

Semantic artefact and ontology services for long-term data interpretation

Clement Jonquet ^[0000-0002-2404-1582]

LIRMM, University of Montpellier, CNRS, Montpellier, France
MISTEA, University of Montpellier, INRAE, Institut Agro, Montpellier, France

Abstract for PV 2023 Conference, 2-4 may 2023, CERN, Switzerland

What would happen if we would be able to preserve data for so long but we would have lost the ways to interpret the knowledge contained in these data? Exactly for the same reason we need to save the specific encoding format and/or software used to read the data (e.g., in which version of PDF a document was encoded and which PDF reader is necessary to open it), we need a way to decouple the model from the data and save this model for future interpretation. As data become more and more complex, the knowledge they incorporate become harder to express. The separation of the data and the model is often made by annotating or describing data with semantic standards spanning from simple metadata schemas to complex and rich ontologies. Indeed, semantic artefacts – a broader term to include ontologies, terminologies, taxonomies, thesauri, vocabularies, metadata schemas and standards – are keys for adopting the FAIR Principles (FAIR Principle I2) [1]. In the processes of archiving and long-term preservation, when used to annotate or describe data, semantic artefacts ensure the data will be stored in a standardized, broadly shared and accessible way that will guaranty long term “interpretation” of these data. Furthermore, semantic artefacts have themselves to be long-term archived and preserved. We therefore argue for two statements:

- (i) Because they are needed for long term interpretation of the data, **we need to develop services to adopt and facilitate semantic artefacts use when archiving data**. For instance, a semantic annotation is a link from an ontology term to a data element, indicating that the data element (e.g., article, experiment, clinical trial, medical record) refers to the concept [2]. These annotations can then be used to build semantic indexes to leverage the knowledge inside the ontologies for better information mining and retrieval.
- (ii) Semantic artefacts are themselves data that need to be archived and preserved with the same arguments as for any other data. Thus, **we need to develop long term repositories for archiving semantic artefacts**. Indeed, the number of these artefact now grow to numbers for which machines are mandatory to index, search and select them. They are spread out all over the Web, in different formats, of different size, with different structures and from overlapping domains. But as any research objects capturing the knowledge of experts at a certain time, for a certain use case, they need to be preserved.

In the context of multiple research projects in different scientific domains (e.g., NCBO, SIFR, D2KAB, OntoCommons, NFDI, FAIR-IMPACT), **we develop an open-source semantic artefact and ontology repository technology called OntoPortal** (<https://ontportal.org>) [3] shared and maintained by multiple communities to ensure cataloguing, harmonized and rich metadata descriptions, and serving the content of semantic artefacts with multiples generic services including for annotation and indexing. These ontology repositories include BioPortal, the primary and historical source of OntoPortal code developed in the biomedical domain, AgroPortal, the reference for agri-food [4] and more. They help address the challenges of sharing semantic artefacts and using them to describe data and enhance the quality of the data by making them more interpretable and interoperable. These semantic artefact catalogues are also designed to archive their content for long term preservation and scientific valorization.

This work has been supported by the D2KAB project (www.d2kab.org – ANR-18-CE23-0017), the French NUMEV-IRT OntoPortal Alliance (ANR-10-LABX-20), the Horizon Europe FAIR-IMPACT project (grant #101057344).

References

1. Jonquet, C.: FAIR data requires FAIR ontologies, how do we do? F1000Res. 7, (2018). ([10.7490/F1000research.1115343.1](https://doi.org/10.7490/F1000research.1115343.1)).
2. Handschuh, S., Staab, S. eds: Annotation for the Semantic Web. IOS Press (2003).
3. Graybeal, J., Jonquet, C., Fiore, N., Musen, M.A.: Adoption of BioPortal’s Ontology Registry Software: The Emerging OntoPortal Community. 13th Plenary Meeting of the Research Data Alliance, Poster session. , Philadelphia, PA, USA (2019).
4. Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé Yeumo, E., Emonet, V., Graybeal, J., Laporte, M.-A., Musen, M.A., Larmande, P.: AgroPortal: A vocabulary and ontology repository for agronomy. Comput Electron Agric. 144, 126–143 (2018). ([10.1016/j.compag.2017.10.012](https://doi.org/10.1016/j.compag.2017.10.012)).