# Recognition and Enrichment of Archival Documents

# D8.9

# Layout analysis and crowdsourcing

Maria Kallio, Lauri Hirvonen

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** | 1 January 2016 / 42 Months |

| | |
|---|---|
| **Distribution** | Public |
| **Contractual date of delivery** | 31.12.2018 |
| **Actual date of delivery** | 03.01.2019 |
| **Date of last update** | 21.12.2018 |
| **Deliverable number** | D8.9 |
| **Deliverable title** | Layout analysis and crowdsourcing |
| **Type** | Report |
| **Status & version** | In progress |
| **Contributing WP(s)** | WP8 Large Scale Demonstrators |
| **Responsible beneficiary** | NAF |
| **Other contributors** | |
| **Internal reviewers** | StAZH; CVL; DUTH; ABP |
| **Author(s)** | Maria Kallio; Lauri Hirvonen |
| **EC project officer** | Christophe Doin |
| **Keywords** | Large Scale Demonstrator, Archives, Reference data, Ground Truth, Handwritten Text Recognition |

## Table of Contents

## Executive Summary

The main objective of this task has been the processing of large amounts of data from the digitized collections of the National Archives of Finland and involving a great number of users who are willing to contribute to the enhancement of the digitized documents. During the project the National Archives of Finland has provided more than 2,8 million images and 2000 pages of Ground Truth of its collections for the use of the READ partners. User involvement has been mainly implemented in the form of small projects, collaboration with MOU-partners and organizing Transkribus courses at Finnish universities. This deliverable describes the results achieved during the third year of the project together with future plans for integrating Handwritten Text Recognition and Probabilistic Keyword indexing as a permanent part of digital services at the National Archives of Finland.

## 1. Introduction

The National Archives of Finland has one of the largest digital collections in the world with circa 80 million images. One of the main goals of the archive is to improve the usability and accessibility of its digital collections. As one of the four large scale demonstrators of READ, the National Archives of Finland have had the chance to test and create workflows for large scale processing and indexing of digital images. During the third year of the project the National Archives of Finland has concentrated on application of layout analysis and text recognition tools of the READ project on two large digitized nineteenth century archives from its collections: Renovated District Court Records and Poll Tax Records. Emphasis has been on ground truth production for handwritten text recognition (HTR) models and planning and execution of a pilot project that will provide full text recognition for circa 635.000 digitized images. The project is planned to be continued during 2019 with an aim of providing a free online access to the digitized images and their machine-readable content.

## 2. Corrective actions on the work plan

The original work plan for the NAF material was to focus on Layout Analysis and user involvement. It was never planned to actually carry out large scale recognition including HTR processing or Keyword

