FIN-CLARIAH 2024/25

D:3.3.1 Enhancing the usability of archival data

Authors:
Antero Holmila
University of Jyväskylä
Ida Toivanen
University of Jyväskylä
Venla Poso
University of Jyväskylä
Tanja Välisalo
The National Archives of Finland
Mikko Lipsanen
The National Archives of Finland

Introduction

The National Archives of Finland is digitising documents across a range of fields, encompassing both historical archives and more contemporary state authority records. Our work focuses on enhancing the accessibility and usability of these documents through the application of machine learning techniques. By automatically recognizing content from unstructured digitized data, we unravel various potential applications for archival data.

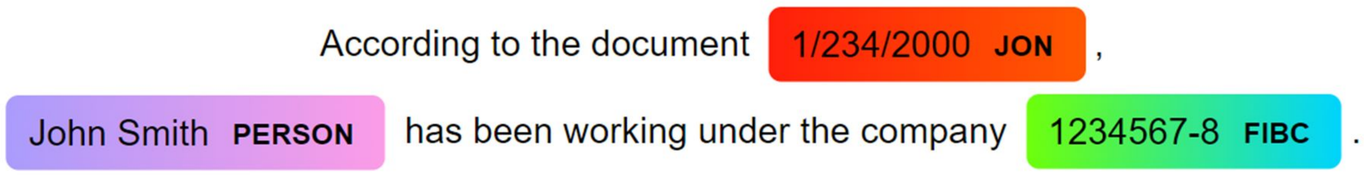
Our research aims to explore both the visual and textual features of digitized archival data. We have developed a named entity recognition (NER) model designed to identify significant entities within archival data. Additionally, we are working on creating a solutions for automatically recognizing various table structures and their contents. This work will build on previously done development with table recognition and HTR in the NAF. The ultimate objective is to benefit end-users in different fields of research and administration by enhancing the usability of archival data.

Archival data

The inherent challenge of archival data is in its heterogeneous nature. The archival data originates from diverse sources and spans various historical periods. The temporal aspect introduces complexities not only due to the evolution of formats and writing styles but also because digitisation practices and image quality vary across documents produced in different eras.

Enriching the archival documents

In order to make digitised archives useful for different user groups, the unstructured archival data needs to be enriched in different ways. Recognising content and creating metadata can help to organise, search and analyse the data. For example, recognising journal numbers (*diaarinumero*) can be used to correctly date the documents.



