A proposal for a contributed talk at Cultural Data Analytics Conference 2023, Tallinn
The role of metadata in interpreting media data as cultural data: ERR case

Andres Kõnno PhD, researcher at Tallinn University, BFM / akonno@tlu.ee
Kais Allkivi-Metsoja MA, researcher at Tallinn University, DTI / kais@tlu.ee

This presentation makes an introduction to the outcomes of a media data analysis project supported by the Estonian Ministry of Education: „Applying automatic language modelling on the ERR archive: the study of coreference in the linked-data based archive queries".

We shall discuss the implications and the results of this project from three different angles: (1) the application of language modeling (BERT) for entity detection (applying the EstBERT language model (open source machine learning framework for NLP) fine-tuned for NER to detect public figures and other individual agencies in cultural domain, cultural institutions and other collective agencies, locations (both Estonian and global) and also list-based subclassification of the main named entity categories (such as relevant culture-related keywords from the Estonian Keywords Database) to generate additional metadata layers to the existing metadata, (2) the modelling of topics (including the conceptualisation of 'cultural topics' in the datafied context) that is necessary to establish a linked database format in ERR's archive and (3) the challenges of identifying the relators between different entities (objects, names, institutions) via syntactical analysis of Estonian language.

The dataset entails approx. 44,000 entries from the culture section of the radio archive, obtained via speech-to-text (input = basic metadata and text, no paragraphs, no annotations).

On the picture: the preliminary prototype of the ERR data linker: