LLMs4Life: Large Language Models for Ontology Learning in Life Sciences

Nadeen Fathallah[1], Steffen Staab[1,2] and Alsayed Algergawy[3,4]

[1]Analytic Computing, Institute for Artificial Intelligence, University of Stuttgart, Stuttgart, Germany
[2]University of Southampton, Southhampton, UK
[3]Data and Knowledge Engineering, University of Passau, Passau, Germany
[4]Institute for Informatics, Friedrich-Schiller-University Jena, Jena, Germany

Abstract

Ontology learning in complex domains, such as life sciences, poses significant challenges for current Large Language Models (LLMs). Existing LLMs struggle to generate ontologies with multiple hierarchical levels, rich interconnections, and comprehensive class coverage due to constraints on the number of tokens they can generate and inadequate domain adaptation. To address these issues, we extend the NeOn-GPT pipeline for ontology learning using LLMs with advanced prompt engineering techniques and ontology reuse to enhance the generated ontologies' domain-specific reasoning and structural depth. Our work evaluates the capabilities of LLMs in ontology learning in the context of highly specialized and complex domains such as life science domains. To assess the logical consistency, completeness, and scalability of the generated ontologies, we use the AquaDiva ontology developed and used in the collaborative research center AquaDiva [1] as a case study. Our evaluation shows the viability of LLMs for ontology learning in specialized domains, providing solutions to longstanding limitations in model performance and scalability.

Keywords

Ontology Learning, Large Language Models, NeOn-GPT, Life Science Domain.

1. Introduction

Ontology learning encompasses tasks such as ontology extraction, ontology generation, or ontology acquisition. It is the automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between the concepts that these terms represent from a corpus of natural language text and encoding them with an ontology language for easy retrieval [1]. Ontology learning in complex domains like life sciences presents a significant challenge. Although Large Language Models (LLMs) have shown promise in automating the generation and enrichment of ontologies [2, 3, 4, 5, 6], their application in highly specialized domains remains difficult and understudied. The inherent complexity of specialized domains such as life sciences, coupled with domain-specific terminologies and data, limits the ability of LLMs to generate ontologies that meet the structural and logical depth required for advanced reasoning. To explore these limitations, we consider the knowledge representation and ontology developed within the collaborative research center AquaDiva. AquaDiva is a large collaborative project encompassing fields such as biology, geology, chemistry, and computer science, all working towards a shared objective of enhancing our understanding of the Earth's critical zone [7]. As the complexity and amount of data collected within AquaDiva increases, there is an increasing necessity to adopt semantic web approaches to standardize data and facilitate its integration and interoperability [8, 9]. To this end, the AquaDiva ontology (*ADOn*) has been developed with 78.840 axioms, 8.892 concepts, and 245 object properties. We leverage the AquaDiva ontology *ADOn* as a use case for evaluating the performance of our method and assessing the structural depth, logical

[1]https://www.aquadiva.uni-jena.de/

EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands

nadeen.fathallah@ki.uni-stuttgart.de (N. Fathallah); steffen.staab@ki.uni-stuttgart.de (S. Staab); alsayed.algergawy@uni-passau.de (A. Algergawy)
0000-0001-7921-034X (N. Fathallah); 0000-0002-0780-4154 (S. Staab); 0000-0002-8550-4720 (A. Algergawy)

consistency, and completeness of the generated ontologies against this established domain-specific resource.

Ontologies are critical for organizing domain knowledge in a structured, reusable way, facilitating scientific research, and supporting advanced data analysis and decision-making. In domains like AquaDiva, an accurate and logically consistent ontology can enable a better understanding of complex ecological processes, enhance data interoperability, and improve scientific communication. Current approaches to ontology learning rely heavily on manual processes, which are labor-intensive and prone to human error. Incorporating LLMs offers the potential to enhance efficiency [3]; however, automating or semi-automating this process with LLMs requires rigorous evaluation to ensure the logical soundness, domain coverage, and adaptability of generated ontologies in complex domains [10]. Evaluating LLMs for their ability to capture deep, complex relationships between concepts is critical, as they often produce simplified structures that lack the depth and coverage necessary for representing domain-specific intricacies. Models like Neon-GPT struggle in ontology learning for domains like life sciences due to insufficient domain adaptation; as a result, they generate ontologies with shallow hierarchies. Inadequate domain adaptation means that LLMs lack sufficient exposure to specialized training data needed to model complex domains like life sciences, leading to generic ontologies. By "shallow hierarchy," we mean that the model struggles to generate deep hierarchical structures because it has difficulty establishing "is a part of" or "is a subset of" relationships, resulting in overly simplistic ontologies with limited subclass depth. Additionally, the vast amount of information required to fully model a domain like AquaDiva often exceeds the constraints on the number of tokens LLMs can generate, leading to incomplete outputs.

In this work, we propose an evaluation-driven approach to improving ontology learning for complex domains such as life science; our approach is evaluated on the AquaDiva ontology. Our methodology evolves through an experimental evaluation pipeline, where we iteratively address the limitations identified when evaluating the generated ontologies. The contributions of this paper are as follows:

- **Extension of the NeOn-GPT Pipeline with Advanced, Domain-Driven Prompt Engineering:** We extend the NeOn-GPT pipeline by introducing advanced prompt engineering techniques driven by domain-specific requirements. This includes re-prompting strategies that iteratively refine the LLM's output, enhancing the depth and hierarchy of the generated ontology. To further improve accuracy and adaptation, we increase the use of few-shot examples and employ advanced role-play prompting using domain-specific personas. Additionally, we develop a domain categorization strategy to handle token limitations, allowing the LLM to manage large, complex domains by breaking them into manageable subsets.
- **Introduction of Ontology Reuse in the NeOn-GPT Pipeline:** We enhance the NeOn-GPT pipeline by introducing ontology reuse, incorporating relevant existing ontological resources to evaluate reuse enhances the quality, depth, and consistency of the generated ontologies.
- **AquaDiva Ontology Case Study:** We evaluate our approach by assessing the structural complexity, depth, and logical consistency of the generated ontologies.

The paper is organized as follows: Section 2 reviews related work on ontology learning with LLMs. Section 3 details our proposed methodology. Section 4 presents the experiments that were conducted and their corresponding results, using the AquaDiva ontology as a case study. Finally, Section 5 concludes the paper and outlines future work.

