much smaller. Because of this the first data set of 500 images contains of microfilm scans from eight different archival units and the second data set of 504 images consists of original scans from nine different units. The data sets were shared within the infrastructure of READ in March 2016.

Microfilm images

- Records of Parish Meetings of Mynämäki 1774-1774 (71 pages)
- Records of Lower Town Court of Turku 1707-1707 (71 pages)
- Letter concepts of Naantali primary school 1870-1874 (71 pages)

- Register of Letters – The Administrative Department for General Management (predecessor of The Ministry of internal affairs) 1810 (66 pages)
- Records of Helsinki Bookbinders` guild 1835-1868 (39 pages)
- Army War diaries – Continuation War – Infantry Battalion 4 (1.company) 1942-1944 (43 pages)
- Death and Burial Records of Hämeenkyrö Parish 1832-1890 (68 pages)
- Provincial accounts Turku and Pori Province Population Registers 1801–1801 (71 pages)

Original images

- The archive of the Court of Appeal in Turku – Estate inventory deeds of Finnish nobility 1867-1867 (56 pages)
- The archive of Trading house J.W. Snellman G: son – Records and deeds of the frigate Toivo 1871-1888 (56 pages)
- Nyland and Tavastehus County Administrative Board – Received Royal Letters 1722-1794 (56 pages)
- Turku and Pori County Provincial account book 1722-1722 (56 pages)
- Governor General´s chancery – List of acts 1809-1825 (56 pages)
- Memoirs of A.F.R. De La Chapelle (1785–1859) (56 pages)
- Records of Renovated Court Books of Kymi jurisdiction 1869-1870 (56 pages)
- Court book of Jurisdictional district of Lower Satakunta 1550-1552 (56 pages)
- Land Tax Register of Nyland and Tavastehus County 1682-1682 (56 pages)

In addition to the two data sets, a playground set of 1,1 million randomly selected images were delivered to UIBK servers in April 2016. These images will become available as Open Data via the ScriptNet dataset.

2.2. Availability of Data Sets

A METS (Metadata Encoding and Transmission Standard) file generator was developed for the transmission of data. METS is a standard for describing documents in a digital library using XML. The generator is designed to parse the Digital Archives of NAF and generate a METS file which can be ingested by Transkribus.

- Future plans include placing a “button” on each archival unit page in The Digital Archives to enable users to conveniently ingest archival units to Transkribus.
- Initially, the button will only be available in limited test use.
- Expected challenges include getting users to create Transkribus accounts, which are necessary for the aforementioned ingestion functionality to work.

2.3. Ground Truth production

It soon became clear that the image quality of the microfilm scans is challenging for image preprocessing and layout analysis; the data set of 500 original scans was chosen to be the first training set of NAF collections. To provide the necessary Ground Truth for the HTR from the data set, a subcontracting agreement was made with Digitexx according to the project policy. Under the agreement, Digitexx had to deliver segmentations and transcriptions from the 500 images by the end of May 2016.

Digitexx maintained the schedule and the whole package was ready in six weeks. However, the actual work began only after that, since the proofreading of the produced Ground Truth turned out to be very time consuming. It appeared that in particular the documents written in 17th and 18th century Swedish had been a real challenge to the transcribers and there were a lot of mistakes to be corrected. Also, the quality of segmentations varied a lot within the document. Clearly the best result was obtained with regular 19th century handwriting and some parts of the document were transcribed very carefully. Proofreading of the first data set was finished in November 2016. Apart from early testing in spring 2016 with the HMM engine, a new test run will be performed on basis of the RNN engine by the end of the year. This will serve as a benchmark test for the NAF collection.

The challenges mentioned above are to be taken into account and in the future, data sets will be chosen in accordance with these facts. In 2017, NAF will focus on processing the registers of renovated court records from the 19th century which form a collection of 500,000 images. These registers are widely used by researchers and therefore it would make sense to provide them also in full-text form. The aim is to utilize the new Transkribus web interface for the production of Ground Truth as well as proofreading. Smaller collections, such as estate inventory deeds of Finnish nobility are also planned to be processed (in year 2/3) with the help of students and researchers.

