A Data-driven Approach for Core Biodiversity Ontology Development

Nora Abdelmageed[1,2], Alsayed Algergawy[1], Sheeba Samuel[1,2] and Birgitta König-Ries[1,2]

[1]*Heinz Nixdorf Chair for Distributed Information Systems*
[2]*Michael Stifel Center Jena*
Friedrich Schiller University Jena, Germany

Abstract

The biodiversity research domain is composed of diverse scientific subdisciplines resting on various conceptual models developed over time, which results in a large number of biodiversity domain ontologies, each representing a part of the domain. On the one hand, these parts overlap to some degree. On the other hand, the meaning of concepts used often depends on the particular interpretation according to the background. In this paper, we propose *BiodivOnto*, a core ontology including a well-defined and limited set of concepts within the biodiversity domain. This core ontology provides a basis for linking different sub-ontologies. To this end, we develop a semi-automatic data-driven approach that uses clear links between domain experts and knowledge engineers. In particular, the proposed method uses the fusion/merge strategy by reusing existing ontologies and is guided by data from several data resources in the biodiversity domain. The used data as a driving force for the proposed approach has been collected from various resources, including tabular data, unstructured data, and metadata extracted from diverse open data repositories.

Keywords

Biodiversity, Knowledge Representation, Core Ontology

1. Introduction

Understanding biodiversity and the mechanisms underlying it is crucial to preserve this important foundation of human well-being. This demands the management and integration of biodiversity data [1]. A large amount of heterogeneous data is collected and generated in biodiversity research, which means integrating these heterogeneous data remains a big challenge. Semantic web in general and ontologies in particular play a vital role in coping with the

S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity, held at JOWO 2021: Episode VII The Bolzano Summer of Knowledge, September 11–18, 2021, Bolzano, Italy

nora.abdelmageed@uni-jena.de (N. Abdelmageed); alsayed.algergawy@uni-jena.de (A. Algergawy); sheeba.samuel@uni-jena.de (S. Samuel); birgitta.koenig-ries@uni-jena.de (B. König-Ries)
https://fusion.cs.uni-jena.de/fusion/members/nora-abdelmageed/ (N. Abdelmageed); https://fusion.cs.uni-jena.de/fusion/members/alsayed-algergawy (A. Algergawy); https://fusion.cs.uni-jena.de/fusion/members/sheeba-samuel (S. Samuel); https://fusion.cs.uni-jena.de/fusion/members/birgitta-konig-ries (B. König-Ries)
0000-0002-1405-6860 (N. Abdelmageed); 0000-0002-8550-4720 (A. Algergawy); 0000-0002-7981-8504 (S. Samuel); 0000-0002-2382-9722 (B. König-Ries)

integration and management of these heterogeneous data by allowing representing the relevant concepts and relations of a considered domain in a machine-readable format [2]. As a result, several domain-specific ontologies have been developed. For example, statistics on BioPortal[1] show that more than 890 ontologies with 13.387.405 concepts have been developed. Several domain ontologies like ENVO[2] and IOBC[3] exist to model specific areas in the biodiversity domain [3]. However, there is a growing need to bridge the more refined biodiversity concepts and general concepts provided by the foundational ontologies. Foundational ontologies span many fields, modeling the basic concepts and relations that make up the world [4]. Core ontologies provide a precise definition of structural knowledge in a specific field that spans different application domains [5]. Hence, core ontologies provide a bridge between the foundational and subdomain ontologies. Several efforts have been made in different domains to represent the basic categories of the domain knowledge using core ontologies. Several approaches exist in the development of core ontologies, including manual and (semi)automatic ways.

In this paper, we present the design of a core ontology, *BiodivOnto* for the biodiversity domain. We use a semi-automatic approach that includes the usage of fusion/merge strategy [6] for the core ontology development. We developed a four-phase pipeline with biodiversity experts and computer scientists involved at different stages. We collected and analyzed a set of heterogeneous biodiversity data sources, including tabular data, unstructured data, and metadata. To extract keywords from the collected data repositories, we used existing ontologies from Bioportal[4] and AgroPortal[5]. We applied biodiversity experts' recommendations to filter the keywords of interest. We generated the core concepts using automated approaches of clustering. The relations between these core concepts are discussed and determined by the domain experts.

