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:	14/10/2022
Time:	10:30-11:30 CEST
Interviewer(s):	Christian Ohmann, Maria Panagiotopoulou, Steve Canham (ECRIN)
Interviewee(s):	Mari Kleemola, Katja Moilanen (CESSDA/TAU-FSD)
RI of interviewee(s):	**CESSDA/TAU-FSD**

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

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

Contextual metadata covers the resources, structures, and processes applied in social sciences and humanities (SSH). Contextual metadata is needed to understand the meaning of the data (e.g. *how was the data collected*, *when was the data collected*, *who funded the data collection* etc.). Social Sciences are primarily dealing with people's opinions and feelings. Specific techniques, such as surveys and interviews are applied. Interestingly, in SSH there is the concept of "paradata"(https://doi.org/10.5281/zenodo.4299076) including information collected during a survey/interview as a "by-product", such as for example the time a person needs to respond to a question. "Paradata" is collected through computer models but such information is not present in the CESSDA catalogue.

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?

In CESSDA, information about "funders" is available and the "research activity" is described at a high level. Information about the "researcher" and "other contributors" such as "data collectors" (e.g. sampling companies) is also available. Contextual metadata describing the research approach is most commonly provided in the form of text (e.g., abstract, linked publication). Some datasets are collected with the use of a detailed protocol but others e.g. resources from statistical agencies or other administrative bodies are structured quite differently. In some cases, "participation criteria" will not exist (e.g. when it comes to internet surveys) and whoever is interested can participate . Nevertheless, "sampling" remains the usual way of selecting participants. The population expected to participate is most of the time quite large e.g. persons aged 18 years and older (*exception: Denmark (18-79), Finland (15-74), South Africa (from 16 years upwards), Suriname (21-74), Norway (18-79) and*

Sweden (18-79 years) (ISSP 2019 https://search.gesis.org/research_data/ZA7600)). This is quite different to clinical trials where detailed inclusion and exclusion criteria apply.

The contextual metadata must be provided by humans, especially by researchers and data curators, which is a bottleneck, especially as DDI and DDI vocabularies are not always "straightforward" to use. With adequate tools, data are collected automatically and effortlessly and they can then feed into the metadata schemas. Such a process would help to reduce the effort and resources for providing structured contextual metadata.

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

In the CESSDA Metadata Model, a lot of contextual metadata elements are applied. This covers, for example, main researcher, organisation, funder, contributors, topics, keywords, time-method, country, area, unit of analysis. The CMM model has many more fields but not all metadata fields can be delivered in the CESSDA Data Catalogue (DDI).

The European Language Social Science Thesaurus (ELSST) is recommended by CESSDA for data discovery across Europe. ELSST is a broad-based, multilingual thesaurus for the social sciences.

The characterisation of background variables (e.g., gender) is not done currently in a way that machines could easily use in SSH. In order to get precise information (e.g., about gender distribution) manual work has to be done and it may be that the data source needs to be searched.

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

See answer to the question above.

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):

Contextual metadata in CMM	Implemented in CESSDA Data Catalogue	Based on
Study Number (Unique archival number)	yes	DDI
Study Title	yes	DDI
Subtitle of the study	no	DDI
Alternative Title of the study	no	DDI
Funder	on progress	DDI
Grant Number	on progress	DDI

Grant Title	**no**	**-**
Principal Investigator / Author	**yes**	**DDI**
Publisher	**yes**	**DDI**
Publication Date (of the dataset)	**no**	**DDI**
Study Version (version number, version date and reason for versioning the study)	**no**	**DDI**
Study PID and the type of the PID	**yes**	**-**
Contributor (e.g. other authors, producer, depositor, distributor, data collector)	**no**	**DDI**
Abstract	**yes**	**DDI**
Study Topic [vocabulary used e.g. CESSDA Topic Classification]	**yes**	**DDI**
Keyword [vocabulary used e.g. ELSST]	**yes**	**DDI**
Time Method [vocabulary used DDI Time Method]	**yes**	**DDI**

Study Area as a Country	**yes**	**DDI**
Universe (desription of the population)	**on progress**	**DDI**
Unit of Analysis [vocabulary used DDI Analysis Unit]	**yes**	**DDI**
Type of Data Source [vocabulary used DDI Data Source Type]	**no**	**DDI**
Sampling Procedure [vocabulary used DDI Sampling Procedure]	**yes**	**DDI**
Mode of Data Collection [vocabulary used DDI Mode of Collection]	**yes**	**DDI**
Data Collection Period (either start + end dates or single date)	**yes**	**DDI**
Data Access (freetext)	**yes**	**DDI**

Data Access Conditions	**no**	**DDI**
Metadata Access Conditions (Study)	**no**	**-**
Metadata Access Conditions (Questions)	**no**	**-**
Type of Instrument (DDI Type of Instrument) (instrument in social sciences is e.g. questionnaire)	**no**	**DDI**
Instrument Language (language of the questionnaire for example)	**no**	**-**
Instrument Source (description of the instrument source when instrument is based on some other instrument(s))	**no**	**DDI**
Instrument PID, type of the PID and URL related to PID	**no**	**-**
Instrument Description	**no**	**DDI**
Title of the document that is related to this study/dataset	**no**	**DDI**
URL of the Document that	**no**	**DDI**

is related to this study/dataset		
Format of the document that is related to this study/dataset	**no**	**DDI**
Publication (e.g. article, book) in which dataset is used	**on progress (there will be bibliographic citation in CESSDA Data Catalogue which not a field in CMM but which is possible to be concatenated from the various fields related to publication - not all of them are copy-pasted into this table)**	**DDI**
Publication - Author	**no**	**DDI**
Publication - Title	**on progress**	**DDI**
Publication - Year of publication	**on progress**	**DDI**
Publication - PID, PID type and PID URL	**on progress (PID and PID type)**	**-**
Publication - URL	**on progress**	**-**
Study Group (e.g. series is one type of the study group) Name	**no**	**DDI**
Study Group Description	**no**	**DDI**
Study Documentation Copyright	**no**	**DDI**
Study Documentation Publication Date	**yes**	**DDI**
Study Documentation Publisher	**yes**	**DDI**
Metadata Access Conditions (all)	**no**	**-**

See answer to the question above.