3. User involvement

In order to lay the groundwork for user involvement and future crowd-sourcing projects, several talks, seminars and workshops were given in Finland in 2016. Also, to find out the needs and interests related to the digital collections of the archives, a user survey was conducted for researchers and other users.

3.1. User survey

The goal of the user survey was to find out what needs researchers had related to the increasingly electronic research process. The survey consisted of eight questions that had the aim of finding out what kind of challenges digitized material poses to carrying out research, and – on the other hand – what kinds of tools researchers need to tackle these challenges. So far, computer-assisted research has primarily offered tools for analyzing printed sources. In cooperation with the READ project, the goal of the National Archives is to

also develop similar tools for studying material written by hand and thus promote the use of digitized material for research purposes.[1]

3.2. Co-operation on organization level

There has been huge national interest in the READ project. The National Archives of Finland have received several contacts from research groups, universities and memory organizations who would like to co-operate with the project. This is for several reasons very positive, but since the result of the HTR is dependent on the amount and quality of training data and NAF have only limited possibilities to produce Ground truth, it is important to collaborate with other Finnish institutions, especially since some of them have already lots of transcribed materials. After several meetings and discussions, the Institute for the languages of Finland (http://www.kotus.fi/en), the Society of Swedish Literature in Finland (http://www.sls.fi/en) and the Finnish Society of Literature (http://www.finlit.fi/en) are starting to produce training data for the READ project. Finnish Society of Literature and the Society of Swedish Literature in Finland have also signed the MoU-agreements with READ.

3.3. Seminars and workshops

The National Archives of Finland has participated on disseminating the project by organizing talks, seminars and workshops during 2016. NAF have organized two short seminars, one for stake holders and one for the employees at the National Archives, talks in national events and meetings, and several READ workshops in co-operation with the Helsinki Centre for Digital Humanities and memory organizations. Dissemination is an integral part of user involvement and groundwork for the future crowdsourcing projects.

3.4. Development of the Transkribus Web interface

A new web user interface to Transkribus is being developed to better facilitate crowd-sourced transcribing. NAF is supporting ULCC in the development of the Transkribus Web interface (WP 4).

[1] The report is attached as a supplement.

4. Supplements

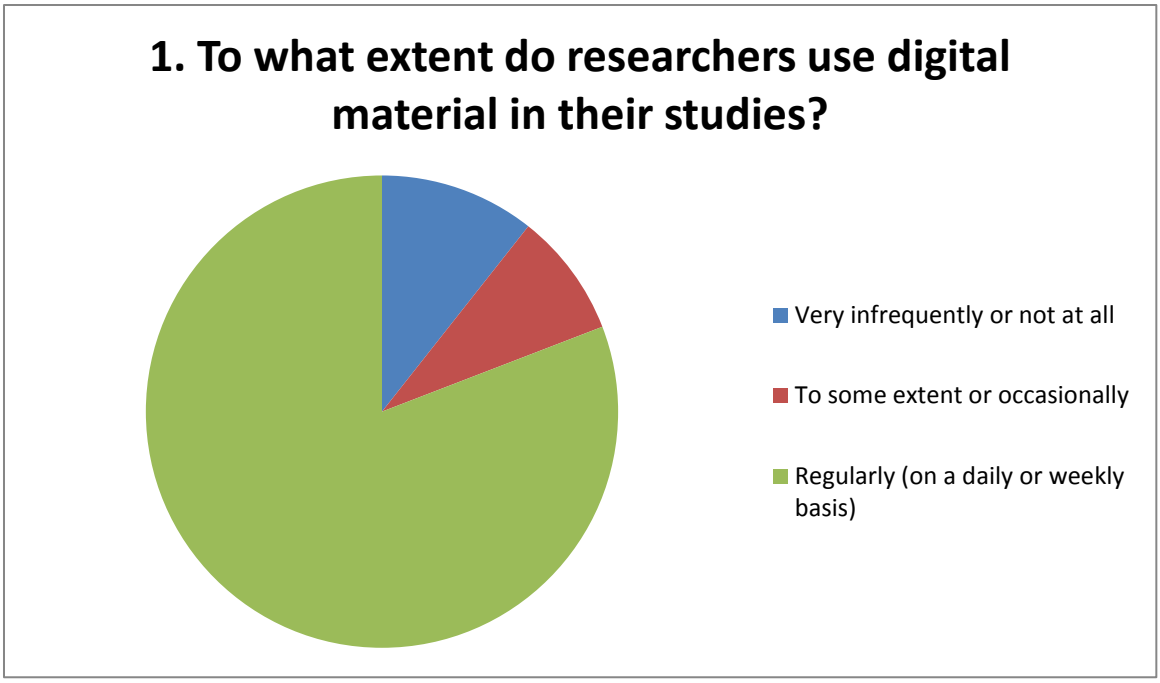
Report on the National Archives user survey regarding the development of electronic research services