The rest of the paper is structured as follows: In Section 2, we discuss related work. We describe the methodology of developing our core ontology in Section 3. We present our evaluation plan and discuss open issues and future works in the development of the core ontology in the biodiversity domain in Section 4. Finally, we conclude in Section 5.

2. Related Work

Biodiversity aims to study the totality and variability of organisms, their morphology and genetics, life history and habitats, and geographical ranges. It is strongly related to ecosystems' services, such as provision of water and food, and climate regulation. Therefore, it is critically important to understand and conserve it properly [1]. Core ontologies provide a precise definition of structural knowledge in a specific field that connects different application domains [7, 8, 5]. They are located between upper-level (foundation) and domain-specific ontologies, defining the core concepts of a specific field. They aim at linking general concepts of a top-level ontology to more domain-specific concepts from a sub-field.

There is a large number of available foundational ontologies [9], such as BFO [10], GFO[11], SUMO[12], PROTON[13] and, etc. At the same time, there is extensive work to formalize

[1] https://bioportal.bioontology.org/, visited on 05.07.2021
[2] https://bioportal.bioontology.org/ontologies/ENVO
[3] https://bioportal.bioontology.org/ontologies/IOBC
[4] https://bioportal.bioontology.org/
[5] http://agroportal.lirmm.fr

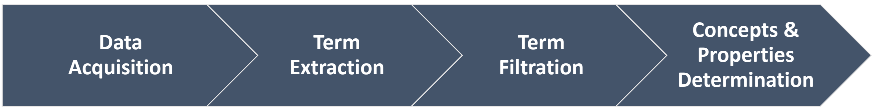


Figure 1: Proposed four-phase pipeline.

knowledge in the biodiversity domain, which results in many domain-specific ontologies. For example, there are 890 ontologies in BioPortal among them ten are titled core ontologies. The core ontology for biology and biomedicine (COB)[6] and the ontology for core ecological entities (ECOCORE)[7] are the only two relevant biodiversity core ontologies. The COB ontology has 73 concepts and 30 relations, while the ECOCORE ontology has more than 2400 concepts. The start of developing both ontologies was in 2020, which indicates a growing interest in developing such core ontologies. However, for both of them, detailed information on how these ontologies have been developed is missing.

A few core ontologies have been introduced in the biodiversity domain; however, several core ontologies developed in other related domains. The work introduced in [14] propose the design of a core ontology to deal with the different types of research activities performed in empirical research, encompassing (physical) sampling, sample preparation, and measurement. SemSur is a core ontology for the semantic representation of research findings[7]. The *GeoCore* ontology has been developed to be used as a core ontology for general use in the geology domain [8]. It makes use of the BFO ontology as an upper-level ontology.

According to [5], core ontologies should combine various features, such as axiomatization, modularity, extensibility, and reusability. Developing a core ontology following these features leads to an elegant way to achieve good interoperability in a complex domain, such as the biodiversity domain. There are different strategies to develop ontologies considering these features, such as fusion/merge and composition/integration strategies[6]. In this work, we use the fusion/merge strategy that builds an ontology by bringing together knowledge from source ontologies.

3. Methodology

The proposed data-driven approach is implemented using the pipeline shown in Figure 1. In the following, we describe main steps of the proposed pipeline.

