Ontology Development for Life Sciences

ANASTASIA PENKOVA, Passau Universität, Germany
YASSIR MAKLOUL, Passau Universität, Germany

The paper introduces us to the important role of ontologies in life sciences, highlighting their necessity for the accessibility of biological data. We describe the ontology development pipeline, which is designed to develop and upgrade ontologies, within life science research, as well. By showing an example, we apply this pipeline to the field of invasion biology, using the hierarchy-of-hypotheses method to build a hierarchy between entities. The resulting ontology offers a base for future work to create more complex relationships between entities, that can become a structured support for data retrieval, further integration, and analysis.

ACM Reference Format:
ANASTASIA PENKOVA and YASSIR MAKLOUL. 2024. Ontology Development for Life Sciences. 1, 1 (February 2024), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Ontologies in the modern world play an important role in organizing, interpreting, and sharing structured information that is understandable by humans and machines. They allow to create the linked data, where many datasets can be connected in a context of semantics. Specifically in life sciences, researchers generate vast amounts of data from experiments, observations, and simulations. Ontologies provide the framework to represent relationships between different biological entities. Usually, to develop an ontology, developers are required to follow the rules of the ontology development pipeline and moderation of experts in interested field.

The structure of this paper is organised in three sections. The first section provides a context for concept in ontology development and references works and information from different papers .....————————. The second section discusses the method we used in order to reproduce an ontology for Invasion Biology. The last section answers the questions from the Q and A session that were asked during the presentations. We conclude by presenting an overview of the work we have done and we provide future directions for our work.

2 RELATED WORK

2.1 Ontologies

2.1.1 *Definition.*

An ontology defines a common vocabulary for a specific domain and includes machine-interpretable definitions of concepts with the relations among them. Ontology in life sciences as a philosophical study became an essential tool for term connections, identifying core concepts, data management, knowledge classification, and integration to collect key information. Essentially, ontology is a multi-directional graph. Many languages and formats are used in ontologies; most

Authors' addresses: ANASTASIA PENKOVA, Passau Universität, Germany, penkov01@ads.uni-passau.de; YASSIR MAKLOUL, Passau Universität, Germany, maklou01@ads.uni-passau.de.

Manuscript submitted to ACM

importantly, OWL (Web Ontology Language), built on top of RDF, provides a richer vocabulary and formal semantics for creating ontologies.

2.1.2 *Background and theoretical framework.*

In the 1970s, the primary focus was on developing methodologies to analyze, represent, and predict outcomes based on large datasets. It influenced the new wave of the Semantic Web. Researchers developed specific and reusable models, namely ontologies. Ontologies became a crucial instrument for describing web semantic entities, their relationships, and categories of things. The Semantic Web is characterized by linked data, such as dates, names, and various properties, supported by technologies such as RDF, OWL, and SPARQL. This framework allows for the creation of data stores, and the development of vocabularies. Semantic Web has become a pivotal infrastructure, extensively applied across many fields including medicine, chemistry, biology, geology, interdisciplinary sciences, etc. [3].

Many existing ontologies have been created manually. Most engineers have used this method of ontology development. However, this approach is notably time-intensive and prone to many errors due to human inability to maintain and update massive datasets. Over the past two decades, researchers have widely utilized ontology development environments (ODEs) such as Protege, Topbraid Composer, Ontostudio, Fluent Editor, VocBench, Swoop, and Obo-edit. Protege is one of the most used ODEs further enhanced by its web version WebProtege [4].

2.2 Use of ontologies

2.2.1 *Semantic search.*

Ontologies play a crucial role in semantic search by providing a structured framework of knowledge. They define a set of concepts and categories within a domain, along with the relationships between these concepts. This structured knowledge helps provide more accurate and relevant search results [3]. This process is divided into 3 main steps:

- Semantic Parse of the natural language query: this intends to translate human language into Abstract Meaning language (AMR) structure.
- Semantic Search: this is used against a collection of ontology annotations and the parsed query is used to search in databases of ontologies.
- Rank: based on match score, return the most relevant ontology annotations and supporting documents. The system evaluates how well each found term matches the user's query. This is done by assessing the closeness of the match between the user's query terms and the ontology terms, as well as considering the context of that terminology. This is also backed up by references to research papers or database entries that support the involvement of the terminology.

Without ontology, the search would lack the depth of understanding of the relationships and hierarchies between biological entities. It would not be able to make the connections that a specific biology ontology provides, such as associating a gene with a specific biological process or molecular function. The search results would be more generic and might require significant user intervention to identify the most relevant and accurate information.

