LLMs4OL 2024: The 1st Large Language Models for Ontology Learning Challenge at the 23rd ISWC
LLMs4OL 2024 Task Overview
https://doi.org/10.52825/ocp.v4i.2480

Published: 02 Oct. 2024

LLMs4OL 2024 Datasets: Toward Ontology Learning with Large Language Models

Hamed Babaei Giglou, Jennifer D'Souza, Sameer Sadruddin, and Sören Auer

TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{hamed.babaei, jennifer.dsouza, sameer.sadruddin, auer}@tib.eu

*Correspondence: Hamed Babaei Giglou, hamed.babaei@tib.eu

Abstract: Ontology learning (OL) from unstructured data has evolved significantly, with recent advancements integrating large language models (LLMs) to enhance various aspects of the process. The paper introduces the LLMs4OL 2024 datasets, developed to benchmark and advance research in OL using LLMs. The LLMs4OL 2024 dataset as a key component of the LLMs4OL Challenge, targets three primary OL tasks: Term Typing, Taxonomy Discovery, and Non-Taxonomic Relation Extraction. It encompasses seven domains, i.e. lexosemantics and biological functions, offering a comprehensive resource for evaluating LLM-based OL approaches Each task within the dataset is carefully crafted to facilitate both Few-Shot (FS) and Zero-Shot (ZS) evaluation scenarios, allowing for robust assessment of model performance across different knowledge domains to address a critical gap in the field by offering standardized benchmarks for fair comparison for evaluating LLM applications in OL.

Keywords: Ontology Learning, Large Language Models, Dataset, LLMs4OL Challenge

1 Introduction

Ontologies have gained a lot of popularity and recognition in the semantic web because of their fine source of semantics and interoperability. The increase in unstructured data on the web has made the automated acquisition of ontology from unstructured text a most prominent research area. Recently, instead of handcrafting ontologies, the research trend is now shifting toward automatic ontology learning (OL) [1]. OL involves automatically identifying terms, types, relations, and potential axioms from textual information to construct an ontology [2].

Looking back to the history of OL research, until early 2002 [3], most OL approaches relied on seed words or existing base ontologies rather than building new ones from scratch. Later in 2003 [4] the natural language processing (NLP) technique showed promise for the extraction of new concepts. However, relation extraction for OL remained still challenging. Also, the prior domain knowledge of the base ontologies still was in the middle of the focus for OL. With progress in the field, in 2006 the concept of *"ontology learning layer cake"* [5] was introduced to organize and describe the different steps involved in the process of ontology learning from the text for real-life application

scenarios. The OL layer cake includes (from the bottom of the cake to the top), Terms, Synonyms, Concepts, Taxonomies, Relations, Rules, and Axioms. This reflects a progression from simpler to more complex and abstract forms, each step building on the results of the previous one. It provides a structured approach to understanding and automating the OL process. Later in 2011, Hazman et al.[6] studied various OL systems and categorized them into two categories, *(1) learning from unstructured data* and *(2) learning from semi-structured data*. They also pointed out that when human-based evaluation is not possible, carrying out five-level evaluations for OL is important, levels such as lexical, hierarchical, contextual, syntactic, and structural levels. Since 2011 and in 2018 survey of [7] showed that a hybrid approach comprising both linguistic and statistical techniques produces better ontologies. However, it is difficult to find the best technique amount approaches due to the domain of the studies. The trend was shifted toward statistical techniques for term extractions, however for relation extraction clustering methods were the most used ones. Moreover, the various evaluations of OL showed that human-based evaluation is the most reliable approach for evaluation.

Considering that most of the approaches in the field were based on statistical approaches or clustering models, the emergence of large language models (LLMs), offered a paradigm shift in OL since their characteristics justify OL as a studied for the first time within LLMs4OL paradigm [8]. One reason for this shift is the LLM's generation capabilities because they are being trained on extensive and diverse text, similar to domain-specific knowledge bases [9]. For the first time, in 2023 the LLMs4OL [8] paradigm was introduced that incorporates LLMs for three important tasks of OL as Term Typing, Taxonomy Discovery, and Non-Taxonomic Relation Extraction. Later, more researchers were involved in the OL tasks from different perspectives [10]–[13].

The current trend in the semantic web reveals a growing interest among researchers in utilizing LLMs [14]. A benchmark dataset is essential to assess the performance of OL approaches, particularly those involving LLMs, in a consistent and comparable manner. Without such benchmarks, it becomes difficult to evaluate progress and compare various methodologies effectively [13]. To address this gap, in this work, we introduce an LLMs4OL paradigm tasks dataset to bridge the gap in benchmark evaluation datasets specifically within the context of OL using LLMs. Our key contribution is the creation of the LLMs4OL dataset, aimed at facilitating consistent evaluation in this emerging field. For the first time, this dataset is introduced in the *"1st LLMs4OL Challenge @ ISWC 2024"* [15], a challenge organized at the prestigious International Semantic Web Conference (ISWC). The primary goal of the challenge is to provide a shared platform for researchers to benchmark their LLM-based OL approaches. By establishing this dataset and launching the LLMs4OL Challenge, we hope to encourage further research and innovation in OL with LLMs, ultimately enabling a more structured and fair comparison of different methods in this rapidly evolving area.