3.1. Data Acquisition

A first and crucial step is collecting and preparing a sufficient and relevant set of data sources from which we can extract core terms in the biodiversity domain. These data sources should be diverse, including structured data (tabular) and unstructured data (publications). To achieve this goal, we have developed a crawling method, as shown in Figure 2. We have considered two important factors during this step: (i) *data resources*, from which data sources will be

[6] http://purl.obolibrary.org/obo/cob.owl
[7] http://purl.obolibrary.org/obo/ecocore.owl

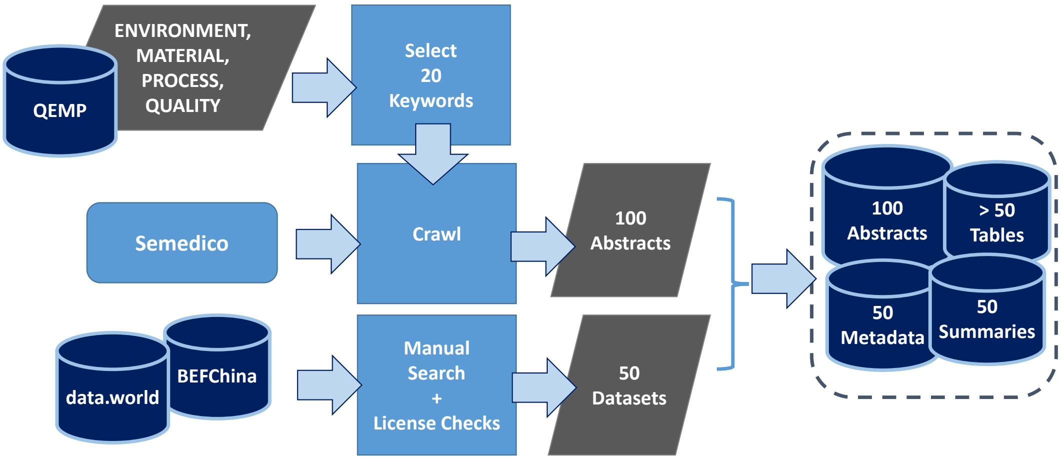


Figure 2: Crawling phase [17].

extracted from and (ii) *a set of keywords* that will be used to query these data resources. For the first point, we consider two well known data portals with very different characteristics (*BEFChina*[8] and *data.world*[9]) to get tabular data. PubMed[10] with more than 32 Million abstracts is deemed to be the data resources for unstructured data. Once identified data resources, the next step is to collect a set of domain-specific keywords that will be used to query these data resources. To this end, we relax a version of the QEMP corpus [15] and a number of keywords, such as *'abundance', 'benthic', 'biomass', 'carbon', 'climate change', 'decomposition', 'earthworms', 'ecosystem'* are selected. The selected set of keywords is used later as input to the Semedico search engine[16] to get relevant publications from PubMed. Among them, 100 abstracts have been chosen, as shown in Figure 2 reflecting the biodiversity domain by applying an iterative manual process for revision and cleaning for the crawled data. The result of this phase is a data repository[11] which contains 100 abstracts, more than 50 tables, some datasets are given by multiple tables and, 50 metadata files. Our selected number of these data sources achieves the balance between biodiversity domain coverage and reasonable human labor time.

3.2. Term Extraction

Once relevant data sources have been collected, the next step is to process them to extract domain-specific terms. To this end, we manually annotated the collected data using GATE tool[12] for document annotation. We have followed the annotation guidelines in [15] making use of the same ontologies and adding more important ontologies and knowledge bases, like *IOBC*,

[8] https://china.befdata.biow.uni-leipzig.de/
[9] https://data.world/
[10] https://www.nlm.nih.gov/bsd/licensee/baselinestats.html
[11] https://github.com/fusion-jena/BiodivOnto/tree/main/data
[12] https://gate.ac.uk/documentation.html

SWEET[13], *ECOCORE*[14], *ECSO*[15], *CBO*[16], *BCO*[17] and the *Biodiversity A-Z* dictionary[18] to cover wider ranges of terms. We also make use of the BioPortal Annotator[19] with the selected ontologies above to fetch the possible annotations for a given term. The extraction and annotation process is not a simple task as it has several challenges to be addressed. On the one hand, some keywords are ambiguous; we could not decide to include them. We keep those keywords in a separate list as *Open Issues*. On the other hand, 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”]. We have enriched the extracted list of terms using other existing resources: 1) annotated keywords in QEMP corpus, 2) keywords from AquaDiva[20] project and 3) soil-related keywords [18]. These existing resources have 578, 222, and 410 keywords, respectively.