2.2.2 *Life science.*

Ontologies in the life sciences have become beneficial tools for the organization and interpretation of large amounts of biological data. They facilitate the integration of data into searchable libraries. They also lead to a decrease in costs and more enhanced flexibility in data categorization in the fields of biology and life sciences. One of the manifestations of the use and importance of ontologies in life science is the Bioportal database. Bioportal is a repository of biomedical ontologies that hosts more than 1,088 ontologies. Defined within these ontologies, are 14,649,134 classes, 36,286 properties, and 81,457,436 mappings. This vast array of ontologies supports a wide range of applications in biomedical research, from basic science to clinical studies, offering researchers the tools needed for sophisticated data analysis and decision-making.

2.3 The General Pipeline of Ontology Development

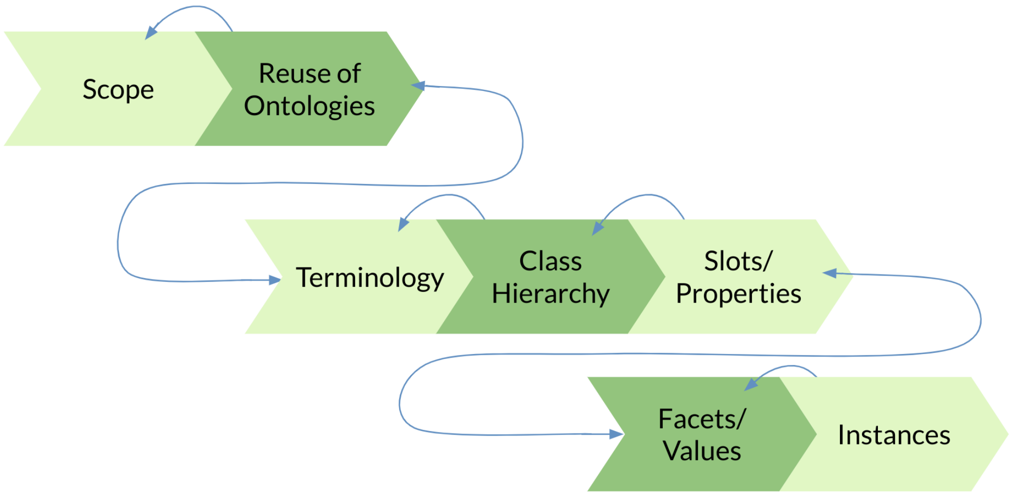
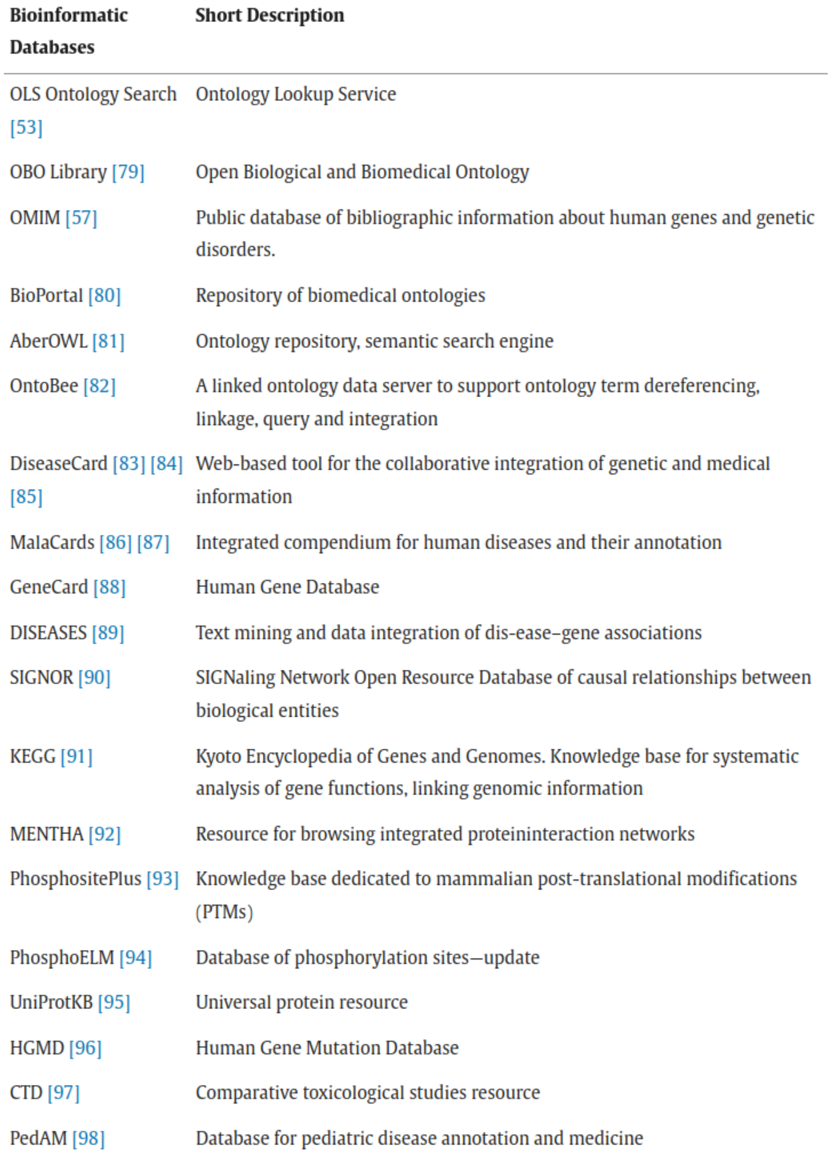


