Title	Qualitative recommendations for depositing LRTs at CLARIN repositories
Version	1.0
Author(s)	Darja Fišer, Jakob Lenardič
Date	8-9-2023
Status	For distribution
Distribution	BoD, NCF, UI, SCCTC
ID	CE-2022-2138

Contents

1. Introduction

This document presents a proposal for a new set of depositing recommendations for the repositories of CLARIN centres. The recommendations focus primarily on the qualitative aspects of the deposited language resources and tools so as to facilitate their reusability within Digital Humanities and Social Sciences research.

The recommendations are presented in Section 2 and concern five basic metadata categories – resource title, language, size, annotation, and free-text description –, which are the most visible and often the most important types of metadata from the perspective of a researcher perusing a repository's catalogue. The guidelines have in part been presented at the 2022 edition of the CLARIN Annual Conference (Lenardič and Fišer 2022: 48–52) and are partially based on the existing guidelines of the CLARIN.SI consortium for the documenting of language resources as well as on the proposals by Odijk (2019: 122–123) for the documenting of language tools.

2. Depositing Guidelines

2.1. Resource Title

We suggest that the title should give a very short description of the resource in addition to the proper name, which can be an acronym. If there is a chance that further versions of the resources will be submitted, the major and minor version should come at the end; see Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0 as an example of such a naming convention.

2.2. Language

Specify any possible and important characteristics of the resource's language(s) that are ambiguous in the language subcomponent of the metadata profile. An example of this is the **directionality of translations** in the case of parallel corpora – if a bilingual corpus contains Slovenian and English texts, it should be specified which language corresponds to the original or translated texts or both. If applicable (e.g., oral history corpora), the **proportion of sources** in the deposited corpora with respect to their language should be clearly indicated.

2.3. Size

The size of the resource should be given in as many sensible categories as possible. If a resource contains **more than one modality** (e.g., audio recordings and their written transcriptions in the case of spoken corpora), provide size for each modality separately. Additionally, if the corpus is **tokenised**, provide both the word and token numbers.

2.4. Resource Annotation

Provide a brief summary of the annotation process, possibly as part of the free-text description if there is no separate submission field for this information. Distinguish between **linguistic** (e.g., tokenisation, sentence segmentation, PoS-tagging, lemmatization, syntactic parsing, named entity recognition) and **non-linguistic levels** of annotation, which are often domain specific (e.g., gender annotation of speakers in parliamentary corpora).

Information on additional subcomponents of the annotation process itself should be provided, such as the **tagset** used for morphosyntactic tagging, the **class** of named entities, and possible **syntactic frameworks** for syntactically parsed corpora (e.g., Universal Dependencies for dependency grammars), and the **tools** used to annotate the corpora. Useful metadata also pertain to **annotation tools, training sets**, and **annotation accuracy**. At the very least, provide a reference where such information is available.

If the resource is only partially annotated or not annotated at all, mention this as well. Mention also if any of the annotations were done **manually**, possibly providing basic information about this as well (e.g., number of annotators, inter- and intra-annotator agreement).

See the Appendix for a more comprehensive overview of possible annotation metadata.

2.5. Free-Text Description

The free-text description should mainly focus on the description of the resource itself rather than on background information such as funding. The resource description should be about half a page in length. See Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1 for an example of a detailed description.

The following subsections provide advice on what to describe for each deposit type separately. Note that the following points need not be described in detail but rather overviewed, especially if more comprehensive descriptions are provided elsewhere, as in published papers.

2.5.1. Corpora

Provide information on the following features: **modality** (spoken, written, visual, etc.), **time period** (based on publication date), **geographic coverage**, **data sampling** (text types and their ratios; text sources and their ratios), **envisaged research domains**, and any important and unique **domain-specific characteristics** (e.g., participants' ages and L1s in learner corpora). Mention whether personal or sensitive data are included, and whether they have been anonymised.

2.5.2. Lexical Resources

If ambiguous, define the aim of the lexical resource and thereby distinguish e.g. morphological lexica for the training of NLP applications from dictionaries aimed primarily at human use. Tailor the description to the type of lexical resource. For **training lexica**, describe the key features of the annotation process (e.g., tagset, manual mark-up) and other structural information (e.g., lemma frequencies). For **dictionaries,** briefly overview how the entries are structured and what kind of grammatical (e.g., morphological information, collocation properties, phonemic transcription) and non-grammatical information (examples of use, contextual features) is presented.

2.5.3. Tools

Provide information on the tool's applicability in terms of the relevant **research domain(s)**, its **distribution and installation requirements**, as well as its **output and input** characteristics, which includes **MIME-types**, **annotation schemata**, and **tagsets**. Furthermore, provide information that's unique to the **functionality**, such as the types and granularity of categories recognized by a named entity recognizer or the types (e.g., sentence-level, document level) and levels (binary, ternary, quaternary, etc.) of sentiment recognized by a sentiment analyser.

2.5.4. Language Models

Consider providing information on the following categories: the **tool** used for building the model, the **dataset** on which the model was trained, the **output** in terms of annotation labels, and the **annotation accuracy.**

Appendix: Resource Annotation Metadata

This appendix provides a comprehensive list on possible metadata components pertaining to annotation. It is primarily based on the *resourceInfo* CMDI profile (the descriptions are taken from there verbatim), as well as on *MDrecord_corpus* and *corpusProfile*.

Metadatum	Component	Description
Annotation Info	Annotation Type	Specifies the annotation level of the resource or the annotation type a tool/ service requires or produces as an output
	Annotated Elements	Specifies the elements annotated at each annotation level (e.g., speaker noise, discourse markers, tokens)
	Annotation Standoff	Indicates whether the annotation is created inline or in a stand-off fashion
	Segmentation Level	Specifies the segmentation unit in terms of which the resource has been segmented or the level of segmentation a tool/service requires/outputs
	Annotation Format	Specifies the format that is used in the annotation process since often the mime type will not be sufficient for machine processing
	Tagset	A name or a URL intended as reference for the tagset used in the annotation of the resource or by the tool/service
	Tagset Language ID	The identifier of the tagset language; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)
	Tagset Language Name	The name of the tagset language; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)
	Conformance to Standards/Best Practices	Specifies the standards or the best practices to which the tagset used for the annotation conforms (e.g., ISO)
	Theoretic Model	Name of the theoretic model applied for the creation or enrichment of the resource, and/or reference (URL or bibliographic reference) to informative material about the theoretic model used
	Annotation Mode	Indicates whether the resource is annotated manually or by automatic processes (automatic, mixed, interactive, manual)
	Annotation Mode Details	Provides further information on annotation process
	Annotation Start Date	The date in which the annotation process has started
	Annotation End Date	The date in which the annotation process has ended
	Inter Annotator Agreement	Provides information on the inter-annotator agreement and the methods/metrics applied
	Intra Annotator Agreement	Provides information on the intra-annotator agreement and the methods/metrics applied
Annotation Manual	Annotation Manual	Provide a reference to or URL for the annotation manual
Annotation Tool	Tool Name	The full name or URL or identifier of the annotation tool
	Training Set	The full name or URL or identifier of the training set for annotation
	Annotation Accuracy	Provide information as to quality of annotation (e.g., score)
Size Per Annotation	Size	Specifies the size of the resource with regard to the Size Unit measurement in the form of a number
	Size Unit	Specifies the unit that is used when providing information on the size of the resource or of resource parts (e.g., tokens, words, hours, etc.)