3.3. Term Filtration

To get the final relevant terms, we have discussed the *Open Issues* list with domain (biodiversity) experts. Based on their votes on each term, we have decided on whether to include it or not. Some keywords are already filtered out manually at this stage. We applied an automatic filtration step for consistency, 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. Figure 3 illustrates the effect of this phase on the original keywords per each data source of our work, where the figure shows that the most significant number of unique keywords is collected using abstracts from PubMed using the Semedico search engine. However, Figure 3 shows that BEFChina has the least number of collected unique keywords. In addition, we have calculated the number of simple and complex keywords as in Figure 4. The used subset of AquaDiva project has only simple keywords, however, the soil-related keywords are only complex. QEMP and our work have a mixture of both, but our work achieves a better balance.

3.4. Concepts and Relations Determination

In this section, we cover how we have reached our core concepts and their interlinks.

3.4.1. Concepts Determination

Given the vast output list from the previous step, we have automatically calculated the intersection among our work, QEMP, and AquaDiva lists. Such intersection yields a narrowed list

[13] https://bioportal.bioontology.org/ontologies/SWEET
[14] https://bioportal.bioontology.org/ontologies/ECOCORE
[15] https://bioportal.bioontology.org/ontologies/ECSO
[16] https://bioportal.bioontology.org/ontologies/CBO
[17] https://bioportal.bioontology.org/ontologies/BCO
[18] https://www.biodiversitya-z.org/
[19] https://bioportal.bioontology.org/annotator
[20] http://www.aquadiva.uni-jena.de/

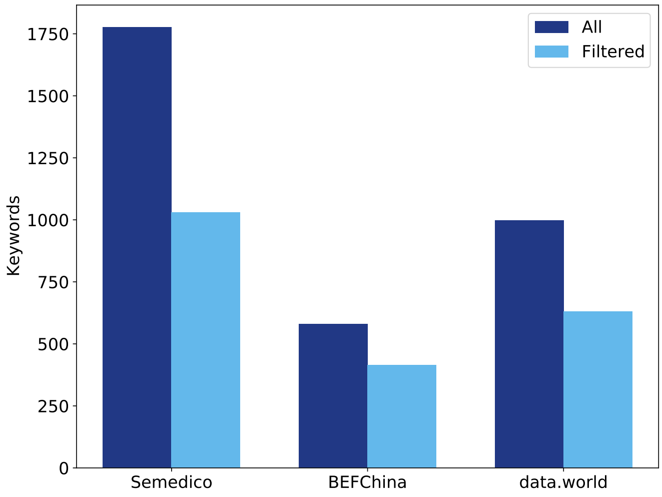


Figure 3: Our extracted keywords vs. external data sources.