The LLMs4OL 2024 dataset addresses three OL tasks, which are known as primitive ontology construction tasks [16]. Considering, L as a lexical entries for conceptual type T, and H_T as a representation of taxonomy of types, and R as a non-taxonomic relations, the LLMs4OL tasks are defined as follows:

- **Task A – Term Typing:** For a given lexical term L, discover the generalized type T.
- **Task B – Taxonomy Discovery:** For a given set of generalized types T, discover the taxonomic hierarchical pairs (T_a, T_b) pairs, representing "is-a" relations.
- **Task C – Non-Taxonomic Relation Extraction:** For a given set of generalized types T and relations R, ddentify non-taxonomic, semantic relations between

types to form a (T_h, r, T_t) triplet, where T_h and T_t are head and tail taxonomic types with $r \in R$.

The LLMs4OL dataset is publicly available on GitHub[1], providing easy access for researchers and practitioners in the field. The paper is organized as follows: Section 2 describes the domains that are being considered for benchmarking LLMs4OI and Section 3 investigates how ontologies are curated for OL. In section 4, we discuss the curated dataset. Finally, we conclude in Section 5

2 Ontological Resources and Domains of the Study

The *LLMs4OL 2024 datasets* support a variety of domains from lexosemantics to biomedical. Such variety supports the comprehensiveness of the studies within the *LLMs4OL 2024 Challenge*. In the following, we detail each ontology within the domains that we used for the construction of the LLMs4OL paradigm tasks dataset.

Lexosemantics. WordNet [17] is a large lexical database of English that serves as a rich ontology for NLP and other applications. It was developed at Princeton University and has become a widely used tool for understanding and representing the relationships between words. WordNet is divided into four main parts of speech, 1) Nouns: Concepts, entities, and objects. 2) Verbs: Actions, processes, or states of being. 3) Adjectives: Descriptive qualities or attributes. 4) Adverbs: Modifiers of verbs, adjectives, or other adverbs. Each part of speech has its own set of synsets and relationships, which helps in distinguishing the different meanings words can have when used in different grammatical contexts

Geographical Locations. The GeoNames [18] Ontology is a formal representation of geographical data that models geographic features, locations, and associated information. It is a crucial part of the Linked Open Data (LOD) cloud, providing a machine-readable format for geographic data to facilitate integration, querying, and sharing of geographic knowledge across different domains. GeoNames contains over 12 million geographical names and 9 million unique features such as cities, countries, rivers, mountains, lakes, etc. This makes GeoNames a rich ontology for further studies of LLMs4OL tasks.

Biomedical. The Unified Medical Language System (UMLS) [19] is a comprehensive biomedical ontology developed and maintained by the U.S. National Library of Medicine (NLM). It integrates various healthcare terminologies, coding systems, and ontologies to create a unified resource that supports NLP, biomedical data integration, and interoperability between different healthcare systems. UMLS Metathesaurus is a large database of biomedical concepts and terms that integrates many existing terminologies and coding systems. It consists of source vocabularies and includes well-known ontologies like SNOMEDCT_US [20], NCI [21], and MEDCIN [22]. The SNOMEDCT_US provides the core general terminology for the electronic health record. However, NCI covers vocabulary for cancer-related clinical care, translational and basic research, and public information and administrative activities. Moreover, the MEDCIN medical terminology encompasses symptoms, history, physical examination, tests, diagnoses, and therapies.

Biological. Gene Ontology (GO) [23] consortium is a major bioinformatics initiative that provides a standardized vocabulary to describe the functions, locations, and processes involving genes and gene products across different species. GO aims to unify

[1] https://github.com/HamedBabaei/LLMs4OL-Challenge-ISWC2024

the representation of gene and gene product attributes, allowing researchers to consistently annotate biological data and make it easier to compare gene functions across organisms. GO provides a hierarchical structure to describe gene products in three key areas such as *Biological Process (BP)*, *Molecular Function (MF)*, and *Cellular Component (CC)*. The BP describes our knowledge of the biological domain in the larger processes accomplished by multiple molecular activities. The CC goes beyond molecular activities and considers only location, relative to cellular compartments and structures. MF describes activities that occur at the molecular level, such as "catalysis" or "transport".

General Knowledge. DBpedia [24] is a crowd-sourced initiative aimed at extracting structured data from content generated across various Wikimedia projects. This data forms an open knowledge graph (OKG) that is accessible to everyone on the Web. The DBpedia Ontology (DBO), as a cross-domain ontology, emerged from a community effort to use Wikipedia's most commonly used infoboxes to create a formal vocabulary for categorizing knowledge for more precise querying and data linking. Wikipedia articles, typically representing specific entities (e.g., people, places, or events), can be classified under one or more of these classes. As a result of this, the ontology is structured as a hierarchy of classes and properties that describe concepts and their relationships, resulting in 768 classes, which form a subsumption hierarchy with around 3,000 properties and contain approximately 4 million instances.

