ARPHA Preprints

PENSOFT

Grant Proposal

Author-formatted document posted on 14/06/2023

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e107873

Improving COVID-19 metadata findability and interoperability in the European Open Science Cloud

Christian Ohmann, Steve Canham, Kurt Majcen, Petr Holub, Gary Saunders, Jing Tang, Tanushree Tunstall, Philip Gribbon, Reagon Karki, Mari Kleemola, Katja Moilanen, Walter Daelemans, Pieter Fivez, Daan Broeder, Franciska de Jong, Maria Panagiotopoulou

ARPHA Preprints *Author-formatted document posted on 14/06/2023*. DOI: https://doi.org/10.3897/arphapreprints.e107873

Title: Improving COVID-19 metadata findability and interoperability in the European Open Science Cloud

Authors: Ohmann, C., Canham, S., Majcen, K., Holub, P., Saunders, G., Tang, J., Tunstall, T., Gribbon, P., Karki, R., Kleemola, M., Moilanen, K., Daelemans, W., Fivez, P., Broeder, D., de Jong, F., Panagiotopoulou, M.

Abstract

This publication details the workplan of the Science Project (SP) “COVID-19 metadata findability and interoperability in EOSC” (short: META-COVID) that is part of the Horizon Europe funded project EOSC Future. The COVID-19 pandemic has generated a huge variety of research activities, studies, and policies across both the life sciences (LS) and the social sciences and humanities (SSH). Useful insights from combining the data and conclusions from these different forms of research are, however, hampered by the lack of a common metadata framework with which to describe them. This is because different scientific disciplines have different ways of organising research activities. For example, the type of the research (e.g., hypothesis testing versus hypothesis generating) and the methodology chosen (e.g., experimental, survey, cohort, case study) are key elements in understanding the data generated and in supporting its secondary use. Another issue to be tackled is the integration of various sources of metadata related to parliamentary and social media metadata. In META-COVID, scientists from the LS and SSH domains gathered to discuss ways in which metadata could go beyond the description of the data itself to include the basic elements of the research process (“contextual metadata”) within the frame of the European Open Science Cloud (EOSC). The main outcomes of the SP will be: i) An inventory of metadata schemas applied across infrastructures and domains; ii) The development of a framework for a metadata model characterising the research approach and workflow across research infrastructures; iii) The application of the framework to selected COVID-19 use cases; iv) The development of an ontology of COVID-19 related topics from parliamentary data and social media.

Key words

EOSC Future; Science Clusters; Science Projects; FAIR principles, Contextual Metadata, Life Sciences, Social Sciences and Humanities, COVID-19

Description

Existing situation: The COVID-19 pandemic has generated a huge variety of research activities, studies and policies across both the life sciences (LS) and the social sciences and humanities (SSH): examples include genomic sequencing, assays of immune response, clinical trials, population health analyses, exploring vaccine hesitancy, investigating the role of social media, public debate and economic analyses of the impact of public policy issues (e.g., lockdown measures, imposed face masking). Potential insights from combining the data and conclusions from these

different forms of research are, however, made more difficult by the lack of a common metadata framework with which to describe them. Even within clusters (e.g., SSHOC [1], EOSC-Life [2]), the metadata landscape is heterogeneous and numerous domain-specific standards are applied, as recently demonstrated by SSHOC [3]. The situation becomes even more complicated when data sharing is performed across broad disciplinary boundaries, as in this SP, which spans life sciences, social sciences, and humanities. Developing widely applicable metadata is a key part of rendering data more valuable, by allowing them to be more easily found and characterised, regardless of the discipline in which they were generated. In the context of the European Open Science Cloud (EOSC), this improves data reuse within and among scientific clusters.