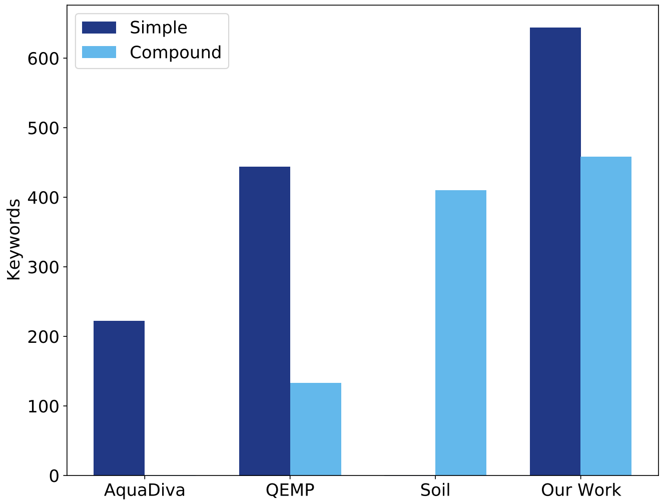
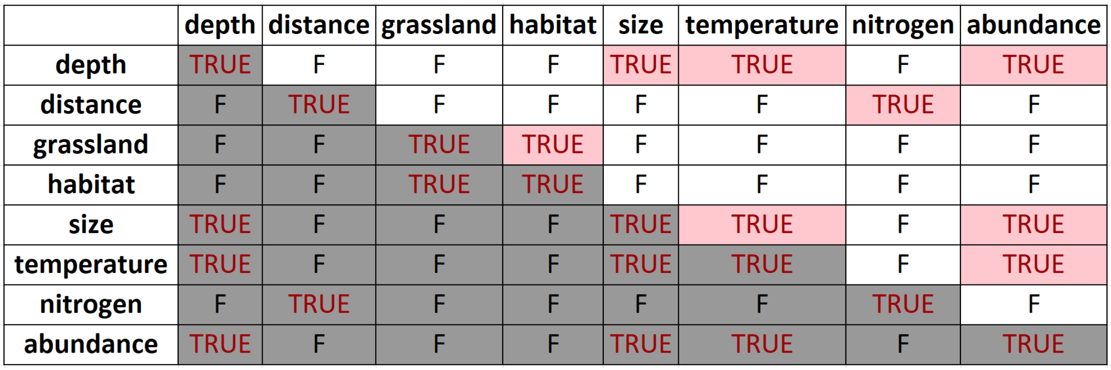


Figure 4: Simple vs. compound keywords in our work and compared to existing data sources.

of keywords which we define as *Seeds Candidates*[21]. For example, *carbon*, *climate*, *composition*, *forest*, *size* and, ... etc. We have considered those 30 terms, as they are the most critical keywords and common among various projects dealing with biodiversity. We have then applied a distance-based clustering technique to assign each of the remaining words to the closest seed. Word embeddings [19], [20], [21] are a good representation for words to capture their semantic meaning. For example, *grassland* is similar to *habitat* in the embedding space, so these pairs of words could be grouped in one cluster. Same case applies for *abundance* and *size*. Word embeddings are commonly used in applications that involve word-word similarity. Seeds and

[21] https://github.com/fusion-jena/BiodivOnto/blob/main/outcome/seeds.md



	depth	distance	grassland	habitat	size	temperature	nitrogen	abundance
depth	TRUE	F	F	F	TRUE	TRUE	F	TRUE
distance	F	TRUE	F	F	F	F	TRUE	F
grassland	F	F	TRUE	TRUE	F	F	F	F
habitat	F	F	TRUE	TRUE	F	F	F	F
size	TRUE	F	F	F	TRUE	TRUE	F	TRUE
temperature	TRUE	F	F	F	TRUE	TRUE	F	TRUE
nitrogen	F	TRUE	F	F	F	F	TRUE	F
abundance	TRUE	F	F	F	TRUE	TRUE	F	TRUE

Figure 5: A sample of seeds WordNet similarity, TRUE has a $threshold >= 0.7$

words are represented by 300D word embedding vectors using word2vec. 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[22] similarity among the remaining seeds, clusters centroids, such that, if the similarity is 0.0, very unique seed, we pick it as a core concept. Figure 5 illustrates a sample of our seeds with WordNet similarity > 0.7. If we have some similarities 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 PATO[22], and SWEET ontologies for looking to a common ancestor *Abstract Seeds*[23]. We have discussed our final list of seeds, *Seeds (Final - Expert)*[24], or core concepts with biodiversity experts. We have based our naming on their recommendation, for example *characteristic* is changed to *trait*.

Figure 6 shows the cluster's members of the Quality core concept. It correctly captures terms with measurements and attributes like width, depth, size, organic nitrogen content, space, and speed. However, it has included non-characteristic terms like tree community and experimental site. The scope of this paper does not yet cover a more detailed and quantitative evaluation. The results of the remaining clusters are available in our GitHub repository[25].