Food. Food Ontology (FoodOn) [25] is a consortium-driven project to build a comprehensive and easily accessible global farm-to-fork ontology about food, that accurately and consistently describes foods commonly known in cultures from around the world. The FoodOn as a food product terminology supports food security, safety, quality, production, distribution, and consumer health and convenience.

Web Content Types. Schema.org [26] vocabulary covers entities, relationships between entities, and actions, and can easily be extended through a well-documented extension model. The schemas are a set of 'types', each associated with a set of properties and the types are arranged in a hierarchy. Overall, schema.org consists of 806 Types, 1476 properties 14 datatypes, 90 enumerations, and 480 enumeration members.

3 Ontology Curation for LLMs4OL Tasks

We curated 6 ontologies comprising a total of 10 datasets for Task A, 6 ontologies for Task B, and 3 ontologies for Task C. The curated ontologies and processes are represented in Figure 1, which involves three steps, each corresponding to the specified tasks. In this section, we provide a brief overview of the curation process.

3.1 WordNet Ontology – Task A

We utilized the WN18RR dataset, as introduced in [27]. For evaluation, we merged the test and validation sets, while the original training set was retained for model training. Additionally, we focused on four specific lexical term types T: nouns, verbs, adverbs, and adjectives. We also incorporated the sentences available in the WordNet dataset as additional context for the terms.

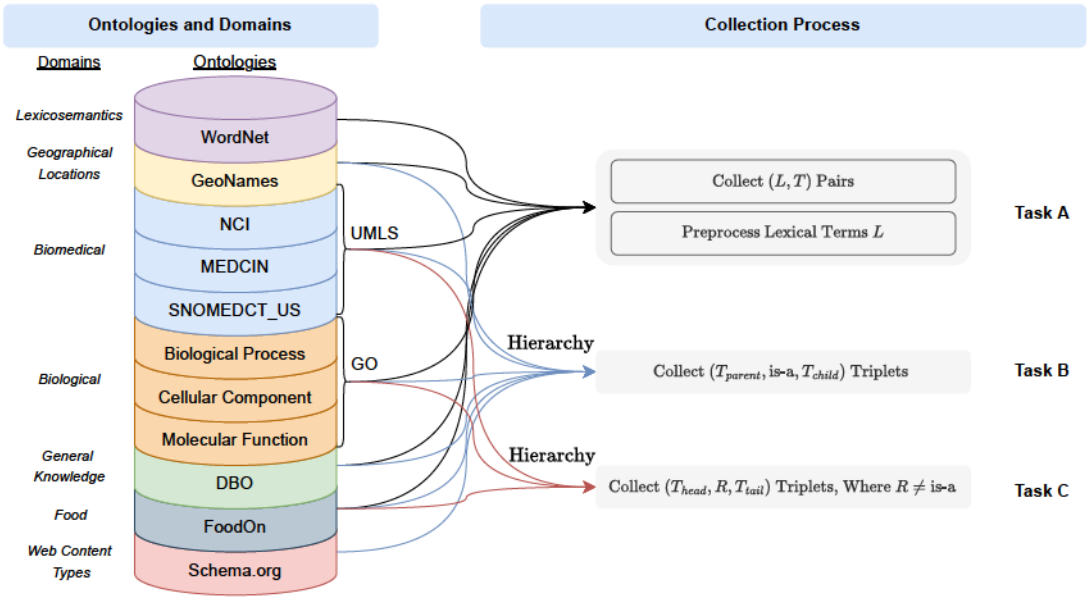

Figure 1. *LLMs4OL 2024 Datasets Curation.*

3.2 GeoNames Ontology – Tasks A and B

The GeoNames ontology encompasses all geographical locations worldwide. To narrow our focus, we first restricted the dataset to locations represented in English letters, resulting in a set of lexical terms L. GeoNames uses *Feature-Code* [28] to categorize and classify various geographic entities. Each location is associated with a *Feature-Code*, which denotes a type of geographical location (e.g. "road", or "populated place" locations). We mapped these *Feature-Code*'s to their corresponding names to create a set T, identifying a total of 680 distinct types within GeoNames. For instance, the term *"Elks Country Club"* with the feature-code *"S.RSRT"* is mapped to its type name, *"resort"*. The resulting (L, T) pairs were then used to create a train-test split based on T, with approximately 10% of the data allocated for testing and the remaining used for training at Task A.

For Task B, we utilized GeoNames *Feature Codes*, which are hierarchically structured to reflect varying levels of granularity in geographic features. These codes are divided into nine primary categories: "Administrative regions," "Hydrographic features," "Area," "Populated places," "Roads and railroads," "Spot features," "Terrain," "Undersea features," and "Vegetation." These categories operate at a higher level within a two-level taxonomy, resulting in 680 pairs with an "is-a" relationship. We then split the data into a 70-30 ratio to create training and test sets.

3.3 UMLS Ontology – Tasks A, B, and C

For generating UMLS sub-ontological sources i.e. NCI, MEDCIN, and SNOMEDCT_US, we considered `umls-2022AB-metathesaurus-full` version of the UMLS and processed the `MRCONSO` files for obtaining the terms that are written in English. Next, we used the following steps for extraction of lexical terms L, their respective types T, and relations:

1. *Filtering Lexical Terms*: For each source (NCI, MEDCIN, SNOMEDCT_US), the dataset is first filtered to extract relationships where both entities in a relationship belong to the specific source being considered. This filtering is done by matching

the source (NCI, MEDCIN, SNOMEDCT_US), ensuring that only triplets from that source are used. The Concept Unique Identifiers (CUIs) of these terms are then stored in a list, representing all the unique CUIs from the source.