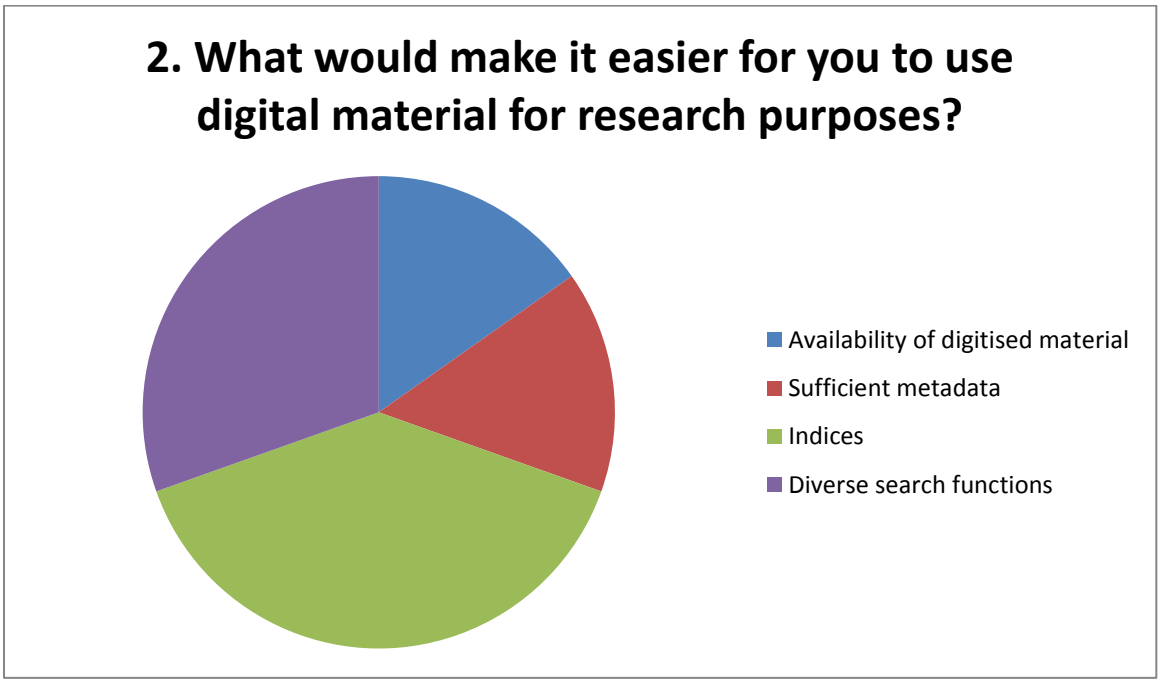
The goal of the user survey of February 2016 was to find out what needs researchers had related to the increasingly electronic research process. Researchers were also given the opportunity to share their preferences in regard to selecting material to be digitised. The survey consisted of eight questions with the aim of finding out what kind of challenges digitised material pose to performing research, and – on the other hand – what kind of tools researchers need to tackle the challenges. The survey results will be used extensively when developing the National Archives' electronic research services, particularly in the EU-funded READ (Recognition and Enrichment of Archival Documents) project. In addition to the National Archives website, survey-related communication was provided through the email lists of universities, and social media. The survey was available from February 11 to 31 March 2016.

The total number of respondents was 43. Almost one-third of them were genealogists, and 65% were postgraduate students, researchers and teachers of different disciplines. The rest were students. History is the discipline of over 70% of respondents, and the share of social science stood at 20%. There also were some respondents who are researchers of archaeology or linguistics.

