Vocabulary Interlinking using Crossovers: Nature FIRST Use Case of Biodiversity*

Albin Ahmeti[1,2,*,†], Robert David[1,3,*,†], Martin Kaltenböck[1,*,†], Artem Revenko[1,*,†] and Jan-Kees Schakel[4,*,†]

[1]*Semantic Web Company, Austria*
[2]*Vienna University of Technology (TU Wien), Austria*
[3]*Vienna University of Economics (WU Wien), Austria*
[4]*Sensing Clues, Netherlands.*

Abstract

Data spaces provide a standardised architecture for trusted partners to exchange data in a secure and sovereign way, supporting certification and governance for business and industry. To implement this data exchange requires interoperability, which is done by applying standardised descriptions of (meta) data in the form of (controlled) vocabularies. IDS leverages Semantic Web technologies to represent these vocabularies using languages like RDF, OWL, SKOS and SHACL. Furthermore, IDS requires vocabularies to conform to FAIR principles (findable, accessible, interoperable, reusable). To achieve these requirements, data spaces need a platform to manage and provide these vocabularies. PoolParty Semantic Suite, a Semantic Middleware Platform, can take the role of an IDS Vocabulary Hub, supporting data modelling based on Semantic Web standards as well as data publication in an open environment. We present a use case from the Nature FIRST research project, where we use PoolParty Semantic Suite to create an interoperable solution for standardised reporting for wildlife and natural habitats based on vocabularies to support ESG goals. We show how interlinking of vocabularies (crossovers) can support automatic inference to satisfy standardisation requirements and enable interoperability between organisations and legal institutions of countries. With this solution, we not only support sustainability for natural habitats, but also bridge different domain standards to help stakeholders to collaborate.

Keywords

data spaces, interoperability, knowledge graphs, biodiversity, data integration, linked open data, FAIR

1. Introduction

1.1. IDS Vocabularies

International Data Spaces (IDS) [1] is an initiative for sharing data in a way that owners can keep the sovereignty with the goal to enable digital economy. IDS provides secure and trusted data exchange between participants, where details and conditions for the sharing of data can be

*The authors appear in alphabetical order.

albin.ahmeti@semantic-web.com (A. Ahmeti); robert.david@semantic-web.com (R. David); matin.kaltenboeck@semantic-web.com (M. Kaltenböck); artem.revenko@semantic-web.com (A. Revenko); jankees.schakel@sensingclues.org (J. Schakel)
0000-0001-8766-4069 (A. Ahmeti); 0000-0002-3244-5341 (R. David); 0000-0001-6681-3328 (A. Revenko); 0000-0002-4619-9313 (J. Schakel)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

specified and enforced y the architecture. IDS is based on standards, specifying the technical architecture itself in the IDS Reference Architecture Model (IDS-RAM), and leveraging Semantic Web technologies for describing the data of data sets. These descriptions are provided as (controlled) vocabularies, which can be anything from lists of controlled terms up to Thesauri with a high expressivity. They are used for providing metadata descriptions for data sets, but also to annotate data and thereby add semantic information. Vocabularies are provided to the IDS ecosystem by IDS Vocabulary Hubs, which are service components in the distributed architecture. Vocabularies of IDS basically use RDF [2] as a standard for knowledge representation and can easily use more expressive languages which build on RDF, like OWL [2] or SKOS [3], which also makes it easy to conform to FAIR (findable, accessible, iteroperable, reusable) principles [4] required by IDS.

1.2. PoolParty Semantic Suite as a Vocabulary Hub

PoolParty Semantic Suite [1] is a semantic middleware platform with graph management as it's central component. PoolParty supports to build and maintain any RDF based vocabularies, like taxonomies, thesauri and ontologies, using a Web UI and it provides a Web API for easy integration into other systems. These features make PoolParty a well-suited candidate to fulfil the role of an IDS Vocabulary Hub.

1.3. Vocabularies as Linked Data

When it comes to semantic descriptions, the Semantic Web standard support the idea of reusability. RDF is an open format, where different data sets can be easily combined and linked together to represent relations between elements of data and to expand the semantic descriptions. The linked data [5] principle is especially encouraged when determining vocabularies for a use case, because existing vocabularies should be reused, complemented and built upon, instead of recreating them for specific cases. However, for certain domains there are different standards which stem from different backgrounds, but which actually model the same or very similar aspects of a domain. Even more so, in some scenarios regarding government reporting, the vocabularies which have to be used are predefined and cannot be chosen by oneself. In the Nature FIRST project, we have such a scenario where different vocabularies for natural habitats and species are used, but need a specific vocabulary for reporting. Using PoolParty's graph management features, we developed a solutions to automatically determine such semantic descriptions by using linked data principles and domain expertise to provide automatic resolution via crossovers. In the following, we describe this Nature FIRST use case and how we developed this solution for vocabulary management, as prerequisite for cross-resource analytics and reporting.