3.4.2. Final Outcome

We have discussed the possible relations that could co-occur among our core concepts. Figure 7 represents our core categories, and domain experts have validated their core links (relations). We have changed the relation between *Quality* and *Trait*, compared to the previous version [17], since we have involved more biodiversity experts, they all agreed on that new relation. 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*)[23] to extract

[22] https://bioportal.bioontology.org/ontologies/PATO
[23] second column in seeds.md file
[24] the last column in seeds.md file
[25] https://github.com/fusion-jena/BiodivOnto/tree/main/outcome/clusters

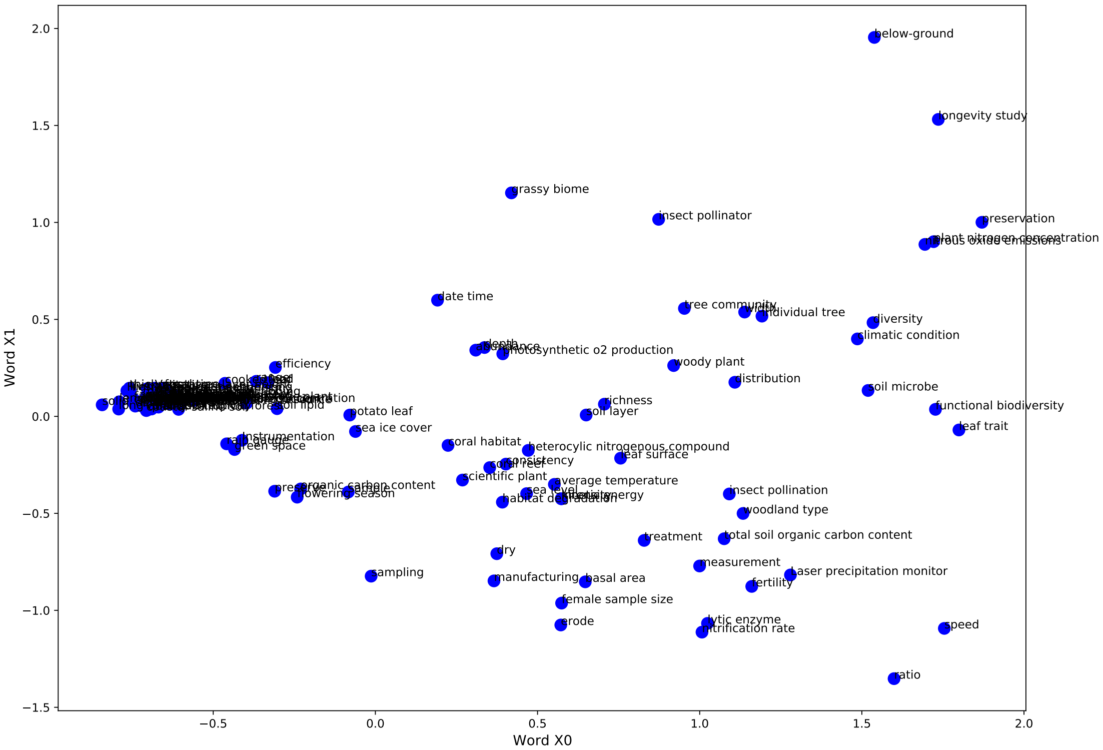


Figure 6: "Quality" cluster in the final iteration. X and Y axis represents the word vectors after the dimensionality reduction.

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 available[26].

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[17].

