Ask your SIS:
Collecting Centre Recommendations on Data Deposition Formats

Piotr Bański, Tomaž Erjavec, Francesca Frontini, Hanna Hedeland, Eliza Margaretha Illig, Neeme Kahusk, Fahad Khan, Karlheinz Mörth, Jan Odijk, Jussi Piitulainen, Christian Thomas, Dieter Van Uytvanck, Menzo Windhouwer, Andreas Witt

CLARIN Annual Conference, September 2021

What we do

The tasks of the CSC are (quoting from the 2019 Bylaws, with non-random highlights):

- to collect, consolidate and prepare for publication in a single place its findings and recommendations related to standards;
- to maintain the set of standards supported by CLARIN and adapt them to new developments within or outside CLARIN;
- to publish and promote the standards supported by CLARIN;
- to develop and implement procedures for the discussion of recommendations and the adoption of new standards;
- to ensure harmonisation of standards between CLARIN ERIC and related initiatives;
- to ensure communication with international standards bodies such as (but not limited to) ISO;
- to advise the BoD in all matters related to standards.

The 2019-2021 cycle and focus

Overall running goal: collect and maintain standards-related information.

Beginning roughly in mid-2019 (CAC in Leipzig), the committee narrowed the initial focus to data format recommendations publish(ed|able) by those centres that offer data-deposition services.

This is a well-defined task that prepares the ground for future activities.

We are at the point where outside contributions become possible and welcome.

In what follows, you will see some of the motivation and a demo.

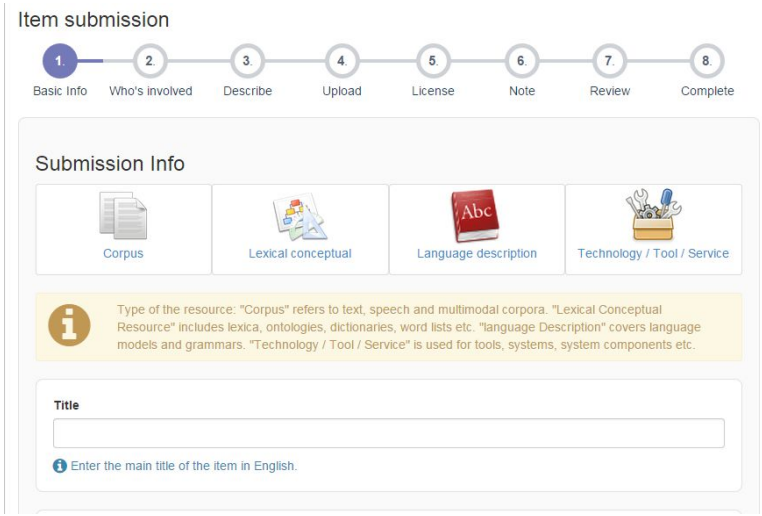
Data deposition approached from the user-side

(Most) B-centres as well as several non-B-centres offer data deposition services.

Accepted data formats depend on the centre profile (e.g. text corpora vs. speech corpora).

What should the user do to gain that info?

Many centres use the “deposition guidelines”, but they often do not provide centre-specific format recommendations (though see CLARIN.SI).



The SIS (Standards Information System) is there to gather it all in one place, in a way that allows the centres to easily update and modify that info, and the user to easily access it in various visualizations.

Motivation for unified format recommendations

- Users: hints on which centre to choose and on the data preparation,
  - but it’s also a matter of outreach
- Centres: among others, it’s an assessment requirement, so why not do that in a uniform way
- CLARIN as an RI needs objective indicators of performance, and the SIS is a way to address one of the KPIs that measures the “percentage of centres offering repository services that have published an overview of formats that can be processed in their repository”

Demonstration

https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

- Note the three levels of recommendation (see the lead text)
- Note the restrictions and sorting
- We have came up with 16 ways in which data can typically be used (+ “other”); we call them Functional Domains
- We record the basic info about formats (not all are clickable yet)
- The system reads centre recommendations from simple “config files”, meant to be updateable easily
- The export uses the same format as the input

What’s next?

- Debug and expand the system (e.g., add more format descriptions)
- Ask centres to participate in updating the recommendations (they are not precise, at this point -- we leave that to the centres themselves)

Some issues:

- (visual) the list is long: should we page the list or provide facets, etc.?
- (operational) can SIS be useful to developers as well, e.g. for gathering info on the recommended media types (note the side menu option)?
- (practical) can the ‘agnostic’ Bagman nevertheless use recommendations expressed in this manner in order to assist the user better?

Useful links

- CSC “visiting card”: https://www.clarin.eu/content/standards
- SIS: https://standards.clarin.eu/sis/ also directly: format recommendations
- Feedback is welcome: https://github.com/clarin-eric/standards/issues
- https://github.com/clarin-eric/standards/wiki/Updating-format-recommendations

Qualified ‘deep’ links are also possible:

recommended-formats-with-search.xq?centre=IDS&level=acceptable

Thanks for your attention!

Please kindly share your feedback

- either via GitHub issues or
- directly with the members of the CSC