2. Nature FIRST Use Case

Climate change is one of the main challenges in the recent decades. The effects of climate change have direct impact on biodiversity where existing ecosystems have been changed, destroyed, or

[1] http://poolparty.biz

newly created [6]. Human actions have impact on ecosystems and induce changes which leads to a decline of biodiversity [7]. In order to address these challenges, among others, the so-called *Environmental, Social and Governance* (ESG) approach and the associated SDGs (Sustainable Development Goals) initiative was established, which provides a collection of concrete objectives for sustainable development. Enterprises will do regular ESG reporting regarding different measures such as carbon footprinting, water use, biodiversity and habitat conservation, in order to counter the effects of climate change and the decline of biodiversity. Furthermore, at EU level, designated areas called Natura 2000 sites have been established for protection and conservation, aimed at safeguarding habitats and species. However, the data regarding these sites, habitats, and species is currently dispersed and isolated, resulting in limited usefulness. Data is provided by different organizations and classification systems such as *EUNIS*[2] and *IUCN*[3] in different structure, format and completeness; with IUCN reporting mainly on threatened species aka. "Red List Species." The problem is further exacerbated due to existence of different versions of habitats alongside the habitat names and codes that have changed over time but in fact are equivalent, namely habitat "Subarctic and alpine dwarf Salix scrub" with code S21 in EUNIS ver. 2021 versus "Subarctic and alpine dwarf willow scrub" with code F2.1 in EUNIS ver. 2012. Domain experts have created spreadsheets maintaining the relationships between habitats – describing how one habitat maps to another using *crossovers*, i.e., if they are equivalent (=), superset (>), subset (<) or overlap (#) to the designated habitats in other versions. Despite those efforts, the data is not linked and contextualized to a larger context, and the semantics of such relations are only known to those experts. In addition, the occurrences of species that are known to exist in habitats are written using latin names as strings (*Ursus arctos*), without further connection to the source of truth species URIs for looking up and dereferencing them as things (https://eunis.eea.europa.eu/species/1568). Similarly, data about sites have connections to habitats and species, and are also of spatial form that are provided in shapefiles along with geometric coordinates. This opens new challenges in terms of querying and performing geo-calculations with polygons, in addition to having relations to habitats and species as triple patterns by using GeoSPARQL. The Nature FIRST research project[4] aims at connecting these individual vocabularies using semantic relations, thereby interlinking sites, species and habitats by using cross-references, so-called *crossovers*, into one *Knowledge Graph* (KG) for biodiversity.

2.1. Methodology

The vocabularies linked together in the KG comprises of habitats, species and Natura 2000 sites. There are various data authorities when it comes to habitat and species data, such as EUNIS and IUCN. The data sources are in different formats, schemas and completeness (c.f. Table 1). In addition, within EUNIS there exist different version of habitats (ver. 2017, 2021) that map to a legacy one (ver. 2012). The requirement is to consolidate the data into a KG, with each version having a crossover link to the source of truth or legacy version. The advantage of this approach is that one can automatically generate report data that is already described using a (source) taxonomy by specifying another (target) taxonomy that is interlinked. Each version

[2]European Nature Information System of the European Environment Agency, https://eunis.eea.europa.eu/
[3]The International Union for Conservation of Nature, https://www.iucn.org/
[4]https://www.naturefirst.info/

has its own description, codes and granularity in terms of relationships in the hierarchy. The entities in the KG are linked using relations based on SKOS with well-defined meaning, namely for habitat mapping `exactMatch`, `broadMatch`, `narrowMatch`, or `closeMatch`. In other cases, we use OWL (object) properties, e.g. `hasDiagnosticSpecies` specifying indicator species for a habitat. We can distinguish three cases when building crossovers:

- EUNIS vs IUCN habitats, species resp., mapping by using the common labels;
- EUNIS habitats with links to different versions by using the expert spreadsheet[5], which uses codes such as =, <, > and #; A SPARQL query generates `skos:exactMatch`, `skos:narrowMatch`, `skos:broadMatch` and `skos:closeMatch` after mappings are run;
- Species mentioned only in latin name that we apply concept annotation via NLP techniques to determine their URI (EUNIS Species taxonomy), using relations such as `:hasDominantSpecies`, `:hasDiagnosticSpecies`, or `:hasConstantSpecies`.