2. *Retrieving Semantic Information*: After identifying the unique CUIs for each source, the next step is to gather semantic information about these CUIs. For each CUI, data from the MRSTY (Metathesaurus Semantic Types) file is used to obtain its Type Unique Identifiers (TUI), Semantic Type Numbers (STN), and Semantic Type Strings (STY). This information is collected and stored in a dictionary that links each CUI to its corresponding semantic types, ensuring that each TUI and STN is consistently associated with only one semantic type.
3. *Conflict Resolution*: During the previous steps, any conflicts—where a TUI or STN might be associated with different semantic types—are checked and reported. Once the consistency of the data is verified, the final hierarchy for each source (NCI, MEDCIN, SNOMEDCT_US) is obtained, which contains mappings from TUIs to their STNs and STYs, along with a list of all unique TUIs and STNs associated with each source, representing the hierarchical structure of entities within that specific source.

Thus, separate datasets for NCI, MEDCIN, and SNOMEDCT_US are created, each capturing the unique semantic relationships and entity types within those sources. For Task A, we only considered CUI's and TUI's to form the task dataset. We split the datasets per source into training and testing sets with a 70-30 ratio. For Tasks B and C, since both datasets are based on the same semantic network, we leveraged this network to extract types along with their relationships. Types with 'is-a' relationships are used for Task B, while non-'is-a' relationships are used for Task C. In both cases, the datasets are split using a 70-30 ratio.

3.4 Gene Ontology – Tasks A, B, and C

For the Term Typing task, we needed to map lexical terms (gene products) to their generalized types, derived from three Gene Ontology (GO) sub-ontologies: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). To collect relevant annotations, we used a Python script to query the GO Lookup Service (GOLR) via the following API: `https://golr-aux.geneontology.io`. The query retrieved annotations containing information such as gene product names (bioentity name), labels (annotation class label), and the associated ontology aspect. The dataset was then grouped by the aspect field, which corresponds to the sub-ontology (BP, CC, or MF), and duplicates were removed. After gathering and preprocessing the data, we created separate datasets for each sub-ontology, organizing gene products (L) and their corresponding types (T). To ensure the quality of the dataset, we applied a frequency threshold of 200, filtering out low-frequency terms, thus reducing noise. Subsequently, the dataset was divided into training and test sets, with a 70-30 split to ensure a robust evaluation of models performing the Term Typing task. The resulting datasets were sufficiently large, with unique term counts for each sub-ontology, ranging from 323 to 792.

For Task B, the objective was to identify hierarchical relationships (i.e., "is-a" relations) between the generalized types from Task A. We used the GO hierarchical structure, which defines relationships as edges between nodes representing different ontology term types. Using the GO ontology file, we extracted nodes and edges from the ontology graphs and then filtered the edges to retain only those that represent "is-a" relations. We then generated pairs of term types representing the child-parent relation-

```
SELECT DISTINCT ?class ?label WHERE {
    ?class a owl:Class;
           rdfs:label ?label .
    FILTER (lang(?label) = 'en')
}
```

Figure 2. *DBO SPARQL query for retrieving leaf classes for task A.*

```
SELECT DISTINCT ?term ?label WHERE {
    ?term a <leaf_class> ;
    rdfs:label ?label .
    FILTER (lang(?label) = 'en')
}
LIMIT 100
```

Figure 3. *DBO SPARQL query for retrieving 100 terms for given leaf class. The* `leaf_class` *is a place holder for replacing it with leaf class and querying for terms.*

ships (sub, obj). These pairs were split into training and test sets based on the unique term types involved, ensuring that no term appeared in both sets.

Finally, for Task C, we curated a dataset of semantic relationships between term types discovered in Task A. The relations are encoded in the GO using properties such as regulates, part of, and occurs in. We parsed the ontology to identify edges representing these relations, using a predefined set of relation mappings. Edges that matched the specified relation types were categorized into training and test sets. Similar to Task B, we ensured that there was no overlap in the relations between the training and test sets. The final dataset for Task C contained 10,538 training triplets and 7,234 test triplets, spanning multiple non-taxonomic relations.

3.5 DBPedia Ontology – Tasks A and B

We have used DBPedia Ontology (DBO) for both Task A and Task B, leveraging the structure and data provided by DBpedia's SPARQL endpoint. The datasets from this ontology has been utilized in a zero-shot setting, meaning it was used exclusively for testing without any prior training. The models were evaluated directly on these unseen tasks, without exposure to any data from the specific domain during training, emphasizing their generalization capabilities for Task A and Task B.

For Task A, we queried DBpedia for leaf classes and their associated terms in English. Leaf classes were identified using the SPARQL query as described in Figure 2, which retrieves all classes with English labels. For each leaf class, we queried up to 100 terms that belong to the class, again filtering for English terms using the SPARQL query provided in Figure 3. The results of these queries were aggregated into terms and their respective types, forming the dataset for Task A.

