Towards a Core Ontology for Hierarchies of Hypotheses in Invasion Biology

Alsayed Algergawy[1], Ria Stangneth[2], Tina Heger[3,4], Jonathan M Jeschke[4,5,6], and Birgitta König-Ries[1,7]

[1] Institute for Computer Science, University of Jena, Germany
{alsayed.algergawy|birgitta.koenig-ries@uni-jena.de}
[2] Institute of Organizational Psychology, University of Jena, Germany
[3] University of Potsdam, Biodiversity Research/Systematic Botany, Germany
[4] Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), Germany
[5] Institute of Biology, Free University of Berlin, Germany
[6] Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Germany
[7] German Centre for Integrative Biodiversity Research (iDiv), Germany

Abstract With a rapidly growing body of knowledge, it becomes more and more difficult to keep track of the state of the art in a research field. A formal representation of the hypotheses in the field, their relations, the studies that support or question them based on which evidence, would greatly ease this task and help direct future research efforts. We present the design of such a core ontology for one specific field, namely invasion biology. We introduce the design of the Hierarchy of Hypotheses (HoH) core ontology to semantically capture and model the information contained in hierarchies of hypotheses created for invasion biology. The proposed core ontology is based on a number of well structured related ontologies, which provide a solid basis for its main entities.

1 Introduction

The work presented in this paper was motivated by efforts by two of its authors, Jeschke and Heger, to advance their field of research, namely invasion biology. This field examines the effects that the introduction of new species has on ecosystems, and which circumstances determine whether a species can establish itself and spread in a new environment. Jeschke and Heger observed that a lack of clear understanding of the hypotheses in this field, their relations, and the evidence supporting or questioning them, considerably hinders scientific progress. Thus, they set out to model their field. This effort resulted in the Hierarchy-of-Hypotheses approach, as shown in Fig. 1, which they applied to sketching possible hierarchies of hypotheses (HoH) for invasion biology [7], (Fig. 2). Overarching ideas branch into more precise, better testable hypotheses at lower levels. This model, however, has not been rooted in formal semantics. It is thus currently not possible to automatically infer new knowledge. We take a first step to closing this gap by defining a core ontology for the field. We believe

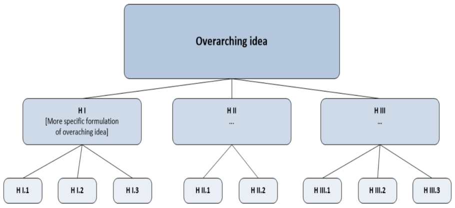

Figure 1: Basic scheme of HoH

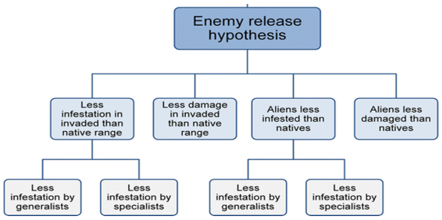
Enemy release hypothesis
Less infestation in invaded than native range
Less damage in invaded than native range
Aliens less infested than natives
Aliens less damaged than natives
Less infestation by generalists
Less infestation by specialists
Less infestation by generalists
Less infestation by specialists

Figure 2: The enemy release hypothesis

that, like the HoH approach, such ontologies are useful in all scientific fields and therefore focus the presentation not on the end result, but on the process. With this, we hope to enable other scientists to develop core ontologies for their fields as well.

In this paper, we propose the design of a core ontology. In general, an ontology is an elegant way to provide tools and methods developing and establishing correct links between data, research questions, and hypotheses towards a more efficient scientific life cycle. Here, we make use of the fusion/merge strategy [10] during the design of the *HoH* core ontology. In particular, a set of collected hypotheses is analyzed and relevant terms are extracted. This set of extracted terms is then used to localize related ontologies that can be reused as a basis for the core ontology design. We employ a module extractor strategy to the selected set of ontologies to reduce the number of selected concepts and properties and to ensure that the core ontology will not contain unneeded concepts making it more complex than necessary. These modules are then combined to form the initial version of the core ontology. Further improvements are made, such as revising the ontology and adding missing concepts.

2 Related work

Core ontologies provide a precise definition of structural knowledge in a specific field that connects different application domains [3,4,5,11]. They are located 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. Even though there is a large body of work making use of ontologies as a formal basis to model different aspects of scientific research, such as [3,4,5], few studies have focused on modeling scientific hypotheses [2,6].

