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:	10/01/2023
Time:	13:00-14:00 CET
Interviewer(s):	Christian Ohmann (ECRIN), Maria Panagiotopoulou (ECRIN)
Interviewee(s):	Gary Saunders (EATRIS), Maddalena Fratelli (EATRIS), Jing Tang (EATRIS)
RI of interviewee(s):	**EATRIS**

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?

EATRIS is fully committed to open science and works on improving the interoperability and accessibility of research data in the field of translational medicine. Best practices are followed to ensure that data and metadata are as FAIR as possible: good documentation, proper curation, and availability in relevant repositories. The suite of policies and guidelines for data and metadata management is put into practice through the different projects in the EATRIS portfolio. Examples of such projects include: EOSC Future[1], REMEDi4ALL[2], and EOSC-Life[3].

In EOSC-Life the MICHA (Minimal information for Chemosensitivity Assays) platform was launched as a data integration platform to FAIRify chemo-sensitivity assays. MICHA is available at https://micha-protocol.org/ for the annotation of drug sensitivity screens for cell lines and patient-derived samples. MICHA aims to address the replicability issue of drug sensitivity assays due to a lack of sufficient annotation and standardisation.

In the field of translational medicine, important contextual metadata remain as lab book notes and are as such not collected, standardised or shared in the most useful, or sustainable, ways. Sometimes, useful contextual metadata is added at the step of the data deposition to relevant repositories as a requirement of the latter. When it comes to which repositories are used in the field and thus which contextual metadata information is collected, it would depend very much on the type of data. For example in genomics there is ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress), the

[1] https://eoscfuture.eu/
[2] https://remedi4all.org/
[3] https://www.eosc-life.eu/

Gene Expression Omnibus (GEO)-NCBI platform[4], the Sequence Read Archive (SRA)-NCBI platform[5], the European Nucleotide Archive (ENA)[6] or sometimes also Zenodo[7] as a more generic research repository. Of course, having the contextual metadata collected in proper format at the beginning of the experiment and not at the end during the data deposition at repositories would be a much more useful practice, but this is how the field currently operates. It is EATRIS policy to promote the reuse of existing metadata and data repositories rather than to reinvent the wheel and create anything new.

From an RI perspective, EATRIS works on catalogues at the meta-level. For example, in REMEDi4ALL there is a WP responsible for gathering suites of IT tools that are used in the drug repurposing pipeline (currently listing >100 of them). In addition, EATRIS is developing toolboxes that include repositories that are favoured by the EATRIS community of users, for example, the multi-omics toolbox (currently in beta testing) developed in the EATRIS-Plus project[8], the innovation management toolbox developed in the EJP RD[9], the Patient Engagement Resource Center also developed in EATRIS-plus, the sensitive data toolbox (beta version) developed in EOSC-Life. These resources detail the resources of best practice in translational medicine processes for key EATRIS use cases.

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

For MICHA, the most important contextual elements for chemo-sensitivity assays have been standardised but each type of experiment is different and standardisation for other types of experiments remains a relevant need. For a typical drug sensitivity screening experiment there is a need to annotate 5 major components: 1) *compounds*: standard InChiKey, name, smiles; 2) *samples*: cell line (e.g. name, provenance, passage number etc.) or human biological sample (age, sex, diagnosis etc.); 3) *reagents* (e.g. format of the assay, detection technology etc.); 4) Experimental design (e.g. (e.g. replicates, tested concentrations, incubation times); 5) *data processing method* (e.g. analysis metric such as DSS or AC50).

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

The depth of the contextual metadata available is dependent on the required fields of the repository platform where this data is deposited. This information can include: What is the biomaterial, what is its origin, what type of assay was used and details on the experimental protocol but the level of detail is very variable and thus often not enough for other researchers to repeat exactly the experiment, causing a reproducibility problem in the field. In some repositories, contextual information around data accessibility is also collected. This is the case for example in SRA and ENA where data access requirements are detailed.

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):

For MICHA, the contextual metadata collected is listed under question 1.1.

---

[4] https://www.ncbi.nlm.nih.gov/geo/
[5] https://www.ncbi.nlm.nih.gov/sra
[6] https://www.ebi.ac.uk/ena/browser/home
[7] https://zenodo.org/
[8] https://eatris.eu/projects/eatris-plus/
[9] https://www.ejprarediseases.org/

GA4GH (https://www.ga4gh.org/) strives for standardisation in the field of genomics. Similarly, there have been standardisation efforts in the reporting of experiments using animal models from the ARRIVE initiative (https://arriveguidelines.org/). Still, nevertheless, FAIRification tools and standards are underused and contextual metadata is collected but in a very variable manner.

