1st Conference on Research Data Infrastructure

Harmonizing RDM

https://doi.org/10.52825/CoRDI.v1i.366

Published: 07 Sept. 2023

Finding a Common Ground for NFDI Terminologies

Proposing I-ADOPT as a NFDI Wide Semantic Layer

Robert Huber[1], Naouel Karam[2,3][https://orcid.org/0000-0002-6762-6417], Oliver Koepler[4][https://orcid.org/0000-0003-3385-4232], and Philip Strömert[4][https://orcid.org/0000-0002-1595-3213]

[1] University of Bremen

[2] Institute for Applied Informatics (InfAI), Leipzig, Germany

[3] Fraunhofer FOKUS , Berlin, Germany

[4] TIB - Leibniz Information Centre for Science and Technology

Terminology Harmonisation in NFDI

Top-level ontologies (TLOs) and mid-level ontologies (MLOs) play a very important role in enabling semantic interoperability between domain-specific ontologies by providing a general structure and common high level entities and relationships for classifying and interlinking domain-specific concepts. A number of such ontologies have been proposed for different purposes. Unfortunately, due to different ontology design patterns, some of these ontologies are not interoperable out of the box. In order to increase the cross-domain interoperability of research data within NFDI and the EOSC, we need to harmonise the used TLO and MLO concepts to a common ground. The Section-Metadata working group Ontology Harmonisation & Mapping was formed to coordinate and guide such an alignment work, by recommending, providing and/or developing mappings, frameworks and tools [1]. We started to analyse which ontologies are used among many NFDI consortia and found that using only one specific TLO & MLO framework will not meet the different ontological requirements. Consequently, we need to provide formal mappings between many commonly used concepts, such as process, information, characteristic or method, defined in a variety of common TLO and MLO as well as SKOS vocabularies and other reference terminologies. Since this will be a complex and labour intensive process, we must also look at less complex solutions for interdisciplinary data integration. One approach could be to focus first on the observational part of research data.

I-ADOPT as a Possible NFDI Wide Interoperable Variables Description Framework

Observational data play a crucial role in many scientific disciplines and the need for combined analysis of findings across disciplines is increasing. Although containing similar information from the semantic point of view, observations are described using a variety of metadata standards and semantic resources, introducing a high level of heterogeneity between data management systems. In order to cope with those issues, members of the Working Group InteroperAble Descriptions of Observable Property Terminology (RDA I-ADOPT WG) proposed an interoperability framework and an ontology to semantically describe observable properties [2]. The I-ADOPT ontology describes a variable (i.e. what is observed, measured, or derived) as a complex concept consisting of at least an entity of interest and its property. Additional components can be used to describe the context as well as constraints on the entities. It is possible to use this framework in a broader context with other representations of observations (e.g. *Characteristic* in OBOE or *study design* in OBI).

In Figure 1, we show the implementation of the framework on a complex variable example “Egg production rate of female Calanus finmarchicus”. The property “rate” is applied to the object of interest “egg production” in the context of the entity “Calanus finmarchicus”. The components of such an I-ADOPT based variable are meant to be mapped to other commonly used terminologies [3], as for example, the Phenotype And Trait Ontology (PATO) to describe properties and constraints. Integrating other terminologies this way, adds a knowledge layer to terminologies that lack such expressivity or can reduce the need for additional instances when more general modelling patterns are used (see Figure 1c). An overall mapping effort reduction is expected, as only the variable components reference terms need to be mapped for interdisciplinary data integration.