For Task B, we queried DBpedia's subclass ('is-a') hierarchy to generate parent-child relationships between taxonomic types. The SPARQL query, as described in Figure 4, retrieved subclass relationships where both parent and child have English labels. The resulting dataset contains hierarchical type pairs of "is-a" relations, with the taxonomic types stored as lists. This dataset serves as the input for our Taxonomy Discovery task.

```
SELECT DISTINCT ?childLabel ?parentLabel WHERE {
    ?child rdfs:subClassOf ?parent .
    ?child rdfs:label ?childLabel .
    ?parent rdfs:label ?parentLabel .
    ?child a owl:Class .
    ?parent a owl:Class .
    FILTER (lang(?childLabel) = "en")
    FILTER (lang(?parentLabel) = "en")
}
```

Figure 4. *DBO SPARQL query for creating "is-a" relationships between taxonomic types for Task B.*

```
PREFIX obo-term: <http://purl.obolibrary.org/obo/>
SELECT ?s ?label ?definition FROM <http://purl.obolibrary.org/obo/merged/FOODON> {
    ?s a owl:Class .
    ?s rdfs:label ?label .
    ?s obo-term:IAO\_0000115 ?definition .
}
```

Figure 5. *FoodOn SPARQL query for extract entity labels and definitions for Task A.*

3.6 Food Ontology - Tasks A, B, and C

For Food ontology (FoodOn), we construct datasets for tasks A, B, and C. All tasks are designed to evaluate models in a zero-shot setting. For Task A, we queried FoodOn to retrieve leaf classes (i.e., specific entity types) and associated terms. The SPARQL query as described in Figure 5 was used to extract entity labels and definitions, ensuring that only classes with English labels were included. The output from this query was processed to assign terms to one of the predefined high-level categories such as "Food", "Environment", "Agronomy", etc. This resulted in a dataset where each term is labeled with its corresponding class type (e.g., "Food", "Plant", etc.).

For Task B on taxonomy discovery, we extracted hierarchical relationships between classes by retrieving *rdfs:subClassOf* relationships from the FoodOn. We used the SPARQL query (presented in Figure 6) to obtain parent-child pairs of classes in English, capturing the taxonomic structure. This resulted in a taxonomy dataset with pairs of parent and child concepts, which we used to evaluate how well models can uncover subclass relationships in a zero-shot context.

For the Task C, we focused on extracting object properties that represent non-taxonomic relations between entities. The Figure 7 SPARQL query was used to retrieve all object properties and their labels from the FOODON ontology. We then applied these relations to extract triples of the form (head entity, relation, tail entity), where each triple represents a non-taxonomic relationship between two entities. This yielded a dataset with various relation types and corresponding triplets, allowing us to evaluate models' performance in predicting non-taxonomic relationships.

3.7 Schema.org – Task B

We also leveraged the Schema.org ontology to generate a dataset for Task B, with a primary goal of extracting hierarchical relations between concepts, enabling the evaluation of how well models can identify 'is-a' relationships within a taxonomy. We ex-

```
PREFIX obo-term: <http://purl.obolibrary.org/obo/>
SELECT DISTINCT ?childLabel ?parentLabel
FROM <http://purl.obolibrary.org/obo/merged/FOODON> WHERE {
    ?child rdfs:subClassOf ?parent .
    ?child rdfs:label ?childLabel .
    ?parent rdfs:label ?parentLabel .
    ?child a owl:Class .
    ?parent a owl:Class .
    FILTER (lang(?childLabel) = "en")
    FILTER (lang(?parentLabel) = "en")
}
```

Figure 6. *FoodOn SPARQL query to obtain parent-child pairs of classes in English for Task B.*

```
FROM <http://purl.obolibrary.org/obo/merged/FOODON> WHERE {
    ?property a owl:ObjectProperty .
    ?property rdfs:label ?propertyLabel .
    FILTER (lang(?propertyLabel) = "en") .
}
ORDER BY ?propertyLabel
```

Figure 7. *FoodOn SPARQL query to extract object properties that represent non-hierarchical relations in English for Task C.*

tracted subclass relationships from the Schema.org taxonomy by processing the ontology. First, we filter out irrelevant concepts by excluding root concept `Thing` or other irrelevant RDF classes like `rdf-schema#Class`. Next, we prepare parent-child pairs by using `subTypeOf` property, where if a child had multiple parents, we split these into separate parent-child pairs. This gave us a list of hierarchical relationships, where each pair represented a child-parent relationship. Finally, to simulate a realistic few-shot scenario, we split the types into training and testing sets. Concepts that appeared in the subTypeOf property were divided into two sets using an 80/20 train-test split. Parent-child pairs were then assigned to the training or testing set based on the parent concepts.