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

Contextual information is well present in the metadata schemas applied in CESSDA. Some gaps identified during the interview would include improving the machine-actionability, including the provision of background information such as gender in a structured way, and introducing PIDs to fields which currently do not have them(as not required by DDI).

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?

- CESSDA has a metadata profile (subset of CESSDA Metadata Model) used for the CESSDA Data Catalogue: https://cmv.cessda.eu/profiles/cdc/ddi-2.5/1.0.4/profile.xml

- The CESSDA Metadata Model (subset of DDI): https://zenodo.org/record/4751455

- User Guide for CESSDA metadata model: https://zenodo.org/record/4672248

As mentioned, the CESSDA Metadata Model is subset of DDI.
DDI documentation:

DDI2.5 (Codebook):
https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation_files/schemas/codebook_xsd/elements/codeBook.html
DDI3.3 (Lifecycle):
https://ddialliance.org/Specification/DDI-Lifecycle/3.3/XMLSchema/FieldLevelDocumentation/

Does your RI provide metadata services (which)?

CESSDA resources can be accessed in the CESSDA Data Catalogue: https://datacatalogue.cessda.eu/

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

Yes, an API to CESSDA resources is implemented (available here: https://api.tech.cessda.eu/). There is also OAI-PMH endpoint for metadata, but only part of the resources are covered (around 50%).. This is still work in progress.

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: CESSDA is compatible with OpenAIRE and there is ongoing work to link CESSDA resources with the OpenAIRE RG. There are some issues to overcome related to PIDs (not always required by DDI) and the way that funder information is presented.

PID graph: N/A

Open Research Knowledge Graph (OKRG): N/A

Any other research (process) graph: N/A

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

As mentioned, the link to OpenAIRE will be established soon but the interviewees believe that the OpenAIRE RG does not cover a useful range of "contextual" metadata for research purposes currently. The OpenAIRE RG is expected to be updated and include more contextual information in the future.

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

- If yes, can you describe the specific contextual data points and possible relations that would be of most value, if available and / or used more widely, and why?

CESSDA sees an added value in upgrading the contextual metadata in the field of SSH. Both CESSDA and the SSH field have already a lot of contextual metadata available but there remain issues around quality and completeness. This is especially the case with many "legacy metadata" which are mainly addressed to "human readers" and are not "machine-actionable". Generating structured metadata requires a lot of "human curator" input but it would offer improved discoverability of resources to the users. There is a need also to include PIDs to certain fields (PIDs are not always required by DDI).

- If no, can you explain in detail why not?

N/A

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

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

The interviewees believe that full compatibility of contextual metadata across different RIs in different scientific fields is unachievable. The **level of interoperability** for contextual metadata that we want to achieve, and that is reasonable to achieve, should be discussed and explored. If it is only at a very high level and abstract, it may not be useful. Mappings between metadata schemas (crosswalks) may help but due to different concepts and meanings, it will remain a complicated task. We should start by checking the meaning of "fields" and try to find "commons" that will allow us to set the limits to the mappings. We should stress that individual RIs have their specific processes and methodologies that cannot be mapped easily to one another.

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: Vocabulary services can possibly be onboarded in EOSC.

Registration in EOSC-catalogue: A registry of vocabularies in EOSC would be a useful source. This is partly covered by FAIRsharing.

Provide EOSC interoperability profile:

This could be relevant for the work of this TSP but it is not clear to the interviewees how the interoperability profile will be helpful in the "day-to-day" work. There are a lot of documents (e.g. deliverables) produced but we have not seen so far an impact on practical work.

Provide input into EOSC-Association task forces:

One interviewee is already a member of an EOSC Task Force (TF FAIR Metrics and Data Quality) and thus aware of the type of input that Task Forces can provide to the EOSC. It would be worth making a link of this TSP with the EOSC Task Force on TF Semantic Interoperability.

Other possibilities within EOSC:

A registry for metadata vocabularies should come from EOSC. FAIRsharing is providing basic links between standards, policies, databases and collections but its exact role within the EOSC environment is unclear. Also, currently, there is too much overlap within EOSC on metadata activities that are not well coordinated. A specific gap is that the EOSC does not provide guidance on which metadata schemas to use and how.

Most work around metadata has a "project-dependent" sustainability model. For example, the DDI-community gets a lot of metadata work done on a voluntary basis from in-kind contributors but the in-kind contributions don't always provide the optimal continuity. Given the fact that a lot of resources are needed for contextual metadata work, we could explore how such work can be better funded and sustained.

Researchers were not so interested in metadata in the past but the culture has changed and the willingness to share data and do "open science" has increased. This may be a chance to improve work on (contextual) metadata.