Minutes of the CLARIN Standards Committee virtual meeting, 2019-12-03

Agenda

1. opening, roll call and role call
2. approving the last meeting's minutes
3. announcing the new, approved bylaws
4. reports on the activity since October (if any)
5. * follow-up on the CSC presentation at CAC: reports by the members' own institutions (Piotr will present the issue)
6. * setting up for March
    a. venue, rough dates, length
    b. what we expect to achieve
    c. another virtual meeting (last week of February?)
7. any other business

Members present

Piotr Bański (Germany)
Tomaž Erjavec (Slovenia)
Francesca Frontini (France)
Hanna Hedeland (Germany)
Neeme Kahusk (Estonia)
Penny Labropoulou (Greece)
Karlheinz Mörth (Austria)
Jan Odijk (Netherlands)
Leif-Jöran Olsson (Sweden)
Jussi Piitulainen (Finland)
Christian Thomas (Germany)
Dieter Van Uytvanck (ERIC)
Andreas Witt (ERIC)

Excused

Fahad Khan
Menzo Windhouwer

Minutes

1. Role call: Piotr is the convener, Christian volunteers to take minutes on the assumption that all can correct and extend them.

2. discussion of CAC meeting minutes

- minutes approved by the committee, Piotr will share them with the BoD and the NCF [update: done on 3.12.19, now in archive]
- Jan (continuing topic mentioned in the minutes): Interoperability committee requests feedback on the content of the interoperability web page: https://docs.google.com/document/d/1mJiGQhReA_Km6DIdiNmPpWbVlL2Yz6Fy0tXIQjy2ZZ8/edit?usp=sharing;
  ⇒ @ALL: please add feedback until end of 2019 in the googledoc (even remark if you have no comments)

Andreas: the web page references CLARIN guides directly. These guides have not been reviewed for years. It may be better not to reference them in a way that may suggest that they are current and endorsed.

Jan shares also the Interim Report Interoperability Committee 2019-11-07

3. announcing the new, approved bylaws

The CSC bylaws were approved by the BoD (on Nov 4th) and are in force now. Find them at https://office.clarin.eu/v/CE-2013-0143-Bylaws-standards-committee_v2.pdf

4. reports on the activity since October (if any)

Piotr asks if the participants wish to share news on activities since October (apart from the activity mentioned by Jan).

Piotr mentions two items from the last meetings: (1) questionnaire, (2) presentation space.

(1) Piotr shared the pilot list of expert's competences with Steven Krauwer back in November. Steven put the issue on the agenda of the next meeting of the KSIC. [update from Steven, 10.12.19: the KSIC wasn't able to meet in late November and needed to reschedule to January; Steven is going to keep us updated with what the KSIC decides]

(2) Dieter has given Piotr editing rights to what was described as the "presentation space" for the CSC (https://www.clarin.eu/content/standards). Piotr initially intended to edit that page with some of the past but relevant items (such as namespace assignment), but realised that other groups, notably the Interoperability Committee, may lay claim to it. Jan: the committee will not claim that page. Penny: it would be great to have such a page, with information for the end-user rather than of any committee-internal nature. Jan: it might make sense to prepare an outline in a google doc, before putting it on the Web.

Christian: is this the place where we want to aggregate the info on formats we gather?

Piotr: yes, first edition scheduled after/at the March meeting

Christian: BTW, what's the status of this https://clarin.ids-mannheim.de/standards/views/recommendation.xq;jsessionid=1p06ktktoial3140t8518gl09l?type=fr space?

[can't recall an answer to this question given during the meeting; Piotr (post-meeting): this is a facility that we can use for publishing the results in an attractive/searchable way, eventually. It requires a revision, either wiping the current data (that would be a shame) or marking them as "unreviewed" and including the records gradually, while adding others; we should have a discussion on the future of this system, once we have the data that we want to publish, so let's please hold on to this question, it's definitely a valid one]

Christian: what's the status of this https://www.clarin.eu/content/standard-recommendations? Apparently "January 2009" -- so we should update/replace it?

Piotr: we may want to replace it, when we have the info (meaning: March). Also, these are recommendations, whereas our task for now is to extract the actual state of affairs and create recommendations on that basis (a.o.).