2. Related Work

In recent years, LLMs have gained significant attention for their ability to enhance various ontology-related tasks, including ontology learning. Studies have demonstrated that LLMs can support the creation, enrichment, and refinement of ontologies, helping to automate traditionally labor-intensive tasks. For instance, Mateiu et al. [2] leverage a fine-tuned GPT-3 to translate natural language into OWL Functional Syntax for ontology enrichment. While the approach reduces the need for manual

intervention, it encounters difficulties in maintaining a deep ontology structure and avoiding the generation of irrelevant axioms. Babaei Giglou et al. [3] propose the LLMs4OL framework, which categorizes OL tasks into three core functions: Term Typing, Taxonomy Discovery, and Non-Taxonomic Relation Extraction. Their work shows that although LLMs can perform ontology learning in a zero-shot setting, fine-tuning is necessary for domain-specific tasks, particularly for complex domains like medicine or food. Similarly, Kommineni et al. [4] developed a semi-automated pipeline using LLMs for Competency Question (CQ) generation and Knowledge Graph (KG) construction. The system reduces human effort but highlights the need for manual validation due to the variability in LLM output and prompt sensitivity. In contrast, Saeedizade and Blomqvist [5] experiment with GPT-4 and other open-source models to generate OWL ontologies from ontological requirements. Their study concludes that while GPT-4 performs well in general ontology tasks, it struggles with more specialized domains, necessitating human intervention to correct errors and ensure the ontology's completeness. Zhang et al. [6] introduce OntoChat, a conversational framework for ontology engineering that uses LLMs to assist in requirement elicitation, CQ extraction, and ontology testing. OntoChat reduces the time and effort required for these tasks, though the authors acknowledge challenges such as LLM hallucination and the need for refinement during the ontology development process.

While these studies show the potential of LLMs in ontology learning, they share common limitations, such as shallow hierarchy generation, token limitations, and insufficient domain adaptation. For example, as highlighted by Mai et al. [10], off-the-shelf LLMs struggle to adapt to specialized terminologies in complex domains, and their performance is limited by token constraints and pre-existing lexical knowledge. Additionally, [10] emphasizes the importance of integrating structured knowledge sources to improve LLM performance on domain-specific tasks. These works provide the foundation for exploring how LLMs can be effectively applied to ontology learning tasks, but further improvements are necessary to address their shortcomings. In comparison, our work improves the structural depth of generated ontologies by employing re-prompting techniques and a keyword categorization strategy to manage token constraints. Additionally, we leverage ontology reuse, incorporating existing ontological structures (such as those from ENVO) to guide the LLM in generating more detailed hierarchies and relationships. Our approach ensures consistency with established domain knowledge while allowing us to generate more comprehensive ontologies, particularly in highly specialized domains like AquaDiva ontology, where token limitations and domain specificity are critical challenges.

3. Methodology

Our approach builds on our previous work with the NeOn-GPT pipeline for ontology learning [11]. Using the NeOn methodology framework, we translate its structured, iterative process into a series of prompts for pre-trained LLMs, ensuring that the generated ontology is both logically sound and aligned with domain requirements. NeOn-GPT demonstrates effective ontology generation in popular domains such as wine ontology. The wine domain is widely recognized, and wine ontology serves as a benchmark in ontology learning, making it likely to have been included in the training data of pre-trained LLMs. However, our empirical experiments show that the current NeOn-GPT pipeline does not perform as well on highly specialized domains, which are often underrepresented in real-world datasets. Domains such as life sciences are scarce in the training data of language models, making them unfamiliar and challenging for the models to handle effectively. This work extends the NeOn-GPT pipeline to address the more complex and specialized domains, such as the life sciences domain, and is evaluated on the AquaDiva ontology; figure 1 shows the steps of the extended pipeline. Such domains require a deeper understanding of domain-specific knowledge that may not be as readily accessible to LLMs due to their complexity and relative obscurity compared to more mainstream fields. Our enhancements enable the pipeline to effectively generate ontologies in intricate domains, significantly advancing ontology learning for niche areas.

Specification of Ontology Requirements. The first phase of the pipeline specifies the ontology requirements using chain-of-thought (CoT) prompting, guiding the model through four logical steps:

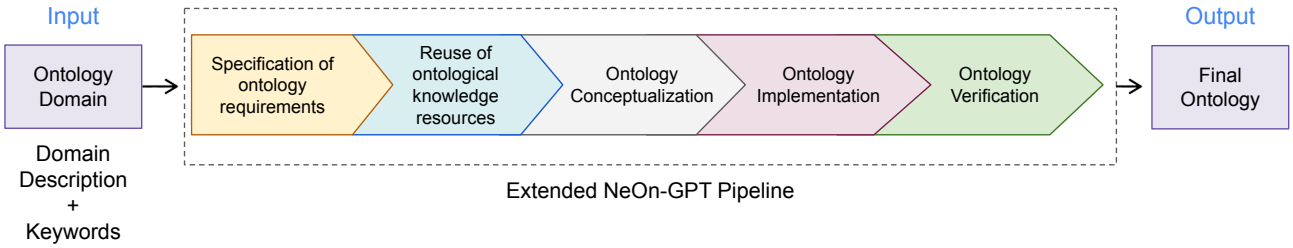

Figure 1: Overview of our proposed methodology to extend the NeOn-GPT pipeline for more complicated domains such as life science domains. The process begins with the ontology domain, incorporating domain-specific descriptions and keywords. The methodology employs a pre-trained LLM to follow a structured sequence of steps: specification of ontology requirements, reuse of ontological knowledge resources, ontology conceptualization, implementation, and verification, producing the final ontology.

defining the ontology's purpose, scope, target group, and functional requirements. To tailor the prompts for the AquaDiva domain, we integrate its description sourced from [12] and manually curated keywords directly into the CoT prompt. Keywords like 'habitat' and related terms such as 'aquifer fungi' and 'aquifer microbes,' sourced from the AquaDiva ontology [1], act as thematic anchors, aligning the model's focus with AquaDiva-specific requirements.

In this work, we refine the role-play persona used in the prompts, building on our previous work [13] that shows that more contextually enriched personas lead to more contextually relevant outputs from LLMs.

The next step involves prompting the LLM to generate competency questions (CQ) for the ontology. To guide the model, we provide seven few-shot examples of competency questions (CQ)—an increase from previous work—including challenging properties such as "SubClassOf," which the LLM previously struggled to generate effectively.

Reuse of ontological knowledge resources. The NeOn methodology traditionally includes a step for reusing relevant ontological and non-ontological resources [14]. However, in our earlier work, this step was omitted due to the challenges of providing large-scale ontologies to the LLM. Input token limitations made it infeasible, and chunking the content caused the model to lose context across different sections, even when chat history was retained. Additionally, when the LLM was prompted to reuse existing resources from its own knowledge base, it often hallucinated, generating non-existent references.

A key contribution of this work is introducing reuse to the NeOn-GPT pipeline based on our observations of the flat hierarchies generated by the LLM. More specifically, we identified a critical limitation in the LLM's ability to generate ontological structures that meet predefined criteria, particularly in terms of aligning with the expected count metrics derived from the gold standard AquaDiva ontology. We reused structural information (count metrics) from the AquaDiva gold standard ontology sourced from [12] to prompt the LLM to improve its structural output. The prompt included specific instructions guiding the model to target predefined counts for various ontology components, such as the number of logical axioms and class, as described in 4.2 (Experiment 2).

Although these adjustments improved the overall structure and alignment with the predefined metrics, the subclass count remained insufficient. The LLM struggled to produce a nuanced, layered hierarchy, often resulting in a flat structure with too few subclasses relative to the number of parent classes. To mitigate this, we introduced a refined prompt designed to increase the subclass count, directing the model to aim for more subclasses. We also re-prompted the model to generate a more detailed hierarchy, interconnected concepts, and robust relationships, as described in 4.2 (Experiment 2) and 4.4 (Experiment 4).

