





BiodivOnto: Towards a Core Ontology for Biodiversity

Nora Abdelmageed^{1,2,3} , Alsayed Algergawy¹ ,
Sheeba Samuel^{1,3} , and Birgitta König-Ries^{1,3} 

¹ Heinz Nixdorf Chair for Distributed Information Systems

² Computer Vision Group

³ Michael Stifel Center Jena

Friedrich Schiller University Jena, Germany

`firstname.lastname@uni-jena.de`

Abstract. Biodiversity is the variety of life on earth which covers the evolutionary, ecological, and cultural processes that sustain life. Therefore, it is important to understand where biodiversity is, how it is changing over space and time, the driving factors of these changes and the resulting consequences on the diversity of life. To do so, it is necessary to describe and integrate the conditions and measures of biodiversity to fully capture the domain. In this paper, we present the design of a core ontology for biodiversity aiming to establish a link between the foundational and domain-specific ontologies. The proposed ontology is designed using the fusion/merge strategy by reusing existing ontologies and it is guided by data from several resources in the biodiversity domain.

Keywords: Biodiversity · Knowledge Representation · Core Ontology.

1 Introduction

The recent IPBES global assessment⁴ foresees a dramatic decline in biodiversity and caused by this a dramatic decline in important ecosystem functions. To preserve biodiversity, research to understand its underlying mechanisms is needed which requires integrated data [6]. An increasing amount of heterogeneous data is generated and publicly shared in biodiversity research. There are also a lot of efforts to semantically describe biodiversity datasets and research outputs. Multiple ontologies, like ENVO⁵ and IOBC⁶, model specific parts of the domain. However, in order to support integrative biodiversity research, there is a growing need to bridge between the more refined biodiversity concepts and general concepts provided by the foundational ontologies.

Core ontologies provide a precise definition of structural knowledge in a specific field that connects different application domains [3, 4, 10]. They are located

⁴ <https://ipbes.net/global-assessment>

⁵ <https://bioportal.bioontology.org/ontologies/ENVO>

⁶ <https://bioportal.bioontology.org/ontologies/IOBC>

in the layer between upper-level (fundamental) and domain-specific ontologies, providing the definition of the core concepts from a specific field. They aim at linking general concepts of a top-level ontology to more domain-specific concepts from a sub-field. Looking at the biodiversity domain, one can observe that existing ontologies tend to model parts of the domain while ignoring related parts. Furthermore, most of them connect directly to one of the existing foundational ontologies, such as BFO⁷ and GFO⁸. This results in a number of challenges, e.g., the same concept can be represented in a different level of abstraction and use in different ontologies.

In this paper, we propose the design of a core ontology for the biodiversity domain using a semi-automatic approach to overcome these problems. We make use of the fusion/merge strategy [9] during the design of the core ontology, where the new ontology is developed by assembling and reusing one or more ontologies. Our design is guided by data from several databases in the biodiversity field. In particular, we develop a four-stage pipeline involving biodiversity experts and computer scientists at different phases. A set of heterogeneous biodiversity data sources is collected and analyzed. We make use of the existing ontologies from Bioportal⁹ and AgroPortal¹⁰ for extracting keywords from the collected data repository. This set of extracted terms is then filtered and revised to construct the final list of keywords. Using automated approaches of clustering and the help of biodiversity experts, we generate the list of core concepts. The links between the core concepts are discussed and determined by the domain experts.

2 Methodology

In this section, we describe the main steps of the proposed pipeline.

Data Acquisition: The aim of this step is to get sufficient data sources from which we can extract relevant terms. To this end, we have developed a crawling method, as shown in Figure 1, considering structured and unstructured data resources. To extract relevant unstructured data, first a relaxed version of the QEMP corpus [7] is used and a number of keywords, such as ‘*abundance*’, ‘*benthic*’, ‘*biomass*’, ‘*carbon*’, ‘*climate change*’, ‘*decomposition*’, ‘*earthworms*’, ‘*ecosystem*’ have been selected. The selected set of keywords is used later as input to the Semedico search engine [1] to get relevant publications from PubMed. Among them, 100 abstracts have chosen, as shown in Figure 1 reflecting the biodiversity domain by applying an iterative manual process for revision and cleaning for the crawled data. To take tabular data into consideration, we have used two well known data portals with very different characteristics (*BEFChina*¹¹ and

⁷ <https://bioportal.bioontology.org/ontologies/BFO>

⁸ <https://bioportal.bioontology.org/ontologies/GFO>

⁹ <https://bioportal.bioontology.org/>

¹⁰ <http://agroportal.lirmm.fr>

¹¹ <https://china.befdata.biow.uni-leipzig.de/>

*data.world*¹²). The result of this phase is a data repository¹³ which contains 100 abstracts, more than 50 tables, some datasets are given by multiple tables and, 50 metadata files.

