# On (CMDI) metadata in CLARIN: the case of annotation

Jakob Lenardič

Institute of Contemporary History

Centre meeting
13 June 2023

# Metadata provision – state of affairs

- In CRF, focus on 3 metadata categories (for corpora): *size*, *licence*, and *annotation*
- Currently, 696 corpora overviewed in CRF
- Current size and licence provision rate: approx. 96% corpora
- Current annotation info. rate: approx. 62% corpora

The rest unannotated or missing metadata?

# Depositing guidelines (1/2)

- 2022 review of depositing guidelines of CLARIN B-centres
- How are depositors instructed to document basic metadata?
- Annotation: 2 out of 23 repositories provide guidelines
- DSpace repositories have no special field for defining annotation

## Qualitative recommendations for describing deposit metadata

https://office.clarin.eu/v/CE-2022-2138-qualitative-depositing-recommendations.pdf

# Depositing guidelines (2/2)

Recommendations for describing annotation

- Distinction between:
  1. Linguistic annotation (e.g., tokenisation, PoS-tagging, lemmatisation)
  2. Non-linguistic annotation (domain-specific, e.g., political parties)
- Additional optional descriptors:
  1. Tagsets, theoretical frameworks (e.g., Universal Dependencies vs. LFG), mode (manual vs. automatic)
  2. Tools used for annotation
- Lack of annotation should be mentioned

# Annotation in CMDI (1/3)

- 431 corpora in CRF are annotated
- 35% (153) records use dedicated CMDI annotation components
- The rest (usually) describe annotation as part of FTD
- The most frequently used profile:
  LINDAT_CLARIN – 125 (29%) of 431 corpora

# Annotation in CMDI (2/3)

- Only 1 corpus avails itself of LINDAT_CLARIN's annotation components:
  *DK-CLARIN Reference Corpus of General Danish*
  - **annotationInfo**
    - **annotationType:** tokenization
    - **annotationType:** sentence and paragraph segmentation
    - **annotationType:** POS-tagging
    - **annotationType:** lemmatization
- The rest typically describe annotation as part of the FTD
- F.e., *Written corpus ccGigafida 1.0*:

  "The corpus is annotated with morphosyntactic descriptions (PoS-tagged) and lemmatised. It is encoded in XML TEI format (Text Encoding Initiative P5)."

# Annotation in CMDI (3/3)

- General correlation between higher metadata detail and use of CMDI components
- Potential cause/factor: DSpace repositories have no field for annotation vs. META-SHARE