[26] https://github.com/fusion-jena/BiodivOnto

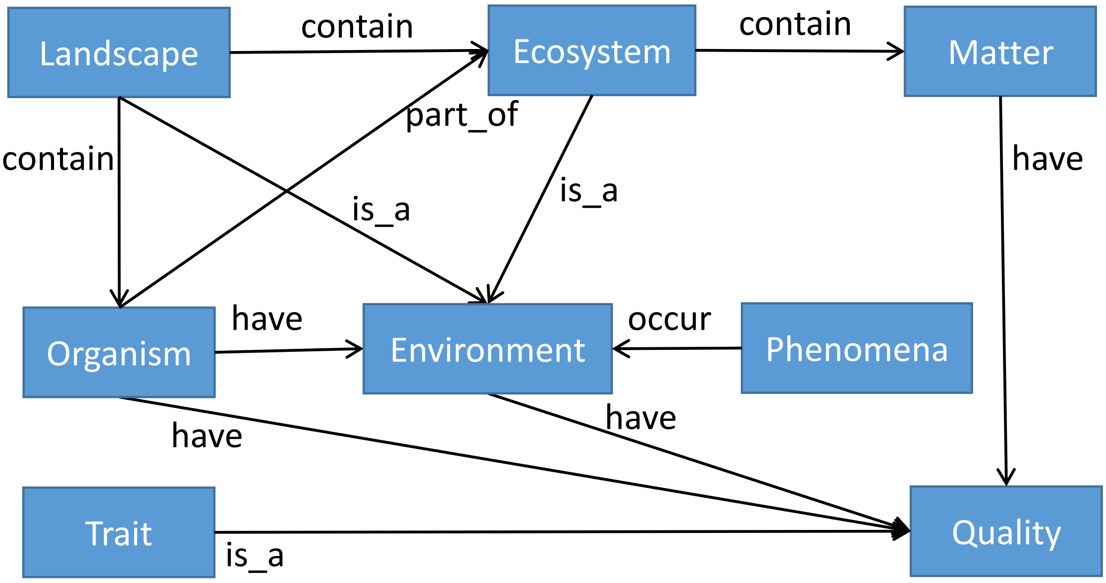


Figure 7: Core concepts and their relations.

4. Discussion and Open Issues

We used a novel data-driven and semi-automatic approach involving both domain experts and computer scientists to develop a core ontology. This approach is different from the traditional approach of developing ontologies manually. We reduce the manual effort of developing core ontology using this semi-automatic data-driven approach. We also extract the crucial concepts from the existing biodiversity domain ontologies to develop our core one. However, there are many open questions regarding the development, quality, and evaluation of our developed core ontology. In the current state, we have determined only the core concepts of *BiodivOnto*. The domain expert at present suggests the relation between the core concepts for the conceptual BiodivOnto core model. We need to determine how the relations between the core concepts could be connected. The relation between core concepts can be determined using the same approach as the core categories are determined. We could reuse the existing properties from the current ontologies to determine the relationship between the core concepts. The other approach is to use the relations validated by the domain experts.

The involvement of domain experts is required for qualitative ontology development. In our methodology, a biodiversity domain expert has been involved in each stage of our pipeline. We have included the other domain experts only after the core concepts creation, only for final evaluation and validation. We have made Quality and Trait be synonyms based on their opinion. Hence, we plan to evaluate the ontology with more domain experts to make the core ontology concrete. The members of each cluster have correctly captured the terms related to the core concept. However, many terms include non-relevant of the core concept. As a result, a detailed and quantitative evaluation is required, in addition to the domain expert evaluation. We also need to compare between data-driven engineering approach for ontology development and manual ontology development using domain experts. In our next phase, we need to bring together the collected modules as an ontology. Currently, it is a conceptual data model with modules from existing ontologies put together. Last but not least, after the complete development of BiodivOnto, we plan to use this model in different biodiversity applications.

5. Conclusions and Future Work

In this paper, we present a semi-automatic approach to build *BiodivOnto*, a core ontology model for Biodiversity domain. Our proposed method makes use of the fusion/merge strategy by reusing existing ontologies and it is guided by data from several data resources in the biodiversity domain. It consists of four steps: data acquisition, term extraction, term filtration and finally, concepts and relation determination.

