Proposal for the EOSC Semantic Interoperability Questionnaire

Supplementary material for the paper:

Converging towards a Semantic Interoperability Framework for EOSC: understanding the different community solutions

Authors affiliation:

- Barbara Magagna, GO FAIR Foundation, Poortgebouw Noord, Rijnsburgerweg 10, 2333 AA Leiden, barbara@gofair.foundation , 0000-0003-2195-3997
- Kurt Baumann, SWITCH, Werdstrasse 2, CH-8021 Zurich, Switzerland, kurt.baumann@switch.ch, 0000-0003-0627-8110
- Romain David, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, romain.david@erinha.eu, 0000-0003-4073-7456
- Thomas Jouneau, Université de Lorraine, Direction de la Documentation, Île du Saulcy, 57000 Metz, France, thomas.jouneau@univ-lorraine.fr, 0000-0001-5986-8128
- Hanna Koivula, CSC - IT Center for Science, Life Science Center Keilaniemi, Keilaranta 14, 02150 Espoo, Finland, hanna.koivula@csc.fi, 0000-0001-5605-9122
- Bénédicte Madon, Escuela Superior de Ingenieros, Universidad de Sevilla/AICIA- Asoc.Invest.Coop.Ind.Andalucia, Camino de los Descubrimientos s/n, Sevilla-41092 , Spain, bcg.madon@gmail.com, 0000-0001-8608-3895
- Wolmar Nyberg Åkerström, Uppsala University, Department of Cell and Molecular Biology, Science for Life Laboratory, SciLifeLab, Box 596, SE-75124 Uppsala, Sweden, wolmar.n.akerstrom@uu.se, 0000-0002-3890-6620
- Milan Ojsteršek, University of Maribor, Faculty of Electrical engineering and Computer Science, Koroška cesta 46, 2000 Maribor, milan.ojstersek@um.si, 0000-0003-1743-8300
- Andrea Scharnhorst, DANS-KNAW, Anna van Saksenlaan 51, 2593HW The Hague, The Netherlands , andrea.scharnhorst@dans.knaw.nl, 0000-0001-8879-8798
- Chris Schubert, Vienna University of Technology (TU Wien), Resselgasse 4, 1040 Wien, Austria, https://orcid.org/0000-0002-4971-2493
- Zhengdong Shi, Université Paris-Saclay, Directorate of Libraries, Information and Open Science, Bâtiment 407, Rue du Doyen Georges Poitou, 91400 Orsay, zhengdong.shi@universite-paris-saclay.fr, 0000-0001-5817-6031
- Letizia Tanca, Politecnico di Milano, Department of Electronics, Information and Bioengineering, Piazza L. Da Vinci 32, 20133, Milano, Italy, https://www.deib.polimi.it/eng/people/details/61038, 0000-0003-2607-3171
- Sadia Vancauwenbergh, UHasselt, Martelarenlaan 42, B-3500, Belgium, sadia.vancauwenbergh@uhasselt.be, 0000-0002-5201-8101
- Lars Vogt, TIB Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30167 Hannover, Germany, lars.m.vogt@googlemail.com, 0000-0002-8280-0487
- Heinrich Widmann, German Climate Computing Center (DKRZ), Bundesstraße 45a, D-20357 Hamburg, widmann@dkrz.de, 0000-0001-9871-2687

Keywords: Semantic interoperability, European Open Science Cloud, FAIR Principles, machine actionability, questionnaire

Introduction

This document provides supplementary material for the paper "Converging towards a Semantic interoperability framework for EOSC: understanding the different community solutions to semantic interoperability" submitted as a contribution to the 2nd Workshop on Ontologies for FAIR and FAIR Ontologies (Onto4FAIR)[1]. Its purpose is to provide additional information and insights about the questionnaire approach proposed to survey the resources utilised by communities in addressing semantic interoperability issues. This document

[1] https://onto4fair.github.io/2023-fois.html

serves as a guide and general framework for implementing the questionnaire. More detailed specifications will be developed during the implementation phase.