Fig. 1. Pipeline of ontology development

The development of ontologies consists of multiple iterative steps. As shown in Figure 1, this iterative process takes multiple steps to conclude one iteration of the whole development of an ontology.

2.3.1 Scope and reuse of ontologies.



Bioinformatic Databases	Short Description
OLS Ontology Search [53]	Ontology Lookup Service
OBO Library [79]	Open Biological and Biomedical Ontology
OMIM [57]	Public database of bibliographic information about human genes and genetic disorders.
BioPortal [80]	Repository of biomedical ontologies
AberOWL [81]	Ontology repository, semantic search engine
OntoBee [82]	A linked ontology data server to support ontology term dereferencing, linkage, query and integration
DiseaseCard [83] [84] [85]	Web-based tool for the collaborative integration of genetic and medical information
MalaCards [86] [87]	Integrated compendium for human diseases and their annotation
GeneCard [88]	Human Gene Database
DISEASES [89]	Text mining and data integration of dis-ease–gene associations
SIGNOR [90]	SIGNaling Network Open Resource Database of causal relationships between biological entities
KEGG [91]	Kyoto Encyclopedia of Genes and Genomes. Knowledge base for systematic analysis of gene functions, linking genomic information
MENTHA [92]	Resource for browsing integrated proteininteraction networks
PhosphositePlus [93]	Knowledge base dedicated to mammalian post-translational modifications (PTMs)
PhosphoELM [94]	Database of phosphorylation sites—update
UniProtKB [95]	Universal protein resource
HGMD [96]	Human Gene Mutation Database
CTD [97]	Comparative toxicological studies resource
PedAM [98]	Database for pediatric disease annotation and medicine

Fig. 2. Summary of bioinformatic databases

This phase is the main step in the process of development. it consists of determining the domain and reusing existing ontologies. It should always be considered to refine and extend the previously developed sources that might be helpful for our task. Figure 2 shows databases for developed ontologies referenced from the paper [4].

The process of development requires a constant integration of scope in further steps of development. Hence, the iterative nature of development in creating relationships between concepts [3].

2.3.2 *Classes, hierarchy, properties.*

In this phase, we should iteratively list terms of the scope of the ontology make classes through identifiers, and state relationships between them [3].

- Enumerating all important terms in the Ontology: this needs the help of domain experts to validate the terminology through their definitions and meanings. Enumerated terminology must fit the context of the domain in question.
- Defining classes and the class hierarchy: it's a process of specifying "is a" relationships between classes, while keeping in mind similar ontologies and their classes.
- Defining slots/properties/attributes of the classes: it's the process of listing the rest of the attributes and relationships between classes.

These classes and relationships can be integrated using an ontology-development environment (ODE) of choice, for example: Protégé.

2.3.3 Constraints and instances. After being able to define the tree of classes and the relationships between them, comes an important step that consists of defining constraints for properties and creating instances of classes [3].

- Constraints/Values/Facets for the properties: it's the process of specifying additional details like cardinality, etc.
- Instances of classes: that includes constant Testing to validate the ontology structure.

The process in General is iterative and should be constant evaluation and integration of terminology so that the ontology meets standards which we will dissect in the coming section.

2.4 OBO standards

OBO sets standards, guidelines, and principles for ontology development in biomedicine, aiming for namely FAIRness (Findable, Accessible, Interoperable, Reusable) among ontologies [2].

- Findable: Data and resources should be easy to find for both humans and machines.
- Accessible: Once found, data should be easily accessible.
- Interoperable: Data should be structured and represented in a way that allows for seamless integration with other datasets. This enables data to be combined and used across various applications and platforms.
- Reusable: Data should be designed for reuse, allowing for different purposes beyond the original intent. It should be well-described, with clear and accessible usage licenses, ensuring that others can use and build upon it.

