CLARIN strives for high-quality metadata (De Smedt et al. 2018)
 However, unequal provision and documentation observed between:
  1. Repositories (McCrae et al. 2015; Cimiano et al. 2020)
  2. Metadata categories (Lenardič and Fišer 2020)
 Qualitative documentation of resources and tools suboptimal (Koolen, Van Gorp, and Van Ossenbruggen 2019)

2 aims for our contribution
  1. Survey of existing guidelines
  2. Proposal for new depositing guidelines
23 repositories certified as CLARIN B-centres (Wittenburg et al. 2019)
B-centres offer LRTs in line with FAIR (Wilkinson et al. 2016)
17 (74%) repositories have guidelines in English

**What we took into account:** the corpus documentation in the 17 English guidelines
The survey

- 5 metadata categories overviewed for the 17 repositories
- Inclusion rates low:
  - **Annotation**
    - 2 (12%) repositories: FIN-CLARIN, Bayerisches Archiv für Sprachsignale
  - **Size**
    - 3 (17%) repositories: CELR, CLARIN.SI, FIN-CLARIN
  - **Language**
    - 5 (29%) repositories: ASV Leipzig, ARCHE, CELR, CLARIN.SI
    - FIN-CLARIN
  - **Resource name**
    - 5 (29%) repositories: same as Language
  - **Free-Text Description**
    - 4 (24%) repositories: ASV Leipzig, CELR, CLARIN.SI, FIN-CLARIN
Proposal for new guidelines

- Qualitatively, existing instructions are lacking in detail
- Documentation should focus on those aspects of the deposit that are important for (re)use in research
Not all depositing systems prompt for annotation metadata unlike size, e.g. DSpace (Smith et al. 2003)

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, syntactic frameworks, named entity classes, etc.
2. Tools used for annotation

Lack of annotation should be mentioned
The guidelines (2/5) – size

- In case of several modalities, provide size for each modality separately (The CLARIN:EL Technical Team 2022)
- Maximally informative: tokens, words, sentences, etc.
- Number of files isn’t a useful descriptor by itself
- If a corpus is tokenised, distinguish nr. of words vs. tokens
Descriptive titles are preferable to non-descriptive ones; for instance *Automatically sentiment annotated corpus AutoSentiNews 1.0* (Bučar 2017) vs. *The HRMM tagger*
- Specify potentially ambiguous language characteristics
- Example 1: bilingual corpora; which language corresponds to the original text?
- Example 2: oral history corpora; language proportion of sources should be clearly indicated
Focus on describing the resource itself rather than background information (e.g., funding, institutions involved)

Useful characteristics to describe:
1. modality (spoken, written, visual)
2. time period of text publication, data elicitation, etc.
3. geographic coverage
4. data sampling (text types and their ratios)
5. domain-specific characteristics
In conclusion

- Depositing guidelines important – minimizing metadata gaps at the stage before publication
- However, recommendations rather than obligatory requirements for new deposits
- Next step: discuss the guidelines with repository admins for possible adoptions/adaptations
References I