Questions and answers

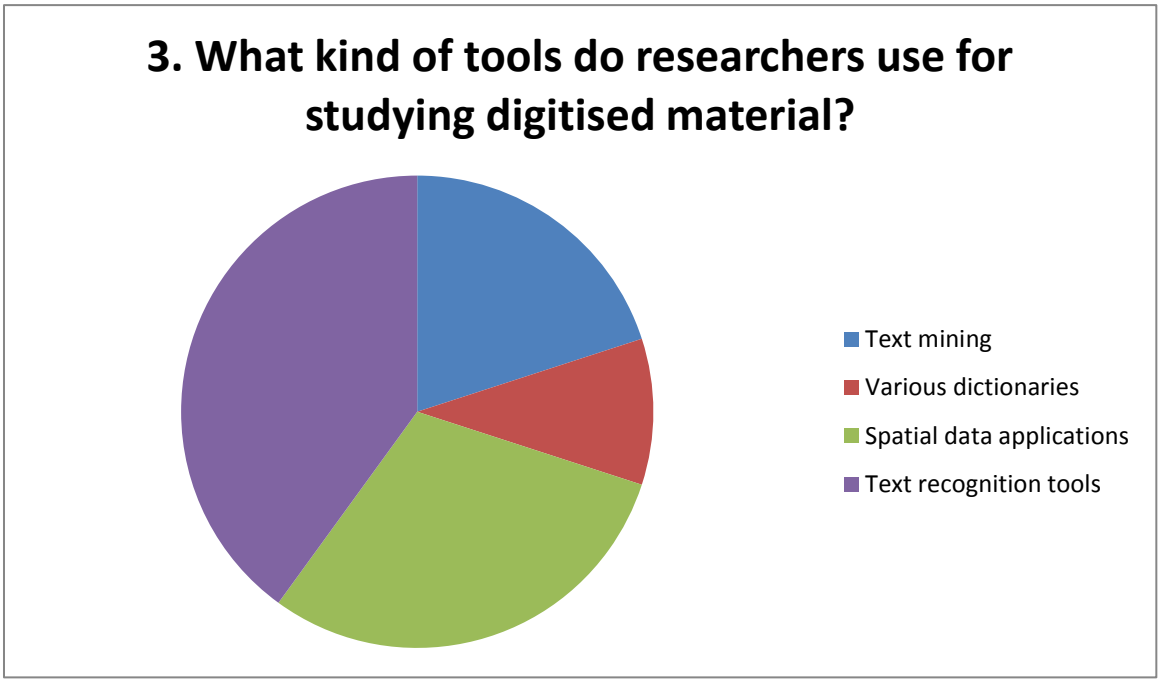
The first question in the survey dealt with the use of digitised material, and the majority of respondents (88%) reported that they use the National Archives' digital archive or other online services of digitised material almost on a daily basis. Only five respondents reported that they only use digitised material very rarely or not at all. The main reason for this was that the material they need is not available in digital format, or its use for research purposes is difficult because of the digital archive's features.