5. follow-up on the CSC presentation at CAC: reports by the members' own institutions

Piotr presents a fragment of a message by Hanna:

"maybe this information could be forwarded as appropriate, since I'm not sure who has been involved in the actual submission for the CTS for their centres and if this is common knowledge in the CSC, but the question of preferred formats is highly relevant for "VIII. Appraisal", with the guidance for R8. including the following questions, for which evidence should be provided for compliance (cf.https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf, p. 19):

- Does the repository publish a list of preferred formats?
- Are quality control checks in place to ensure that data producers adhere to the preferred formats?
- What is the approach towards data that are deposited in non-preferred formats?

This means centres who passed the CTS with a higher level of compliance for this requirement should have this list somewhere (or there might be more information in the CTS assessment)."

In short: it appears that making information about accepted formats (in English) is part of the certification requirements.

Piotr's speculation on the basis of Hanna's earlier work on extracting this information from centres, and on the basis of browsing the individual sites: presumably, many centres have addressed this requirement by pointing to the standards list on the CLARIN pages, although that document is not maintained and far too general for the needs of any particular centre

(https://www.clarin.eu/node/2320) -- this is, incidentally, the same page that Christian has asked about under a different alias (https://www.clarin.eu/content/standard-recommendations). In other words, many centres performed what appeared to be an honest and straightforward action: point to the CLARIN recommendations. Since we are about to update these recommendations, we have to bear in mind the stability of these links. But this ties in with the CTS requirements mentioned above: each centre should publish its own information on what formats this very centre accepts. Our job, in the first stage, will be to extract information from these statements and generalize it across CLARIN in nifty ways.

Later on during the meeting Piotr comes back to this treatment of CTS requirements and asks the committee if the CSC should pursue this in any formal way (e.g. through the Centres committee). Dieter: there seems to be no need for this, given our progress and goals. The issue should resolve itself naturally.

The following centres have published explicit information on the formats that they accept:

1. https://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html
2. https://cocoon.huma-num.fr/exist/crdo/formats.htm
3. http://fedora.clarin-d.uni-saarland.de/ressources/AcceptedFormats.en.pdf
4. https://corpora.uni-hamburg.de/hzsk/en/corpus-hosting
5. https://www.mpi.nl/corpus/html/lamus2/apa.html
6. https://arche.acdh.oeaw.ac.at/browser/formats-filenames-and-metadata#formats
7. https://facile.cines.fr/
8. https://repos.ids-mannheim.de/Formate-Liste.html

Piotr: the first 7 have been taken from https://www.clarin.eu/content/standards-and-formats#formats (this is what Dieter has worked with for his KPI-based research). The AGD project at IDS Mannheim was also represented there, but now, in order to substantiate the suggestion upcoming in the next point, Piotr has asked his colleagues, Denis Arnold and Bernhard Fisseni, to prepare a unified list of formats that the IDS accepts, and this is now offered under (8) in the above list. Piotr wishes to mention that Denis and Bernhard have prepared this despite each of them having been struck by illness, and to thank both of them for managing to get this done in time for the current meeting. This takes us to the next point:

Piotr would like to suggest that we as the CSC should at least reach 100% among the centres that our committee members represent. This is (a) doable within a relative short time (witness the IDS delivery) and (b) would provide a firmer basis for our work in March, and (c) would meet the BoD goal of increasing the relevant KPI (in other words, this is not for the sake of the CSC alone, but also for the sake of CLARIN as such).

Piotr asks if the committee agrees to this suggestion. The committee agrees.

Jan: I do not represent one centre as such, but I am the National Coordinator for NL. Will request this from the centres. When can they expect to know about what information precisely needs to be provided?

Piotr: that would create an inverse loop that we do not want at this stage, because we would like to formulate these expectations in March, when we have talked, a.o., about what parameters need to be mentioned ("preferred/discouraged", etc.). For now, the request that we would like to pose to "our" centres is: on the basis of the actual everyday practice, please state what formats are accepted (for which kinds of data). I think it is best to point to the 8 centres that have done that (cf. the list above) and ask to follow any of these examples. This is to make it as easy on the persons that prepare this information as possible. And please bear in mind that, while this comes as a request from the CSC, we are actually riding upon two horses here: (a) the KPI statistics gathered by the BoD and (b) the CTS certification requirements.

Jan: accepts this reply.

Francesca: agrees that centres decide about this on the basis of their profile, and we look at the result and work on this basis.

[post-meeting remark given some correspondence that ensued afterwards: it may be sensible to limit our effort at this stage to "our" centres, where we can talk to the relevant colleagues directly, and simply "get this done", in a way that is satisfactory to those colleagues, rather than us or CLARIN in general, in time for the March (actually Feb) meeting. Otherwise, we risk falling in that reverse loop, whereby our request is met with a counter-request for specific directions. We don't want to give specific directions at this stage -- we'd rather see what the actual practice is, and generalize from that.]