3 The core ontology for HoH

To create a core ontology for the hierarchies of hypotheses (HoH) developed for invasion biology [8], we focus on the following issues:

Scenario. Invasion biology is concerned with the question why some species are able to establish and spread in an area where they have not evolved. Over time, the research community has developed several major hypotheses and empirical studies have been performed to test them. Since each hypothesis has been formulated in a general way, suggesting its validity across study systems (e.g., in terrestrial as well as aquatic habitats and across taxonomic groups), empirical tests apply a variety of approaches and produce a wealth of not always consistent results. The Hierarchy-of-Hypotheses (HoH) approach has been introduced as a tool for disclosing the complexity of this research. In an HoH, a major hypothesis can be depicted as a hierarchy of increasingly specific formulations of a broad idea. By assigning empirical tests to sub-hypotheses, it becomes clear that each of them is only addressing a specific aspect of the overall idea. The HoH approach has been applied to twelve major hypotheses in invasion biology [8]. Empirical evidence has been used to assess the validity of these major hypotheses and their sub-hypotheses. So far, however, this has been done manually. A formal representation of the twelve hypotheses and the respective HoHs could provide the basis for future computer-aided updates and expansions. Also, it would allow to reveal the different meanings oftentimes connected to terms, and thus avoid miscommunication and misinterpretation of results.

Strategy. To model the complex structure of knowledge in the hierarchy of hypotheses in the domain of invasion biology, we adopt the fusion/merge strategy [10], where the new ontology is developed by assembling and reusing one or more ontologies. To this end, the proposed pipeline starts by processing the description of each hypothesis extracting relevant terms (with the help of domain experts). Each term can be a noun, verb, or an adjective/adverb. Nouns can be simple or complex nouns. The **Biotic Resistance Hypothesis**, e.g., states that `"An ecosystem with high biodiversity is more resistant against exotic species than an ecosystem with lower biodiversity"`. Analyzing this hypothesis, the terms `"ecosystem, biodiversity, species"` can be extracted and identified as main entities of this hypothesis. In order to model the meaning of the hypothesis, additional entities not mentioned in the definition of the hypothesis need to be added. For example, in this domain lower and higher biodiversity are viewed as either related to the number of observed species, or to some index calculated for a specific area within a location. So, we add the `"number of species, indices, area, location"` entities to the set of extracted terms from the hypothesis, as shown in Fig. 3. After including `"area"`, species can be described as native or exotic species based on their relationship to area. In general, the outcome of this phase are 45 (noun) terms from 12 different hypotheses. We should mention that we consider the extraction of simple and compound terms, e.g. `"invasion"` and `"invasion success"`. After that we make use of the BioPortal API[8] to look for relevant ontologies that cover the set of extracted terms. We selected the National Center for Biomedical Ontologies (NCBO) BioPortal, since its deployment, it has evolved

[8] `http://bioportal.bioontology.org/`

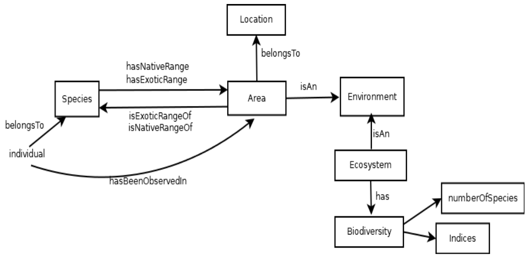

Figure 3: Entities and relations extracted from the Biotic Resistance Hypothesis

to become the prevalent repository of biological and biomedical ontologies and terminologies [9].

Several challenges arise during this step. First, the same term can be differently represented in several ontologies and we have to select the most suitable one. For example, the term `"ecosystem"` has been found in 21 ontologies representing different pieces of the domain. The `ecosystem` concept is defined in the environmental ontology (*ENVO*[9]) as an environmental system that includes both living and non-living components, while it is defined within the Interlinking Ontology for Biological Concepts (*IOBC*[10]) as an ecological system. Also, in the three invasion biological hypotheses where this term is a main term, it has three different meanings. Another challenge concerning the design of the core ontology is that it needs to satisfy a number of requirements as mentioned in [11]. After having a set of ontologies, for each term we extracted the set of corresponding concepts from different ontologies along with their URIs, labels, and definitions (if they exist). We then asked our domain experts to validate this selection. For example, the term `"species"` exists in 32 different ontologies, but our experts selected only two ontologies that align with the intended meaning. The term `"enemy"` exists in two ontologies, but none of them matches our requirements. Thus, we had to define our own concept. After settling on a number of ontologies to be adopted according to the fusion/merge strategy, we applied a module extractor to each ontology to elicit smaller partitions from the selected set of ontologies [1] containing only relevant concepts and those needed to connect them. Finally, these set of partitions were combined and merged to form the initial version of the new ontology.