Questionnaire design

In the same manner as for recording the resources used for implementing the FAIR Principles in the FIP (FAIR Information Profile) approach (Magagna et. al., 2022), the Data Stewardship Wizard (DSW)[2] will be used as an interactive interface to guide the interviewee through the questionnaire. The advantage of this service is twofold. It provides a human-readable interface, which means that users can understand its aims and functionalities intuitively and interact with it. It also allows for a machine-readable output for a better comparability of the results. The result of the survey will document the resources used or planned to be used by a community to support semantic interoperability. For convenience, these resources will be referred to as FAIR Supporting Resources (FSRs). The assumption is that resources supporting FAIR also encompass resources facilitating semantic interoperability. The aggregation of resources utilised by a community is referred to as a Semantic Interoperability Profile (SIP) of that community.

A simple ontology will serve as a semantic backbone for defining the concepts and the relations needed to set up the questionnaire in the DSW. A Community is a collective of researchers collaborating in projects, initiatives or research infrastructure, aiming to set up a common strategy to implement a community specific Semantic Interoperability Framework. The community might have different use cases determined by their research activities (e.g., clinical trials, literature reviews, observational studies, model simulations, surveys, etc.), producing specific data types (e.g., text, images, data streams) for which it can provide specific Semantic Interoperability Profiles. For each specific question a community specifies the utilization of FSRs through SIP declarations.

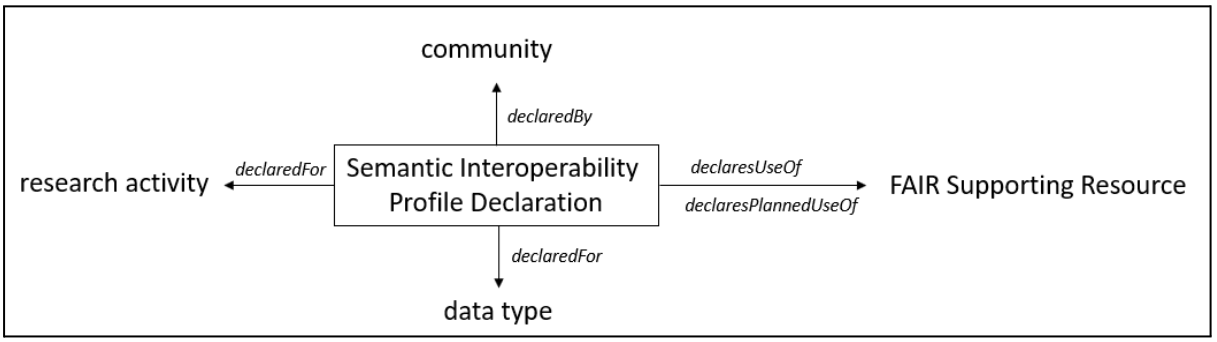

Figure 1: Conceptual model for the Semantic Interoperability Declaration

The aim is to describe FSRs as nanopublications to provide a globally unique, persistent and resolvable identifier and machine-readable representation based on a specific metadata schema. A nanopublication is a smallest unit of publishable information, with associated provenance, expressed as a knowledge graph that is formal and machine-interpretable[3].

[2] https://ds-wizard.org/
[3] https://nanopub.net/

The metadata schema for describing a FSR will reuse existing ontological properties and will be implemented as a nanopublication template. FSRs encompass different resource types like data managing practices, services, specifications, and training materials. They support the interoperability of different artefact types like semantic artefacts, metadata, crosswalks or term mappings. Accordingly, the schema will include the following fields (with * standing for mandatory metadata) :

- ☐ Name of the resource*
- ☐ Short description*
- ☐ Type of resource*
- ☐ Supported artefact type*
- ☐ Provided function* (see list below)
- ☐ URL of specification it is based on (e.g. schema)
- ☐ URL to the online resource, if existing

Functions provided for artefacts:

- Creation
- Editing
- Publication
- Registration
- Quality Validation
- FAIRness assessment
- Semantic search

Nanopublication descriptions will be prepared for all known FAIR Supporting Resources as drop-down options in the DSW tool using as source information the outcomes of related initiatives preceding this work[4]. Having examples in the drop-downs for each question helps to make the requested answers more understandable and accelerates the answer time, making the interview more efficient. In case a resource is not already referenceable, the nanopublication can be created in the Wizard environment; once created, it will be immediately available for selection in the drop-down lists of the answer field for the corresponding question. As the intention is to keep the interview time to a minimum, it will be more likely that the interviewer will provide the nanopublication description after the interview, based on the information given by the interviewee.

