Towards preserving Biodiversity using Nature FIRST Knowledge Graph with Crossovers

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

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

Abstract

Preserving biodiversity, encompassing species and their habitats, is gaining significant attention and becoming a central concern, alongside the focus on climate change. Climate change directly impacts biodiversity and is a prominent aspect of Environmental, Social, and Governance (ESG) criteria. At the 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. To address this issue, we introduce our work on a *Knowledge Graph* (KG) for biodiversity, known as *Nature First KG*. This KG aims to connect various data silos, including information about sites, species, and habitats, through cross-references called *crossovers*. Combining it with a digital twin, we empower recommender use cases such as: preventing human-wildlife conflicts, facilitating species reproduction, and combating illegal poaching to name a few.

Keywords

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

1. Introduction

Climate change is one of the main challenges that has preoccupied mankind in the recent decades. The effects of climate change have critically altered ecosystems and biodiversity all around the world, changes in ecosystem range and distribution, ecosystem composition, local species extinctions or mass mortality events of plants and animals have been observed [1]. Human-induced land cover change has led to environmental impacts such as the decline of biodiversity and other ecosystem services[1]. Similar negative results have been observed with anthropogenic change of land use, i.e., replacing nature with architectural buildings for humans to live in enclosed spaces, as reported in this survey [2]. In order to tackle the issue – in the broad level of climate change – the United Nations (UN) have identified a number of goals to be achieved on the topic of *Environmental, Social and Governance* (ESG).

Furthermore, at the European Union (EU) level, designated areas called Natura 2000 sites

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece

[*] Corresponding author.
albin.ahmeti@semantic-web.com (A. Ahmeti); jankees.schakel@sensingclues.org (J. Schakel); robert.david@semantic-web.com (R. David); artem.revenko@semantic-web.com (A. Revenko)

[1] Land cover accounts – an approach to geospatial environmental accounting, European Environment Agency, https://www.eea.europa.eu/themes/landuse/land-accounting

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 habitat taxonomies alongside 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. For a better presentation, we have summarised all the discussed problems and challenges in Table 1.

In this paper, we present our work towards a *Knowledge Graph* (KG) for biodiversity in the context of Nature FIRST research project[4]—dubbed *Nature FIRST KG*—that connects silos of data, namely sites, species and habitats by using cross-references, so-called *crossovers*. The imported data is heterogeneous, ranging from shapefiles to tabular data, which are then mapped, integrated and consolidated in a KG; afterwards the entities in KG are linked using relations based on crossovers, constituting *Linked Open Data* (LOD). The crossover relations are based on SKOS relations with well-defined meaning, namely for habitat mapping `exactMatch`, `broadMatch`, `narrowMatch`, or `closeMatch`; in other cases we use a bespoke OWL (object) property e.g., `hasDiagnosticSpecies` specifying indicator species for a habitat. The relation from a site to a habitat also contains the coverage information in percentage that has motivated us to use the RDF*[5] (RDF-star) data model for representing the percentages in the relation itself in a compact way akin to property graphs. We summarize our contributions:

- Provide a KG that semantically links disparate information, allowing to traverse and get new insights powering new use cases pertaining to biodiversity;
- Methodology for creating relations from crossovers in KG;
- Publish and Consume KG using LOD frontends, graph view and SPARQL endpoint to comply with FAIR principles.

[2] European Nature Information System of the European Environment Agency (EUNIS/EEA), https://eunis.eea.europa.eu/

[3] The International Union for Conservation of Nature, https://www.iucn.org/

[4] https://www.naturefirst.info/

[5] https://www.w3.org/2021/12/rdf-star.html

Problem	Authority sources	Challenges
Silo-ed data	EUNIS/EEA, IUCN habitats & species	completeness, mapping
Dataset versions	EUNIS/EEA habitats	mapping, semantics
String occurrences	EUNIS/EEA sites, habitats, species	entity extraction
Shape files	EUNIS/EEA sites	GeoSPARQL computations
Natura 2000 sites	EUNIS/EEA sites	reified statements

Table 1
Problems and challenges summarized.

2. Methodology

The data ingested 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 2). 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 report data that is already described using a taxonomy by specifying another taxonomy that is interlinked. Each version has its own description, codes and granularity in terms of broader relationships in the SKOS hierarchy. It is worth mentioning that Red List Species contained the taxonomic rank (`Species->Genus->Family->Order->Class->Phylum->Kingdom`), whereas EUNIS only contained 'genus' as a parent relationship. In both cases, `skos:broader` relationships were created in order to create the hierarchy.

As seen from Table 2, some of the data is already in RDF, while others are non-RDF and need to be transformed using ETL (Extract-Transform-Load). For the transformation, we used UnifiedViews [3] tool that is able to do transformation of tabular data (CSV, XLS) using respective Data Processing Units (DPUs). Each DPU contains a logic where one can configure the mappings of how each column is mapped to a property in the ontology. Regarding the shapefiles, we used GeoTriples [4] application that generates RML[6] mappings from the provided Natura2000 shapefiles[7]. We can distinguish three cases when building crossovers:

- EUNIS vs IUCN habitats, species resp. by using the common labels (latin names);
- EUNIS habitats with links to different versions by using the expert spreadsheet[8], 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`.