There are certainly metadata schemas (such as DataCite, or the Data Documentation Initiative-DDI) that can describe the concrete outputs of research, e.g., papers and datasets, but relatively little work has been done on finding a metadata schema for the research itself. The problem is that different disciplines have vastly different ways of organising research activities, for instance because of differences in funding models and mechanisms, or in requirements for approval, and thus differences in how and when research is split into discrete activities and labelled. In addition, research efforts take place at a variety of scales, have varying requirements for pre-published protocols, use a huge range of different methodologies and workflows, and may even draw upon different underlying assumptions. Research design, approach, strategy and method are heavily influenced by the researchers' epistemology and research philosophy [4]. The type of the research (e.g., hypothesis testing versus hypothesis generating), the methodology chosen (e.g., experimental, survey, cohort, case study) and the research methods applied (e.g., type of sampling) are of major importance in understanding the data generated, and thus in supporting any secondary use of that data. Another issue to be solved is the integration of various sources of metadata related to parliamentary and social metadata. Consequently, metadata should go beyond a description of the data itself to include the basic elements of the research process ("contextual metadata").

Objectives: A legitimate question emerges from these observations. Despite the enormous heterogeneity described above, is there a way to describe research activity, methodologies, and workflows, consistently across disciplines and in a way that can be understood by non-specialists? Such a schema would provide a useful context to the more basic metadata description of research outputs and could be used to support findability within the EOSC, potentially including planned and ongoing activity as well as completed projects. If available, such a schema could considerably strengthen the Findability of FAIR data objects (F2; Data are described with rich metadata) [5]. The objective of the Science Project (SP) "COVID-19 metadata findability and interoperability in EOSC" (short: META-COVID) is to develop a framework for a metadata model, characterising the research approach and workflow across research infrastructures and to apply it to use cases in COVID-19 research.

Compliance to criteria developed by EOSC Future:

Eligibility: The SP is open to participation by research communities from life sciences and social sciences and humanities.

Contribution to EOSC: The SP will provide a framework for a metadata model, able to characterise the research approach/activity within and between different research infrastructures and domains. Its implementation (which is not foreseen within the SP due to time limitations) will considerably strengthen the FAIR findability of digital objects within the EOSC through richer metadata. A proposal for a “research activity profile” for use within the EOSC interoperability framework will be made to support a more consistent approach to metadata describing “research activity” and related “contextual metadata”.

In addition, an ontology of COVID-19 related topics as discussed in parliamentary data and social media data will be developed and integrated into EOSC as a resource. This ontology will be modeled after established coding systems across multiple scientific disciplines, such as the Comparative Agendas Project for political agenda issues [6] and the ICD-11 coding system for classification of diseases [7].

Quality: The research to be performed will bring together for the first time senior researchers experienced in metadata schemas, ontologies and classifications from the clusters of EOSC-Life and SSHOC with the aim to work on a common framework for defining research strategies and workflow segments across the domains of life sciences and social sciences and humanities. The SP will build upon existing standards and best practices and include a first application of the framework in use cases in COVID-19 research.

Relevance: Dedicated to COVID-19, the work is highly relevant for the EU focus area on infectious diseases and will support the European Commission’s (EC) mission on health. The deliverables of this SP (especially the inventory on metadata schemas in use) will provide input to other projects (e.g. BY-COVID [8]) as they will describe a baseline of standards that need to be recognised in this space.

Implementation, Plan of work

Task 1

The SP will start by inventorying the metadata schemas used in each discipline as reported by each participating Research Infrastructure (RI) (ECRIN [9], BBMRI-ERIC [10], EATRIS [11], CESSDA [12], CLARIN [13], etc.) and then investigate and report whether there have been mapping exercises for cross-RI domains or even for cross-cluster domains before and use this "inventorying" material as a basis for a framework to characterise research approach and workflow across infrastructures and clusters. This work will take existing ontologies, standard classifications and controlled vocabularies into consideration. Input for this work will come from the specification of standard workflow fragments (as suggested by the Canonical Workflow Frameworks

for Research (CWFR) [14]) and harmonised provenance concepts that can be used cross domains and infrastructures. Another focus will be structural elements of research design, such as the typology of research methods, where standardised classifications within clusters are available, which need to be mapped across infrastructures and domains [15, 16]. This is of major importance for COVID-19, where evidence on effective interventions can only be generated, if relevant studies from different areas can be discovered, accessed and linked (e.g. vaccination studies, registry data from intensive care units, public health policies as masks or lockdown and lab data on mutants). Task 1 will also cover a status report on ontologies/classifications for parliamentary data and social media with a focus on COVID-19.