Throughout the questionnaire, explicit questions are addressed related to specific resources used and reasons that lead to these decisions, and if no resource is used, the difficulties and challenges that the interviewee's community has experienced on this specific topic. If the question is not understood, we provide explanations and reformulate the question, in order for it to be better understood (a process referred to as an iterative test of understanding). Review the proposed pattern for each question, using the example of metadata schemas used:

---

[4] https://fair-impact.eu/synchronisation-force, https://www.fairsfair.eu/fair-semantics-interoperability-and-services-0, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02315001v1/file/RR-Ontology-metadata-survey-June2018.pdf

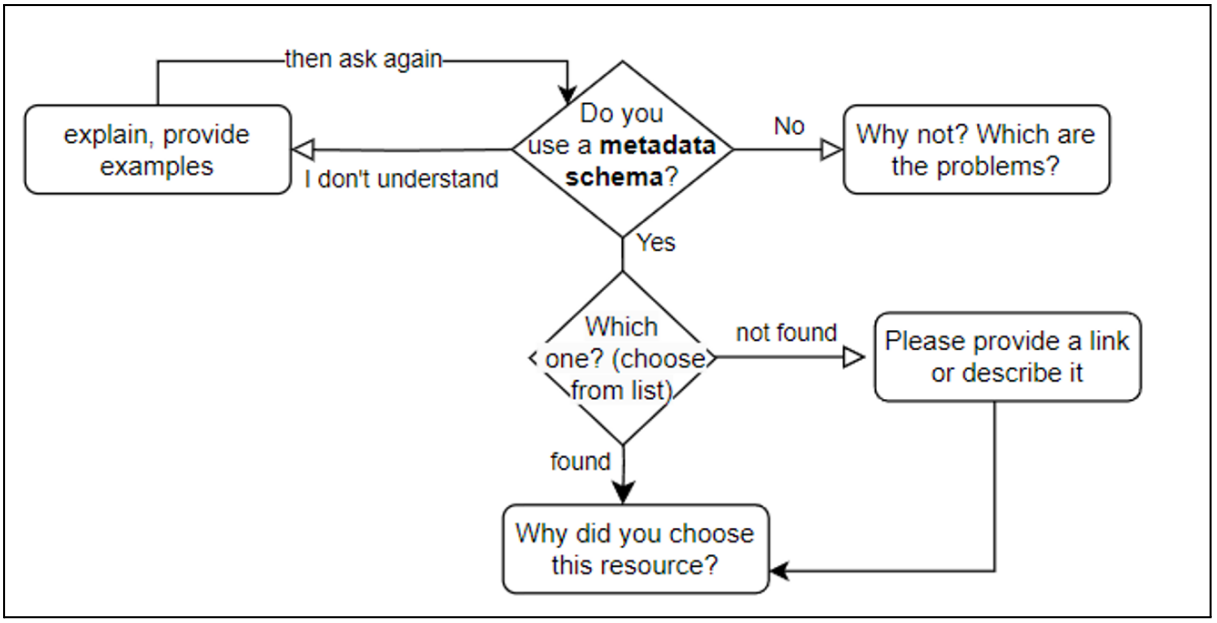

Figure 2: Design of the questionnaire, using the example of metadata schemas used

Following this pattern, both machine-readable and comparable outputs as well as rich-text descriptions about the reasons, problems and challenges described by the expert can be provided.

Questionnaire outline

The following section gives an overview about the topics addressed in the questionnaire. These questions are grouped in different sections related to general information, metadata, and semantic interoperability.

The questionnaire includes Boolean options (yes/no), information requiring a weblink (URL), open questions the answers of which should be provided by literals (like names, or considerations including reasons why no resources are used), or questions that require choosing from a list of items (in most cases from resources described as nanopublications). All questions should be answered, but providing detailed answers is optional. Per question the use of multiple FSRs can be specified. For any question the singular form is used, but it entails that multiple answers can be provided.