2.5 Invasion biology (INBIO) ontology

2.5.1 *The Hierarchy-of-Hypotheses (HoH).*

The Hierarchy-of-Hypotheses (HoH) approach is a method used by Algergawy et al. in their work [1]. It's a method to break down broad ideas or major hypotheses into more specific aspects of hypotheses. By organizing these hypotheses into a hierarchy, researchers can better understand the different components that contribute to the overall idea.

Although the HoH approach has been applied manually in invasion biology, there's a call for formal representation.

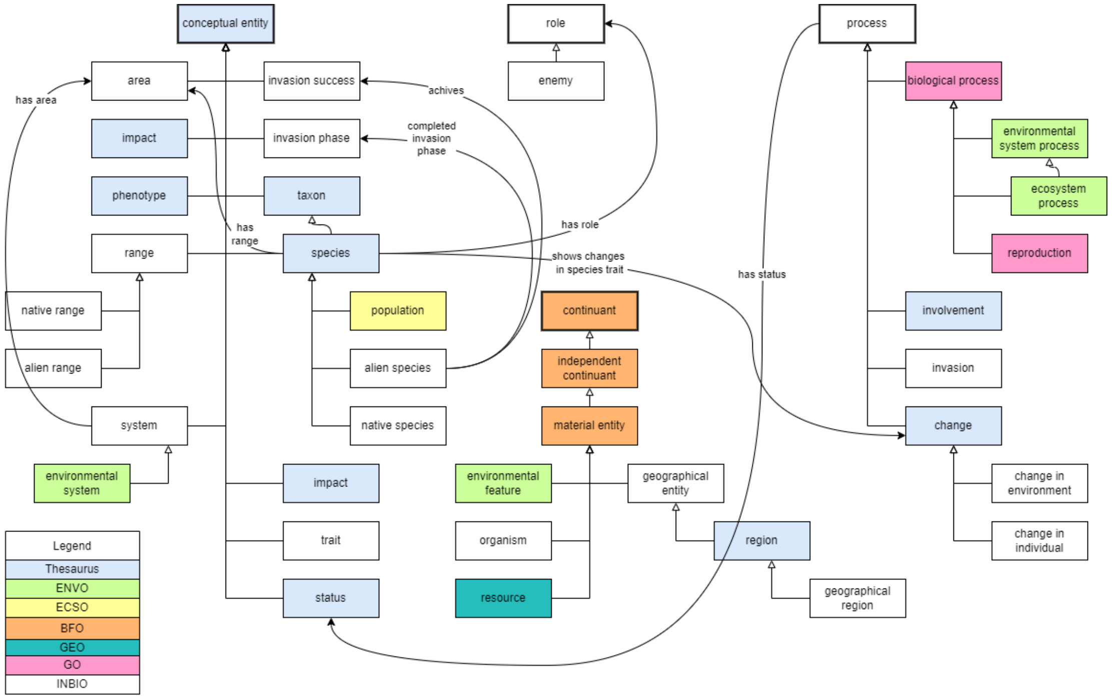


Fig. 3. Ontology tree for invasion biology

2.5.2 *Application of HoH in Invasion Biology.* Based on 11 important hypotheses in invasion biology, the authors of "Towards a Core Ontology for Hierarchies of Hypotheses in Invasion Biology" [1] have developed an ontology for the domain of Invasion biology. Figure 3 references the final ontology tree of the domain proposed by these authors.

3 DISCUSSION

Our work during this seminar was to reproduce the (INBIO) tree while considering only the following steps of ontology development: scope, ontology reuse, terminology, and hierarchy.

3.1 Scope

With the suggestion of Professor Algergawy, we chose the domain “Invasion Biology” since the professor and his colleagues worked on its development, and we found it also interesting to reproduce it.

Invasion biology has generated numerous hypotheses and theories over time, many of which come from scientific research and observations in the field of ecology and biology. These hypotheses are often developed through a combination of empirical studies, experiments, and theoretical modeling.

Algergawy et al. [1] suggested the approach Hierarchy-of-Hypotheses. It is a method used to break down major hypotheses into more specific formulations.

By organizing these hypotheses into a hierarchy, researchers can better understand the different components that form the foundation of the overall idea. We have 11 major hypotheses in the invasion biology field.

Here is an example of how we can break down the hypothesis into a logical graph. We can identify entities and relationships between them and model this kind of structure.