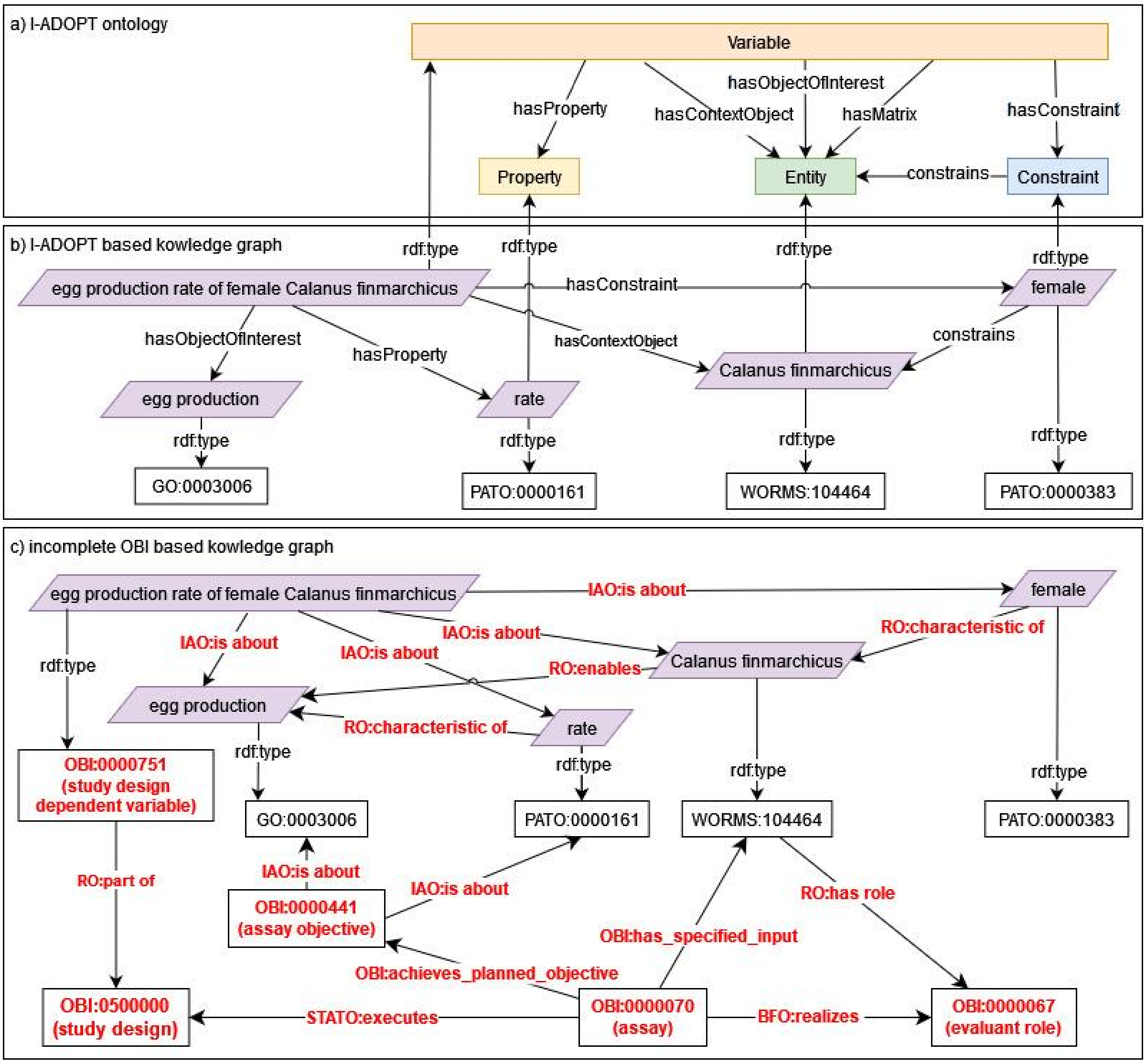

Figure 1: Semantic differences between the I-ADOPT and OBO Foundry framework

In the context of NFDI4Biodiversity [4], we are implementing the I-ADOPT framework to enable interoperability between what is observed or measured in different contexts in order to facilitate search and data integration between multiple data provider partners like among others PANGAEA [5]. The semantic model of Schema.org’s *Dataset* type is connected through the *variableMeasured* property to the I-ADOPT model. All necessary terminologies for the I-ADOPT components are collected in our terminology repository and service [6] together with their mappings [7].

I-ADOPT Annotator

A widespread adoption of such a framework depends on the automation of variable annotation. Yet, mapping I-ADOPT's variable components to external terminologies is not a trivial task. It is especially difficult with ontologies that cover a much broader domain and use a more complex axiomatisation, as concepts of such ontologies might possibly be mapped to I-ADOPT's *entity* as well as *property* or *constraint* concept. This makes it hard to decide automatically with high confidence which mapping fits the actual context, as it requires more complex algorithms. For highly specialised taxonomies, like WORMS (the World Register of Marine Species) [8] or ITIS (the Integrated Taxonomic Information System) [9], on the other hand, an automatic and high confidence based mapping to I-ADOPT is rather simple.