4 Dataset Statistics

The LLMs4OL 2024 dataset is designed to support the benchmarking of ontology learning models, with a total of 19 datasets distributed across three core tasks: Task A - Term Typing, Task B - Taxonomy Discovery, and Task C - Non-Taxonomic Relation Extraction. The largest proportion of data is allocated to the Term Typing task, given its fundamental role in associating terms with predefined types, which lays the groundwork for downstream OL processes. Moreover, Taxonomy Discovery and Non-Taxonomic Relation Extraction tasks are more specialized, focusing on hierarchical and non-hierarchical relationships, respectively. This balanced yet task-specific distribution ensures that models are tested across diverse, real-world learning scenarios.

Task A - Term Typing. Task A datasets as described in Table 1 covers both few-shot (FS) and zero-shot (ZS) evaluation phases across multiple domains. The GeoNames (A.2 FS) is the largest dataset, with over 8 million training samples and 702 thousand testing samples, making it highly significant for large-scale geographic term

Table 1. *LLMs4OL 2024 datasets – TASK A - TERM TYPING – domains and evaluation phases. "FS" refers to the Few-Shot testing phase dataset containing train and test sets, But "ZS" refers to the Zero-shot testing phase evaluation dataset containing only test sets.*

Dataset	***Domain***	***Train***	***Test***	***Types***
A.1 (FS) - WordNet	lexicosemantics	40,559	9,470	4
A.2 (FS) - GeoNames	geographical locations	8,078,865	702,510	680
A.3 (FS) - UMLS - NCI	biomedical	96,177	24,045	125
A.3 (FS) - UMLS - MEDCIN		277,028	69,258	87
A.3 (FS) - UMLS - SNOMEDCT_US		278,374	69,594	125
A.4 (FS) - GO - Biological Process	biological	195,775	108,300	792
A.4 (FS) - GO - Cellular Component		228,460	126,485	323
A.4 (FS) - GO - Molecular Function		196,074	107,432	401
A.5 (ZS) - DBO	general knowledge	-	44,724	484
A.6 (ZS) - FoodOn	food	-	18,087	12

Table 2. *LLMs4OL 2024 datasets – TASK B - TAXONOMY DISCOVERY – domains and evaluation phases. "FS" refers to the Few-Shot testing phase dataset containing train and test sets, But "ZS" refers to the Zero-shot testing phase evaluation dataset containing only test sets. "Size" refers to ground truth "is-a" pairs.*

Dataset	***Domain***	***Train***		***Test***	
		Size	***Types***	***Size***	***Types***
B.1 (FS) - GeoNames	geographical locations	476	477	204	212
B.2 (FS) - Schema.org	web content types	1,070	2,062	364	728
B.3 (FS) - UMLS	biomedical	74	76	45	51
B.4 (FS) - GO	biological	33,703	25,372	5,753	6,621
B.5 (ZS) - DBO	general knowledge	-	-	742	762
B.6 (ZS) - FoodOn	food	-	-	30,240	25,631

typing. Moreover, UMLS (A.3 FS) provides detailed biomedical data across three sub-ontological sources such as NCI, MEDCIN, and SNOMEDCT_US, each with a large number of types crucial for specialized medical term categorization. The GO (A.4 FS) dataset, particularly the *"Biological Process (BP)"* subset, offers terms and types with the highest variety of types up to 792. DBO (A.5 ZS) and FoodOn (A.6 ZS) are important zero-shot datasets, to study the generalization of fine-tuned models.

Task B - Taxonomy Discovery. Task B dataset statistics are covered in Table 2, showcasing 6 datasets from different domains. The GeoNames (B.1 FS), Schema.org (B.2 FS), and UMLS (B.3 - FS) are relatively small in terms of training examples but represent unique domains (geographical locations, web content, and biomedical). Similarly, a zero-shot dataset DBO (B.5 FS) has small examples for testing which plays a real-world scenario to study the generalization of models, when they are fine-tuned. Moreover, GO (B.4 FS) stands out with over 33,703 training samples and the highest variety of types 25,372, making it key for biological taxonomy discovery. And FoodOn (B.6 ZS) is significantly large with 30,240 test samples and 25,631 types, focusing on the evaluation of the generalization of models in finding taxonomies in the food domain.

Task C - Non-Taxonomic Relation Extraction. Task C consists of 3 datasets, as shown in Table 3, the datasets for this task are few in comparison to task A and B datasets. The UMLS (C.1 FS), despite its moderate size, holds significance in biomedical relation extraction, focusing on multiple relation types. GO (C.2 FS) shows an imbalance in relation types, with 5 relations for training but only 2 relations for testing.

Table 3. *LLMs4OL 2024 datasets – TASK C - NON-TAXONOMIC RELATION EXTRACTION – domains and evaluation phases. "FS" refers to the Few-Shot testing phase dataset containing train and test sets, But "ZS" refers to the Zero-shot testing phase evaluation dataset containing only test sets. "Size" refers to ground truth (h, r, t) triplets.*

Dataset	Domain	Train			Test		
		Size	Types	Relations	Size	Types	Relations
C.1 (FS) - UMLS	biomedical	3,030	121	33	2,611	111	15
C.2 (FS) - GO	biological	10,538	10,901	5	7,234	14,065	2
C.3 (ZS) - FoodOn	food	-	-	-	7,086	7,298	26