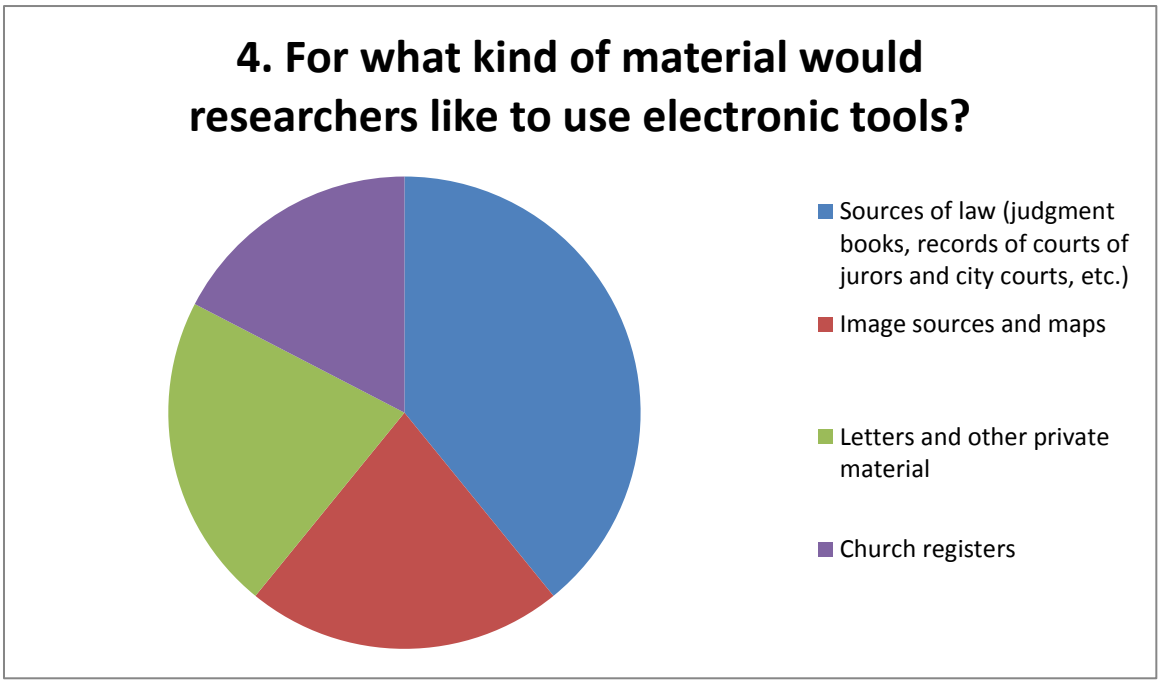
The second question dealt with things that make it easier for researchers to use digital material for research purposes. Four factors clearly stood out from the responses. Almost

half of respondents (18) specified different indices as the most useful feature. They referred to name and location name indices as well as various material-specific content indices. Almost the same number of respondents (14) would like digitised material to include diverse search functions so that it would be easier to exactly find the material they are looking for. The respondents also focus on the availability of digitised material. Many of them would like material from different eras to be more extensively made freely available. The respondents also considered sufficient metadata to make it easier for them to use digitised material for research purposes.

