SCL Test Science Projects for EOSC Future (WP6.3):

COVID-19 metadata findability and interoperability in EOSC (META-COVID)

Structured answer sheet for semi-standardised interviews

Date:	29/11/2022
Time:	10:00-11:00 CET
Interviewer(s):	Christian Ohmann (ECRIN), Maria Panagiotopoulou (ECRIN)
Interviewee(s):	Daan Broeder (CLARIN), Walter Daelemans (UA), Pieter Fivez (UA)
RI of interviewee(s):	**CLARIN**

1. Objective aspects of the use of contextual metadata in the RI's domain

1.1 What does 'contextual metadata' mean to your RI?

How is the RI organising its services and tasks? What does that mean for contextual metadata that are directly applied within and by the RI?

Metadata for language resources and tools exist in a multitude of formats. Often these descriptions contain specialised information for a specific research community (e.g. TEI[1] headers for text). To overcome this dispersion CLARIN has initiated the **Component MetaData Infrastructure (CMDI)**. The CMDI provides a framework to describe and reuse metadata blueprints. Description building blocks ('components', which include field definitions) can be grouped into a ready-made description format (a 'profile').

In order to promote the reuse and sharing of 'components' and 'profiles', the CMDI **Component Registry** was created. A web application (https://catalog.clarin.eu/ds/ComponentRegistry/#/) allows metadata modellers to browse through all existing components and profiles and to create new ones, with the possibility of including existing components. The Component Registry is open to anyone to read components. Submitting new components can only be done by accredited experts to guarantee that only correct and proven components are ready for reuse by others. Despite that, component and profile proliferation is still a problem for CLARIN and reusing existing components is something that should happen more.

More detailed documentation of the Component Registry can be found here: https://www.clarin.eu/content/component-registry-documentation.

[1] TEI: Text Encoding Initiative, More info: https://en.wikipedia.org/wiki/Text_Encoding_Initiative

The CLARIN **Concept Registry** (available here: https://concepts.clarin.eu/ccr/browser/) is used to create Concept Links in the CMDI.
Researchers who wish to make their resources available through CLARIN should map and transform their metadata to CMDI. CLARIN is providing some components that act as "mapping" themselves such as from OLAC or from Dublin Core.

In what concerns 'contextual metadata', if these play a key role in language resources description then one should be able to find such in 'components' that express this. Inspection of the components learns that contextual information is usually distributed over several components in a profile.

Preliminary list of elements of contextual metadata applied within and by the RI:

Language materials are collected in different ways. For instance, there is the term "data elicitation" or "data elicitation circumstances". Some researchers and some technical staff are using this metadata field and they provide values such as "postal elicitation", "telephone elicitation", "naturalistic observation" etc. Secondly, when experiments are performed there is a field "experiment type" with values such as "Wizard-of-Oz", where the researcher is using some kind of an intermediary contraption or screen to talk to the subject. Another example is the "contextual" information collected on a subject that speaks, writes or uses sign language: "What is this person's mother tongue?", "Is he/she bilingual?", "What is the educational level of the parents?" etc. There are plenty of metadata values available for describing a speaker or an assigner.

Looking at language technology, Natural Language Processing (NLP) and machine learning there has been highlighted the issue of replicability and it is expected that over the following years there will be a trend of collecting more "contextual metadata" mainly to allow other researchers to replicate studies. 2 elements of contextual metadata collected are: i) *the data collection protocol* and ii) *the data annotation guidelines*. A lot of contextual information can be found also in the "GitHub code". This does not only include the code used but also the methodology applied and the "run of the programme" and how it has been divided into "train" and "test". The level of technical detail found in the GitHub repository is usually not present in scientific publications.

In CLARIN, *data collection protocols* are available as resources pointed to in the CMDI metadata. However, the standardisation of these protocols remains a problem (maybe they are standardised at a research team or institute level but no wider levels of standardisation apply). The findability of the protocols is also an issue as in many cases they are not issued with the proper metadata.

A first step of metadata standardisation in the deep learning community is the requirement of completing a questionnaire in Hugging face (e.g. https://huggingface.co/datasets/clips/VaccinChatNL) with an explicit description of the dataset that is becoming available through the platform. For answering the questionnaire there is no standardised vocabulary used.

The need for "standardisation" in research activity in the field of NLP is exemplified by the recent addition of "Frequently Asked Questions" in the appendix of publications from influential actors in the field (such as Jason Wei from Google Brain).

What kind of contextual metadata are used in the domain represented by the RI:

1.2 What elements of contextual metadata of the resources/digital objects are modelled in the metadata schemas applied at your research RI (research organisations, researchers, services, projects, funders, etc.)?

List of elements of contextual metadata that are modelled at the RI with a reference to the metadata schema used (ask whether the contextual metadata element is already applied by the RI or whether it is foreseen but not yet implemented):

Common elements of metadata as well as a minimum set of metadata is not available due to the heterogeneity of the domain.