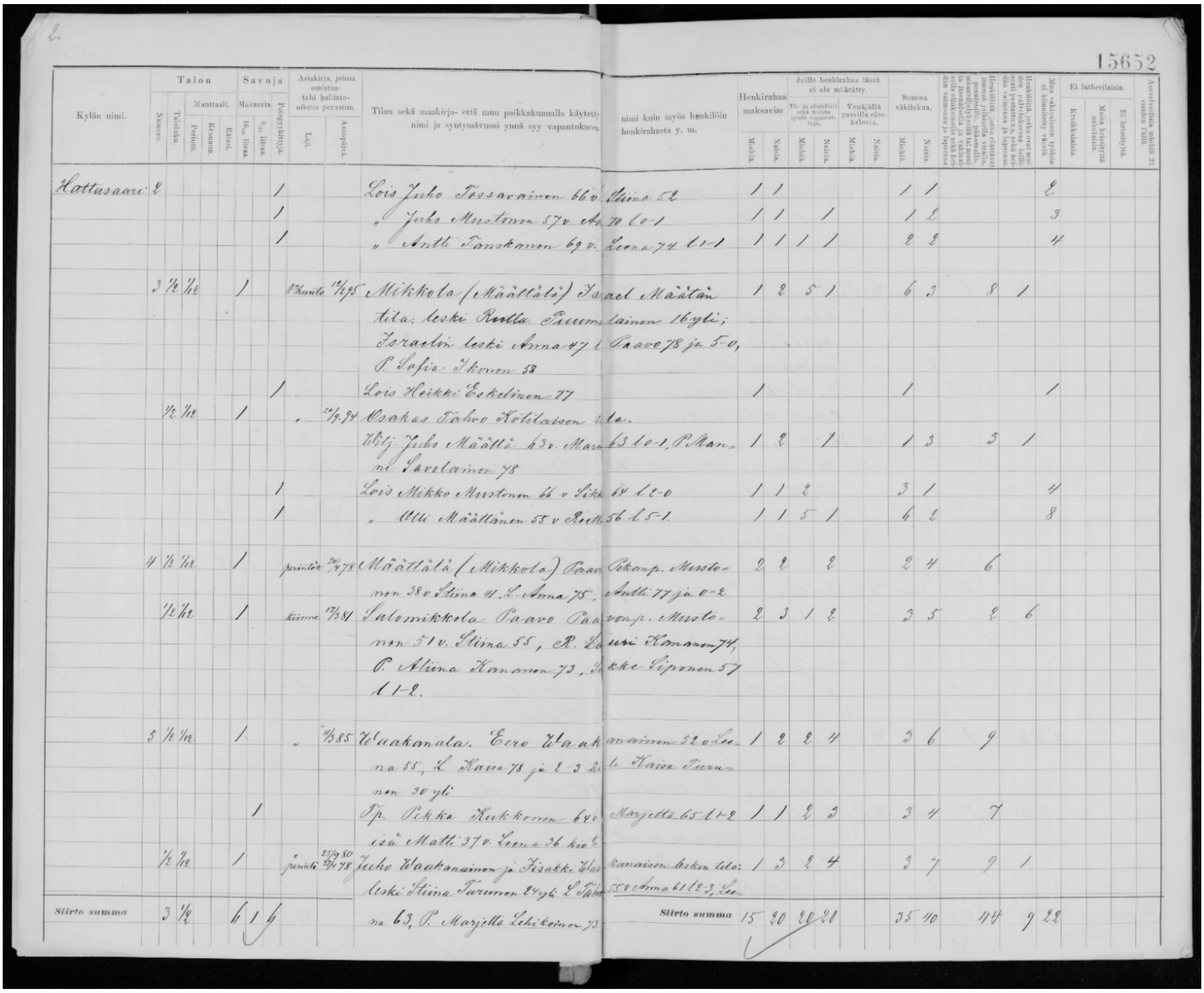
Spotting. However, after year 2 of the project it turned out that it would not make much sense to continue with this goal. There are mainly three reasons which need to be mentioned here:

First of all, after the success of the line finder tool from URO it became clear, that no further or specific development is needed to process millions of pages from the NAF collection. However to process these pages just to proof that the baselines are detected with high accuracy of about 97% was not feasible. Secondly, it turned out that certain types of documents, such as the Tax Records with their table structure and their detailed information within cells often just marked with ditto signs or numbers, would have required a much higher manual effort than foreseen in the project (and focusing on the table records from Passau made more sense from a research point of view).

But the most important reason was the decision of the National Archives to not only go for a "demonstrator" but for a "real" implementation including HTR and Keyword Spotting and to invest extra money to set up a first version of a search interface based on READ technology. During 2018 it became therefore clear that not only the HTR processing will be a core part of the NAF demonstrator, but also to set up a fully featured KWS search interface. We adapted therefore the work plan and will deliver a first version at the end of the project (06/2019) and the final version already as part of the READ-COOP at the end of 2019.

# 3. The processing of large amounts of data

## 2.1 The Poll Tax Records collection



2.