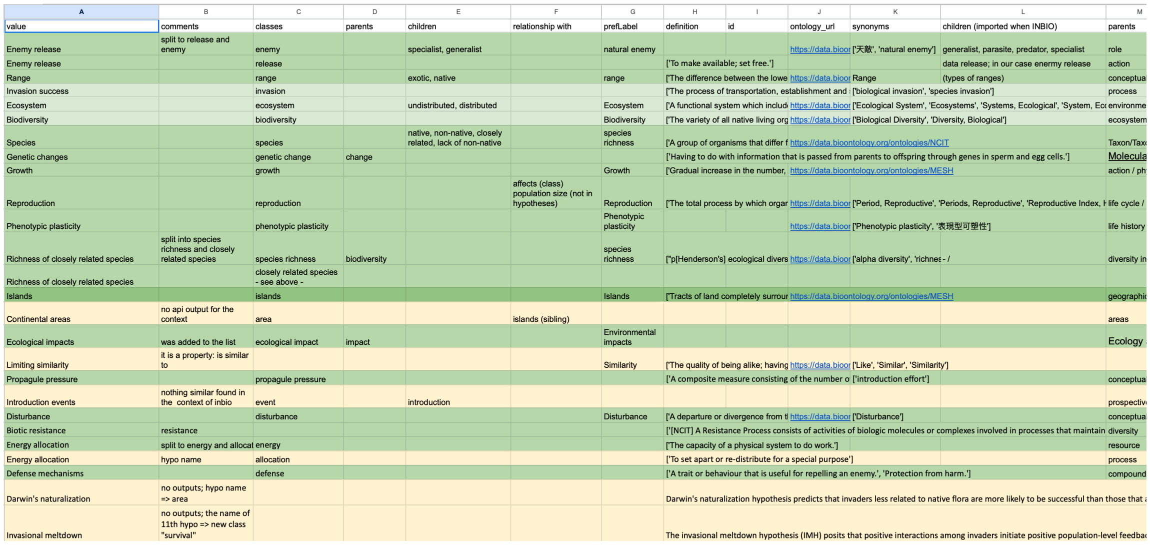
For example, we can take the second hypothesis: "An ecosystem with high biodiversity is more resistant against non-native species than an ecosystem with lower biodiversity."

In this sentence, we can identify entities, such as ecosystems, biodiversity, and species. We can also identify some relationships and parameters: ecosystem has biodiversity; species can be non-native, therefore they can be also native; biodiversity can be high or low. The entities become edges and relationships become nodes in graphs to visualize the developing ontology.

Identification of entities is a routine task that can be automatized by using LLMs. By comparing different LLMs in the market, we decided to implement it by using OpenAI API. It gave us the list of 27 entities found in the hypotheses.

3.2 Reuse of ontologies

For this step, we used the API of the Bioportal. It has many various ontologies that can become an extension of others. Of course, if we have entities like "species", other ontology like NCIT already has a lot of information about species, synonyms, and hierarchy that we add to our ontology. We fetched definitions from different ontologies and checked which definitions fit best. This can be seen through colors in Figure 4.



A	B	C	D	E	F	G	H	I	J	K	L	M
value	comments	classes	parents	children	relationship with	prefLabel	definition	id	ontology_url	synonyms	children (imported when INBIO)	parents
Enemy release	split to release and enemy	enemy		specialist, generalist		natural enemy			https://data.bioor	['天敵', 'natural enemy']	generalist, parasite, predator, specialist	role
Enemy release		release					['To make available; set free.']				data release; in our case enermy release	action
Range		range		exotic, native		range	['The difference between the lowe		https://data.bioor	Range	(types of ranges)	conceptua
Invasion success		invasion					['The process of transportation, establishment and :			['biological invasion', 'species invasion']		process
Ecosystem		ecosystem		undistributed, distributed		Ecosystem	['A functional system which includ		https://data.bioor	['Ecological System', 'Ecosystems', 'Systems, Ecological', 'System, Ec		environme
Biodiversity		biodiversity				Biodiversity	['The variety of all native living org		https://data.bioor	['Biological Diversity', 'Diversity, Biological']		ecosystem
Species		species		native, non-native, closely related, lack of non-native		species richness	['A group of organisms that differ f		https://data.bioontology.org/ontologies/NCIT			Taxon/Tax
Genetic changes		genetic change	change				['Having to do with information that is passed from parents to offspring through genes in sperm and egg cells.']					Molecula
Growth		growth				Growth	['Gradual increase in the number,		https://data.bioontology.org/ontologies/MESH			action / ph
Reproduction		reproduction			affects (class) population size (not in hypotheses)	Reproduction	['The total process by which orgar		https://data.bioor	['Period, Reproductive', 'Periods, Reproductive', 'Reproductive Index, H		life cycle /
Phenotypic plasticity		phenotypic plasticity				Phenotypic plasticity			https://data.bioor	['Phenotypic plasticity', '表現型可塑性']		life history
Richness of closely related species	split into species richness and closely related species	species richness	biodiversity			species richness	["p[Henderson's] ecological divers		https://data.bioor	['alpha diversity', 'richnes	- /	diversity in
Richness of closely related species		closely related species - see above -										
Islands		islands				Islands	['Tracts of land completely surrour		https://data.bioontology.org/ontologies/MESH			geographic
Continental areas	no api output for the context	area			islands (sibling)							areas
Ecological impacts	was added to the list	ecological impact	impact			Environmental impacts						Ecology
Limiting similarity	it is a property: is similar to					Similarity	['The quality of being alike; having		https://data.bioor	['Like', 'Similar', 'Similarity']		
Propagule pressure		propagule pressure					['A composite measure consisting of the number o			['introduction effort']		conceptua
Introduction events	nothing similar found in the context of inbio	event		introduction								prospectiv
Disturbance		disturbance				Disturbance	['A departure or divergence from tl		https://data.bioor	['Disturbance']		conceptua
Biotic resistance	resistance						['[NCIT] A Resistance Process consists of activities of biologic molecules or complexes involved in processes that maintain					diversity
Energy allocation	split to energy and allocat	energy					['The capacity of a physical system to do work.']					resource
Energy allocation	hypo name	allocation					['To set apart or re-distribute for a special purpose']					process
Defense mechanisms		defense					['A trait or behaviour that is useful for repelling an enemy.', 'Protection from harm.']					compound
Darwin's naturalization	no outputs; hypo name => area						Darwin's naturalization hypothesis predicts that invaders less related to native flora are more likely to be successful than those that a					
Invasional meltdown	no outputs; the name of 11th hypo => new class "survival"						The invasional meltdown hypothesis (IMH) posits that positive interactions among invaders initiate positive population-level feedbac					