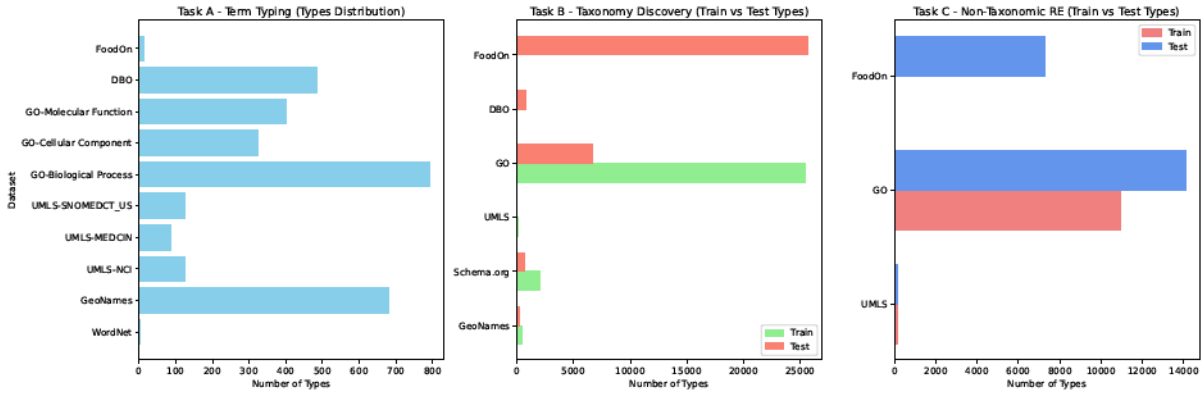

Figure 8. *LLMs4OL datasets type distributions in train and test sets.*

FoodOn (C.3 ZS), with 7,086 test samples and 26 relations, highlights the complexity of non-taxonomic relations in the food domain.

Types Distributions. The Figure 8 highlights the complexity of the datasets across tasks. In Task A (Term Typing), the GeoNames and GO-Biological Process datasets stand out with the highest number of types, while WordNet and FoodOn have relatively fewer types, indicating simpler classification challenges. For Task B (Taxonomy Discovery), the Schema.org and GO datasets show a large number of types in both train and test phases, suggesting their complexity, while FoodOn features a high number of test types despite having no training data, making it a challenging zero-shot task. Lastly, in Task C (Non-Taxonomic Relation Extraction), the GO dataset shows a significant increase in types from train to test, and FoodOn again presents a large number of types and relations, reinforcing its difficulty in a zero-shot setting.

5 Conclusion

In this paper, we introduced the LLMs4OL 2024 dataset, designed to advance the field of OL by leveraging the capabilities of LLMs. The dataset encompasses three core tasks— Task A - Term Typing, Task B - Taxonomy Discovery, and Task C - Non-Taxonomic Relation Extraction—across seven distinct domains, providing a comprehensive benchmark for evaluating LLMs in diverse semantic and structural contexts. By focusing on these tasks, we aim to push the boundaries of OL and enhance the development of models capable of processing unstructured text into formalized knowledge representations. The dataset also reflects real-world challenges such as class imbalance and domain-specific variations, which are crucial for the development of robust, generalizable models. Furthermore, its integration into the LLMs4OL Challenge

at the 23rd International Semantic Web Conference (ISWC) 2024 aims to foster community engagement and encourage the exploration of novel approaches to OL.

Moving forward, this dataset and its benchmarks will provide researchers with a foundational resource to explore the intersection of LLMs and OL, promoting further innovations in knowledge extraction, classification, and relation discovery. We believe that the LLMs4OL 2024 dataset will serve as a key catalyst in the ongoing evolution of OL and its practical applications across a variety of domains.

Data Availability Statement

The datasets supporting this article are publicly available and can be accessed via Zenodo at https://doi.org/10.5281/zenodo.13851373, or through the GitHub repository: https://github.com/HamedBabaei/LLMs4OL-Challenge-ISWC2024.

Authors Contributions

Hamed Babaei Giglou: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data Curation, Writing - Original Draft, Writing – Review & Editing, Visualization.
Jennifer D'Souza: Conceptualization, Methodology, Investigation, Resources, Supervision, Project administration, Funding acquisition, Writing – Review & Editing, Visualization.
Sameer Sadruddin: Methodology, Resources, Data Curation.
Sören Auer: Conceptualization, Methodology, Review & Editing, Supervision, Project administration, Funding acquisition.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The 1st LLMs4OL Challenge @ ISWC 2024 jointly supported by the NFDI4DataScience initiative (DFG, German Research Foundation, Grant ID: 460234259) and the SCINEXT project (BMBF, German Federal Ministry of Education and Research, Grant ID: 01IS22070).

References

[1] A. Maedche and S. Staab, "Ontology learning," in *Handbook on Ontologies*, S. Staab and R. Studer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 173–190, ISBN: 978-3-540-24750-0. DOI: 10.1007/978-3-540-24750-0_9. [Online]. Available: https://doi.org/10.1007/978-3-540-24750-0_9.

[2] A. Konys, "Knowledge repository of ontology learning tools from text," *Procedia Computer Science*, vol. 159, pp. 1614–1628, 2019.