Time-frame: Piotr has heard from his colleagues that when they hear a date, they conceptualize it as that day plus a week, SO I suggest that we ask the colleagues at our centres to provide the data by mid-February, so that all the customary date (re-)calculations do not harm our meeting schedule. The committee agrees.

Piotr asks the committee about additional issues concerning the lists of accepted formats: format (layout) of the presentation, name of the page (uniform?), language(s) of the presentation.

Format: should be simply a list of what is being used by the resp. centers currently, potentially by analogy to one of the statements linked from 1-8. Anything that feels right to them -- we're asking about actual practice, not for promises on what a given centre might support.

Dieter: as far as names of the respective pages are concerned, it will be very hard to make them uniform (individual centres choose their own CMSes etc.)

Piotr: language(s) of the presentation: for now, any (local) language is natural; but since CLARIN is an international initiative, an English version should probably be offered.

Dieter: this is actually clearly stated in the certification requirements: an English version should exist. The language required is so simple that translations should be a trivial issue.

Penny: converters should be offered for non-preferred formats.

Dieter: better not to conflate the two issues, converters are definitely useful but outside of the immediate scope of our task.

An offshoot of the discussion: non-preferred formats

Note: the statements below started out as a side thread that the half-baked convener was unfortunately quite oblivious to (beyond replying to Christian's question that we *are not* going to dismiss "non-preferred" formats, but rather treat this as one of the parameters -- just like many of the centres listed in 1-8, if not all of them, do). Since this has involved three of the participants (even well after the meeting has concluded) and contains useful data, Piotr suggests that we keep this in the minutes, as part of documentation that we may want to keep handy. Piotr thinks that the original reason for this thread branching off was a reinterpretation of a question *quoted* by Hanna (quoted in turn by Piotr in the Notes section) as a question that Piotr posed to the participants of the meeting. The question was "What is the approach towards data that are deposited in non-preferred formats?", and it is part of the set of questions that appear in the CTS Guidance ch. VIII (link above).

- BBAW's perspective: if worth it, will be curated/converted into preferred format! [e.g. docx=>XML (DTABf), so DOCX is not preferred, but accepted (under certain circumstances)] https://clarin.bbaw.de/de/kuration/ (English version https://clarin.bbaw.de/en/curation/)
- SBX's perspective: Make it simple and base the order of priority for readiness on frequency of materials flowing in. Reasons clear on preferences like BBAW states.
- HZSK's perspective: Since (human) resources are not infinite and data curation of audiovisual data is very costly (since it's very time-consuming and requires expert knowledge), in most cases, researchers need to make sure data is in the preferred formats, but we can give advice and assist in finding further support for conversion and curation tasks. There are guidelines (https://corpora.uni-hamburg.de/pdf/Leitfaden_Aufbereitungsaufwand_und_Nachnutzbarkeit_von_Korpora.pdf) to decide whether to accept legacy data.

6. setting up for March

a. venue, rough dates, length

   ⇒ venue: Utrecht,

   ⇒ rough dates: End of March, after March 16 [Update: doodle has been created and distributed on the CSC mailing list; the selected dates are 26-27.03]

   ⇒ length: lunch to lunch

b. what we expect to achieve

   ⇒ @ALL: each partner/center make list of formats preferred, discouraged etc. (as discussed above) -- this is a prerequisite for the meeting

   ⇒ we need to generalize the information across several parameters that we first need to agree on, we need to think about whether anything obvious is missing, and we should end the 2nd day with the first unified list of formats accepted (with all possible provisos) by CLARIN centres; note that this is not necessarily the same as the list of CLARIN-endorsed standards, but it's a significant step ahead

   ⇒ we need to make at least an initial decision on how we want the list published

c. another virtual meeting (last week of February?)
   ⇒ [Update: doodle is at https://doodle.com/poll/x5i8tnhzzvhwvagk Please fill it in **until January 3rd**]

7. any other business

Piotr asks Dieter if it is still the case that we need to collect information about member activities, for the purpose of reporting to the the BoD. Dieter: given the current level of activity, there is no need for extra reporting work.

Jussi informs the CSC: Started/on-going in FIN-CLARIN to transform our data to other formats, with Parla-CLARIN TEI as a goal for parliamentary data.

[Meeting adjourned, with thanks to all present.]