Fig. 4. Reuse of other ontologies

3.3 Terminology

We filtered out unnecessary results from other ontologies and marked green the terms we will use in our ontology which can be seen in figure 4.

3.4 Class hierarchy

By using OpenAI API and other ontologies we found some relationships between the terms we chose. For example, the ecosystem is a class and its subclasses are distributed ecosystems and undistributed ecosystems; species is a class and its subclasses are non-native and native species. This type of work is iterative and needs the experts' opinion in a specific field like invasion biology.

After several iterations we came up with a possible hierarchy of classes: Figure 5.

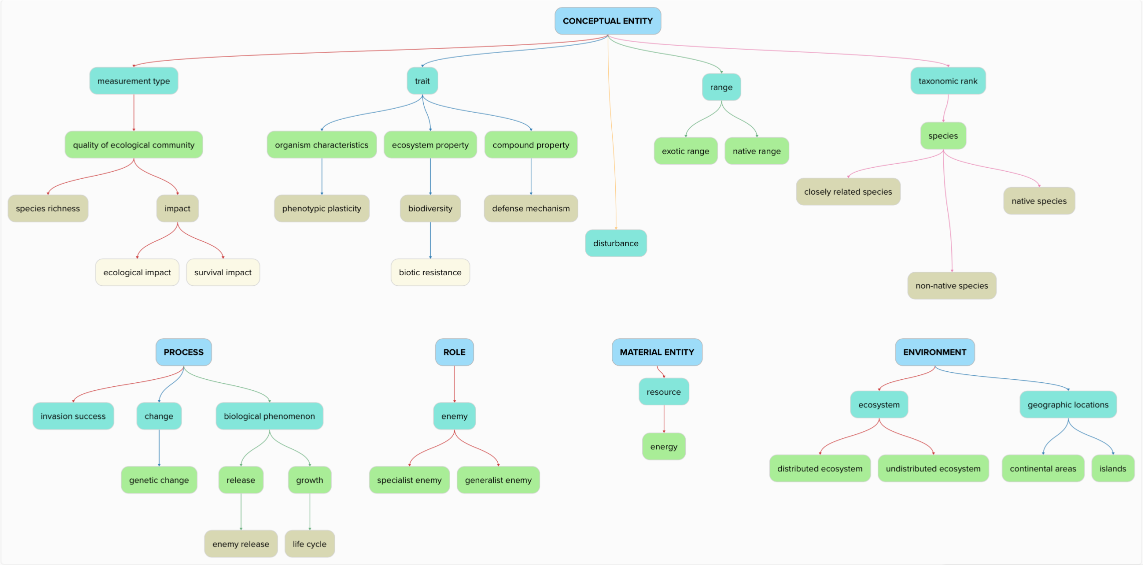
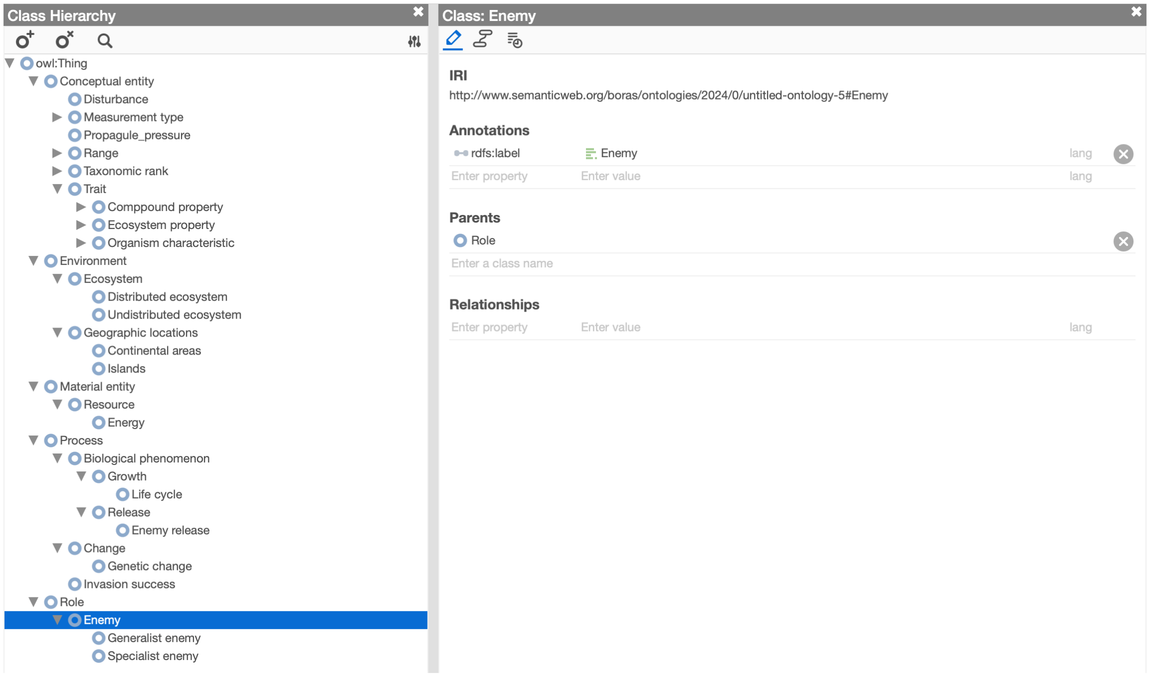