To address the challenges of integrating large ontologies into the LLM, we manually extracted examples from the Environment Ontology (ENVO) [15], a highly relevant resource for the AquaDiva

[1] https://www.aquadiva.uni-jena.de/

ontology, which is the target domain in this case. This integration allowed us to assess how reusing established ontological resources improves the generated ontology's hierarchical depth, interoperability, and relevance within the broader ontological ecosystem. In this step, the ENVO example is provided to the LLM as part of the prompt, with specific instructions on how to reuse its structure to enrich the generated ontology, as described in 4.5 (Experiment 5).

Ontology Conceptualization Ontology conceptualization and conceptual modeling begin with extracting entities and relationships, facilitated through few-shot prompting. Here, the LLM is guided to identify and extract entities and relationships directly from the previously generated competency questions (CQ). In this work, we modify the original prompt by incorporating domain-specific examples to tailor this process to the AquaDiva ontology. All few-shot examples are manually extracted from the AquaDiva ontology. Here's one of the few-shot examples injected in the prompt:

```
"cq1": "What measurement is associated with an observation?"
Entity: ["Observation", "Measurement"]
Property: ["hasMeasurement"]
```

Following this, we prompt the model to construct a comprehensive conceptual model for the entire ontology, represented as subject-relation-object triples from the extracted entities and relations. Here's an example of the generated triples:

```
Observation -- hasMeasurement --> Measurement
```

Ontology Implementation. In the ontology implementation phase, the original pipeline prompts the LLM to utilize the previously generated triples to create a complete ontology serialized in Turtle syntax. The original prompt emphasizes fundamental ontological constraints in the prompt, such as the correct application of Turtle syntax. The prompt also includes cautionary advice on ontology consistency and common pitfalls, such as proper prefix usage and the declaration of ontology prefixes. It provides a clear roadmap for the LLM to follow.

Next, the pipeline applies formal modeling to capture the domain's complexities and relationships within the ontology, ensuring it is logically consistent and capable of supporting advanced reasoning. This process involves prompting the LLM to introduce data properties and adjust domain and range settings accordingly, using a few-shot prompting technique. In this work, we tailor the original prompt to include domain-specific examples; here are some of the few-shot examples used:

```
:hasMeasurementValue rdf:type owl:DatatypeProperty, owl:FunctionalProperty ;
    rdfs:domain :Measurement ;
    rdfs:range xsd:float ;
    rdfs:label "has measurement value"@en .
```

Subsequently, the formal modeling process includes object properties, such as inverse, reflexive, transitive, symmetric, and functional properties. The prompts are structured to ensure that these object properties are meaningfully integrated across the entire ontology rather than being limited to isolated snippets. Building on insights from our empirical experiments, in this work, we implemented syntax consistency restrictions and cautionary advice on ontology consistency and common pitfalls across all prompts, not just the initial one. This decision arose from observations that inconsistencies often appeared in later stages of ontology generation when restrictions were only applied initially. Additionally, we observed the tendency of the LLM to adopt Occam's razor approach when correcting inconsistencies during the **Ontology Verification** phase, often leading to the omission of crucial properties or classes. To counter this, we guided the LLM to generate a syntactically correct and logically consistent ontology from the outset, reducing the need for later corrections.

To enhance the ontology's usability and readability, the original prompt pipeline includes instructions for enriching entities and relationships with natural language descriptions and adding essential metadata such as IRIs, labels, and versioning information.

Additionally, a few-shot prompt populates the ontology with real-world instances, grounding the ontology in practical data and facilitating knowledge discovery. We adapt these examples to align with our specific domain. Here is a sample of the few-shot examples used:

```
:Exbio_Antibodies rdf:type :Company, owl:NamedIndividual ;
    rdfs:label "Exbio Antibodies"@en .

:Becton_Dickinson_BD_Biosciences rdf:type :Company, owl:NamedIndividual ;
    rdfs:label "Becton Dickinson (BD Biosciences)"@en .
```

To further improve the structural hierarchy, depth of the ontology, and ontology coverage, we employed re-prompting. Corrective re-prompting is a technique in prompt engineering where an LLM is asked the same question again to improve the quality of its responses using error-related feedback [16, 13]. We use this approach to improve the initial output from the LLM. This iterative process involved prompting the LLM again to refine and extend the ontology, specifically emphasizing increasing the subclass count and creating a more layered hierarchy, as described in 4.4 (Experiment 4).

Ontology Verification Following ontology generation, the NeOn-GPT workflow employs RDFLib for syntax validation, HermiT and Pallet reasoners for consistency checking, and a custom-built pitfall detection module for identifying common ontology issues. Errors and inconsistencies identified by these tools are used to prompt the LLM for corrections, ensuring the ontology is both syntactically sound and logically coherent.

4. Experiments and Results

In this section, we present a series of experiments and their corresponding results to evaluate the LLM's performance before and after updating the NeOn-GPT pipeline; the generated ontologies are evaluated in terms of logical consistency and structural depth. Our objective is to assess how the proposed updates impact the LLM's ability to generate ontologies for complex life science domains, specifically using AquaDiva ontologies. The AquaDiva ontology domain encompasses the study of groundwater ecosystems, integrating hydrogeology, microbial ecology, geochemistry, karst systems, and environmental science. This ontology supports the annotation and standardization of diverse datasets related to subsurface habitats. The current AquaDiva ontology has 78.840 axioms, 8.892 concepts, and 245 object properties [12]. All experiments are conducted using GPT-4o [17]. The results of these experiments are discussed to illustrate the improvements achieved through the updated pipeline. Our code base is publicly available for research and development purposes, accessible at: https://github.com/NadeenAhmad/NeOn-GPTAquaDivaOntology. It includes all details about the prompts, interactions with LLMs, and the methodology used in our approach.

4.1. Experiment 1: Baseline NeOn-GPT (AquaDiva)

In this experiment, we applied the original NeOn-GPT pipeline without any of the enhancements introduced in this work. The only modification was the inclusion of 222 domain-specific keywords alongside the textual descriptions in the input to the LLM to compensate for the anticipated scarcity of relevant training data related to AquaDiva. This allowed us to evaluate the LLM's performance in its original configuration when applied to a complex life sciences domain.

4.1.1. Results of Experiment 1: Baseline NeOn-GPT (AquaDiva)

In this experiment, we evaluated the LLM-generated AquaDiva ontology against the AquaDiva gold standard ontology [12]. While the LLM successfully captured key concepts such as 'aquifers' and 'microbial communities', the ontology remained overly simplistic, with sparse hierarchy, and lacked the complexity needed for advanced ecological modeling. The metrics and class hierarchy from the newly generated ontology are shown in Figure 2. Compared to the gold standard, the LLM-generated

ontology included 176 classes (significantly fewer than the 8,892 in the gold standard) and only 44 object properties, which resulted in the omission of crucial relationships and subclass hierarchies. The absence of equivalent and disjoint classes, as well as a reduced set of logical axioms (323 versus 16,303 in the gold standard), further limited its ability to accurately represent the interactions within ecological systems.

