GO BUILD Workshop on "Semantic Interoperability of Metadata for Cross-Domain Research of the Future"

Summaries of the Discussions

Group "FAIR Semantic Artefacts"

In order to define the semantic artefacts we have to clarify first which type of semantic artefact it will be (a simple vocabulary, a taxonomy or an ontology, which is not only machine-readable but also allows for higher grades of reasoning).

To enable interoperability in the FAIR sense, some steps have to be done first:

1. Data should be analysed: Which kind of data do we have? Which features characterize them?
2. Metadata should be captured and data exposed via these metadata (both old and new datasets, which would have to be validated).

In this process all our choices have to be made explicit. This is crucial to reach interoperability as we document the choices we made and the reasoning behind them (versioning of metadata across the time) as if our approach was centralized / decentralized. We should shift the complexity from the metadata to standards and tooling, which would make data exchange simpler.

The practical creation of a semantic artefact takes some known steps:

1. Definition of requirements;
2. Design of the artefact that meets the requirements;
3. Implementation and deployment;
4. Evaluation.

Group "Tools and Software"

The amount of software tools for FAIR data exchange is ever increasing - forcing the community to face new challenges and scenarios. Especially as tool developers need to address several stakeholders at once: Researchers the tools are aimed for as well as the heterogeneous world of standards, APIs and frameworks.

In the Tools WG we identified several requirements that alleviate the burden while implementing new tools and foster sustainability of long term support. Starting with the individualisation of PIDs and an increase in the usability for researchers to allow for more flexible usage, PID services should support researchers with the FAIR principles. This continues with not being limited as a researcher to a single tool or metadata model. Automatic conversions should be available, which could be supported by well-known approaches in computer science such as Blockchain, Machine Learning or Crowd-Sourcing. The former would support the exchange by supporting provenance while the last two help with filling in gaps in metadata or ontologies. To further enhance long term support, standards need to be established and strengthened by the community to allow for easy adoption both by centralised institutions such as the RDCs as well as the developers and researchers. Here, the FAIR principles can provide an additional level of support when extending them to a meta-level.

Making information about researchers and tools themselves FAIR supports the overall goal towards an Open Science community.

Group "Guidelines and Policies"

Our group considered whether high-level guidelines could support greater Semantic Interoperability of Metadata for Cross-Domain Research of the Future, and if so, what sort of guidelines would be necessary (and for whom?).

We first considered the target of such guidance. One option would be to create Guidelines for organisations/institutions hoping to become more 'FAIR', specially, to promote semantic interoperability. In this scenario, one would probably need to consider subsets of recommendations or guidelines, for different stakeholders. For example, one set would be for key decision-makers, i.e. heads of department, senior figures who set the priorities and direction of the organisation; then another set for researchers, who typically struggle to provide metadata in robust ways; then another set perhaps for data managers (people with technical expertise). The messages would need to be different for each of these. Generally speaking, however, such guidelines would need to illustrate the value propositions behind FAIR, elucidating what would an organisation/institution gain from paying closer attention to semantic interoperability. This would need to be relatively high level, in order to 'sell' the concept to non-experts.

The group also considered guidance which could be generated by GO FAIR INs aimed at their research communities (such guidance would, of course, need to be made available by the GO FAIR office, in order to support cross-domain collaboration). Some of the points discussed in terms of contents of such guidance are really recommendations for the GO FAIR office to give to the INs, whereas others are more guidelines for the INs to present to the wider world.

For instance, the group discussed guidelines on more technical issues. These would take the format of 'to make your data searchable, you need to do X Y and Z'. GO FAIR / the INs could provide guidance on where to 'log' your metadata - both cross-domain (e.g. re3data) and domain-specific (thus we would need to agree on the optimum 'home' for each domain). The INs could outline use cases and recommend specific semantic resources: The points of convergence could then be mapped.