Task 2

In the next step, a framework for a metadata model, characterising the research approach and workflow across research infrastructures will be developed. The intention would be to use an iterative process of consensus building amongst experts drawn from different disciplines in the life sciences and social sciences and humanities. Specific studies drawn from COVID-19 research will be used to inform the process and to test the final result in a use case. The work carried out in the SP will build on the Research Data Alliance (RDA) COVID-19 Guidelines and recommendations for data sharing (where ECRIN was involved, [17]) and other standardising approaches (WHO's guidance on COVID-19, FAIRsharing [18], OpenAIRE [19]).

Within Task 2, the following questions will be discussed:

1) What does "contextual metadata" mean to you/your RI?

2) Are the metadata of resources/digital objects you hold in your research infrastructure linked to a research graph (e.g. PID graph, OpenAIRE Research Graph, Open Knowledge Research Graph, Scholexplorer, CERIF)? Are there gaps in the use of "contextual metadata"?

3) Is it planned to make a link to any of the research graphs mentioned above?

4) What elements of the research entities/research artefacts are modeled in the metadata schemas applied at your research infrastructures (research organisations, researchers, services, projects, funders, etc.)?

5) A specific focus should be given to the artefact "research activity". Here the question is: How is "research activity" modeled in your RI (which level of granularity, domain, common elements, interoperability layer)?

6) What services, protocols, APIs are implemented to harvest contextual metadata of your metadata schemas?

7) How could interoperability for “contextual metadata” between research infrastructures be improved?

8) What is the best organisational framework for moving this work forward, within EOSC in particular? Would it make sense to provide a specific EOSC Profile for contextual metadata?

Based on these discussions, the possibility of providing an interoperability layer for “contextual metadata” between research infrastructures will be investigated, a metadata model that can apply across research infrastructures will be developed and demonstrator solutions will be demonstrated for the COVID-19 use cases (Task 3).

In addition, an ontology of COVID-related topics from parliamentary data and social media will be developed, providing a societally relevant categorization to which subtopics of a diverse set of scientific fields can be easily linked. This wouldn't be restricted to only labels from the Comparative Agendas Project, but would include, for example, links to medical ontologies such as the ICD-10 or SNOMED-CT as well as identifiers of public policy issues such as lockdowns and imposed face masking etc. This approach is supposed to facilitate the integration of various sources of metadata so that COVID-19 related societal issues, which we categorize to the extent that they are automatically detectable from parliamentary data and social media data, can be easily traced to specific scientific concepts from diverse scientific fields. This would allow us to systematically investigate public attitudes towards COVID-19 related public health measures. This work will be provided by TEXTUA, a core facility of the University of Antwerp (UA) which provides scalable text mining solutions to researchers from any scientific discipline.

The work by UA describes the broader implicit ontologies, foci and concerns of the pandemic, as seen in legislatures, political debates and social media. This approach will be compared to the scientific ontologies, foci and concerns around COVID-19, as represented by the contextual metadata employed by scientists. It will be explored whether these two ways of “looking at COVID-19” are broadly similar or whether they suffer from having different priorities, with different topics emphasised. In this context, the following questions will be discussed:

- How can scientific metadata and the informal “public metadata” implicit in broader social debates be brought closer together?
- Does this offer an opportunity for science and scientists to learn how to better align themselves with public concerns, and / or is it an opportunity to identify gaps in public understanding?
- Does this offer opportunities for reducing the suspicion of science and thus increase future compliance with public health measures (e.g., mask wearing, vaccination)?