Despite including important concepts like "Aquifer" and its subclasses (e.g., "Fractured Rock Aquifer," "Karst Aquifer") and environmental factors such as "Aquifer Vulnerability" and "Biogeochemical Cycle," the generated ontology lacked the relational depth necessary to describe the interactions between these entities. This omission impacts the ability to model and reason about the relationships, such as the specific types of microbial interactions within these environments. Additionally, the reduced number of object properties, individuals, and data properties (13 individuals and 26 data properties in total) further simplified the representation, preventing complex ecological relationships and taxonomic structures from being expressed. The simplified logical framework, including only 6 SubClassOf axioms, made it difficult to support detailed ecological queries, significantly reducing its utility for in-depth reasoning about environmental and biological phenomena.

4.2. Experiment 2: Count Metric-Guided Prompts (AquaDiva)

In Experiment 2, we addressed the limitations identified in 4.1 (Experiment 1) by revising the prompt pipeline to incorporate explicit count metrics from the AquaDiva gold standard, such as the number of classes (8,892) and object properties (245). The prompt instructed the LLM to align with these metrics, emphasizing a subclass count of at least n-1 (where n is the total number of classes) to address shallow hierarchies observed in 4.1 (Experiment 1).

4.2.1. Results of Experiment 2: Count Metric-Guided Prompts (AquaDiva)

The ontology generated in Experiment 2 demonstrated significant improvements over the initial version, exhibiting a more interconnected structure with increased density and a more layered hierarchy. As shown in Figure 3, this version includes 342 classes and 795 axioms, a notable increase from Experiment 1 but still lower than the expected 8,892 classes and 78,840 axioms in the AquaDiva gold standard ontology. This iteration represents a broader range of concepts and relationships, demonstrating more alignment with the domain's complexity. For instance, the ontology now contains 108 SubClassOf axioms and 103 EquivalentClasses axioms, improving upon 4.1 (Experiment 1), resulting in a more layered hierarchy with more subclass levels like (e.g., "HydroChemistry" – "SubClassOf" → "Geological Chemistry" – "SubClassOf" → "Earth Science"). However, the ontology still falls short in certain areas, particularly in object property count, which remains at 8 compared to the expected 245 in the gold standard. This can be partially attributed to GPT-4o's limitations, including its 4096-token output limit [17], which restricts the amount of content generated in a single response. Additionally, the LLM's mathematical limitations, particularly in precise counting tasks [18], likely contribute to the discrepancies in class and axiom counts. Moreover, some redundancy persists, with overlapping object properties such as "interact with" and "interacts with," indicating a need for further refinement.

4.3. Experiment 3: Merging Ontologies (AquaDiva)

In Experiment 3, we merged the ontologies generated from Experiments 1 and 2. Merging ontologies is a widely accepted approach in ontology engineering, as it can improve coverage, coherence, and the overall quality of the resulting ontology by combining complementary strengths from different sources [19]. By merging ontologies, we can address gaps that may exist in individual outputs and ensure a more comprehensive and robust knowledge representation. We utilized the RDFLib library to merge these ontologies; the ontology from 4.2 (Experiment 2), which contains richer concepts and better hierarchical structures, was used as the foundation. Unique and complementary elements from 4.1 (Experiment 1), particularly object properties and relationships, were incorporated to enhance depth and coverage.

4.3.1. Results of Experiment 3: Merging Ontologies (AquaDiva)

The ontology generated in Experiment 3 shows a notable improvement in several key metrics, as shown in 4. The total axiom count increased to 1,479, and the object property count rose to 50, compared to the lower counts observed in earlier experiments. These increases suggest that merging ontologies effectively captured a broader set of relationships and axioms, resulting in a more comprehensive ontology. However, while the merged ontology shows progress, some limitations persist. The class count, now at 500, is significantly improved compared to previous versions but remains below the gold standard AquaDiva ontology. There are still discrepancies between the generated metrics and the expected counts, particularly regarding data and annotation properties, where further refinement is needed to match the complexity of the domain. While the object property count has risen, it still falls short of the expected 245, indicating that additional adjustments are necessary to fully capture the domain's complexity. Notably, the ontology has made significant strides in logical consistency, with 713 logical axioms, and includes 114 SubClassOf axioms. This improved structure provides better support for defining relationships such as hierarchical taxonomies and equivalence between classes (e.g., "Aquatic Fungi" = "Aquatic Microorganism = Fungi"). Yet, despite these advances, the number of disjoint classes (109) still lags, impacting the ontology's ability to distinctly differentiate between overlapping or mutually exclusive categories, which is critical for accurate environmental modeling.

4.4. Experiment 4: Re-prompting & Advanced Role-play Prompting (Habitat)

In previous experiments, we tried to avoid exceeding the output token limitation of LLMs by instructing the model in each prompt to print only the new parts of the ontology. We manually aggregated these responses into a single ontology to prevent the model from regenerating the entire ontology repeatedly. However, this approach alone was not enough to fully overcome the token limitation issue. In Experiment 4, we addressed the limitations of using GPT-4o for generating a comprehensive ontology due to the model's output token constraints. Rather than attempting to generate the entire AquaDiva ontology at once, we instructed the LLM to categorize 222 AquaDiva-specific keywords into distinct groups, resulting in 22 categories. Inspired by the improvements observed in 4.3 (Experiment 3), where merging outputs led to a more interconnected ontology, we envision that this categorization will help generate better ontologies for each category. Ultimately, merging these individual ontologies will result in a larger and more comprehensive representation of the AquaDiva ontology.

We selected the "Habitat" category due to its ecological significance in AquaDiva. By developing an ontology for this category, we aimed to create a detailed and accurate representation that could serve as a model for expanding to other categories. To improve the quality and precision of the generated ontology, we applied several enhancements to the prompt pipeline in 4.2 (Experiment 2). First, we provided the LLM with a detailed description of the Habitat category along with relevant keywords, ensuring a richer domain-specific context as input. Additionally, we increased the number of few-shot examples from three examples to seven examples, tailoring them to the specific concepts within the Habitat domain. Furthermore, we refined the role-play persona used in the prompts, building on our findings that show enriched personas can yield higher-quality outputs [13]. We refined the role-play persona to represent an expert aquatic ecologist, leveraging domain knowledge to guide the model more effectively. The persona provided detailed instructions on how to structure the Habitat ontology with rich ecological context, ensuring domain relevance.

To iteratively refine the generated ontology, we applied re-prompting, asking the model to enhance the hierarchical depth and align with predefined metrics after the initial output. For example, a prompt to increase the subclass count to at least n-1, where n refers to the total number of classes, while other prompts address shallow hierarchies.

4.4.1. Results of Experiment 4: Re-prompting & Advanced Role-play Prompting (Habitat)

The results from Experiment 4 show that while directing the LLM's attention to the 'Habitat' category allowed for a more concentrated development of the ontology, certain limitations remain evident.