15652

| Kylän nimi. | Talon Numero. | Talon Taloluku. | Manttaali. Perintö. | Manttaali. Kruunun. | Manttaali. Rälssi. | Savuja Maksavia 16 litraa. | Savuja Maksavia 8 litraa. | Savuja Poisjyrhittyjä. | Asiakirja, johon omistus- tahi hallinto-oikeus perustuu. Laji. | Asiakirja ... Antopäivä. | Tilan sekä maakirja- että muu paikkakunnalla käytetty nimi ja syntymävuosi ynnä syy vapautukseen | nimi kuin myös henkilöin henkirahasta y. m. | Henkirahaa maksavia: Miehiä. | Henkirahaa maksavia: Naisia. | Joille henkirahaa tässä ei ole määrätty. Yli- ja alaikäisiä sekä muista syistä vapautettuja. Miehiä. | Yli- ja alaikäisiä ... Naisia. | Venäjällä passilla oleskelevia. Miehiä. | Venäjällä passilla oleskelevia. Naisia. | Summa väkilukua. Miehiä. | Summa väkilukua. Naisia. | Henkilöitä, jotka elättävät itsensä julkisella viralla, pensionilla, pääomalla, maanviljelyksellä tai muulla itsenäisellä ja vakinaisella elinkeinolla sekä heidän vaimonsa ja lapsensa | Henkilöitä, jotka ovat muiden palveluksessa laillisesti pestattuina, sekä heidän vaimonsa ja lapsensa | Muu vakinaiseen työhön ei kiinnitetty väestö | Ei lutherilaisia. Kreikkalaisia. | Ei lutherilaisia. Muita kristittyjä uskolaisia. | Ei lutherilaisia. Ei kristittyjä. | Asevelvollisia miehiä 21 vuoden iällä. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hattusaari | 2 | | | | | | | 1 | | | Lois Juho Tossavainen 66 v. | Stiina 52 | 1 | 1 | | | | | 1 | 1 | | | 2 | | | | |
| | | | | | | | | 1 | | | " Juho Mustonen 57 v. An | 70 l 0-1 | 1 | 1 | | 1 | | | 1 | 2 | | | 3 | | | | |
| | | | | | | | | 1 | | | " Antti Tanskanen 69 v. | Leena 74 l 1-1 | 1 | 1 | 1 | 1 | | | 2 | 2 | | | 4 | | | | |
| | 3 | 1/2 | 1/12 | | | 1 | | | Khunto | 12/2 95 | Mikkola (Määttälä) Is<br>tila: leski Reetta Tuum<br>Israelin leski Anna 47 l<br>P. Sofia Ikonen 58 | rael Määtän<br>lainen 16 yli;<br>Paavo 78 ja 5-0, | 1 | 2 | 5 | 1 | | | 6 | 3 | 8 | 1 | | | | | |
| | | | | | | | | 1 | | | Lois Heikki Eskelinen 77 | | 1 | | | | | | 1 | | | | 1 | | | | |
| | | 1/2 | 1/12 | | | 1 | | | " | 20/9.94 | Osakas Tahvo Kotilaisen tila.<br>Vilj. Juho Määttä 63 v. Mari<br>ne Savolainen 78 | 63 l 0-1, P. Man- | 1 | 2 | | 1 | | | 1 | 3 | 3 | 1 | | | | | |
| | | | | | | | | 1 | | | Lois Mikko Mustonen 66 v Sikk | 64 l 2-0 | 1 | 1 | 2 | | | | 3 | 1 | | | 4 | | | | |
| | | | | | | | | 1 | | | " Olli Määttänen 55 v Reett | 56 l 5-1. | 1 | 1 | 5 | 1 | | | 6 | 2 | | | 8 | | | | |
| | 4 | 1/2 | 1/12 | | | 1 | | | perintök. | 20/4 78 | Määttälä (Mikkola) Paav<br>nen 38 v Stiina 41, L. Anna 75, | Pekonp. Musto-<br>Antti 77 ja 0-2 | 2 | 2 | | 2 | | | 2 | 4 | 6 | | | | | | |
| | | 1/2 | 1/12 | | | 1 | | | kiinne | 17/3 81 | Salomikkola Paavo Paa<br>nen 51 v. Stiina 55, R. La<br>P. Aliina Komonen 73, Ju<br>l 1-2. | vonp. Musto-<br>uri Komonen 74,<br>kke Siponen 57 | 2 | 3 | 1 | 2 | | | 3 | 5 | 2 | 6 | | | | | |
| | 5 | 1/2 | 1/12 | | | 1 | | | " | 11/3 85 | Waakanala. Eero Waak<br>na 55, L. Kaisa 78 ja 2-3 ä<br>nen 30 yli | anainen 52 v Lee-<br>li Kaisa Turu- | 1 | 2 | 2 | 4 | | | 3 | 6 | 9 | | | | | | |
| | | | | | | | 1 | | | | Tp. Pekka Kukkonen 64 v<br>isä Matti 37 v. Leena 36 kio | Marjetta 65 l 1-2 | 1 | 1 | 2 | 3 | | | 3 | 4 | 7 | | | | | | |
| | | 1/2 | 1/12 | | | 1 | | | perintö | 25/9 80<br>24/4 78 | Juho Waakanainen ja Iisakki Waa<br>leski Stiina Turunen 24 yli L. Taht<br>na 63, P. Marjetta Lehikoinen 73 | kanaisen lesken tila:<br>55 v Anna 61 l 2-3, Lee- | 1 | 3 | 2 | 4 | | | 3 | 7 | 9 | 1 | | | | | |
| Siirto summa | | 3 | 1/2 | | | 6 | 1 | 9 | | | | Siirto summa | 15 | 20 | 20 | 21 | | | 35 | 40 | 44 | 9 | 22 | | | | |

