Minutes of the CLARIN Standards Committee virtual meeting
Date: 2020-03-09, 12:00 CET

Agenda

Members present
- Excused

Minutes

Notes
- 3. Announcements
- 4. Information on the accepted formats
  - 4.1. State of play for early December 2019:
  - 4.2. New lists since December 2019
  - 4.3. Notes/remarks
- 5. parameters for the unified list

Agenda

(see the Notes section for links and handout-style info)

1. opening, roll call and role call
2. approving the last meeting's minutes
3. *brief* announcements:
   a. Parthenos Standards Survival Kit
   b. ISO proposed work item 24613-6 (LMF-6: syntax and semantics)
   c. ISO proposed work item 24613-7 (LMF-7: inflectional morphology)
   d. intended revision of ISO MAF (Morpho-syntactic Annotation Framework)
   e. other ongoing projects in ISO TC37 SC4
   f. QUEST
4. *** primary topic: reports(/discussion) on the lists of accepted formats from our centres**
5. * establishing the parameters for the unified list (**initial steps**)
6. any other business

Members present

Piotr Bański
Tomaž Erjavec
Francesca Frontini

Hanna Hedeland
Fahad Khan
Penny Labropoulou
Jan Odijk
Jussi Piitulainen
Christian Thomas

Excused

Neeme Kahusk
Dieter Van Uytvanck
Menzo Windhouwer

Minutes

1. Role call
Piotr chairs the meeting, no one able to take the minutes

2. Minutes from 2019-12-03
Piotr asks for approval of the minutes of the previous meeting; the participants approve.
[**Update**, March 9th: a copy of the minutes is now at https://office.clarin.eu/v/CE-2020-1636-Minutes-CSC-2019-12-03.pdf]

3. Announcements (see Notes.3 for background info)
Piotr informs the CSC about the Standardization Survival Kit (SSK) from the Parthenos project; Piotr participated in the work and attempted to have it : Francesca asks how permanent/ephemeral the Parthenos setup is.

LMF-6 (to become ISO 24613-6): Francesca introduces briefly the goals of the ISO LMF-6 proposed work item and mentions a potential primarily hands-on workshop co-located with and following this year's CAC in Madrid, designed both to present the standard proposal to the CLARIN researcher audience and to gather feedback and use cases necessary to define various application profiles (the current state of work is going to be presented in advance, so that the participants can do their homework and present the results).
The workshop proposal is nearly ready to be submitted, but given the current events, it will remain on hold until the situation clears enough to proceed (or not).

LMF-7: Piotr provides brief info, there are also plans to treat LMF-7 similarly to LMF-6, i.e. present it to a wider CLARIN audience and gather feedback

ISO CQLF-2 "Corpus Query Lingua Franca" (ISO 24623-2; project leaders: Stefan Evert and Piotr Bański): under the CD ballot since the end of March 2020.

ISO MAF "Morphosyntactic Annotation Framework" (ISO 24611): new work item proposal is under ballot currently. The suggestion is to thoroughly revise the document and provide TEI encoding for it as the normative serialization. Proposed project leaders: Laurent Romary and Piotr Bański.

Hanna presents the basic info on the QUEST project, where she conducts a survey of standards and formats for audiovisual annotated language data in use within CLARIN centres and other relevant organisations.

4. Information on the accepted formats from our centres (see Notes.4 for background info)

Piotr thanks the members for the effort towards having our centres publish information on the formats that they accept (Jan and Hanna have also worked on this on a wider scale: across the NL centres and among CLARIN developers, respectively).

Jan: My request for information among the NL centres was not successful at all, except for a small update by Paul Trilsbeek for MPI. DANS promised to start working in this, but they are not a CLARIN Centre yet (though they claim they are going to be one). I reminded all and gave them till March 26 (Utrecht meeting)

Christian comments on the information published by the BBAW, and specifically on the need to bear in mind the distinction between formats that the given centre *wants* to accept and the formats that it *has* to accept (and then convert to a/the preferred format). Pasted from a comment: “That's what I wanted to discuss:
one answer could be exactly this: we only accept what we work with ourselves, = TEI-XML/DTABf (and CMDI). another answer would be to update this page {https://clarin.bbaw.de/de/kuration/ | https://clarin.bbaw.de/en/curation/} and list formats we (do not encourage but) have curated in the past and would (reluctantly, if we have the resources) curate in the future. The preferred formats will stay and only be TEI-XML/DTABf (and CMDI)”

[Christian (in a later comment in the googledoc): I shall find out, why there is no mention of "the CLARIN-D/WebLicht Text Corpus Format (TCF)", https://www.clarin.eu/category/glossary/tcf, here https://clarin.bbaw.de/en/repo/ (or here: https://www.clarin.eu/content/standards-and-formats!)
Piotr (also a post-meeting reply): let us keep this as a potential point for a future meeting. My hunch is that it’s rather good to keep TCF as a CLARIN-internal exchange (tool-)format, rather than have people from the outside try to learn it. BBAW offers a TEI-to-TCF converter created by Brian Jurish -- this seems a very reasonable transition strategy.]