To better understand the required resource, examples {example1, example2, …} are provided in curly brackets. In the Glossary, at the end of the document, explanations for specific terms are provided.

General information:

- Please provide your name, email, and role you have in the community you represent.
- Please provide the name, a brief description, a link to the website, and the research area {biodiversity, geophysics, healthcare…} of the community you are representing.
- What is the research activity {surveys, observations, clinical trials,..} that led to the data your community generates?

- Which type of data {tabular, text, images, data streams...} is created by this research activity?
- Is this data published? Y/N
  If yes:
    - Does your community collect datasets from different sites? Y/N
    - Does your community offer a common repository for your data? Y/N
        - If yes, which data repository?
        - If not, which external data repository is used?
- Do you offer a solution for achieving semantic interoperability to your community? Y/N
    - If yes: Please point to a document describing it.
    - If not: Let's talk about specific features required to enable semantic interoperability in your community.

Metadata-related information:

- Does your community describe the created data in a common way using metadata? Y/N
  If yes:
    - Does your community have guidelines for creating and maintaining metadata records?
    - Is this metadata used to increase the findability of the data? Y/N
    - On which generic metadata schema {DCAT, CERIF,...} is your metadata for discoverability of the datasets based?
    - Do you use mappings and crosswalks to link different schemes? Y/N
      If yes:
        - How do you represent the mappings and crosswalks?
        - Are the mappings and crosswalks accessible through an API? Y/N
          If yes: Which API is used?
    - Do you use other community-specific metadata schemas? Y/N
        - If yes: Which community-specific metadata schema?
    - Which is the representation language {XML, RDF,..} used for these metadata?
    - Does your community use a common API to retrieve your metadata from the web? Y/N
        - If yes: Which API is used?
    - Do you expose the metadata together with your data in the same repository? Y/N
        - If not: Which metadata repository do you use?
    - What is the usage licence {CC0, CC-BY 4.0, …} of the metadata?
    - Which metadata editor tool {CEDAR[5], Describo[6], ..} does your community use?
        - Does this tool allow the use of semantic artefacts for the entry of metadata? Y/N
    - Does your community provide quality validation services for the provided metadata in terms of completeness, use of common semantics, accuracy, consistency, …? Y/N

[5] https://cedar.metadatacenter.org/
[6] https://uts-eresearch.github.io/describo/

        - If yes, which metadata quality validation service?

Semantic-interoperability-related information:

- Which semantic artefact {ENVO, SNOMED, …} is used to describe the (meta)data?
- Does your community create semantic artefacts? Y/N
  If yes:
    - Which community semantic artefact is developed?
    - Which editing service or tool for semantic artefacts {Protégé, VocBench, …} is used?
    - Which governance model is used for the evolution of the semantic artefacts?
    - Is metadata provided for these semantic artefacts? Y/N
        - If yes, which metadata schema for semantic artefacts {OMV, DCAT, …} is used?
    - What usage license {CC0, CC-BY 4.0, …} is used for these semantic artefacts?
    - Are terms mapped to concepts from other semantic artefacts? Y/N
        - If yes, which term mapping representation {tabular format, community specific model, SSSOM, …} are used?
        - If yes, is there a governance model for the term mappings? Y/N
            - If yes, which governance model for term mappings is used?
        - If yes, are you using a mapping repository? Y/N
            - Which mapping repository {EMBL-EBI Ontology Xref Service (OxO)[7], Metadata Schema and Crosswalk Registry (MSCR[8]), …}?
    - Do you upload your semantic artefacts in a dedicated public repository? Y/N
        - If yes, which repository for semantic artefacts {BioPortal, OLS, …} is used?
    - Is your community using other specific services for semantic artefacts? Y/N
      If yes:
        - SPARQL endpoint? Y/N -> Which?
        - Service for finding exact matches? Y/N -> Which?
        - Service for mapping terms? Y/N -> Which?
        - Service for assessing the FAIRness of semantic artefacts? Y/N -> Which?
        - Service for annotating datasets with semantic concepts? Y/N -> Which?