Fig. 5. Hierarchy of classes

Afterward, we formatted this hierarchy to OWL format, the most used and convenient format for ontologies. We used the software Protege which is very helpful to visualize and edit ontologies: figure 6.



Class Hierarchy

- owl:Thing
  - Conceptual entity
    - Disturbance
    - Measurement type
    - Propagule_pressure
    - Range
    - Taxonomic rank
    - Trait
      - Comppound property
      - Ecosystem property
      - Organism characteristic
  - Environment
    - Ecosystem
      - Distributed ecosystem
      - Undistributed ecosystem
    - Geographic locations
      - Continental areas
      - Islands
  - Material entity
    - Resource
      - Energy
  - Process
    - Biological phenomenon
      - Growth
        - Life cycle
      - Release
        - Enemy release
    - Change
      - Genetic change
    - Invasion success
  - Role
    - Enemy
      - Generalist enemy
      - Specialist enemy

Class: Enemy

IRI
http://www.semanticweb.org/boras/ontologies/2024/0/untitled-ontology-5#Enemy

Annotations

rdfs:label	Enemy	lang
Enter property	Enter value	lang

Parents

Role

Enter a class name

Relationships

Enter property	Enter value	lang

Fig. 6. Protege

3.5 Future work

Our outcome is not the final version. It has still many more iterations and steps. In our ontology, we just created the primitive classes and their hierarchy, but they also need to have properties, various values, and instances. It is also important to mention that some of the steps are not strictly sequential; rather they form an iterative process. This flow involves not only the execution of routine tasks that can be made automatically but also the collaboration and brainstorming sessions with experts in the relevant domain. Our goal here was to structure the provided information.

4 Q/A

In this section, we dissect the multiple questions that we answered in the seminar presentations:

4.1 Concept presentation

The following questions concern the first presentation that we have done regarding the literature and concept in ontology development:

Q- Explain the color scheme in the Excel sheet for terminology.

A- Color schemes in the Excel sheets are used for organizational purposes. These different colors indicate various statuses of terms, such as unreviewed, validated, or unfound. This scheme can vary depending on the project.

Q- How the set of terms are identified and validated?

A- Identification is a process of many iterations. Terms are usually identified through a combination of expert input, literature review, and existing databases. Validation often involves expert review and checks for adherence to ontological principles.

Q- What are the challenges to meeting the requirements between the real-world use cases and the developed ontology?

A- One of the challenges is to keep the ontology updated with the latest scientific findings and to ensure interoperability with other ontologies and databases. "FAIR" principles check ontology's quality and its robustness towards challenges [2].