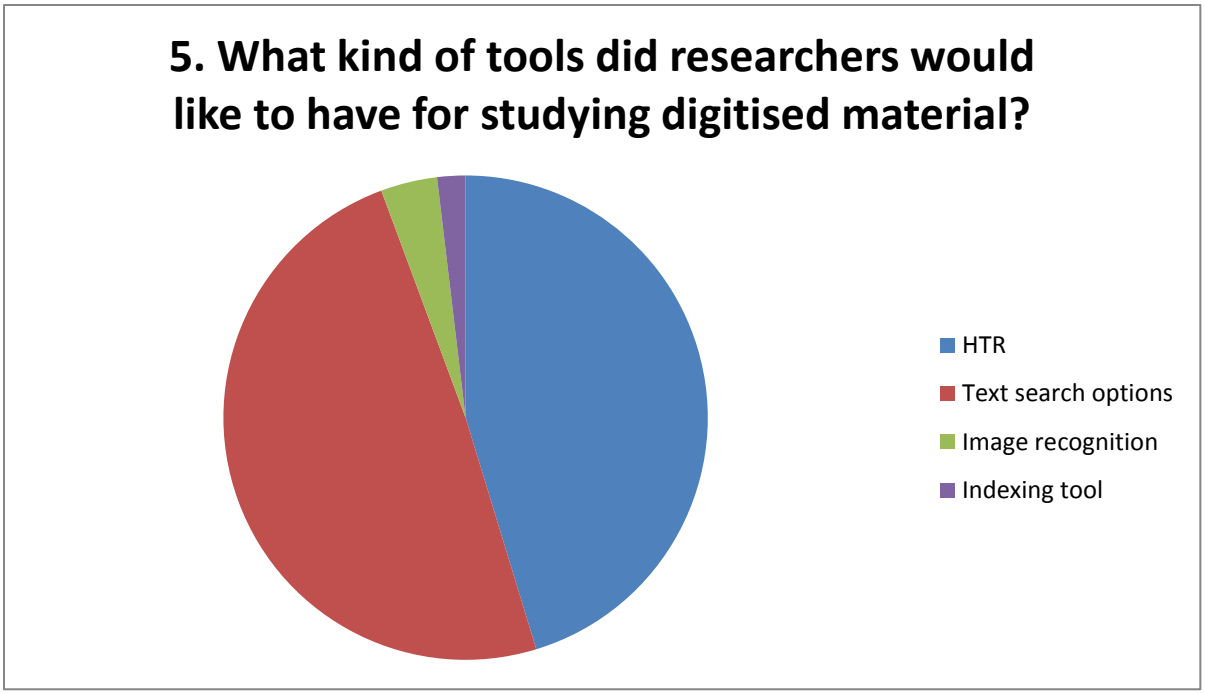


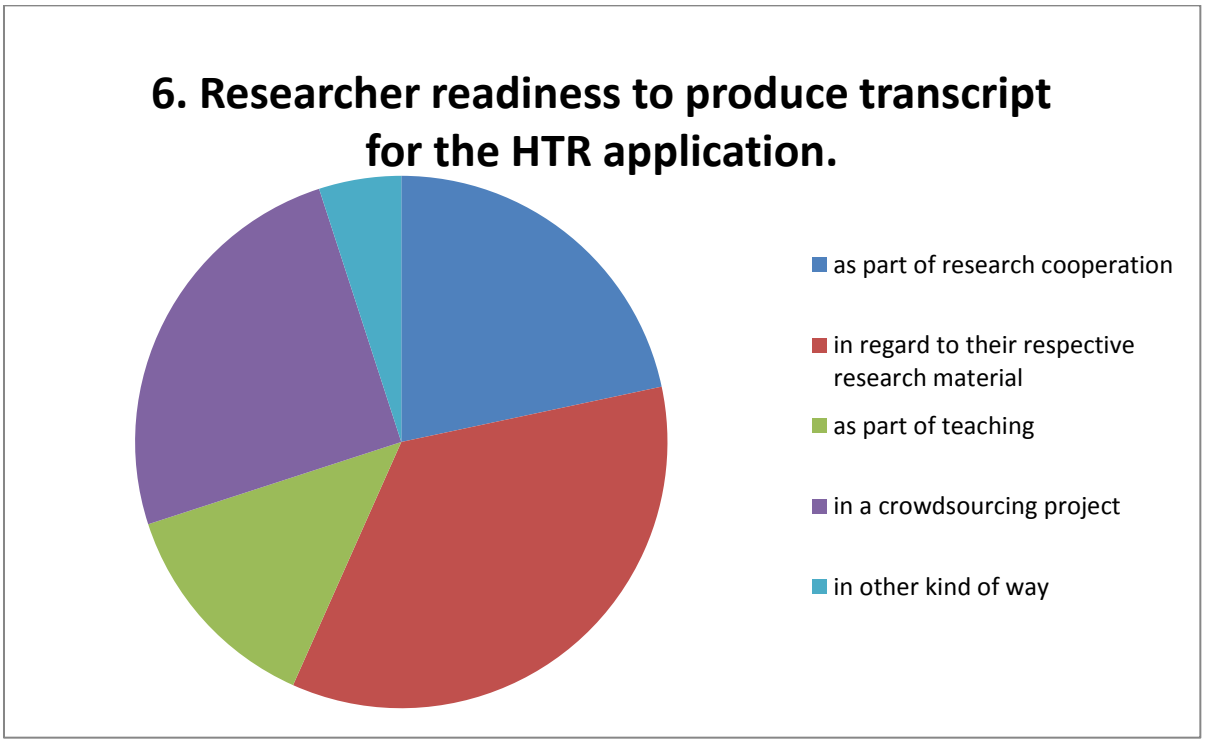
In the third question, researchers were asked to specify different kinds of tools that they currently use for studying digitised material. Actual tools in use include applications for text mining, as well as spatial data and georeference tools. However, most respondents specified tools that they would like to utilise. The tools include various text recognition tools and interactive dictionaries. Surprisingly, the responses indicate that only a few respondents had experience in using electronic research tools, or were even aware of the opportunities provided by them.