Task 3

Use case A:
There are numerous research areas for COVID-19, which must be taken into consideration. This covers basic research, epidemiology, public health and clinical management, diagnostics, therapeutics, vaccines, technologies and socioeconomic responses. In this use case, the “research activities” from the different domains covered by the SP will be linked via the metadata model for “contextual metadata” to better support identifying, collating, appraising and summarising the body of research evidence for COVID-19 across research infrastructures and domains.

Use case B:
This is a use case to validate that the newly developed ontology will automatically label parliamentary data and social media data. This will allow for an interdisciplinary analysis of the interactions between parliamentary sessions and the public debate about COVID-19 related topics, as reflected on prominent social media platforms.

Task 1 will be based on a survey dedicated to the partners of the SP. To build consensus in Task 2 and Task 3, the following working approach is proposed:

- Initiation of an SP working group with regular meetings.
- Mobilisation of the various metadata/research graph “owners” and of the RDA interest group working in the research graph area for contribution to the discussion.
- Survey and/or semi-structured interviews with the participating research infrastructures. Target is to examine and describe in detail the use of “contextual metadata” within the schemas used in the RIs and to clarify their conception of “research activity”.
- Preparation and consent of a report entitled “Conceptions of “research activity” and associated “contextual metadata” across selected life and social sciences RIs”.
- Preparation of a report on a societally relevant ontology, characterizing resources linked to parliamentary data and social media.
- Discussion of ways how “research activity” metadata could be made more consistent. A possible strategy could be to identify core elements plus optional domain-specific “supplementary sections”.
- Development of a pilot linked to COVID-19 to demonstrate interoperability between research infrastructures related to “contextual metadata”/”research graphs”/”research activity”.

- Proposal for a “research activity profile” for use within the EOSC interoperability framework to support a more consistent approach to metadata describing “research activity” and related “contextual metadata”.

Deliverable 1: Inventory of metadata schemas applied across infrastructures and domains

Deliverable 2: i) Development of a framework for a metadata model characterising the research approach and workflow across research infrastructures;
ii) Development of an ontology of COVID-related topics from parliamentary data and social media

Deliverable 3: Application of the framework to the COVID-19 use cases

Use of resources: Needed are senior experts from the RIs, experienced in the field under study. There is a need to use ontology and terminology services.

What are the demands of the SP from EOSC Future platform: No major demands in terms of storage volume, HPC power and composability services from the EOSC Future platform. Interactions with the EOSC Interoperability Framework team are expected.

What the SP brings to EOSC Future platform: The SP will provide the basis for improved services of the EOSC Future platform by setting the scene for FAIRer data discoverability of digital objects cross research infrastructures (F2). The benefit of the approach will be demonstrated through a use case linked to COVID-19.

Partners: From the SSHOC cluster: CESSDA’s Linked Third Party Finnish Social Science Data Archive (FSD)/Tampere University (TAU), CLARIN ERIC and its Linked Third Party University of Antwerp (UA). From the EOSC-Life cluster: ECRIN, BBMRI-ERIC, EATRIS, EU-OPENSCREEN.

Impact

Strategic: A metadata framework, characterizing the research process and associated workflow cross research infrastructures and domains will be available after the end of SP. If implemented, it will strengthen the FAIRification of digital objects in the life sciences and SSH by increasing findability of data. The use case of COVID-19 research will be the focus of this SP.

Scientific/User communities: The scientific and user communities will profit from better discoverability of digital objects for complex secondary analysis. This will be achieved by adapting and extending existing metadata models.

Societal/Economic: It is expected that the SP will support more efficient and effective generation of evidence out of existing information, which has considerable financial and social consequences. The focus of the SP in COVID-19 will facilitate complex

multidisciplinary research in the field and increase preparedness for future pandemics by linking studies across domains.

EU Policies: It is expected that upcoming EU policies on FAIRness of data will take the experiences from this SP into consideration. Especially the lessons learned for future pandemic preparedness and improved findability of studies cross-domains and using the EOSC.