The ontology is shown in Figure 5 shows part of the ontology metrics and class hierarchy. The ontology generated has a total of 630 axioms, 275 logical axioms, and 75 classes. Although these metrics indicate progress, particularly in the object property count (47), they still fall short in several areas. For instance, there is only a single DisjointClasses axiom and 3 EquivalentClasses axioms, reflecting an incomplete structure in terms of class relationships. Moreover, the number of SubClassOf axioms (44) remains insufficient for a fully detailed hierarchical structure, and the ontology still lacks comprehensive disjointness and equivalence axioms, which are crucial for distinguishing and relating different categories within the ecological domain.

4.5. Experiment 5: Reuse (Role)

In Experiment 5, we generated an ontology for the "Role" category within the AquaDiva ontology with the same enhancements done to the pipeline in 4.4 (Experiment 4). To address the lack of hierarchical depth observed in previous experiments, we improved the subclass structure by incorporating an ontology reuse strategy using a detailed example manually extracted from the ENVO ontology. This reuse example demonstrated an extensive hierarchy of classes and subclasses, utilizing a visual structure with arrows to represent increasing levels of subclass specificity. This model served as a guide for the LLM, ensuring that each class in the Role ontology would have a well-defined hierarchy of subclasses, thereby enhancing the overall depth and complexity of the ontology. Here's a simplified portion of the example provided in the prompt for reuse to illustrate the hierarchical structure:

```
-> biological_process
--> biodegradation
--> cellular process
---> cellular metabolic process
----> cellular alkane metabolic process
----> photosynthesis
```

In this format, each arrow represents increasing levels of subclass hierarchy, starting from broad categories like "biological_process" and moving down to more specific entities such as "cellular_process." This reuse example, manually curated from ENVO, helped guide the LLM in generating deeper subclass hierarchies and producing a more layered structure in the Role ontology.

4.5.1. Results of Experiment 5: Reuse (Role)

The role ontology generated in Experiment 5 demonstrates notable strengths, particularly in its axiom count, which includes 969 axioms, and class count, which includes 118 classes. These metrics accurately represent the relationships within the 'Role' domain, showcasing the ontology's potential for supporting complex reasoning tasks. Additionally, the inclusion of 57 individual instances suggests a more comprehensive and practically applicable ontology, contributing to its overall depth and usability for modeling in the AquaDiva ontology. The ontology is shown in Figure 6, which highlights part of the ontology metrics and class hierarchy.

One of the key improvements in this experiment was the significant increase in the subclass count compared to 4.4 (Experiment 4), with the ontology now containing 86 subclasses. This improvement was achieved through the incorporation of the manually extracted ENVO example, which provided a structured reuse of existing hierarchical ontologies. The enhanced subclass hierarchy contributed to a more layered and detailed ontology, addressing prior limitations in the structural depth observed in earlier experiments.

However, despite these strengths, the ontology still exhibits significant limitations; while there is some improvement in logical consistency (e.g., 17 EquivalentClasses), the ontology remains underdeveloped in terms of disjoint class distinctions, with only 10 DisjointClasses axioms. This gap weakens its logical coherence and reduces its robustness for reasoning tasks. Moreover, the broad and generic nature of some classes within the 'Role' category, such as "Biological Role" or "Chemical Role," raises concerns

about the potential inclusion of overly generic terms that could dilute the ontology's focus and reduce its utility in the specific context of the AquaDiva ontology.

4.6. Experiment 6: Reuse of domain-specific examples (Carbon & Nitrogen Cycling)

In Experiment 6, we generated an ontology for the Carbon and Nitrogen Cycling domain, building on the lessons learned from previous experiments. Earlier attempts demonstrated that the reuse of existing ontological resources can significantly improve terminology generation and result in a more complex and layered hierarchy. This was evident in the increase in the number of classes and subclasses from 4.4 (Experiment 4) to 4.5 (Experiment 5). Motivated by these findings, we selected the Carbon and Nitrogen Cycling domain to evaluate how the reuse of an example with domain-specific terminology could further enhance the ontology generation process.

This experiment incorporates all the improvements added to the original NeOn-GPT prompt pipeline. First, we continued using the advanced role-play persona from 4.4 (Experiment 4) and 4.5 (Experiment 5) to maintain contextual relevance. For this domain, we provided a detailed description along with domain-specific keywords to guide the model's understanding. Additionally, we increased the number of few-shot examples, tailoring them to the Carbon and Nitrogen Cycling domain. To ensure logical consistency and structural depth, we implemented syntax and consistency restrictions at all stages of ontology generation prompts, reducing the need for later corrections related to missing properties or classes.

We enhanced the reuse of existing ontological resources. Instead of using broader, generic examples, we incorporated specific components manually extracted from ENVO that closely align with the Carbon and Nitrogen domain. This targeted reuse approach provided a clearer structure, ensuring the ontology reflected accurate hierarchical depth, interconnected concepts, and detailed relationships. A portion of the reuse example included classes like "carbon atom" and "nitrogen atom" and their corresponding subclasses, organized into multiple levels of hierarchy.

Here's a simplified portion of the example provided in the prompt for reuse to illustrate the hierarchical structure:

```
-> carbon_atom
--> carbon-13_atom
--> carbon-14_atom
-> dissolved_carbon_atom_in_environmental_material
--> dissolved_carbon_atom_in_soil
--> dissolved_carbon_atom_in_water
```

This example, manually curated from ENVO, demonstrated the expected level of hierarchy, with each class and its corresponding subclasses represented by increasing levels of specificity.

4.6.1. Results of Experiment 6: Reuse of domain-specific examples (Carbon & Nitrogen Cycling)

The Carbon and Nitrogen Cycling ontology developed in Experiment 6 shows significant improvements in capturing complex biochemical processes. Key entities such as "Carbon Fixation," "Nitrogen Transformation," and "Methanogenesis" are accurately modeled, with 157 classes and 63 object properties, including 13 functional, 10 symmetric, and 10 transitive properties, enabling detailed representations of interactions like fixing nitrogen and releasing methane.

A major improvement is the hierarchical depth, with 130 SubClassOf axioms, enhanced by reusing domain-specific components from the ENVO ontology. This includes entities such as "ammonia oxidation" and "biogeochemical cycle", reflecting a richer, more structured subclass hierarchy. The ontology also includes 1,169 axioms, 455 of which are logical, offering a more detailed representation of processes like "CO2 Fixation" and "Trace Gas Production".

Despite these advancements, the ontology still has only 11 EquivalentClasses and 16 DisjointClasses, limiting its ability to fully capture equivalent biochemical processes and distinctions between exclusive pathways like "carbon sequestration" and "carbon release".

4.7. Comprehensive Ontology Performance Overview

This section presents a comparative analysis of the generated ontologies in terms of precision and concept similarity across the six experiments conducted. We use the AML (AgreementMakerLight) ontology matching system [20] to automatically align and match concepts between the generated ontologies and the gold standard ontologies, the AquaDiva Ontology and the ENVO ontology. For each matched concept, AML produces a similarity score, indicating the degree of semantic overlap between the two concepts. We use these matched concepts to calculate the following evaluation metrics:

- Number of entities in the LLM-generated ontologies that match entities in the Gold standard ontology.
- Concept similarity evaluates how semantically similar the matched concepts are with the gold standard ontology concepts, calculated by averaging the individual similarity scores for all matched concepts.