Regarding URI management — by following the Linked Data principles [5] — we reused source authority URIs, e.g., http://eunis.eea.europa.eu/habitats/409 and in cases where we ought to

[6] https://rml.io/specs/rml/
[7] https://www.eea.europa.eu/data-and-maps/data/natura-14/natura-2000-spatial-data
[8] https://www.eea.europa.eu/data-and-maps/data/eunis-habitat-classification/eunis-habitat-classification-review-2017

generate the URI, we made sure that it conforms to our Linked Data frontend so that it becomes dereference-able, e.g., https://sensingclues.poolparty.biz/HabitatClassificationScheme/237.

3. Nature FIRST KG

In Table 2 is shown the current snapshot of Nature First KG. Per each project (taxonomy) are given the stats such as the input data, number of total concepts, the crossovers with respect to other projects, and the total number of relations with respect to other projects. The relations are only materialized in direct relationships without storing the inverse relations - as also seen from the 'no value' (-) for #1 EUNIS Species.

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 2
Nature First KG in numbers.

The Linked Data frontend can be used to browse the projects and is accessible here[9], whereas the SPARQL endpoint for a specific project, e.g. for 'EUNIS Habitats 2012' can be accessed here[10]. Moreover, the graph visualisation for all the projects is accessible using GraphViews application[11]. The aforementioned URIs ensure that the approach complies with FAIR.

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 included to specify the percentage of habitat that the site contains using RDF*. We provide such a query 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[12].

```
PREFIX geonames: <https://www.geonames.org/ontology#>
PREFIX site: <https://sensingclues.poolparty.biz/SiteOntology/>
PREFIX : <https://sensingclues.poolparty.biz/Natura2000Site/>

SELECT ?label (SUM(?percentage) as ?sum) (group_concat(?percentage) as ?cnt)
WHERE
{
  :AT1101112 geonames:nearby ?sites .
  <<?sites site:siteHasHabitat ?label>> site:percentageCover ?percentage .
```

[9] https://sensingclues.poolparty.biz/

[10] https://sensingclues.poolparty.biz/PoolParty/sparql/Habitats

[11] https://sensingclues.poolparty.biz/GraphViews/

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

```
}
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. On top of this, one can use SPARQL *query federation* using SERVICE keyword in order to query different SPARQL endpoints and join results based on common variables. This system combined with a digital twin [6] is useful as it provides observations and reasoning that can be leveraged in order to prevent a human-wildlife conflict.

4. Conclusions & Future work

We have created a first version of Nature FIRST KG that can be used to power different use cases that pertain biodiversity, addressing the problems reported in Table 1. We plan to enrich our KG and ingest new sources that are related to site conservation, threats, treatment actions and plans. Similarly, we are planning to add *Ecological Networks* [7] as a backbone to our KG in order to perform different reasoning tasks. This infrastructure will be used to build recommender systems that predict the movement of *Ursus arctos* and other relevant species in the context of Nature FIRST research project. We will also study related knowledge graphs on biodiversity such as Ozymandias[13] and relevant parts of Wikidata, in order to reuse, link and query those KGs in conjunction with the Nature FIRST KG.

Acknowledgments

We thank Boris Hinojo and Emil Zegers for their assistance and constructive feedback.

References

[1] 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.

[2] 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.

[3] T. Knap, P. Hanecák, J. Klímek, C. Mader, M. Necaský, B. V. Nuffelen, P. Skoda, Unifiedviews: An ETL tool for RDF data management, Semantic Web 9 (2018) 661–676. URL: https://doi.org/10.3233/SW-180291. doi:10.3233/SW-180291.

[4] K. Kyzirakos, D. Savva, I. Vlachopoulos, A. Vasileiou, N. Karalis, M. Koubarakis, S. Manegold, Geotriples: Transforming geospatial data into rdf graphs using r2rml and rml mappings, Journal of Web Semantics 52-53 (2018) 16–32. URL: https://www.sciencedirect.com/science/article/pii/S1570826818300428. doi:https://doi.org/10.1016/j.websem.2018.08.003.

[5] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers, 2011. URL: https://doi.org/10.2200/S00334ED1V01Y201102WBE001. doi:10.2200/S00334ED1V01Y201102WBE001.

[6] K. de Koning, J. Broekhuijsen, I. Kühn, O. Ovaskainen, F. Taubert, D. Endresen, D. Schigel, V. Grimm, Digital twins: dynamic model-data fusion for ecology, Trends in Ecology and Evolution (2023). doi:10.1016/j.tree.2023.04.010.

[7] G. Torta, L. Ardissono, L. L. Riccia, A. Savoca, A. Voghera, Representing ecological network specifications with semantic web techniques, in: D. Aveiro, J. L. G. Dietz, J. Filipe (Eds.), Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - (Volume 2), Funchal, Madeira, Portugal, November 1-3, 2017, SciTePress, 2017, pp. 86–97. URL: https://doi.org/10.5220/0006573500860097. doi:10.5220/0006573500860097.

[13] https://ozymandias-demo.herokuapp.com