Since the qualitative evaluation is done by a domain expert. Our future plan considers involving more domain experts. In addition, a quantitative evaluation of our approach, for example, the quality of the automatically created clusters. Moreover, after the complete development of BiodivOnto, we plan to use it in various Biodiversity applications.

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)), Anahita Kazem (German Centre for Integrative Biodiversity Research (iDiv)) and Jitendra Gaikwad (FUSION, Friedrich Schiller University Jena), as the domain experts.

References

[1] L. M. R. Gadelha, et al., A survey of biodiversity informatics: Concepts, practices, and challenges, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 11 (2021). URL: https://doi.org/10.1002/widm.1394. doi:10.1002/widm.1394.

[2] R. Studer, V. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data & Knowledge Engineering 25 (1998) 161 – 197. URL: http://www.sciencedirect.com/science/article/pii/S0169023X97000566. doi:https://doi.org/10.1016/S0169-023X(97)00056-6.

[3] V. Senderov, K. Simov, N. Franz, P. Stoev, T. Catapano, D. Agosti, G. Sautter, R. A. Morris, L. Penev, Openbiodiv-o: ontology of the openbiodiv knowledge management system, Journal of biomedical semantics 9 (2018) 1–15.

[4] N. Guarino, Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy, volume 46, IOS press, 1998.

[5] A. Scherp, C. Saathoff, T. Franz, S. Staab, Designing core ontologies, Applied Ontology 6 (2011) 177–221.

[6] H. S. Pinto, J. P. Martins, Ontologies: How can they be built?, Knowledge and information systems 6 (2004) 441–464.

[7] S. Fathalla, S. Vahdati, S. Auer, C. Lange, SemSur: A core ontology for the semantic representation of research findings, in: SEMANTICS, 2018.

[8] L. F. Garcia, et al., The GeoCore ontology: A core ontology for general use in geology, Computers & Geosciences 135 (2020).

[9] C. Trojahn, R. Vieira, D. Schmidt, A. Pease, G. Guizzardi, Foundational ontologies meet ontology matching: A survey, Semantic Web (2021).

[10] R. Arp, B. Smith, A. D. Spear, Building ontologies with basic formal ontology, Mit Press, 2015.

[11] H. Herre, General formal ontology (gfo): A foundational ontology for conceptual modelling, in: Theory and applications of ontology: computer applications, 2010, pp. 297–345.

[12] I. Niles, A. Pease, Towards a standard upper ontology, in: Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, 2001, pp. 2–9.

[13] I. Terziev, A. Kiryakov, D. Manov, et al., Base upper-level ontology (bulo) guidance, SEKT deliverable 1 (2005).

[14] P. M. Campos, C. C. Reginato, J. P. A. Almeida, Towards a core ontology for scientific research activities, in: ER, 2019.

[15] F. Löffler, N. Abdelmageed, S. Babalou, P. Kaur, B. König-Ries, 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, 2020, pp. 4557–4564.

[16] E. Faessler, U. Hahn, Semedico: A comprehensive semantic search engine for the life sciences, in: Proceedings of ACL 2017, System Demonstrations, Association for Computational Linguistics, 2017, pp. 91–96. URL: https://www.aclweb.org/anthology/P17-4016.

[17] N. Abdelmageed, A. Algergawy, S. Samuel, , B. König-Ries, Biodivonto: Towards a core ontology for biodiversity (2021).

[18] V. Udovenko, A. Algergawy, Entity extraction in the ecological domain–a practical guide, BTW 2019–Workshopband (2019).

[19] Y. Goldberg, O. Levy, word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, arXiv preprint arXiv:1402.3722 (2014).

[20] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[21] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv:1802.05365 (2018).

[22] T. Pedersen, S. Patwardhan, J. Michelizzi, et al., Wordnet: Similarity-measuring the relatedness of concepts., in: AAAI, volume 4, 2004, pp. 25–29.

[23] E. Faessler, F. Klan, A. Algergawy, B. König-Ries, U. Hahn, 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, volume 10180, 2016, pp. 114–118.