Table 1
Precision and Concept Similarity Scores for Generated Ontologies with AquaDiva Ontology

Experiment	Number of Matched Entities with AquaDiva	Average Similarity Score with AquaDiva
Experiment 1 – Baseline NeOn-GPT (AquaDiva)	17	0.896
Experiment 2 – Count Metric-Guided Prompts (AquaDiva)	66	0.894
Experiment 3 – Merging Ontologies (AquaDiva)	80	0.874
Experiment 4 – Re-prompting & Advanced Role-play Prompting (Habitat)	16	0.898
Experiment 5 – Reuse (Role)	56	0.905
Experiment 6 – Reuse of domain-specific examples (Carbon & Nitrogen Cycling)	65	0.859

The results in Tables 1 and 2 reveal that despite the LLM-generated ontologies not fully capturing the breadth and depth of domain-specific knowledge as comprehensively as the gold standard ontologies (AquaDiva and ENVO), the aligned entities demonstrate exceptionally high similarity scores across all experiments. This suggests that the generated concepts closely reflect the established domain knowledge, as evidenced by scores consistently approaching or exceeding 0.85. Moreover, the number of matched entities increases across experiments, indicating that improvements in our LLM prompt engineering techniques and pipeline refinements lead to a progressively better alignment with the domain-specific ontologies. This trend shows that the generated ontologies evolve to incorporate a broader range of relevant entities while maintaining high conceptual similarity to the gold standard, underscoring the potential for LLM-based approaches in complex ontology generation tasks.

5. Conclusion and Future work

This work extends the NeOn-GPT pipeline to enhance ontology learning in complex domains, such as life sciences, by addressing the limitations of LLMs in generating deep and well-structured ontologies. Our approach leverages advanced prompt engineering, ontology reuse, and iterative refinement to

Table 2
Precision and Concept Similarity Scores for Generated Ontologies with ENVO Ontology

Experiment	Number of Matched Entities with ENVO	Average Similarity Score with ENVO
Experiment 1 – Baseline NeOn-GPT (AquaDiva)	8	0.877
Experiment 2 – Count Metric-Guided Prompts (AquaDiva)	57	0.969
Experiment 3 – Merging Ontologies (AquaDiva)	60	0.885
Experiment 4 – Re-prompting & Advanced Role-play Prompting (Habitat)	13	0.800
Experiment 5 – Reuse (Role)	54	0.886
Experiment 6 – Reuse of domain-specific examples (Carbon & Nitrogen Cycling)	51	0.884

tackle challenges like shallow hierarchies and token constraints, as demonstrated in the AquaDiva case study. We conclude that complex domains, such as those in life sciences, require additional contextual information in prompts and carefully curated examples for reuse. Currently, this process relies on manual efforts to extract relevant examples and domain-specific knowledge; the quality of those examples can be significantly improved with input from domain experts. In future work, we also aim to explore automating this process through Retrieval-Augmented Generation (RAG), integrating external domain-specific resources dynamically to reduce the reliance on manual intervention. Additionally, we plan to evaluate the complete AquaDiva ontology by systematically generating and integrating each domain category using the finalized pipeline, with a focus on refining consistency in relationships and ensuring the ontologies fully capture the intricacies of specialized domains like AquaDiva.

References

[1] A. Maedche, S. Staab, Ontology learning for the semantic web, IEEE Intelligent Systems 16 (2001) 72–79. URL: https://doi.org/10.1109/5254.920602. doi:10.1109/5254.920602.

[2] P. Mateiu, A. Groza, Ontology engineering with large language models, in: 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2023, Nancy, France, September 11-14, 2023, IEEE, 2023, pp. 226–229. URL: https://doi.org/10.1109/SYNASC61333.2023.00038. doi:10.1109/SYNASC61333.2023.00038.

[3] H. B. Giglou, J. D'Souza, S. Auer, Llms4ol: Large language models for ontology learning, in: The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I, volume 14265 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 408–427. URL: https://doi.org/10.1007/978-3-031-47240-4_22. doi:10.1007/978-3-031-47240-4_22.

[4] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An LLM supported approach to ontology and knowledge graph construction, CoRR abs/2403.08345 (2024). URL: https://doi.org/10.48550/arXiv.2403.08345. doi:10.48550/ARXIV.2403.08345. arXiv:2403.08345.

[5] M. J. Saeedizade, E. Blomqvist, Navigating ontology development with large language models, in: The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I, volume 14664 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 143–161. URL: https://doi.org/10.1007/978-3-031-60626-7_8. doi:10.1007/978-3-031-60626-7_8.

[6] B. Zhang, V. A. Carriero, K. Schreiberhuber, S. Tsaneva, L. S. González, J. Kim, J. de Berardinis, Ontochat: a framework for conversational ontology engineering using language mod-

els, CoRR abs/2403.05921 (2024). URL: https://doi.org/10.48550/arXiv.2403.05921. doi:10.48550/ARXIV.2403.05921. arXiv:2403.05921.
[7] K. Küsel, K. U. Totsche, S. E. Trumbore, R. Lehmann, C. Steinhäuser, M. Herrmann, How deep can surface signals be traced in the critical zone? merging biodiversity with biogeochemistry research in a central german muschelkalk landscape, Frontiers in Earth Science 4 (2016) 32.
[8] A. Algergawy, H. Hamed, B. König-Ries, Towards scientific data synthesis using deep learning and semantic web, in: The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers 18, Springer, 2021, pp. 54–59.
[9] A. Algergawy, H. Hamed, S. Thiel, B. König-Ries, Towards semantic annotation for scientific datasets, in: The Semantic Web: ESWC 2024 Satellite Events: May 26–30, 2024, ???? URL: https://api.semanticscholar.org/CorpusID:269758799.
[10] H. T. Mai, C. X. Chu, H. Paulheim, Do llms really adapt to domains? an ontology learning perspective, CoRR abs/2407.19998 (2024). URL: https://doi.org/10.48550/arXiv.2407.19998. doi:10.48550/ARXIV.2407.19998. arXiv:2407.19998.
[11] N. Fathallah, A. Das, S. De Giorgis, A. Poltronieri, P. Haase, L. Kovriguina, Neon-gpt: A large language model-powered pipeline for ontology learning, in: The Extended Semantic Web Conference, 2024.
[12] A. Algergawy, H. Hamed, S. Thiel, B. K¨onig-Ries, Towards semantic annotation for scientific datasets, ESWC Posters and Demos (2024).
[13] A. Das, N. S. Fathallah, N. Obretincheva, Navigating nulls, numbers and numerous entities: Robust knowledge base construction from large language models, in: KBC-LM/LM-KBC@ ISWC, 2024.
[14] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The neon methodology framework: A scenario-based methodology for ontology development, Applied Ontology 10 (2015) 107–145. URL: https://doi.org/10.3233/AO-150145. doi:10.3233/AO-150145.
[15] P. L. Buttigieg, N. Morrison, B. Smith, C. Mungall, S. Lewis, Environment ontology (envo), http://obofoundry.org/ontology/envo.html, 2021. Accessed: 2024-09-12.
[16] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, S. Tellex, Planning with large language models via corrective re-prompting, in: NeurIPS 2022 Foundation Models for Decision Making Workshop, 2022.
[17] OpenAI, Hello gpt-4o, https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-18.
[18] S. Frieder, L. Pinchetti, A. Chevalier, R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, J. Berner, Mathematical capabilities of ChatGPT, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html.
[19] J. De Bruijn, M. Ehrig, C. Feier, F. Martín-Recuerda, F. Scharffe, M. Weiten, Ontology mediation, merging and aligning, Semantic web technologies (2006) 95–113.
[20] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The agreementmakerlight ontology matching system, in: On the Move to Meaningful Internet Systems: OTM 2013 Conferences: Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings, Springer, 2013, pp. 527–541.