All guidance could include a recommendation for researchers / groups to contact the GO FAIR office in the first instance for advice on issues such as:

- Which standards to use for greater semantic interoperability
- Where to deposit one's data;
- How to describe / list metadata;

in order to make it Findable (on the grounds that GO FAIR INs, collectively, should be ideally placed to build consensus on this sort of information).

An obvious but very important message in any guidelines directed at would-be data generators and researchers should be to avoid creating new standards at all costs, and instead, first seek the advice from GO FAIR / the INs and consult sites such as DublinCore, as a minimum. If it IS subsequently deemed necessary to create new ontologies and standards, one recommendation would be to involve a wider range of stakeholders representing a good cross-section of one's domain. It would also be helpful here to point people towards sensible resources on how to create new ontologies / standards.

Further Results

Coordination and methodology

Coordination is needed to define roles between the various organizations able to provide the guidelines (GO FAIR, GO FAIR INs, RDA, ERICs, RIs, etc.). Coordination is also required for the methodology used to establish and communicate the guidelines (use cases, recommendations, trainings). Should the high-level guidelines be the ones of GO FAIR? Should they be provided by GO FAIR or / and RDA or on ´high / generic / specific level?

- Show best practices vs. worst practices.
- Provide 'Trainings' on guidelines / good practices → GO TRAIN?
- Collect use cases: but who is in charge of collecting this (GO FAIR? Discovery IN started in Porto (OSFair 2019) to collect use cases and search tools in a spreadsheet:
    - Exemplary researcher use case from medicine area: A patient has a disease - which ontologies, medications, treatment fit to this disease?
- Collect use cases from all research domains.
- Don't reinvent the wheel, but first look what is already available.
    - But where is the overview about all the existing wheels?
    - In this context wheels are semantic resources and tools for:
        - Searching, generating, providing metadata (generic GUI's);
        - Generating ontologies, vocabularies by researchers (without IT skills).
    - Maybe GO FAIR's implementation matrix can help here. Is GO FAIR responsible for collecting the wheels = semantic resources? → See the Leiden workflow: Is it a guideline?
- OPERAS will provide practical solutions for FAIR repositories harvesting in the SSH field: Ensure that there's a minimal FAIR set (identifiers, interoperable metadata, protocols for harvesting, clear licenses).

Adapt to specific audiences

The workshop's participants also found that the guidelines should be adapted to specific audiences (decision makers, researchers, data stewards) and that a multi-target approach could create a virtuous cycle of knowledge and practices sharing amongst:

- Decision makers (PIs, CEOs, etc);
- Researchers (people who produce the data);
    - In the stage of data producing metadata should already be generated, but scientists need clear and simple guidelines and easy-to-use UIs to do this in the right way;
    - Contact an expert from an existing list; e.g.
        - To get an expert to create / implement a (research specific) ontology. Likely the best answer is a search portal / repository of ontologies?
    - Find the right search tool / portal / repository for my requirements:
        - https://repositoryfinder.datacite.org/
        - Re3data.org
        - https://ddrs-dev.dariah.eu/ddrs/;
- Data managers / stewards; e.g.
    - Guidelines for data providers of OpenAire : https://guidelines.openaire.eu/en/latest/data/index.html
    - Guidelines for data providers of EUDTA-B2FIND : http://b2find.eudat.eu/guidelines/index.html

At the same time recommend, support and propagate OAI-PMH as harvesting tool and check DataCite, or at least Dublincore, as a minimal metadata schema.

Content perimeter of the guidelines

As to the content of the guidelines, the participants envisioned various possibilities linked to the both previous questions on coordination and audiences: Guidelines for generic FAIRification, technical recommendations for data deposit, centralization by GO FAIR of information on all available repositories and tools, etc.

Regarding the technical guidelines the participants highlighted the importance to know where to put (domain- specific) metadata and suggested to use re3data.org to find the appropriate catalogue / repository fitting to one's data and requirements. GO FAIR was seen as potentially being able to support this by, e.g. finding the right tools and repositories.