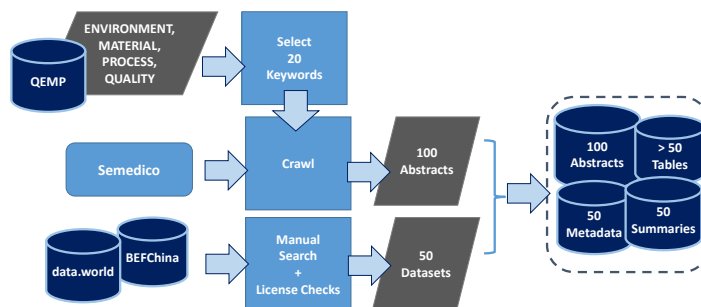


Fig. 1: Crawling phase

Term Extraction: Once we have the data repository, the next step is to extract domain-specific terms. To this end, we manually annotated the collected data following the annotation scheme in [7] making use of the same ontologies and adding more important ontologies knowledge bases, like *IOBC*, *SWEET*¹⁴, *ECOCORE*¹⁵, *ECISO*¹⁶, *CBO*¹⁷, *BCO*¹⁸ and the *Biodiversity A-Z* dictionary¹⁹ to cover wider ranges of terms. During the extraction process, several challenges have been addressed. Our main challenge is the handling of compound words. For example, *photosynthetic O2 production* is expanded into the following keyword list: [“photosynthetic”, “O2”, “O2 production”, “photosynthetic O2 production”]. Finally, the extracted list of terms has been enriched using other existing resources: 1) annotated keywords in QEMP corpus, 2) keywords from AquaDiva²⁰ project, and 3) soil related keywords [11].

Keywords Filtration: To get a final list of relevant terms, we applied an automatic filtration step, where we normalized keywords to be case insensitive and in a singular form. Furthermore, we manually revised the final list of keywords to exclude spelling mistakes. At the end of this step, we have 1107 unique keywords, which is 1.8x of QEMP corpus in size and covers a broader range of Biodiversity.

¹² <https://data.world/>

¹³ <https://github.com/fusion-jena/BiodivOnto/tree/main/data>

¹⁴ <https://bioportal.bioontology.org/ontologies/SWEET>

¹⁵ <https://bioportal.bioontology.org/ontologies/ECOCORE>

¹⁶ <https://bioportal.bioontology.org/ontologies/ECISO>

¹⁷ <https://bioportal.bioontology.org/ontologies/CBO>

¹⁸ <https://bioportal.bioontology.org/ontologies/BCO>

¹⁹ <https://www.biodiversitya-z.org/>

²⁰ <http://www.aquadiva.uni-jena.de/>

Concepts and Relations Determination: Given the huge output list from the previous step, we have automatically calculated the intersection among our work, QEMP and AquaDiva lists. This yields a narrowed list of keywords which we define as *Seeds* as they are the most important keywords and are common among various projects dealing with Biodiversity. We have then applied a distance-based clustering technique with the objective to assign each of the remaining words to the closest seed. Seeds and words are represented by 300D word embedding using word2vec [5]. Our selected metric is the cosine similarity. Afterwards, we have manually revised the created clusters multiple times. For each revision iteration, we check how the remaining keywords are grouped, discuss the results with Biodiversity experts, and modify the selected seeds by tending to more general concepts. In the last iteration, we performed the WordNet [8] similarity among the remaining seeds, clusters centroids, such that, if the similarity is 0.0, very unique seed, we pick this seed as a core concept. In case of having some similarity with other seeds, we have checked BioPortal for those seeds and have picked the common ancestor for them. In the previous step, we have used PATO²¹, and SWEET ontologies for looking to a common ancestor. We have discussed our final list of seeds or core concepts with Biodiversity experts. Finally, we discussed the possible relations that could co-occur among our core concepts. Figure 2 represents our core categories and their core links (relations) as been validated by domain experts. Each category has a set of terms as a result of the clustering algorithm. To implement the fusion/merge strategy, we make use of the ontology modularization and selection tool (*JOYCE*) [2] to extract relevant modules from each category. Table 1 shows the results of this process. The next step is to combine (merge) the set of modules in each category to get a core ontology representing the category. All the resources related to the design of the core ontology as well as the current preliminary results are publicly available²².