Acknowledgments

A. Algergawy' work has been funded by the *Deutsche Forschungsgemeinschaft (DFG)* as part of CRC 1076 AquaDiva (Projectnumber 218627073).

6. Appendix A: Figures

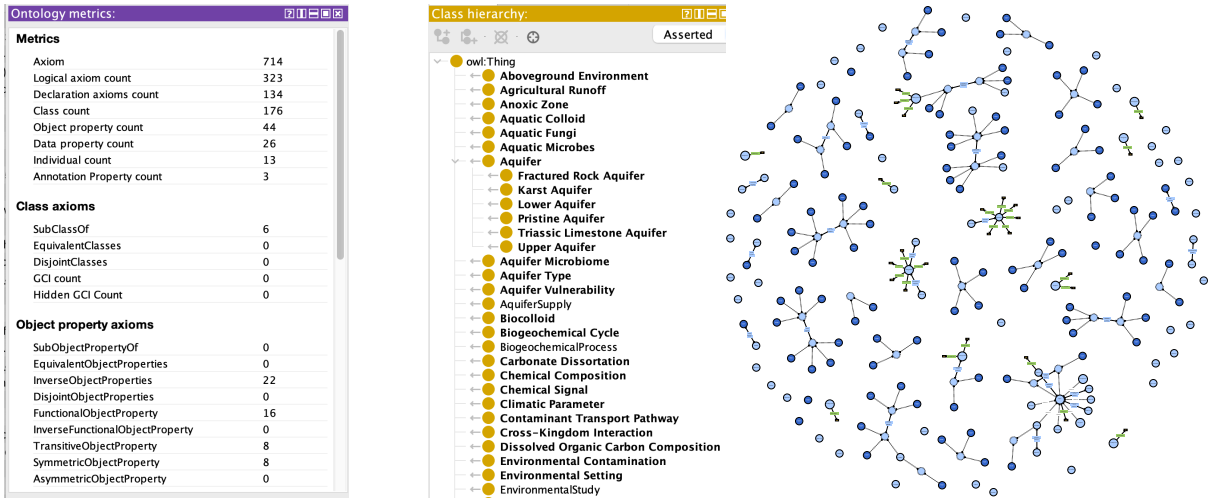

Figure 2: Visualization of the AquaDiva ontology generated from Experiment 1. The left side presents ontology metrics. The center panel shows a portion of the hierarchy of classes and their relationships (visualized using Protégé), while the right side features a structural network representation of the ontology generated using WebVOWL 1.1.7.

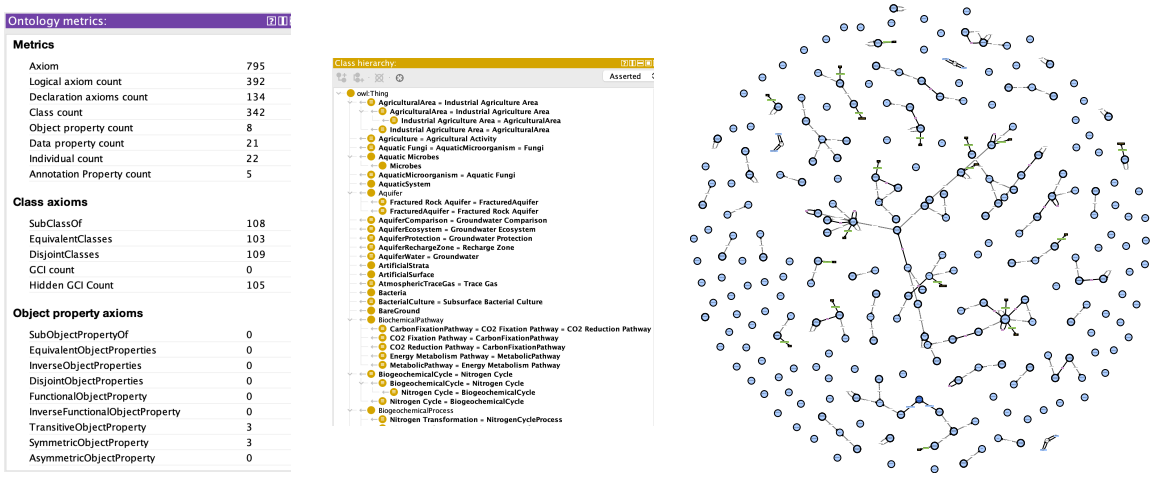

Figure 3: Visualization of the AquaDiva ontology generated from Experiment 2.

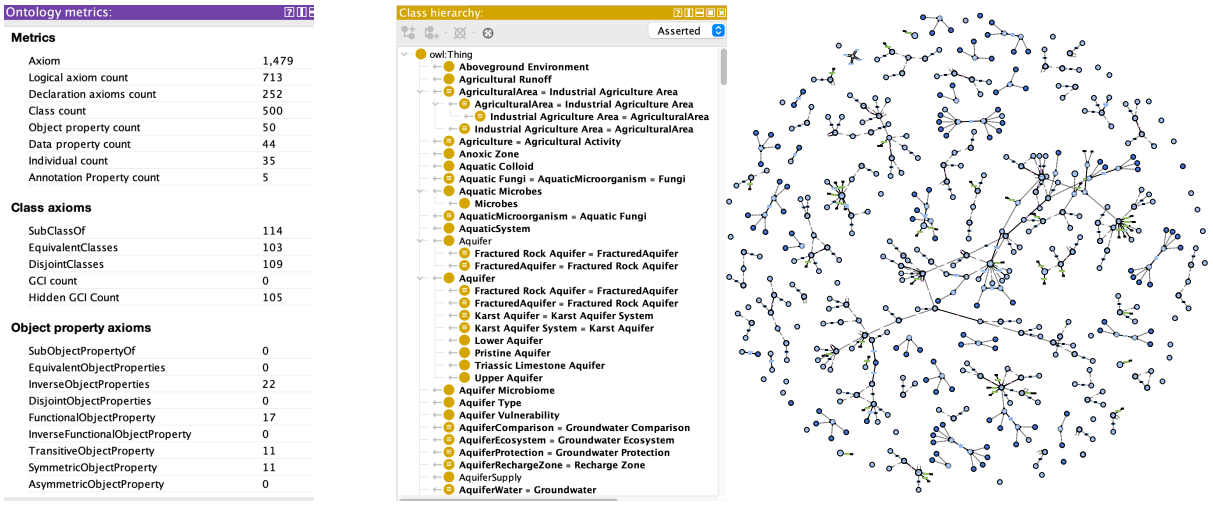

Figure 4: Visualization of the AquaDiva ontology generated from Experiment 3.