Engagement plan

Target groups: 1) Researchers conducting health-related research and interested in the application and use of metadata schemas in their research activity; 2) Researchers from the life sciences cluster (EOSC-Life RIs, LS-RIs), the social sciences and humanities cluster (SSHOC RIs) and e-infrastructures involved in the work on metadata schemas as well as initiatives and projects dealing with metadata conception and usage; 3) EOSC stakeholders (e.g. task forces and working groups) focused on metadata and interoperability aspects.

SP key concept: The key concept is to see if contextual metadata can be developed that is usable cross-domain.

The SP contributes to the important issue of “contextual metadata” and is expected to deliver:

- Awareness of the importance of contextual metadata, characterising the “research activity”
- Clarification of the concepts behind contextual metadata for RIs and their stakeholders
- Status on the use of contextual metadata in the RIs (metadata schemas used, APIs, link to research graphs)
- Proposal for a metadata framework on contextual metadata cross RIs
- Application of the metadata framework in selected use cases (COVID-19)

As a result, the RIs involved will:

- Better understand the importance of contextual metadata for mapping between RIs via use cases
- Get an overview on schemas used for contextual metadata in the RIs and how to access and map the schemas
- Learn from use cases how to handle issues with contextual metadata
- Develop further use cases building upon a framework for contextual metadata

The work will be done via semi-structured interviews with experts from participating RIs on the use of contextual metadata. Protocol and interviewer guide are openly available [20, 21].

In the interviews possible links to EOSC will be explored. From the results of the interviews a simple and practical applicable framework for contextual metadata will be provided and summarised in a discussion paper. The discussion paper will be presented to the EOSC community and a wider audience in a workshop (see workshops) and will lead to a publication.

Specific attention will also be given to considering which COVID-19 data resources are available and metadata harmonisation would be needed for construction of a COVID-19 observatory application that would allow researchers to visualise COVID-19 data and information originating from different research infrastructures in an integrated way and browse these in different dimensions e.g. time and geography. This should result in a short feasibility study document.

Work in the SP foresees exchanges with:

- EOSC Future WP3 partners
- EOSC Taskforce on Semantic Interoperability
- FAIRcore4EOSC [22]
- FAIRsharing

Dissemination measures

- Scientific publications: A scientific paper summarising the results from the semi-structured interviews with the SP partners and the resulting metadata framework will be provided (study protocol and interviewer guide already published on ZENODO [20, 21]). A scientific paper describing the outcomes of the work on social media/parliamentary data (University of Antwerp) will also be published.

- Conferences: C. Ohmann, S. Canham (ECRIN), K. Moilanen, M. Kleemola (CESSDA): Bridging scientific domains with metadata: CESSDA and ECRIN European DDI Users conference, 28 November – 1 December 2022, Paris.

  More conferences, as identified by the project partners. A poster summarizing the work of META-COVID will be created and published in Zenodo to facilitate dissemination from partners.

- Workshops: Workshop on contextual metadata (with invited speakers from FAIRsharing, OpenAIRE, COVID-19 data portal [23]) performed during the EOSC-Life 3rd Annual General Meeting on 22 March 2022.

  Workshop with EOSC Future partners, projects and initiatives involved in metadata work (e.g., OpenAIRE, FAIRcore4EOSC, EOSC task force), representatives from life sciences RIs (ECRIN, EATRIS, BBMRI, EU-OPENSCREEN), Social sciences and humanities (CESSDA, CLARIN) and ELIXIR to discuss the results from the META-COVID SP (April 2023).

FAIR metadata workshop at ECCB 22: COVID-19 use case presented.

- Demonstration to other communities: The results from the SP will be used as input to the ECRIN-CESSDA use case and the ECRIN-BBMRI use case and feed into the BY-COVID project.

  The SP will make a link to the FAIRcore4EOSC project and here, especially to the MSCR (EOSC metadata schema and crosswalk) (via CLARIN).

  A link will be made to the EOSC Task Force “Semantic Interoperability” and other working groups involved in this field.

- Education and training events: The SP has initiated discussion with EOSC Future WP9 to discuss the conception of a training event focused on the conceptual linkage of services across RIs and conceptual metadata alignments. The type of training event will be discussed on the basis of the needs of the research communities involved.