Outcome. Applying the proposed strategy to the given set of hypotheses resulted in six core concepts in the *HoH* domain, as shown in Fig. 4, where each concept has one or more associated concepts. This set of concepts is used to select a set of related ontologies that maximize the coverage of these concepts. The six core concepts together with the associated concepts deliver the basis for semantically modelling twelve major hypotheses in invasion biology. Since these twelve hypotheses are well-known in the research field and regarded as important potential explanation for biological invasions, this core ontology delivers an

[9] `http://purl.obolibrary.org/obo/ENVO_01001110`

[10] `http://purl.jp/bio/4/id/200906003112410894`

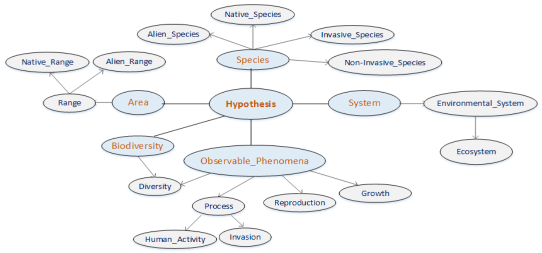

Figure 4: Core concepts in the HoH domain

important first step towards semantically modelling the research field of invasion biology.

All the resources related to the design of the HoH core ontology as well as the first versions of the ontology are accessible online at `https://github.com/fusion-jena/HoH_Core_Ontology`

Acknowledgments. This work has been funded by the *Deutsche Forschungsgemeinschaft (DFG)* as part of CRC 1076 AquaDiva. Additional support has been received from the German Federal Ministry of Education and Research BMBF within the Collaborative Project "Bridging in Biodiversity Science - BIBS" (funding number 01LC1501).

References

1. A. Algergawy and B. König-Ries. Partitioning of bioportal ontologies: An empirical study. In *SWAT4LS*, 2019.
2. A. Callahan, M. Dumontier, and N. H. Shah. HyQue: evaluating hypotheses using semantic web technologies. In *Journal of biomedical semantics*, 2011.
3. P. M. Campos, C. C. Reginato, and J. P. A. Almeida. Towards a core ontology for scientific research activities. In *ER*, 2019.
4. S. Fathalla, S. Vahdati, S. Auer, and C. Lange. SemSur: A core ontology for the semantic representation of research findings. In *SEMANTICS*, 2018.
5. L. F. Garcia, M. Abel, M. Perrin, and R. dos Santos Alvarenga. The GeoCore ontology: A core ontology for general use in geology. *Computers & Geosciences*, 135, 2020.
6. D. Garijo, Y. Gil, and V. Ratnakar. The disk hypothesis ontology: Capturing hypothesis evolution for automated discovery. In *K-CAP Workshops*, 2017.
7. T. Heger, A. T. Pahl, Z. Botta-Dukát, F. Gherardi, C. Hoppe, I. Hoste, K. Jax, L. Lindström, P. Boets, S. Haider, et al. Conceptual frameworks and methods for advancing invasion ecology. *Ambio*, 42(5):527–540, 2013.
8. J. M. Jeschke and T. Heger. *Invasion Biology: Hypotheses and Evidence.* CABI Wallingford, 2018.
9. M. A. Musen, N. F. Noy, N. H. Shah, P. L. Whetzel, C. G. Chute, M.-A. Story, B. Smith, and the NCBO team. The national center for biomedical ontology. *Journal of the American Medical Informatics Association*, 19(12):190–195, 2012.
10. H. S. Pinto and J. P. Martins. Ontologies: How can they be built? *Knowledge and Information Systems*, 6(4):441–464, 2004.
11. A. Scherp, C. Saathoff, T. Franz, and S. Staab. Designing core ontologies. *Applied Ontology*, 6(3):177–221, 2011.