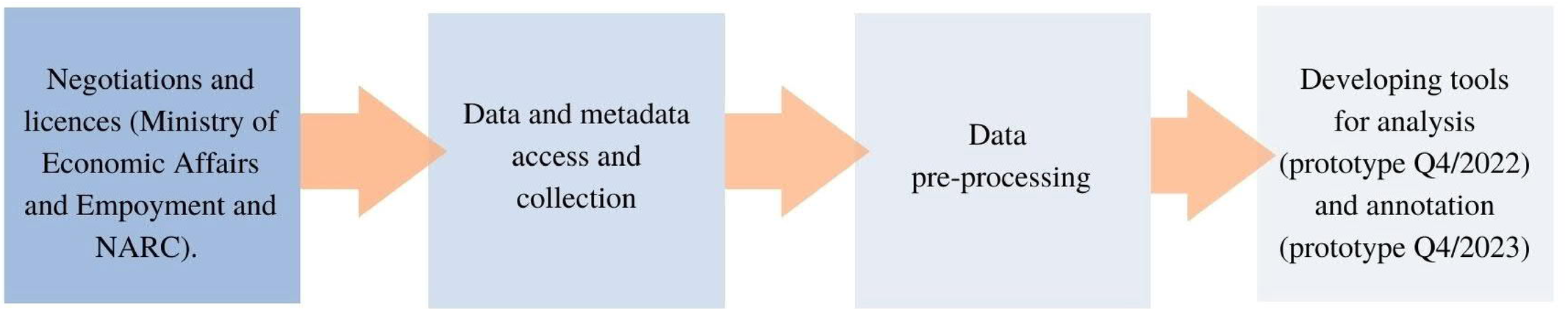
WP 3.2 AI solutions to better use National Archives mass digitisation services

Jari Ojala
Faculty of Humanities and Social Sciences
University of Jyväskylä
Antero Holmila
Department of History and Ethnology
University of Jyväskylä
Tanja Välisalo
Faculty of Humanities and Social Sciences
University of Jyväskylä
Ida Toivanen
Faculty of Humanities and Social Sciences
University of Jyväskylä
Venla Räsänen
Faculty of Humanities and Social Sciences
University of Jyväskylä

The amount of archive data is growing exponentially. We develop AI based solutions to help researchers in various fields to effectively use massive digitized datasets from the National Archives of Finland.

As the National Archives of Finland (NARC) have started the mass digitization in 2019, WP3.2 aims at creating tools for analysis and annotation of the massive data produced in the process. We will apply named entity recognition (NER) technology in tool development.

Simultaneously, we develop best practices for accessing and transferring the data. Building a detailed description of the possible restrictions and complexities of the process, we aim to offer time-saving procedures for researchers who hope to use digitized data from the National Archives.



Roadmap

OCR has been used for the data produced in the mass digitization process. We are using data from the Ministry of Economic Affairs and Employment of Finland for the pilot case. The data is accompanied by metadata. Possible needs for pre-processing the data will be identified as they appear during the process.

We expect to have three different outcomes from this pilot case. First, we will develop a tool for structural analysis of the mass digitized data to identify e.g. different content types. Next, we will create an annotation tool based on named entity recognition (NER) technology. This will enable identifying and annotating e.g. persons, organizations, events, and locations. These tools are developed for documents written in Finnish, particularly taking into consideration the properties of digitized archive material of different eras. Finally, we collect information about the best practice for accessing and transferring data from NARC to research organizations. Tool development will involve benchmarking the existing text analysis tools.

Outcomes

1. Analysis tool for identifying document structures
2. Annotation tool for identifying named entities from archive data.
3. Best practices for accessing and transferring data from NARC to research organizations

Challenges

How to identify images and other non-textual content from the data?