Q- How can ontologies help bridge the gap between the different research data in life sciences?

A- Ontologies provide structure to represent and link diverse data. It offers standardized terms and relationships. These enable more effective data integration, sharing, and analysis across different studies and disciplines.

Q- what are the IRIs?

A- IRIs are used in ontologies to uniquely identify entities (like classes, properties, and individuals). They are similar to URLs but can include a wider range of characters, accommodating various languages and scripts.

Q- what ontology kit techniques are used to ensure quality control?

A- Techniques include consistency checking, use of controlled vocabularies, peer review, and validation against real-world data.

Q- on slide 20, what are these complex hypotheses about?

A- The hypotheses are the chosen source for the authors algergawy et al. in their work [1]. It's a basis for terminology in invasion biology and the relationships between them.

Q- What about testing and evaluating the developed ontology?

A- That's obligatory during the different phases of development. This involves checking for logical consistency and ensuring alignment with domain knowledge.

4.2 Outcome presentation:

The following questions concern the second presentation that we have done regarding the outcome of ontology development:

Q- How do you come to the list of main terms from LLM outcome?

A- To that end, we used the API of "OpenApi" to list terminology based on the 11 hypotheses mentioned as a base for invasion biology.

Q- Do you try to look up in bioPortal term by term or by a combination of terms?

A- We looked up the Bioportal term by term, but it's recommended to combine terms to get a better context output from the API. This latter we opted to do this manually.

Q- Since hierarchical structures are used, can XML be used as well?

A- Yes, XML can be used to represent hierarchical structures, although it's more common to use formats like OWL (Web Ontology Language) for ontologies, as they provide more sophisticated tools for expressing relationships and constraints.

Q- What are the colors meaning in the Excel sheet (slide 27)?

A- Same question/answer from the concept presentation.

Q- Can you tell more about how the terminology analysis process happened?

A- We have done the work without the help of experts and this was done by checking fitting definitions, synonyms, and ontology context from the Bioportal database.

Q- Can we present the hierarchical structure in other formats rather than owl and visualize them?

A- Yes, hierarchical structures can also be represented in formats like XML, JSON, or visually through diagrams and tree structures.

Q- what will be the next step?

A- refer to the "Future work" section.

Q- Are those results generated by you (page 37)?

A- All results are of our production except the INBIO diagrams we have shown. We envisioned reproducing the ontology of Invasion biology and comparing it with the works of INBIO ontology [1].

4.3 Bonus:

The following is a bonus question we have come up with regarding OBO principles:

Q- What's the difference between findable and accessible?

A- An ontology can be findable but not accessible, same, it can be reusable but not interoperable which demands a level of combination with other ontologies not just by itself.

5 CONCLUSION

In resume, our journey to experiment with the Hierarchy-of-Hypotheses approach, mixed with the help from OpenAI API and existing ontologies in the Bioportal, has been successful. We managed to get our ideas into the OWL format, and with Protege, a popular tool for this kind of work, following the steps of the ontology development pipeline, we could develop a good base for an ontology for invasion biology.

Ontologies are very important for making the search more precise because they lay out information in a clear, organized way. This helps search engines get a sense of what's being asked, making it easier to pull up the right information. Our ontology is not perfect and finished yet. It's still a work in progress that needs more details like properties, values, and specific examples. But, we've got a solid start, and it's a great base to build on.

REFERENCES

[1] Algergawy, A., Stangneth, R., Heger, T., Jeschke, J. M., and König-Ries, B. Towards a core ontology for hierarchies of hypotheses in invasion biology. In *European Semantic Web Conference* (2020), Springer, pp. 3–8.

[2] Matentzoglu, N., Goutte-Gattat, D., Tan, S. Z. K., Balhoff, J. P., Carbon, S., Caron, A. R., Duncan, W. D., Flack, J. E., Haendel, M., Harris, N. L., Hogan, W. R., Hoyt, C. T., Jackson, R. C., Kim, H., Kir, H., Larralde, M., McMurry, J. A., Overton, J. A., Peters, B., Pilgrim, C., Stefancsik, R., Robb, S. M., Toro, S., Vasilevsky, N. A., Walls, R., Mungall, C. J., and Osumi-Sutherland, D. Ontology development kit: a toolkit for building, maintaining and standardizing biomedical ontologies. *Database 2022* (Jan. 2022).

[3] Noy, N., and Mcguinness, D. Ontology development 101: A guide to creating your first ontology. *Knowledge Systems Laboratory 32* (01 2001).

[4] Panzarella, G., Veltri, P., and Alcaro, S. Using ontologies for life science text-based resource organization. *Artificial Intelligence in the Life Sciences 3* (2023), 100059.