2.2. Nature FIRST KG

Table 1 lists the interlinked vocabularies of the Nature First KG. For each vocabulary, we list the given stats such as the input data, number of total concepts, the crossovers to other vocabularies, and the total number of crossover relations within these vocabularies. We use PoolParty's

No #	Project (Taxonomy)	Input data	# Concepts	Crossovers	# Crossovers
1	EUNIS Species	RDF	315316	-	-
2	EUNIS Habitats 2012	RDF	7495	#1 #10	38306 ; 388
3	EUNIS Habitats 2017	XLS	2214	#1 #2	1777 ; 2231
4	EUNIS Habitats 2021	XLS	3558	#1 #2	4869 ; 3765
5	Habitats Annex I	XLS	264	#4	586
6	General habitats	XLS	54	-	-
7	IUCN Species	RDF	15139	#1	2655
8	IUCN Habitats	CSV	252	-	-
9	Natura 2000	CSV, shapefile	27054	#1 #6	240790 ; 139802
10	Corine Land Cover	RDF	65	-	-

Table 1
Nature First KG in numbers.

LOD publishing component, the Linked Data Frontend, to provide the KG based on FAIR principles at [6]. The vocabularies of the KG can be browsed and queried, For each vocabulary, there is a SPARQL endpoint to access it, e.g. for 'EUNIS Habitats 2012' can be accessed here[7]. Moreover, the graph visualisation for all the projects is accessible using PoolParty's GraphViews application at [8].

We created explicit `geonames:nearby` relationships between Natura 2000 sites that in addition have relations to EUNIS species and 'General habitats' via the ontological relationships `:siteHasSpecies` and `:siteHasHabitat` resp. Moreover, the percentage coverage has been

[5]https://www.eea.europa.eu/data-and-maps/data/eunis-habitat-classification/eunis-habitat-classification-review-2017
[6]https://sensingclues.poolparty.biz/
[7]https://sensingclues.poolparty.biz/PoolParty/sparql/Habitats
[8]https://sensingclues.poolparty.biz/GraphViews/

included to specify the percentage of habitat that the site contains using RDF*. We provide such a query (prefixes omitted) in the following that combines `nearby` relations and percentage of habitats in RDF*, which computes the TOP 5 largest habitats that are close to `:AT1101112` area[9].

```
SELECT ?label (SUM(?percentage) as ?sum) (group_concat(?percentage) as ?cnt)
WHERE
{
  :AT1101112 geonames:nearby ?sites .
  <<?sites site:siteHasHabitat ?label>> site:percentageCover ?percentage .
}
group by ?label order by desc(?sum) limit 5
```

Similarly, one can exploit `:siteHasSpecies` relations in order to build recommender systems that can predict *Ursus arctos* movement in respect to sites, based on preferred habitats and species, leveraging observations and reasoning in order to prevent human-wildlife conflict.

3. Conclusions & Future work

The example of the Nature FIRST KG shows how different vocabularies can be semantically interlinked using crossovers to create more expressive and interoperable solutions for describing data. By using these crossovers for inference, we can automatically generate report data that is described using a source vocabulary by specifying another interlinked target vocabulary. Furthermore, it enables use cases, like recommender scenarios, by providing expanded semantic descriptions and reasoning via crossovers. This crossover approach was developed as part of the Nature FIRST project and aims to support the protection of wildlife, biodiversity and sustainability for natural habitats. However, the approach is a general one and can be applied to other domains as well, thereby enriching existing descriptions for data sets based on vocabularies and increasing interoperability for data spaces and beyond.

References

[1] B. Otto, M. t. Hompel, S. Wrobel, International Data Spaces, Springer Berlin Heidelberg, Berlin, Heidelberg, 2019, pp. 109–128. URL: https://doi.org/10.1007/978-3-662-58134-6_8. doi:10.1007/978-3-662-58134-6_8.

[2] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (2022) 71:1–71:37. URL: https://doi.org/10.1145/3447772. doi:10.1145/3447772.

[3] D. Allemang, J. Hendler, Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL, 2 ed., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

[4] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL: https://doi.org/10.1038/sdata.2016.18. doi:10.1038/sdata.2016.18.

[5] Linked Data, Association for Computing Machinery, New York, NY, USA, 2020. URL: https://doi.org/10.1145/3382097.3382103.

[6] H.-O. Pörtner, D. Roberts, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama, D. Belling, W. Dieck, S. Götze, T. Kersher, P. Mangele, B. Maus, A. Mühle, N. Weyer, Climate Change 2022: Impacts, Adaptation and Vulnerability Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 2022. doi:10.1017/9781009325844.

[9] https://sensingclues.poolparty.biz/PoolParty/sparql/Natura2000Site

[7] C. Pruski, D. S. Hensel, The Role of Information Modelling and Computational Ontologies to Support the Design, Planning and Management of Urban Environments: Current Status and Future Challenges, Springer International Publishing, Cham, 2022, pp. 51–70. URL: https://doi.org/10.1007/978-3-031-03803-7_4. doi:10.1007/978-3-031-03803-7_4.