[**Update**, March 30th: Dieter has kindly added the links to https://www.clarin.eu/content/standards-and-formats . Many thanks to everyone involved (!) and let us please continue by having the remaining centres publish this information.]

5. Parameters for the unified list

Piotr goes through the content of section Notes.5 below.

6. Any other business
Piotr mentions issues raised by Linda Stokman (CLARIN Office) in connection with the upcoming f2f meeting in Utrecht. Note: this point became moot due to the pandemic and the consequent cancellation of the meeting.

Piotr mentions a reply from the KSIC concerning the CSC initiative concerning the database of expert competences and promises to forward the message to the CSC mailing list. [Update: forwarded on March 10th]

Notes

This section provides the context for the minutes above. Most of it was created before the meeting, as a background/context for the proposed agenda, and got trimmed/modified for the minutes.

3. Announcements

These announcements are not meant for discussion at the telco (unless the CSC wishes otherwise) but rather as information that the members can access here at any point. They do not directly influence our March task, but are something that the CSC should be notified about, and this just happens to be the best opportunity.

3a. Parthenos Standardization Survival Kit: A collection of research use case scenarios illustrating best practices in Digital Humanities and Heritage research:

General link: https://www.parthenos-project.eu/portal/ssk-2
Direct link: http://ssk.huma-num.fr/#/
YT: https://youtu.be/JVJ1dUDpX5I
GitHub: https://github.com/ParthenosWP4/SSK

Relevance for CLARIN: SSK can (and should) reference some of the CLARIN resources. Suggestion: let us all have a look at the SSK *in the coming months* and think of potential new scenarios, especially if they align with the interests of one or more centres. We can then devote some time to discussing them (and about how to delegate the work on them across our centres) at a later telco.

3b LMF-6 (Lexicon Markup Framework: syntax and semantics). Project leader: Francesca Frontini (FR). Stage: early, gathering use cases from the research community.

3c LMF-7 (inflectional morphology). Project leader: Benoît Sagot (FR: AFNOR)

3d revision of ISO-MAF (general updates, TEI serialization). Project leaders: Piotr Bański, Laurent Romary. Ballot expected: [date tba, 12 weeks before the June ISO Conference]

3e state of LMF 1-5, CQLF-2 (ballot expected: date tba, 12 weeks before the June ISO Conference)

3f QUEST: Standards and relevant formats for audiovisual annotated language data (reporter: Hanna Hedeland, URL: https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html)

4. Information on the accepted formats

4.3. Notes/remarks

At the December virtual meeting, the CSC decided to make sure that the centres represented by its members will all provide information on the formats actually accepted by them, on a dedicated page (or in a dedicated section of a page), by analogy to any of the lists in 1-8 above (or in any other way that is deemed reasonable by the persons responsible for data ingest). We are not imposing any template at this point -- merely asking about the actual practice at the given centre.
Hint: replies such as “XML”, “HTML” or “TEI” are probably too general -- unless the given centre really accepts all that these abbreviations may refer to.
We have agreed to ask the colleagues responsible for this to provide this information until mid-February.

5. parameters for the unified list

(We’re going to take the first stab at formulating them, so that the Utrecht meeting can be more concrete.)

General program for Utrecht: start small, start with the core = what _is_
From there, we can move on to what _should be_, in the second step (after Utrecht)

Piotr's **minimalistic expectations** of the Utrecht meeting:

- organize the formats as Dieter did for the KPIs and look for the cut-off points (that should be quick)
- agree on the parameters for use in description of formats (see below)
- decide what to recommend for handling "special issues" (see below)
- publish this under clarin.eu

Potential **parameters** that we need to start the Utrecht meeting with:
(this list is meant as food for thought -- please munch on it until we meet :-))

- status of the format in the field (international standard, local standard, tool format, ...)
- “age” of the format/standard (newly introduced, established, retiring, …)
- preference by the centre: (preferred, unproblematic, acceptable)
- directionality (ingest, availability/export, internal)
- document/data types (e.g. documentation vs. transcription @HZSK need different formats)

Special issues:

- multiple MIME types or extensions for a single format
- some format references are far too broad (“XML”...)
- centres _want_ the data, offering various curation strategies -- should that be part of the main picture or the periphery?