CLARIN Standards Committee virtual meeting

1.07.2020, 14:00-16:00 CEST

CSC Zoom space: https://clarin.zoom.us/j/222532527

Meeting ID: 222 532 527

Zoom dial-in numbers: https://clarin.zoom.us/u/adaxILtQe

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. Where we are now: a review of our recent extensions to the "KPI spreadsheet"
4. Next steps
5. Any other business
    a. Doodles for August and September
    b. Info on the freshly concluded ISO Conference

Members present

Piotr Bański	DE
Tomaž Erjavec	SI
Francesca Frontini	FR
Hanna Hedeland	DE
Neeme Kahusk	EE
Fahad Khan	IT
Karlheinz Mörth	AT
Jan Odijk	NL
Jussi Piitulainen	FI
Christian Thomas	DE
Dieter Van Uytvanck	ERIC

Excused

Penny Labropoulou	GR
Menzo Windhouwer	ERIC
Andreas Witt	ERIC

Notes

[please consider referencing from the minutes rather than copying]

2. The last meeting's minutes

https://docs.google.com/document/d/1rd_DvVnXQMMOUOUn-Eg6Icugg43_VZiN8c7VHjRqid4/edit?usp=sharing

The KPI spreadsheet is at Formats and Mimetypes

4. Next steps

- How do we abstract away from the data provided by centres to the categories in the spreadsheet? (example: PCM-WAV in AGD vs. ".wav" in the spreadsheet)

5.a Doodles

- August doodle: https://doodle.com/poll/27bt4fg8xf5s7hqd (please fill in **until 06.07**)
- September doodle: https://doodle.com/poll/mhnec3xi764mssnw (please fill in **until 02.08**); note that this is in the second week of September, rather than the first

5.b ISO docs available for consultation

The yearly ISO Conference has just concluded and brought several upcoming standards to a state where they can be shared with CLARIN (under the TC37SC4-ERIC liaison) in expectation of feedback.

- The CQLF-2 (CQLF Ontology) document will be in a shareable state next week; in connection with a GitHub repository that is going to rely on community feedback; Piotr will notify the Committee and share it with the interested members
- The LMF-3 (Etymological extension) and LMF-4 (TEI serialization) documents are in a shareable state as well. Details forthcoming.
- SC4 has approved of a new work item: a revision of the MAF (Morphosyntactic Annotation Framework) standard. So far, only a proof-of-concept TEI serialization is available and nearly in a shareable state. Piotr will notify the Committee when it becomes worthwhile to look at it and provide feedback (time-frame: ca. next week)
- LMF-6 (Syntactic and semantic extensions) may be getting close to a shareable state -- Francesca may be able to say more
- LMF-7 (Inflectional morphology) will probably be in a shareable state relatively soon (the project got delayed). Piotr will investigate and notify the Committee.

Please note that the ISO documents can be shared under liaison while they are in the process of preparation, and with the expectation that the editors of the relevant document will receive feedback from the interested experts with whom the document got shared.

[Update: the documents are/will be available from the semi-restricted page at https://trac.clarin.eu/wiki/ISO_TC37SC4_Standards_in_preparation]

Meeting Minutes, CSC telco - July 1st, 2020, 14:00-16:00

1. Roll call

Members present

Piotr Bański	DE
Tomaž Erjavec	SI
Francesca Frontini	FR
Hanna Hedeland	DE
Neeme Kahusk	EE
Fahad Khan	IT
Karlheinz Mörth	AT
Jan Odijk	NL
Jussi Piitulainen	FI
Christian Thomas	DE
Dieter Van Uytvanck	ERIC

Excused

Penny Labropoulou	GR
Menzo Windhouwer	ERIC
Andreas Witt	ERIC

Piotr chairs, Hanna takes the minutes.

2. The Last Meeting's Minutes

The KPI spreadsheet, Formats and Mimetypes, should be linked in the minutes.

The committee approves the last meeting's minutes, CSC telco - June 4th, 2020 - minutes. [update: available at https://office.clarin.eu/v/CE-2020-1704-Minutes-CSC-2020-06-04.pdf]

3. KPI Spreadsheet Update

Where we are now: a review of our recent extensions to the "KPI spreadsheet"

Technical and Administrative Issues

The colour grey, which showed which formats were only used once in the initial version, is now removed. Hanna suggests conditional formatting.

We should add the URLs to the format lists directly in the spreadsheet column headings and the spreadsheet has to reflect the documents linked in the column heads, no other information should be added/deleted.

Everybody can invite themselves in the sharing settings to enable a proper version history.

Progress and problems

Most columns have been filled in/reviewed, there were some problems that prevented people from doing this.
Christian finishes UdS (unclear who was responsible) and his centres until the next meeting (BBAW had technical issues).
ORTOLANG (as an example) was not very responsive in making their generic recommendations explicit. We can provide examples to the centres to help them create the explicit lists we need.

4. Next steps

The committee encountered several problems in how to interpret the task at hand. They were discussed, partly with conclusions.

Use Case and Findings

We were supposed to only focus on “data formats”, but this is also not completely clear. Apart from annotation data e.g. dialectologists need KML. Schemas and formats for settings and visualization/analysis (e.g. DTD, XSD, XSLT or ELAN settings files) are already partly in the list belonging to a category called “Text/Markup related” and can be included if recommended. We decided to not consider tagsets and similar standards yet.

We decided (again) to focus on recommended formats only, we want to reflect what centres want, not what they have to accept because users bring these formats. The summarizing column “support”, should maybe rather be called “recommendation”.

We need to find out what low recommendation/support numbers (1) mean, either by asking the centres or by publishing/disseminating summarized lists/reports and hope for reactions.

Granularity/Specificity

The degree of detail varies between centres when it comes to format definitions, we need to map information provided by the centres to the more coarse-grained categories in the spreadsheet (e.g. (L)PCM-WAV vs. WAV). The problem of varying format parameters and varying degrees of specification arises both with non-resource-type-specific formats such as plain text (e.g. encoding), PDF, audio or video (various quality related measures), and for formats used specifically for language resources. Other institutions are already defining and maintaining standards and best practices for archiving the non-resource-type-specific formats. The CSC can focus on the resource-type-specific aspects (e.g. audio quality necessary for ASR) and on resource-type-specific formats.

When it comes to resource-type-specific formats, TEI and its dialects is the main example, it would be helpful to be able to group e.g. individual TEI dialects or other similar formats in the list. Jan prepares a solution to include more information on hierarchies/families of formats in a separate tab of the spreadsheet, this is partly reflected in the Image/Raster vs. Image/Vector categories already.

Recommendations for Format Lists

Based on the discussion and the problems encountered, the CSC can provide centres with some advice on how to create format lists:

- include a timestamp/version number
- use English language (in addition)
- define used terms like preferred/recommended/accepted
- use an appropriate degree of detail (XML is usually not specific enough) and recommend best practice format parameters (e.g. plain text in UTF-8 where possible).

5.a Doodles

Everybody needs to fill in:

- August doodle: https://doodle.com/poll/27bt4fg8xf5s7hqd (please fill in until 06.07)
- September doodle: https://doodle.com/poll/mhnec3xi764mssnw (please fill in until 02.08); note that this is in the second week of September, rather than the first

----The end---