In the fourth question, respondents were asked to specify material that they would like to study with electronic tools. A total of 30% of respondents would like to have new kinds of instruments for studying digitised sources of law in particular. Many of these respondents stated that their research process is like looking for a needle in a haystack because looking up individual cases in judgment books, for example, always requires the processing of huge chunks of material. The respondents also would like to have tools for studying letters and photographs. Genealogist responses highlighted the willingness to utilise new technology for interpreting church registers.



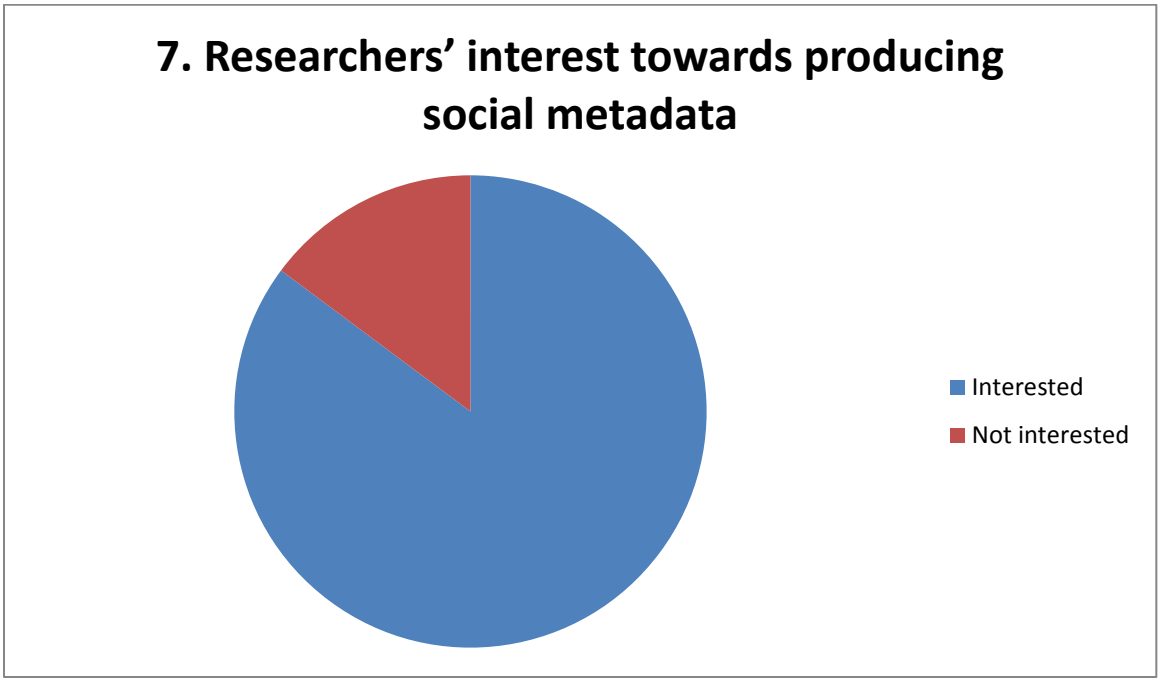
In the fifth question, researchers specified what kind of tools they would like to have for studying their material. The HTR application (i.e. automatic recognition of text written by hand) and various text search options were given to them as examples. In fact, most respondents specified one of these two; the text search option was slightly more frequently mentioned. The responses indicate major interest towards the recognition of text written by hand. However, many respondents also expressed their doubts over the functionality of the technology. Some respondents were also interested in the opportunities provided by computer-assisted image recognition.