The Poll Tax Records collection consists of records of head tax produced between 1634 and 1975. In addition, to their use as tax registers, the records served as census records and were the basis for the official population until the independence of Finland in 1917. The records from nineteenth century contain some 1.5 million pages, which are freely available from the digital archives of NAF. The project making the Poll Tax Records available through HTR started with careful documentation of the records and creating the Ground Truth.

The Poll Tax Records were produced according to hearth, and originally included only the name of the head of the family, while other persons were recorded by numbers. Later on, during the eighteenth century, the Crown ordered that all members of a given family had to be recorded by name and birth year in the records. The records were written in table form from the beginning. Numerous regulations were given during the eighteenth and nineteenth centuries in order to normalise the tables and to add new rows to the records. The Poll Tax Records were also used for reporting the official census of Sweden, a practice continued during the nineteenth century in the Grand Duchy of Finland. The records were written in Swedish until 1880s or 1890s when administration gradually started to use Finnish in written records. The collection of the poll tax was discontinued in 1924, but the records were continued as census records until 1975.

The Poll Tax Records at the National Archives are organised by the large administrative regions or departments into yearly series. Each year is covered by a number of volumes depending on the size of the department. Entries in each volume are organised by cities or in the country side by hundreds, which were civil administrative regions typically encompassing one parish. The records of each hundred were organised by village and farms / houses sharing one hearth, while the cities were divided into districts and buildings. The entries for each family contain at least some of the following information: name of the farm, occupation or social class of the head of the family, fore- and surname / patronym and year of birth of the head of the family followed by the names and birth years of his family members. Family relations are most typically mentioned for nuclear families, but not for extended family.

HTR model for the Poll Tax Records Collection is based on 650 transcribed pages, which are subdivided into 500 pages of ground truth, 100 pages for validation and 50 pages for testing the model. Records from the second half of the nineteenth century were selected for the model, because uniform printed tables were used for record keeping. This should help the table recognition. The model will be extended to cover the first half of the century later on.