For PANGAEA, we have made first efforts to map the assignment of terms in its parameters [10] to I-ADOPT. To this end, an existing annotation service [11] is currently extended to use rules, ontologies, and grammar to decide how to map a term recognized in a parameter to I-ADOPT. Figure 2 shows a possible output of the service as JSON. In the example, the term 'Calanus finmarchicus' is recognized as 'ContextObject' because all WORMS entries are considered to be of type *entity* and since the precise I-ADOPT role (e.g. object of interest, matrix or context object) cannot be distinguished the most generic role 'ContextObject' is listed here, 'egg production rate' is recognised as a *property* based on the identification of the term *rate* which is listed in PATO as a direct child of a term that is a subclass of *quality* and in contrast the PATO term *female* is considered as a *constraint* since it has a larger graph distance to a PATO entry which is a subclass of a quantity.

```
{
  "parameter": "Calanus finmarchicus, egg production rate per female, wet",
  "dim_match": {},
  "text_match": [
    {
      "fragment": "Calanus finmarchicus",
      "start_offset": 0,
      "end_offset": 20,
      "term": [
        {
          "id": 1053596,
          "name": "Calanus finmarchicus",
          "score": 103.663475,
          "terminology": "WoRMS, Aphia 1.0",
          "description_uri": "https://www.marinespecies.org/aphia.php?p=taxdetails&id=104464",
          "iadopt_type": "ContextObject"
        }
      ]
    },
    {
      "fragment": "egg production rate",
      "start_offset": 22,
      "end_offset": 41,
      "term": [
        {
          "id": 1073136,
          "name": "rate",
          "score": 46.096287,
          "terminology": "PATO",
          "description_uri": "http://purl.obolibrary.org/obo/PATO_0000161",
          "iadopt_type": "Property"
        }
      ]
    },
    {
      "fragment": "female",
      "start_offset": 46,
      "end_offset": 52,
      "term": [
        {
          "id": 1074133,
          "name": "female",
          "score": 46.831738,
          "terminology": "PATO",
          "description_uri": "http://purl.obolibrary.org/obo/PATO_0000383",
          "iadopt_type": "Constraint"
        }
      ]
    },
```


Figure 2: JSON output of Param-Annotator

Conclusion

Our aim within the NFDI working group and related international initiatives is to provide both a high level linking of the NFDI ecosystem terminologies through the mapping of TLOs and MLOs as well as concentrate our efforts on specific ones which could rapidly lead to interoperabilability across disciplines.

Competing interests

The authors declare that they have no competing interests.

Funding

Part of this work has been done within the NFDI4Chem and NFDI4Biodiversity project, both funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme (grant no. 441958208 & 442032008).

Acknowledgement

We want to thank all participants of the Section Metadata Ontology Harmonisation & Mapping Working Group for their feedback and input to this work.

References

1. I. Anders, T. Arera-Rütenik, S. Arndt, R. Baum, N. Betancort, I. Blümel, C. Busse, A. Daniel, F. Engel, L. Ghiringelli, S. Hachinger, H. Israel, N. Karam, A. Kranz, R. Lenz, D. Linke, T. Petrenko, L. Rossenova, D. Schulz, C. Weiland, H. Weiland, C. Wiljes, M. Özaslan, N. Kockmann. (2022). Ontology Harmonization and Mapping - Working Group Charter (NFDI section-metadata) (1.0). Zenodo. https://doi.org/10.5281/zenodo.6726519
2. B. Magagna, I. Rosati, M. Stoica, S. Schindler, G. Moncoiffe, A. Devaraju, J. Peterseil, R. Huber: The I-ADOPT Interoperability Framework for FAIRer Data Descriptions of Biodiversity. JOWO 2021.
3. S. Schindler, G. Moncoiffé. (2022). List of Terminologies. https://i-adopt.github.io/terminologies/list/all. Accessed 2023-04-26.
4. GFBio e.V.. (2023). NFDI4Biodiversity Homepage. https://nfdi4biodiversity.org. Accessed 2023-04-26.
5. PANGEA, Data Publisher for Earth & Environmental Science. (2023). https://pangaea.de. Accessed 2023-04-26.
6. Karam, N.; Fichtmueller, D.; Gleisberg, M.; Becker, F.; Tolksdorf, R.; Müller-Birn, C.; Paschke, A. & Güntsch, A. (eds.) 2014 onwards. The Terminology Service of the German Federation for Biological Data (GFBio) - Service of semantic technologies in scientific environments. - http://terminologies.gfbio.org. Accessed 2023-04-26.
7. N. Karam, A. Khiat, A. Algergawy, M. Sattler, C. Weiland, M. Schmidt: Matching biodiversity and ecology ontologies: challenges and evaluation results. Knowl. Eng. Rev. 35: e9 (2020).
8. WoRMS Editorial Board (2023). World Register of Marine Species. Available from https://www.marinespecies.org at VLIZ. Accessed 2023-04-26. https://doi.org/10.14284/170
9. Retrieved [April, 26, 2023], from the Integrated Taxonomic Information System (ITIS), www.itis.gov. https://doi.org/10.5066/F7KH0KBK
10. M. Diepenbroek et al (2017): Terminology supported archiving and publication of environmental science data in PANGAEA, Journal of Biotechnology 261 (2017) 177-186, https://doi.org/10.1016/j.jbiotec.2017.07.016
11. PANGAEA Web Services Server. (2023). ws.pangaea.de/param-annotator. Accessed 2023-04-26.