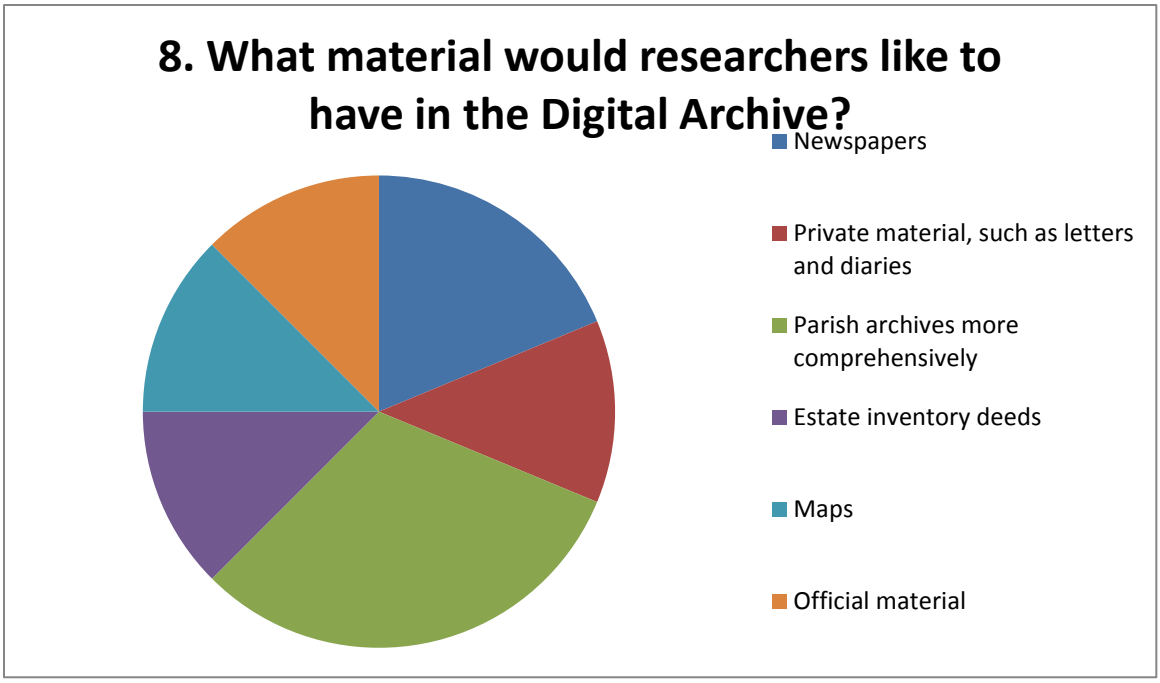


At the moment, the HTR application of READ must be separately taught to read each user's personal handwriting. Therefore, the application requires transcript to be able to recognise different styles of handwriting. In the sixth question, researchers were asked if they would

be willing to produce this background material with some preconditions. The chart above shows the ready alternatives. In principle, all respondents were willing to produce transcript, and most of them would expressly do that in regard to their respective research material. Crowdsourcing projects and research cooperation were also considered to be possible ways of producing text.



The seventh question dealt with researchers' interest towards indexing their respective research material or enriching metadata. Their approach towards producing social metadata was very positive, and many responses highlighted its advantages. However, according to respondents, the method of implementation should be as simple as possible and usability should already be focused on at the planning stage. One of the researchers with a negative view pointed out that work of this kind is not unfortunately recognised in the academic world.



In the eighth, i.e. the last question, researchers were asked to specify material that they would particularly like to be available in the Digital Archive. Surprisingly many respondents were hoping for newspaper material; libraries are primarily responsible for storing it. In fact, many respondents reported that they regularly use the digital newspaper archive of the National Library. However, the majority of all responses were related to parish archives, especially church registers. The respondents' wish is that material exempted from the limitation of use would be more quickly converted into digital format. Private material, such as diaries and letters, were the second most frequently preferred type of material to be included in the Digital Archive. The grounds for this included the modern trend in history research, where researchers are more and more interested in everyday phenomena. In addition to the above, various official material, maps and estate inventory deeds from recent history were mentioned by some respondents. Many respondents were also interested in the idea of obtaining material stored in Sweden to the Digital Archive of the National Archives.

Summary

The responses given in the user survey regarding the development of the National Archives' electronic research services help explain the challenges related to source material of history and social studies. The responses show that the amount of digitised material is extensive,

but there are only a limited number of tools for analysing it. Research themes have changed, too, and there is an increasing interest towards studying phenomena related to the daily lives of people or to recent history. New kinds of approaches and opportunities provided by new technology set new challenges to the entire research process. So far, computer-assisted research has primarily offered tools for analysing printed sources. In cooperation with the READ project, the goal of the National Archives is to also develop similar tools for studying material written by hand, and thus promote the use of digitised material for research purposes. The intention is to integrate new digital research tools as part of the National Archives' electronic services, which means that they hopefully will better cater for the needs highlighted by researchers.