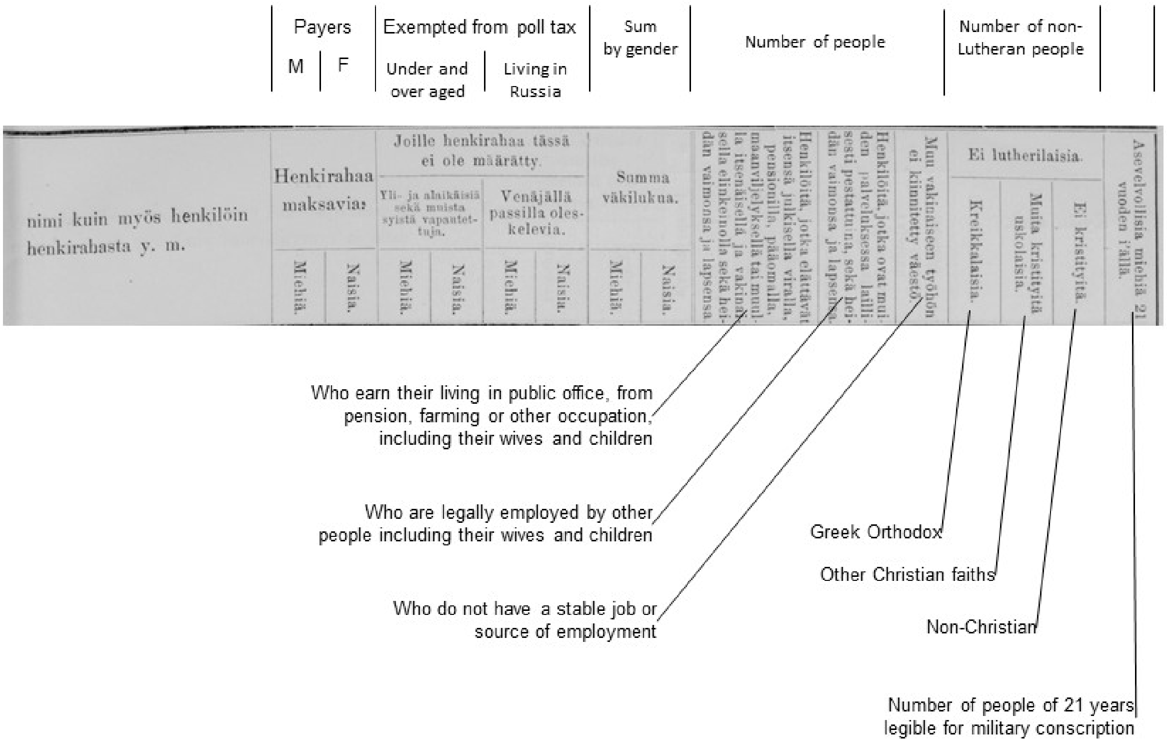
The pages were selected from the poll tax registers written in Swedish from 1845 to 1895. Basis for selection was use of printed form for the poll tax records as printed templates would be consistent in size and easier to produce table templates for further development of the HTR model. Images were selected as equally as possible from eight Finnish regions in order to provide transcripts for as large are of Finland as possible. In addition, the selection was spread as equally across the latter half of the nineteenth century as possible. This was somewhat curtained by time when printed forms were introduced in different parts of the century, which happened most often from 1860s onwards, and language switch from Swedish to Finnish that occurred during 1880s and 1890s depending on the region.

The ground truth contains 400 transcribed pages from 1860s to 1880s with 120 pages from 1860s, 140 pages from 1870s and 140 pages from 1880s. 1840s were covered by fifteen pages, 1850s by fifty transcribed pages and the early years of 1890s by thirty five pages. The model should thus be most representative for the third quarter of the nineteenth century, but it should also work for the second quarter and 1890s as the script used in the poll tax accounts is quite stable throughout the century. Depending on the quality and availability of digitized record books samples for transcription were selected as sequences of three, five or ten pages. Most often sequences were of five pages selected from a single record book. If there were only a few digitized books for a particular decade, for example poll taxes for 1860 and 1865, ten pages were selected from a single poll tax register, mostly in a sequence of five pages, but in some cases as 10 pages. For the earlier registers fifteen pages in sequences of three or five pages were selected. The transcribed pages were selected mostly in arbitrary fashion, but attempt was made to select samples from different parts of each region so that no single locality would be over represented. In some cases hand drawn forms were selected for the ground truth to provide more material for training of HTR model. The hand drawn forms were similar in execution to printed ones in the number of columns and headers of the columns with the only difference being drawn by hand. An attempt was made to select images produced by overhead scanners, but digitized microfilms were also used when the image quality was good.