Not even “language” because there have been examples in the past of studies conducted in only one country (the Netherlands), in which case the researchers were not providing information on the “language” considering it self-explanatory (that it will be Dutch). This of course can become problematic when metadata from different studies in different countries need to be integrated. Similarly, Emily Bender proposed what we call the #BenderRule highlighting that the language of the data that scientists work on should be stated and not implied.

There is work within CLARIN towards approved metadata components that should always be there but this is a complicated task. Especially providing metadata for data collected retrospectively is both difficult and expensive. The META-SHARE infrastructure (https://aclanthology.org/L12-1647/) made additional standardisation efforts in the field that could be looked into.

Are there contextual metadata elements, which are important but not used in your RI (gaps)?

N/A as highly dependent on the study in question.

1.3 What services, protocols, standards, APIs are implemented in your RI to support harvesting of contextual metadata from outside (e.g., public or non-public API)?

Which metadata standards/schemas/protocols are used in your RI?

CLARIN harvests metadata from its centres using OAI-PMH API methods.
In collaboration initiatives CLARIN also harvests metadata for example from Europeana (https://www.europeana.eu/en), focusing on Europe’s digital cultural heritage domain.

There might be a couple of additional projects that contribute also to the harvested metadata displayed in the Virtual Language Observatory (https://vlo.clarin.eu/?1)

Does your RI provide metadata services (which)?

Component MetaData Infrastructure (CMDI): https://www.clarin.eu/content/component-metadata
Component Registry: https://catalog.clarin.eu/ds/ComponentRegistry/#/
Concept Registry: https://concepts.clarin.eu/ccr/browser/
Virtual Language Observatory: https://vlo.clarin.eu/?1

Are APIs implemented and used to support metadata harvesting of contextual metadata from outside?

APIs for the CMDI are also available. Although CMDI is the “metadata framework” and not contextual metadata as such.

1.4 Are the contextual metadata used in your RI already linked to a research process graph or is it planned to do so?

Are you familiar with research (process) graph approaches?

Yes

Which type of research (process) graph is already in use in your RI or planned to be used?

OpenAIRE: No current mapping to the OpenAIRE RG but maybe in the future.

Some individual CLARIN centres are being harvested by OPENAIRE but this is currently an "individual centre" decision and is not applicable to all CLARIN centres. This is especially true for centres that have double allegiance and are both part of CLARIN and DARIAH.

PID graph: N/A

Open Research Knowledge Graph (ORKG): N/A

Any other research (process) graph: The Dutch CLARIAH project (https://www.clariah.nl/) did work that is close to a "research graph". The CMD2RDF service was created to allow connection with the growing Linked Open Data (LOD) world, and facilitate experiments within CLARIN merging CMDI with other, RDF based, information sources (https://portal.clarin.nl/node/4226).

Is or will the research (process) graph implemented or to be implemented in your RI cover your elements of contextual metadata adequately?

N/A.

Some potentially interesting information on contextual metadata can be found in the DMPs. Despite these not always being standardised. An example of DMP standardisation is the OPENAIRE ARGOS tool.

2. Opinion-based and subjective views of the interviewees about use and potential value of contextual metadata in their scientific domain

2.1 Do you believe that a greater generation and use of contextual metadata would be valuable enough to justify the additional effort that would likely be involved?

Yes/no/undecided

It is seen as necessary, especially to improve research replicability and reliability. In NLP there is currently a trend of higher generation of contextual information and this is expected to improve in the following years.

Do you think that your opinion is also covering the stakeholders of your RI?

Yes.

2.2 From your viewpoint how could interoperability for contextual metadata between RIs be improved?

If we aim at having a unique research metadata component we will not succeed. We need to understand and respect the differences in each discipline. Mappings between vocabularies can be created as needed to improve interoperability when common metadata elements are not sensible.

Nevertheless, the disciplines still have a lot to learn from one another and exchanging best metadata practices and giving recommendations to one another for improvement is very relevant. Currently, the scientists and infra specialists are missing this “metadata overview” and how “research activity” or “context” is handled across disciplines.

The interviews are a good starting point but we should get explicit as a next step and maybe propose metadata attributes and values and see how these can overlap across different domains. Maybe even perform a consensus exercise by scoring the different proposals.

A “community approved” metadata schema would be of great value for other EOSC projects.

2.3 What could be the best organisational framework for moving this work forward within EOSC?

Integrating into EOSC core services: N/A

Onboarding to EOSC: N/A

Registration in EOSC-catalogue: N/A

Provide EOSC interoperability profile: N/A

Provide input into EOSC-Association task forces: N/A

Other possibilities within EOSC: If we succeed in obtaining consensus across disciplines on common metadata elements these should be published and made also widely available e.g. in FAIRsharing or in the RDA metadata registry. Also disseminated widely within EOSC and to EOSC related projects. Becoming explicit on interoperable vocabularies would be a powerful outcome.