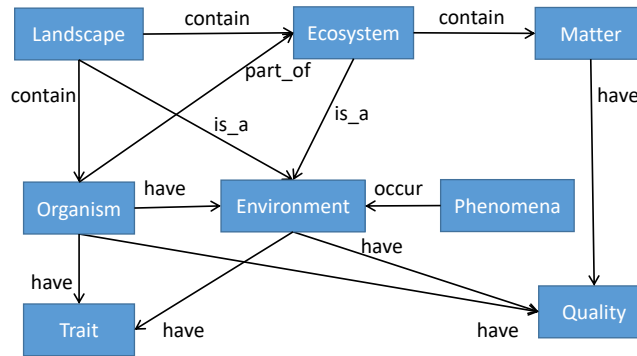


Fig. 2: Core concepts and their relations.

²¹ <https://bioportal.bioontology.org/ontologies/PATO>

²² <https://github.com/fusion-jena/BiodivOnto>

Category	Ontology Modules	Terms sample inside category
Environment	ENVO, ECOCORE, ECSO, PATO	groundwater, garden
Organism	ENVO ECOCORE, ECSO, BCO	mammal, insect
Phenomena	ENVO, PATO, BCO	decomposition, colonization
Quality	ENVO, PATO, CBO, ECSO	volume, age
Landscape	ENVO	grassland, forest
Trait	BCO	texture, structure
Ecosystem	ENVO, ECOCORE, ECSO, PATO	biome, habitat
Matter	ENVO, ECSO	carbon, H2O

Table 1: Core concepts in existing ontologies with examples.

Acknowledgments

The authors thank the Carl Zeiss Foundation for the financial support of the project “A Virtual Werkstatt for Digitization in the Sciences (K3, P5)” within the scope of the program line “Breakthroughs: Exploring Intelligent Systems for Digitization” - explore the basics, use applications”. Alsayed Algergawy’ work has been funded by the *Deutsche Forschungsgemeinschaft (DFG)* as part of CRC 1076 AQUADIVA. Our sincere thanks to Tina Heger (Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB)) as the domain expert.

References

1. Faessler, E., Hahn, U.: Smedico: A comprehensive semantic search engine for the life sciences. In: Proceedings of ACL 2017, System Demonstrations (Jul 2017)
2. Faessler, E., Klan, F., Algergawy, A., König-Ries, B., Hahn, U.: Selecting and tailoring ontologies with JOYCE. In: Knowledge Engineering and Knowledge Management - EKAW 2016 Satellite Events, EKM and Drift-an-LOD, Bologna, Italy, November 19-23, 2016, Revised Selected Papers. vol. 10180, pp. 114–118 (2016)
3. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: SemSur: A core ontology for the semantic representation of research findings. In: SEMANTICS (2018)
4. Garcia, L.F., et al.: The GeoCore ontology: A core ontology for general use in geology. *Computers & Geosciences* **135** (2020)
5. Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
6. Jr., L.M.R.G., et al.: A survey of biodiversity informatics: Concepts, practices, and challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **11**(1) (2021)
7. Löffler, F., Abdelmageed, N., Babalou, S., Kaur, P., König-Ries, B.: Tag me if you can! semantic annotation of biodiversity metadata with the qemp corpus and the biodivtagger. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4557–4564 (2020)
8. Pedersen, T., Patwardhan, S., Michelizzi, J., et al.: Wordnet: Similarity-measuring the relatedness of concepts. In: AAAI. vol. 4, pp. 25–29 (2004)
9. Pinto, H.S., Martins, J.P.: Ontologies: How can they be built? *Knowledge and information systems* **6**(4), 441–464 (2004)
10. Scherp, A., Saathoff, C., Franz, T., Staab, S.: Designing core ontologies. *Applied Ontology* **6**(3), 177–221 (2011)
11. Udovenko, V., Algergawy, A.: Entity extraction in the ecological domain—a practical guide. BTW 2019–Workshopband (2019)