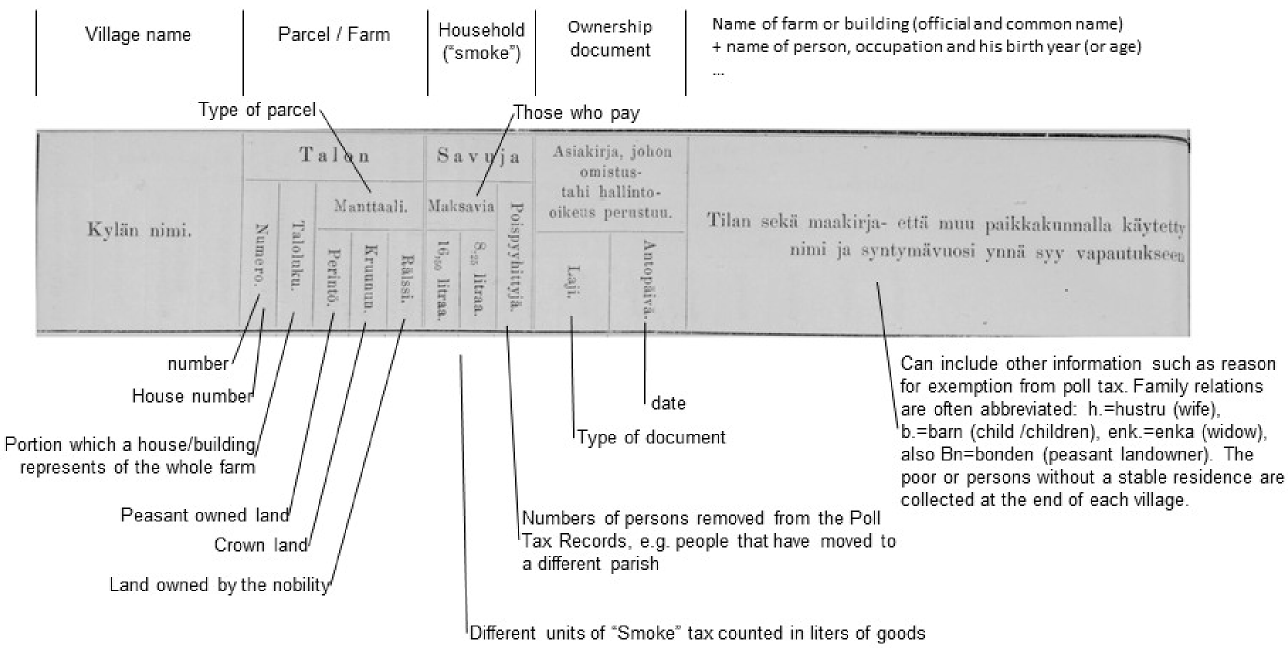
A data set of 100 pages was selected for validation of the ground truth. Five batches of twenty pages were arbitrary selected from Poll Tax Records that had been digitized with overhead scanners. An attempt was made to ensure that the images did not contain blank pages and that the text in the middle of the page is readable and not obscured by the spine of the book as the records have been bound in volumes of more than thousand pages. The validation data set contains following images: 1860 U:52 pages 241-253 and 357-362; 1865 Va:79 pages 1602-1621; 1870 K:70 pages 800-819; 1880 Ou:87 pages 1525-1544 and 1885 T:151 143-162. The pages from 1885 are from the town of Pori in order to provide validation data for records of cities.