    If not: Does your community participate in the creation of a semantic artefact? Y/N ->

    - If yes: please specify the semantic artefact.
    - If yes: please specify the maintenance group coordinate the creation of the semantic artefact.
- Are the data providers of the data catalogue using a common data model?
    - If yes, please specify the data model used.
- Does your community integrate data based on different data models? Y/N

[7] https://www.ebi.ac.uk/spot/oxo/
[8] https://faircore4eosc.eu/eosc-core-components/metadata-schema-and-crosswalk-registry-mscr

        - If yes, is a crosswalk between those data models used? Y/N
            - If yes, which crosswalk?
- Do you provide provenance information about your data? Y/N
    - If yes, does this include information about the original source and all the processing steps (i.e. lineage information)? Y/N
        - If yes, which provenance model {PROV, community specific, …} is used?
    - If yes, does this provenance metadata include quality information? Y/N
- Are you using a portal for exposing your FAIR data to enable machines to operate on them? Y/N
    - If yes, which access point is used?

Glossary:

In this glossary, terms are defined as discussed between the authors of the paper only for the purpose of this questionnaire without any claim of completeness or consensus within EOSC SI TF.

Crosswalk: set of rules that defines how (meta)data elements or attributes (i.e., slots) from one schema or format can be aligned and mapped to (meta)data elements or attributes in another schema or format that share the same constraints and thus share the same semantic role (as defined in the main paper).
FAIR: Findable, Accessible, Interoperable and Reusable.
FAIR Supporting Resource (FSR): a resource used by a community to facilitate the process of FAIRification of (meta)data and ultimately also semantic interoperability.
Machine actionability: machine-interpretable bit sequences for which operations have been specified in symbolic grammar such as logical reasoning for OWL-based (meta)data and other rule-based operations.
Metadata Schema: A schema is a logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality (obligation level) of values (based on ISO 23081.1 Terms and Definitions)
Nanopublication: is a specific form of machine-readable information that represents small, self-contained units of information. It is typically structured using standardised formats expressed in RDF (Research Description Format) and Named Graphs.
Semantic artefact: a machine-actionable formalisation of a conceptualisation enabling sharing and reuse by humans and machines. These artefacts may have a broad range of formalisation, from a loose set of terms, taxonomies, thesauri to higher-order logics, and include the concepts/terms/classes constituting these. Moreover, semantic artefacts are serialised using a variety of digital representation formats, e.g., RDF Turtle, OWL-RDF, XML, JSON-LD (Le Franc et al., 2022). The SI Task Force discussed whether metadata schemas could be considered as semantic artefacts. It was agreed that this is true only when they are machine readable which is not always the case. For this reason and also because the FAIR Principles explicitly refer to those concepts separately because metadata needs to use vocabularies to be FAIR, it was decided to provide separate definitions for both terms.
Semantic Interoperability (SI): the ability of computer systems to transmit data with unambiguous, shared meaning… It ensures that the precise format and meaning of

exchanged data and information is preserved and understood throughout exchanges between parties. In other words, ‘what is sent is what is understood (Corcho et al., 2021).
Semantic Interoperability Profile (SIP): a list of FAIR Supporting Resources chosen by a community to support semantic interoperability of (meta)data.
Term mapping: Mappings of terms defined in different semantic artefacts sharing the meaning and identifying the same referent (i.e., the object designated by the term).

References

Corcho, O., Eriksson, M., Kurowski, K.et al., European Commission, Directorate-General for Research and Innovation. *EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture*, Publications Office, 2021, https://data.europa.eu/doi/10.2777/620649
International Organization for Standardization. (2017) ISO 23081-1: Information and documentation - Records management processes - Metadata for records - Part 1: Principles
Magagna B, Schultes EA, Suchánek M, Kuhn T (2022) FIPs and Practice. *Research Ideas and Outcomes* **8**: e94451. https://doi.org/10.3897/rio.8.e94451
Le Franc Y., Bonino L., Koivula H., et al., (2022). D2.8 FAIR Semantics Recommendations Third Iteration (V1.0). Zenodo. https://doi.org/10.5281/zenodo.6675295