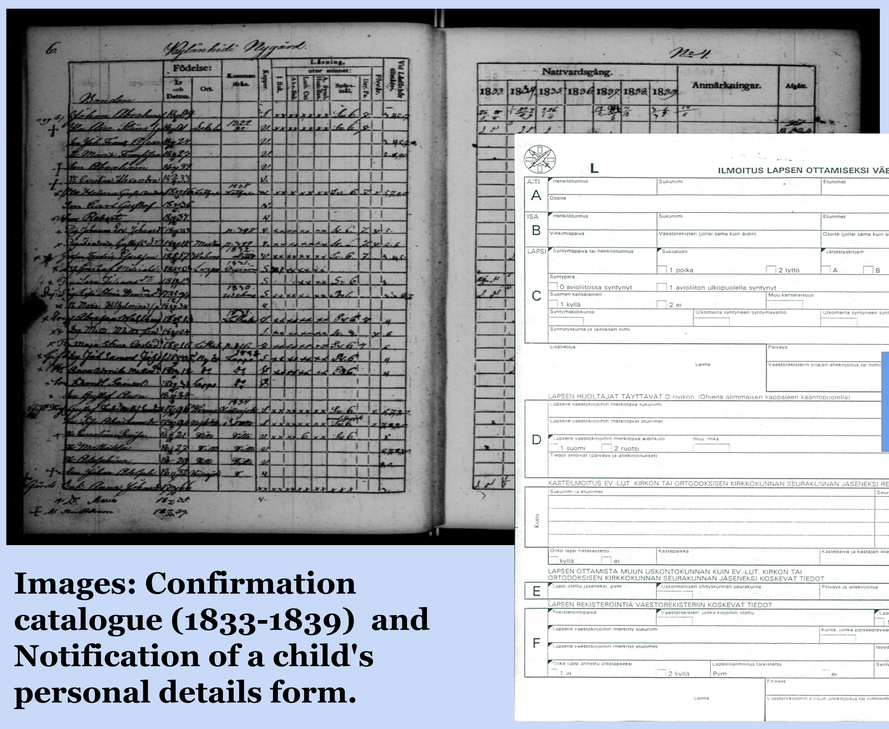
Task 1: Named entity recognition

In the NER model development we concentrated on two main challenges: 1) on reducing the effect of OCR noise and 2) idiosyncrasies of formal language in mass digitised state authority data. First version of named entity recognition model developed for the archival data is available in Huggingface, but we will publish the final model in the fall with annotation guidelines.

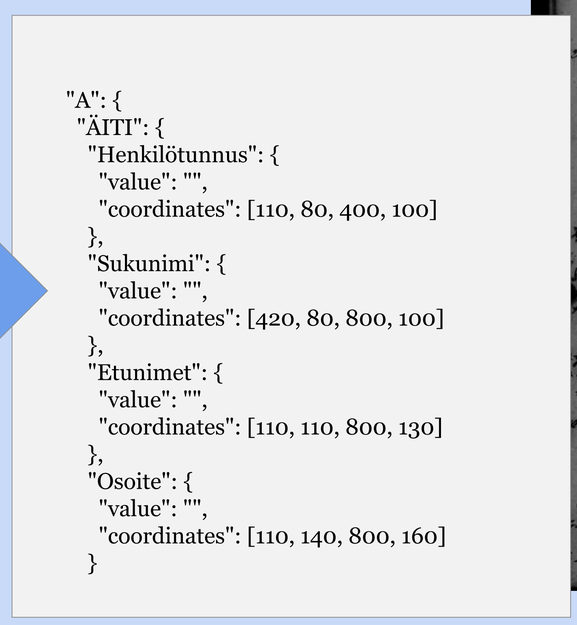
huggingface.co/Kansallisarkisto/finbert-ner

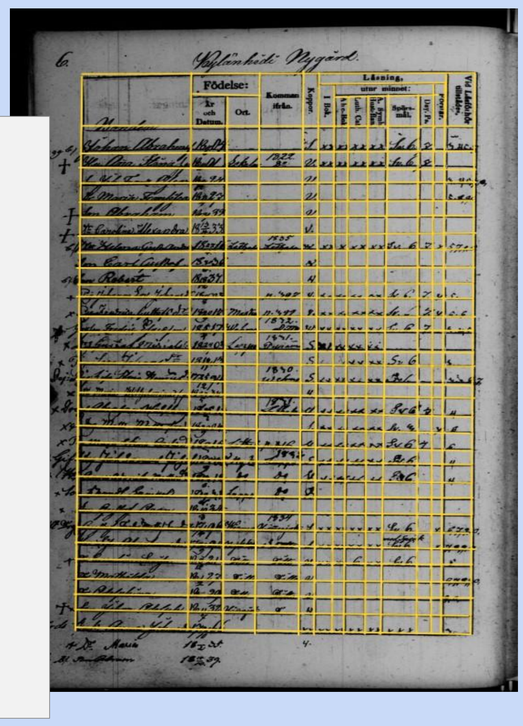
Task 2: Table recognition

Starting with testing existing HTR and table recognition models and then focusing on researching the possibilities of **multimodal approaches** to the archival data, we aim to find sustainable solutions for recognising tables and their content from both historical and modern archival data.



Images: Confirmation catalogue (1833-1839) and Notification of a child's personal details form.





Collaboration

This work will be conducted in cooperation with the National Archives of Finland and the University of Jyväsylä.