A HTR model for tabular Poll Tax Records was produced from the Ground Truth during summer of 2018. Unfortunately the results were quite bad as the character error rate of the model was circa 30%. After investigation we concluded that there were numerous polygon to text errors on certain pages of the training data, which caused drastically worse results, and led to a model that did not have good recognition rate. The records contain large amounts of short abbreviations and cells with numbers, which cause noise for the HTR process. In addition, the HTR model was trained without table understanding tools and the idea is to continue the work by testing these tools in cooperation with NaverLabs. The whole collection of  1,5 million images of 19th century Poll Tax Records have been uploaded to Transkribus and we plan to process them after having finished the Court Records Collection.

# Poll Tax Records


Payers
M
F
Exempted from poll tax
Under and over aged
Living in Russia
Sum by gender
Number of people
Number of non-Lutheran people
nimi kuin myös henkilöin henkirahasta y. m.
Henkirahaa maksavia:
Joille henkirahaa tässä ei ole määrätty.
Yli- ja alaikäisiä sekä muista syistä vapautettuja.
Venäjällä passilla oleskelevia.
Summa väkilukua.
Miehiä.
Naisia.
Miehiä.
Naisia.
Miehiä.
Naisia.
Miehiä.
Naisia.
Henkilöitä, jotka elättävät itsensä julkisella viralla, pensioonilla, pääomalla, maanviljelyksellä tai muulla itsenäisellä ja vakinaisella elinkeinolla sekä heidän vaimonsa ja lapsensa.
Henkilöitä, jotka ovat muiden palveluksessa laillisesti pestattuina, sekä heidän vaimonsa ja lapsensa.
Muu vakinaiseen työhön ei kiinnitetty väestö.
Ei lutherilaisia.
Kreikkalaisia.
Muita kristityitä uskolaisia.
Ei kristityitä.
Asevelvollisia miehiä 21 vuoden i'ällä.
Who earn their living in public office, from pension, farming or other occupation, including their wives and children
Who are legally employed by other people including their wives and children
Who do not have a stable job or source of employment
Greek Orthodox
Other Christian faiths
Non-Christian
Number of people of 21 years legible for military conscription

**Poll Tax Records**


Village name
Parcel / Farm
Household ("smoke")
Ownership document
Name of farm or building (official and common name)
+ name of person, occupation and his birth year (or age)
...
Type of parcel
Those who pay
Kylän nimi.
Talon
Numero.
Taloluku.
Manttaali.
Perintö.
Kruunun.
Rälssi.
Savuja
Maksavia
16,50 litraa.
8,25 litraa.
Poispyyhittyjä.
Asiakirja, johon omistus- tahi hallinto-oikeus perustuu.
Laji.
Antopäivä.
Tilan sekä maakirja- että muu paikkakunnalla käytetty nimi ja syntymävuosi ynnä syy vapautukseen
number
House number
Portion which a house/building represents of the whole farm
Peasant owned land
Crown land
Land owned by the nobility
date
Type of document
Different units of "Smoke" tax counted in liters of goods

Can include other information such as reason for exemption from poll tax. Family relations are often abbreviated: h.=hustru (wife), b.=barn (child /children), enk.=enka (widow), also Bn=bonden (peasant landowner). The poor or persons without a stable residence are collected at the end of each village.

Numbers of persons removed from the Poll Tax Records, e.g. people that have moved to a different parish

**Smoke**: a tax term for farm houses or parts of farms, whose occupants did not own the building or the land, but typically paid rent to the main land owner.

## 2.2 The Court Records Collection

In 2018 NAF continued the production of GT for District courts' notification records. Because of the size of the renovated district court record collection, it was decided that for the testing of the HTR only the notification records from the 19th century would be used. Records from 1850's to 1870's are being used for ground truth generation. All of the records are in Swedish. From the districts where digitized records between 1850's and 1870's already existed, the criteria for records were the quality of the digitization and the records' uniform appearance. From each district that filled the requirements stated above, 20 pages of the records were chosen for ground truth generation, excluding blank pages. The records chosen present as large geographical distribution as possible. 2018 GT includes 660 pages of records, 500 of which will be used for generating the HTR model, 100 for validation, and 60 for testing. Transcription of the records was done by a service provider and afterwards corrected by NAF staff members.