Contextual information about the funders of research that resulted to the data is not typically collected in translational research. The information about the organisation that generated the data is available, but this is due to the fact that repositories collect such details for the uploaders. Information about the project that generated the data or collaborating partners in the data generation is not typically collected. This information is provided in the paper, at the time of publication.

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

Providing a detailed research protocol is a contextual element that should accompany the data in translational research but this is not always the case. An identified gap concerns "data access and reuse" metadata relating to ELSI, such as for what applications data can be reused, how can potential reusers access this data etc. There have been attempts to include such information, for example from GA4GH groups but a standardised and automated way to collect "data access and reuse" metadata is currently missing.

The EATRIS work within EOSC Future aims to create research graphs specific to translational medicine. This activity started by looking at what services would make sense to be linked from a research perspective and how to link them. The platform supporting this work is Neo4j (https://neo4j.com/).

One problem is that in most cases FAIRification tools are not used as much as they could be by the relevant scientific communities, and to reverse this situation more support would be needed from funders and publishers as they have the "carrots and the sticks". For example, if research is supported financially by a certain funder, this funder should try to impose standardisation and FAIRification.

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?

Dependent on the data type and selected repository for the data deposition.

Does your RI provide metadata services (which)?

MICHA is a service already available to the community. Different toolboxes are currently under development in EU projects and will become available such as the multi-omics toolbox, the innovation management toolbox, the Patient Engagement Resource Center, the EOSC-Life sensitive data toolbox. In addition, within EOSC Future WP6 a research graph specific to translational medicine is currently being developed in collaboration with EU-OPENSCREEN.

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

APIs for MICHA are available and detailed here: https://api.micha-protocol.org/.

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: N/A. Might become relevant later when the research graph for translational medicine will be ready.

PID graph: N/A. The majority of the MICHA components have already sort of permanent identifiers, for example for compounds the compound ID and Smiles ID. Each type of assay also has standardised identifiers, more in the form of a vocabulary, so each assay is associated with a specific term and this term is unique to describe the specific assay. Thus, separate permanent identifiers within the system are not applied.

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?

Within EOSC Future WP6 there is currently ongoing development of a research process graph focused on the field of translational medicine and looking into relevant contextual metadata addition.

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?

The unmet need for standardisation is what led to the development of MICHA but more support is needed from funders and publishers to convince users that for addressing the reproducibility issue in the field, the metadata collected for each experiment need to be better standardised.

One of the EATRIS core services is consortium building and a lot of efforts have been put in over the past years to help create consortia for grant applications and the requests with which the RI was faced often focused on: How do we find people involved in x/y/z research type or how can we find people that used the x/y/z service. Mapping these elements will facilitate the operations of the RI.

As previously mentioned also, having the research protocols available and information on “data access and reuse” is also seen as important for achieving FAIR in practice.

- 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?

Yes.

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

EATRIS has joined the EU-AMRI[10] together with BBMRI and ECRIN as it strongly believes that cooperation between these 3 medical Research Infrastructures can accelerate scientific developments in the field of health. This does not only concern the "business development" perspective but aims to address the concrete needs of the community of users. For example, when looking into what the EATRIS users are trying to achieve, often services from more than one RI are required. But apart from intra-science cluster coordination, inter-science cluster coordination is also seen as important. Mapping the relationships between Research Infrastructures and the services that each one offers and how these can be linked or are already linked is seen as a useful exercise for the future. Regarding inter-cluster linking, EATRIS is currently in discussions with CERN in the frame of a joint project looking into building AI solutions for stroke. The inter-cluster INFRAEOSC projects are also a milestone in improving coordination and interoperability across scientific fields.

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

EATRIS is an observer in the EOSC association and has representation in different task forces. Metadata work for EATRIS is also targeting the different INFRAEOSC Calls.

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: Having a Task Force or Working Group where the ESFRIs are represented and provide feedback on the different aspects of the EOSC development is seen as a good way of integrating what happens on the individual RI level with what happens on the EOSC level.

Other possibilities within EOSC:

The RIs have concrete services and tools that could be offered to the EOSC. In the life sciences cluster, the RIs have already catalogued such resources in FAIRsharing[11] and other relevant catalogues in their field of activity. The current requirement of having to onboard separately such services and tools to the EOSC portal is seen by many as a duplication of efforts. A good way forward could be for the EOSC to develop mechanisms of cooperation with such catalogues and harvest the information to semi-automatise the onboarding process.

---

[10] https://eu-amri.org/

[11] https://fairsharing.org/