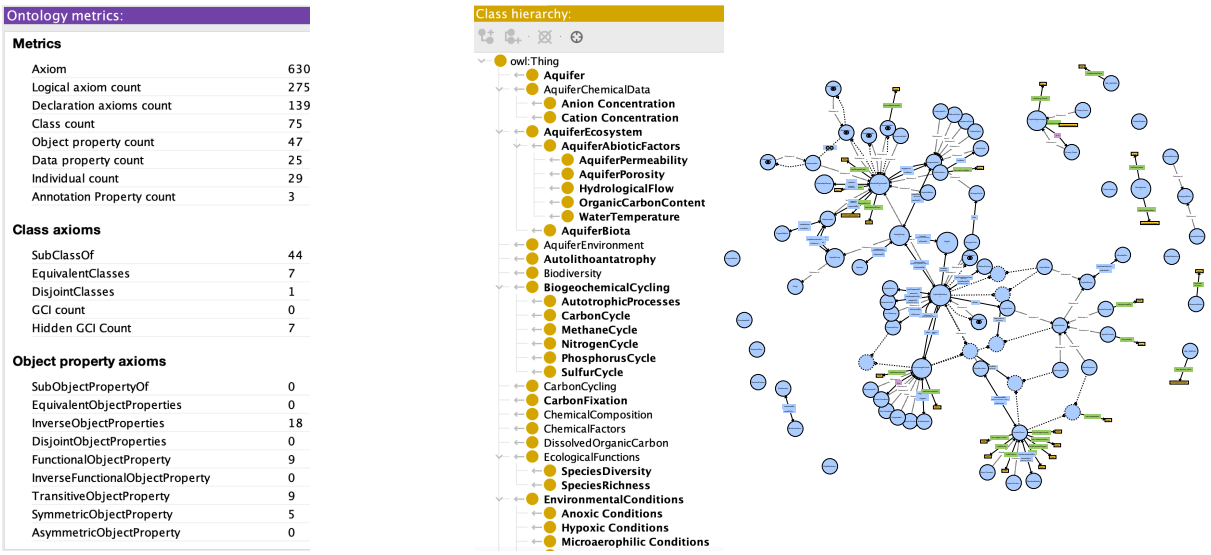

Figure 5: Visualization of the Habitat ontology within the AquaDiva ontology generated from Experiment 4.

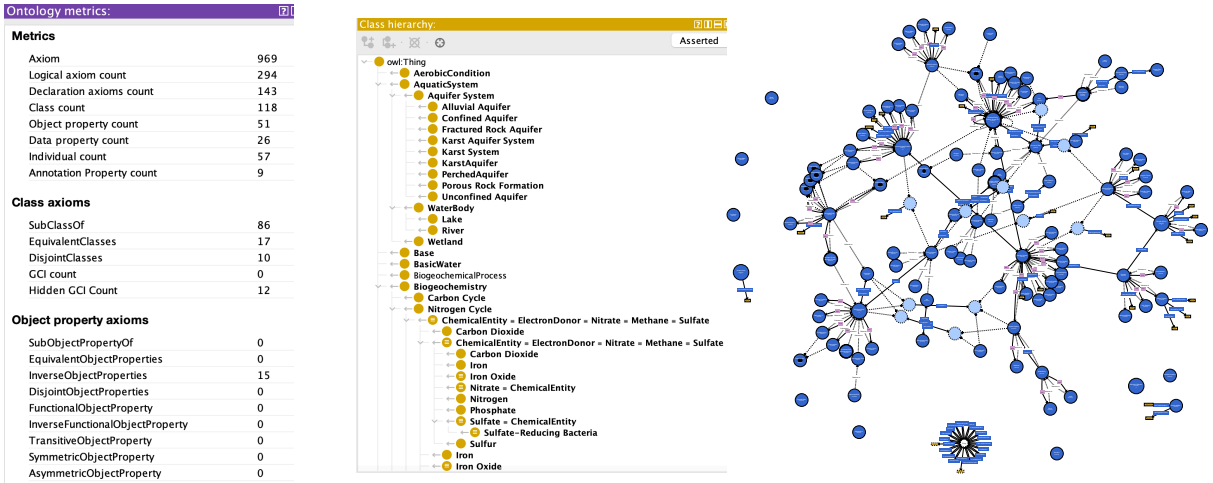

Figure 6: Visualization of the Role ontology within the AquaDiva ontology generated from Experiment 5.

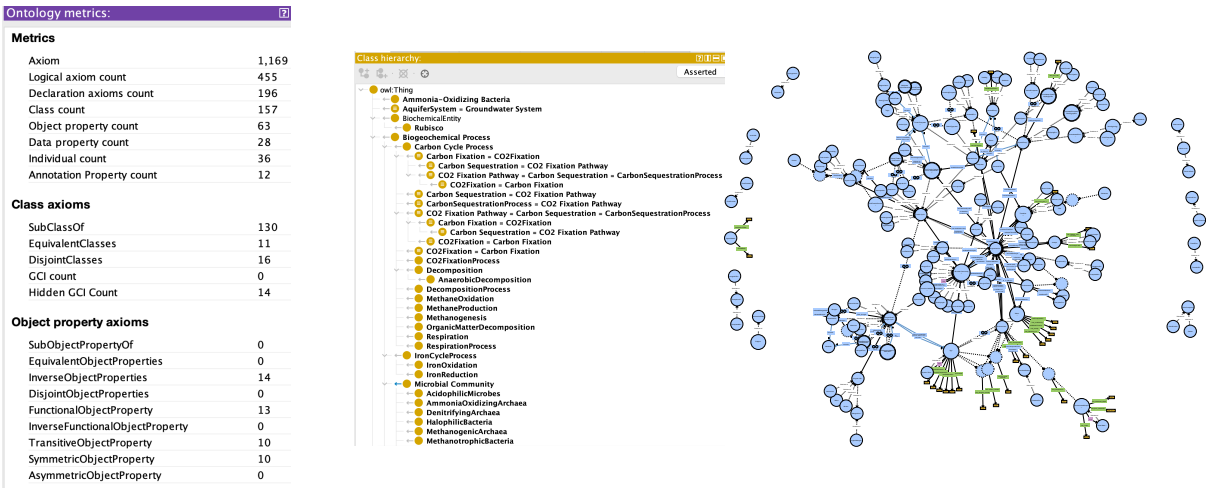

Figure 7: Visualization of the Carbon and Nitrogen Cycling ontology within the AquaDiva ontology generated from Experiment 6.