DARIAH-EU

CLARIN
Common Language Resources and Technology Infrastructure

FIN-CLARIAH

Jussi Piitulainen
Jyrki Niemi
Erik Axelson
Ute Dieckmann
Mietta Lennes
Sam Hardwick
Tommi Jauhiainen
Krister Lindén
FIN-CLARIN / University of Helsinki

Martin Matthiesen
FIN-CLARIN / CSC – IT Center for Science

WP 1.1 Text processing and annotation environments

SANITIZE

Ensure that all characters are *sane* UTF-8.
Ensure that source format is otherwise all right.

Publish downloadable source version.

SEGMENT

First segment source data into paragraph elements inside text elements so that the elements carry any available information as attributes.

Then segment the running text inside paragraph elements into sequences of tokens (word forms, punctuation marks, and such) inside sentence elements.

Result format is VRT, of Corpus WorkBench, to be further annotated and published in Korp. (There *can* be other structural elements, too.)

ANNOTATE

Develop and use tools adapted for VRT (*with named fields*) to annotate data with
- various base forms
- morpho-syntactic descriptions of word forms
- syntactic dependency relations (moving to TNPP and UD2)
- name classes (FiNER)
- sentiment annotations (positive, neutral, negative; Suomi24)
- language identifiers (HeLI-OTS).

Always preserve any previous annotations and structure.

CONFIGURE

Map corpus structure and annotations to the Korp user interface.

Publish searchable version in Korp.

PACKAGE

Put corresponding data in downloadable archive files.

Publish downloadable VRT version.

Need to develop conversions to other download formats.

MANAGE RIGHTS

Data may have owners who want to have some control.

Data may contain personal information that may need protection.

Map access conditions to CLARIN PUB, ACA, RES, with decorations.

MAINTAIN METADATA

Assign persistent identifiers.

Keep track of corpus versions, both data and annotations.

Publish descriptions in META-SHARE, harvested to CLARIN VLO and other services.

UNIVERSITY OF HELSINKI

DARIAH-FI

KIELIPANKKI
The Language Bank of Finland

FIN-CLARIN