- Networking: Networking will cover EOSC-related projects and other initiatives relevant for the discussion of contextual metadata e.g. EOSC-Life, EOSC4Cancer [24], FAIRcore4EOSC.

Acknowledgements

This Science Project is part of the EOSC Future Project, WP6.3, co-funded by the EU Horizon Programme call INFRAEOSC-03-2020 – GA 101017536.

References

[1] https://sshopencloud.eu/, access on 30/05/2023

[2] https://www.eosc-life.eu/, access on 30/05/2023

[3] Kleemola, M. SSHOC metadata interoperability aspects. Presented at the Realising the European Open Science Cloud - Towards a FAIR Research Data Landscape for the Social Sciences, Humanities and Beyond (RealisingEOSC), 2020, Zenodo. http://doi.org/10.5281/zenodo.4279855

[4] Al-Ababneh, M.M.: Linking ontology, epistemology and research methodology. Science & Philosophy. 2020; 8: 75-91

[5] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzales-Beltran, A., Gray, A.J.G., Groth P., Goble, C., Grethe, J.S.,

Heringa, J., ‘t Hoen, P.A.C., Hoft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone M.E., Mons, A., Packer, A.I., Persson, B., Rocca-Sera LaFlamme, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Thao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3 (1): 160018 (2016) https://doi.org/10.1038/sdata.2016.18

[6] https://www.comparativeagendas.net/pages/master-codebook, access on 30/05/2023

[7] https://www.who.int/standards/classifications/classification-of-diseases, access on 30/05/2023

[8] https://by-covid.org/, access on 30/05/2023

[9] https://ecrin.org/, access on 30/05/2023

[10] https://www.bbmri-eric.eu/, access on 30/05/2023

[11] https://eatris.eu/, access on 30/05/2023

[12] https://www.cessda.eu/, access on 30/05/2023

[13] https://www.clarin.eu/, access on 30/05/2023

[14] Tobi, H., Kampen J.K.: Research design: the methodology for interdisciplinary framework. Qual Quant 2018; 52: 1209-1225

[15] Boiten, J.W., Ohmann, C., Adeniran, A., Canham, S., Cano Abadia, M., Chassang, G., Chiusano, M.L., David, R., Fratelli, M., Gribbon, P., Holub, P., Ludwig, R., Mayrhofer, M.T., Matei, M., Merchant, A., Panagiotopoulou, M., Pireddu, L. Sanchez Pla, A., Schlünder, I., Tsamis, G., Wagener, H.: EOSC-LIFE WP4 TOOLBOX. Toolbox for sharing of sensitive data - a concept description. Zenodo. http://doi.org/10.5281/zenodo.4483694 (2021)

[16] The CWFR Group, Hardisty, A, Wittenburg, P (eds): CWFR-position-paper-v3.doc (version 2). https://osf.io/2cy86/ (2020) Accessed 20 June 2021

[17] RDA COVID-19 Working Group. Recommendations and Guidelines on data sharing. Research Data Alliance, 2020. DOI: https://doi.org/10.15497/rda00052.

[18] https://fairsharing.org/, access on 30/05/2023

[19] https://www.openaire.eu/, access on 30/05/2023

[20] Ohmann, Christian, Canham, Steve, & Panagiotopoulou, Maria. (2022). Protocol of a qualitative study to characterise the contextual metadata and workflows in selected research infrastructures. Zenodo. https://doi.org/10.5281/zenodo.7025319

[21] Ohmann, Christian, Canham, Steve, & Panagiotopoulou, Maria. (2022). Interview guide for a qualitative study to characterise the contextual metadata and workflows in selected research infrastructures. Zenodo. https://doi.org/10.5281/zenodo.7025502

[22] https://faircore4eosc.eu/, access on 30/05/2023

[23] https://www.covid19dataportal.org/, access on 30/05/2023

[24] https://eosc4cancer.eu/, access on 30/05/2023