[3] Y. Ding and S. Foo, "Ontology research and development. part 2-a review of ontology mapping and evolving," *Journal of information science*, vol. 28, no. 5, pp. 375–388, 2002.

[4] M. Shamsfard and A. Abdollahzadeh Barforoush, "The state of the art in ontology learning: A framework for comparison," *Knowl. Eng. Rev.*, vol. 18, no. 4, pp. 293–316, Dec.

2003, ISSN: 0269-8889. DOI: 10.1017/S0269888903000687. [Online]. Available: https://doi.org/10.1017/S0269888903000687.

[5] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology learning from text: methods, evaluation and applications*. IOS press, 2005, vol. 123.

[6] M. Hazman, S. R. El-Beltagy, and A. Rafea, "A survey of ontology learning approaches," *International Journal of Computer Applications*, vol. 22, no. 9, pp. 36–43, 2011.

[7] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi, "A survey of ontology learning techniques and applications," *Database*, vol. 2018, bay101, Oct. 2018, ISSN: 1758-0463. DOI: 10.1093/database/bay101. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/bay101/27329264/bay101.pdf. [Online]. Available: https://doi.org/10.1093/database/bay101.

[8] H. Babaei Giglou, J. D'Souza, and S. Auer, "Llms4ol: Large language models for ontology learning," in *The Semantic Web – ISWC 2023*, T. R. Payne, V. Presutti, G. Qi, *et al.*, Eds., Cham: Springer Nature Switzerland, 2023, pp. 408–427, ISBN: 978-3-031-47240-4.

[9] F. Petroni, T. Rocktäschel, P. Lewis, *et al.*, *Language models as knowledge bases?* 2019. arXiv: 1909.01066 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1909.01066.

[10] B. Zhang, V. A. Carriero, K. Schreiberhuber, *et al.*, "Ontochat: A framework for conversational ontology engineering using language models," *arXiv preprint arXiv:2403.05921*, 2024.

[11] V. K. Kommineni, B. König-Ries, and S. Samuel, "From human experts to machines: An llm supported approach to ontology and knowledge graph construction," *arXiv preprint arXiv:2403.08345*, 2024.

[12] M. J. Saeedizade and E. Blomqvist, "Navigating ontology development with large language models," in *European Semantic Web Conference*, Springer, 2024, pp. 143–161.

[13] R. Du, H. An, K. Wang, and W. Liu, *A short review for ontology learning: Stride to large language models trend*, 2024. arXiv: 2404.14991 [cs.IR]. [Online]. Available: https://arxiv.org/abs/2404.14991.

[14] H. Khorashadizadeh, F. Z. Amara, M. Ezzabady, *et al.*, *Research trends for the interplay between large language models and knowledge graphs*, 2024. arXiv: 2406.08223 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2406.08223.

[15] H. Babaei Giglou, J. D'Souza, and S. Auer, "Llms4ol 2024 overview: The 1st large language models for ontology learning challenge," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[16] A. Maedche and S. Staab, "Ontology learning for the semantic web," *IEEE Intelligent systems*, vol. 16, no. 2, pp. 72–79, 2001.

[17] G. A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[18] *Geonames geographical database*, 2023. [Online]. Available: http://www.geonames.org/.

[19] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D267–D270, Jan. 2004, ISSN: 0305-1048. DOI: 10.1093/nar/gkh061. eprint: https://academic.oup.com/nar/article-pdf/32/suppl_1/D267/7621558/gkh061.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkh061.

[20] National Library of Medicine (US), *US Edition of SNOMED CT*, http://www.nlm.nih.gov/research/umls/Snomed/us_edition.html, Bethesda, MD, 2013.

[21] National Cancer Institute (US), *NCI Enterprise Vocabulary Services (EVS)*, https://www.cancer.gov/research/resources/terminology, Bethesda, MD, 2015.

[22] Medicomp Systems, Inc., *MEDCIN*, http://www.medicomp.com/index_html.htm, Chantilly, VA, 2004.

[23] S. Carbon and C. Mungall, *Gene ontology data archive*, version 2024-01-17, Zenodo, Jan. 2024. DOI: 10.5281/zenodo.10536401. [Online]. Available: https://doi.org/10.5281/zenodo.10536401.

[24] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web*, K. Aberer, K.-S. Choi, N. Noy, *et al.*, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 722–735, ISBN: 978-3-540-76298-0.

[25] D. M. Dooley, E. J. Griffiths, G. S. Gosal, *et al.*, "FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration," *NPJ Science of Food*, vol. 2, p. 23, Dec. 2018. DOI: 10.1038/s41538-018-0032-6. [Online]. Available: https://www.nature.com/articles/s41538-018-0032-6.

[26] P. F. Patel-Schneider, "Analyzing schema.org," in *The Semantic Web – ISWC 2014*, P. Mika, T. Tudorache, A. Bernstein, *et al.*, Eds., Cham: Springer International Publishing, 2014, pp. 261–276, ISBN: 978-3-319-11964-9.

[27] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, *Convolutional 2d knowledge graph embeddings*, 2018. arXiv: 1707.01476 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1707.01476.

[28] GeoNames, *Geonames feature codes*, https://www.geonames.org/export/codes.html, 2024.