The first HTR model created from Porvoo data yielded results with CER of 12% on average (2017). In April 2018 we used the same data with an addition of hundred pages and managed to improve the character error rate to 9.02 percent. Both models were based on materials from the region of Porvoo in Southern Finland from 1850s. The model works surprisingly well with similar material. We tested it with court records from the region Yddre in Eastern Sweden from 1872, which were digitized by the digital archive of the Swedish National Archives. The model produced a character error rate of circa 12 to 15 percent on the Swedish material, which is remarkable as it had been trained only on Finnish material from twenty years earlier. With HTR+ the current model has CER of below 4% and it is likely to get even better after training with all the existing GT. The whole collection of Notification Records has been already transmitted to the UIBK team for further preparations of the processing pipeline.

*HTR models trained for NAF Court Records collection*

| **NAF GT Court records** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Title** | **ID** | **Training** | **Test** | **CER test set** | **CER train set** | **epoch** | **Train size per epoch** | **learning rate** | **Noise** |
| NAF Court Record test1 | 6458 | 120 | 14 | 11.21% | 4.01% | 120 | 10000 | 1-e3 | both |
| Combined court records M1 | 6436 | 139+ID 3279 (462) | 3 | 12.80% | 11.92% | 120 | 10000 | 2e-3 | both |
| Porvoo court records test 3 | 3279 | 112+ ID 3056 (323) | 6 | 9.04% | 10.18% | 175 | 10000 | 1e-3 | both |
| NAF Court Records 1853–1855 model | 2461 | 141 | 21 | 11.42% | 9.74% | 120 | 10000 | 1e-3 | both |
| Court records test 2 (htr+) | 8726 | 330 | 10 | 2.87% | 3.32% | 200 | | | |

## 2.3 Large scale implementation in the future

The National Archives of Finland has requested funding of 250 000 euros from the Finnish Ministry of Education and Culture for 2019 in order to integrate Handwritten Text Recognition as part of the digital services of the National Archives. The goal is to create a sustainable workflow for processing large collections not only at the National Archives of Finland but also in other archives, as it is intended to create general practices for carrying out similar projects. In addition to processing the data, the purpose is to build a user interface for using and searching of text recognized collections in cooperation with the READ-COOP. The main goal of the implementation is to provide the whole 19th century collection of District Court Records in computer readable and searchable form through the digital services of the National Archives. In the first phase the collection of 635 000 pages of Notification Records of District Court Records will be processed during 2019.

## 3. User involvement and networking

The National Archives of Finland has actively carried out its task of involving users who are willing to contribute to enhancement of digitized documents in Transkribus. At the same time it has built a sustainable network among Finnish and Nordic archives who are working on their own projects using the HTR technology. During the current year NAF participated on making a Massive Open Online Course of Digital Cultural Heritage at University of Helsinki presenting the possibilities of Transkribus and HTR technology on archival collections. In addition NAF organized a Transkribus course for the Master Programme in Cultural Heritage where students were transcribing Finnish War Diaries of Second World War. Thanks to the course, the CER of recognizing the War diaries improved down to 7% which is an excellent result for such variable collection. The course will be running again in the next academic year and NAF is planning to enhance the recognition of the War diaries on larger scale.

NAF has also supported the MOU -partners with their own projects sharing information on Ground Truth production and training HTR models. Cooperation has been active especially with the Society of Swedish Literature in Finland, Finnish Literature Society and Institute of Languages of Finland. For this reason, NAF plans to set up a working group on 19th century Swedish records together with the aforementioned institutions. Interest in Handwritten Text Recognition has increased in general and several National Archives in the Nordic and Baltic countries are starting their own projects with Transkribus and NAF has been able to present its own results as an example which will hopefully help to implement large scale projects in the future.