Automated Ontology Evaluation: Evaluating Coverage and Correctness using a Domain Corpus

Antonio Zaitoun
University of Haifa,, Israel
azaitoun@campus.haifa.ac.il

Tomer Sagi
Aalborg University,, Denmark
tsagi@cs.aau.dk

Katja Hose
Aalborg University,, Denmark
TU Wien,, Austria
katja.hose@tuwien.ac.at

ABSTRACT

Ontologies conceptualize domains and are a crucial part of web semantics and information systems. However, re-using an existing ontology for a new task requires a detailed evaluation of the candidate ontology as it may cover only a subset of the domain concepts, contain information that is redundant or misleading, and have inaccurate relations and hierarchies between concepts. Manual evaluation of large and complex ontologies is a tedious task. Thus, a few approaches have been proposed for automated evaluation, ranging from concept coverage to ontology generation from a corpus. Existing approaches, however, are limited by their dependence on external structured knowledge sources, such as a thesaurus, as well as by their inability to evaluate semantic relationships. In this paper, we propose a novel framework to automatically evaluate the domain coverage and semantic correctness of existing ontologies based on domain information derived from text. The approach uses a domain-tuned named-entity-recognition model to extract phrasal concepts. The extracted concepts are then used as a representation of the domain against which we evaluate the candidate ontology's concepts. We further employ a domain-tuned language model to determine the semantic correctness of the candidate ontology's relations. We demonstrate our automated approach on several large ontologies from the oceanographic domain and show its agreement with a manual evaluation by domain experts and its superiority over the state-of-the-art.

CCS CONCEPTS

• **Information systems → Web Ontology Language (OWL)**; • **Computing methodologies → Natural language processing**.

KEYWORDS

ontology, natural language processing, BERT, knowledge engineering

ACM Reference Format:
Antonio Zaitoun, Tomer Sagi, and Katja Hose. 2023. Automated Ontology Evaluation: Evaluating Coverage and Correctness using a Domain Corpus. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543873.3587617

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587617

1 INTRODUCTION

An ontology is a collection of concepts and relations. Each concept is unique and is often characterized by multiple attributes. Ontologies usually describe a single domain and are used to abstract and formally define the semantic meaning of concepts in a domain and the relations between them [21]. In data integration, ontologies can play an important role by unifying and relating data elements under concepts despite having different schemas [11]. For example, the concepts of *address* and *residence* can be placed under the concept of *location*, i.e., an address/residence is a location. Then, during the integration of the two data sources, the fields *address* and *residence* are mapped to a common concept - *location*. Another example from the domain of oceanography is *Nutrients*. Nutrients refer to the amount of dissolved inorganic macronutrients in seawater such as *Silicate* or *Phosphate*. Similar to the *location* example, both concepts will be grouped under a common ancestor - *Nutrients*. Thus, ontology-based data integration/access (OBDI/OBDA) [10, 49] requires the existence of an ontology that encompasses the knowledge domains of the datasets being integrated.

The construction of an ontology remains a challenge, and it is often painstakingly done manually by domain experts [2]. One of the main challenges with manually constructed ontologies is their inability to adapt to other tasks. Not only do they contain subjective knowledge that may be incompatible with the task at hand, but they may also lack important concepts and relationships required for the specific data integration task or even just contain errors [20, 45]. Therefore, reusing manually constructed ontologies requires an evaluation step. More precisely, the evaluation of the relevance and coverage of the set of concepts contained in the ontology and the semantic correctness of the relations between these concepts with respect to the domain.

Computer-generated ontologies are potentially far more robust in terms of size and scope as they are based upon a comprehensive review of the domain and require repeated evidence for each proposed concept and relation [35]. However, these auto-constructed ontologies tend to lack nuanced information (such as definitions, constraints, or axioms), and are limited in the type of generated relations. To take advantage of both the utility and nuance of manually constructed ontologies and the robustness of automated methods, one requires an automated method to evaluate and correct ontologies. In this paper, we suggest an automated ontology evaluation framework using a representative domain model to address this need.

Existing approaches of automated ontology evaluation with respect to a domain [6] [14] are limited in two respects. The first is concept extraction from the domain, a required step that generates

the pool of concepts to which the ontology under evaluation is compared. Existing methodologies utilize Part-of-Speech (POS) models that can only determine single-word terms. Multi-word phrases, such as *Air Temperature*, will be split into separate concepts (*Air*, *Temperature*). Moreover, when evaluating ontological relationships, existing approaches only consider their semantic relations with respect to external sources such as thesauruses, dictionaries, and vocabularies, all of which are general-purpose and are not representative of the domain.

To address these gaps we propose a novel automated evaluation framework able to both determine the semantic correctness of relationships between concepts as well as the completeness of an ontology with respect to a particular domain. To achieve this, the framework utilizes a language-model-based representation of the domain to serve as an authoritative source of truth. Our approach utilizes a Named-Entity-Recognition (NER) model, which can not only pick up multi-word phrases but also label their types. We employ a pre-trained bi-directional transformer-based language model BERT [13], which serves as an auto-generated representation of the domain. We demonstrate our method over the oceanographic domain and show how the evaluation generates useful and actionable insights that can be used to improve the evaluated ontology. We further evaluate the language model and show it to be a good representation of the domain, comparable to human experts.

The remainder of this paper is organized as follows. Section 2 provides preliminary definitions, and Section 3 reviews previous work. In Section 4, our proposed automated evaluation method is described in detail. In Section 5, we demonstrate our evaluation method over three ontologies, and in Section 6, we perform a meta-evaluation of the method. Finally, Section 7 presents our conclusions and directions for future work.

2 BACKGROUND AND PRELIMINARIES

As defined by Gruber [21], "An ontology is an explicit specification of a conceptualization". The representation is made through a collection of concepts and relations between them. Formally:

Definition 2.1 (Ontology). Let C be a set of concepts, let R be a set of relations and let A be a set of relation associations such that $A \subseteq \{r(x,y) | \forall r \in R, \forall x \in C, \forall y \in C\}$ then an *Ontology* O is a triple $O :=< C, R, A >$

Ontologies oft describe a single domain and are used as the definitive source for the semantics of concepts in that domain. Ontologies are a crucial part of web semantics and information systems as they capture representations of the domain such that machines can interpret them. Such interpretations are mostly required in tasks such as information retrieval [37, 46, 50], data integration [17, 27, 52]), and knowledge alignment [9, 24]. It is important to note that ontologies may also encompass additional knowledge, such as constraints, axioms, instances, and properties [21] but were not explicitly stated in the definition for simplicity. When evaluating an ontology, we use the term *concept family*, comprised of a parent concept and a set of direct child concepts, formally:

Definition 2.2 (Concept Family). Let $O =< C, R, A >$ be an ontology and let $ISA \in R$ be one of its relationships, then CF is *concept family* $\iff$ $CF =< C_p, C_s >$ where $C_p \in C, C_s \subseteq C, \forall c \in C_s \implies ISA(c, C_p)$

Ontology construction by domain experts is a labor-intensive task. Thus, several (semi-)automated methods were suggested using rule-based approaches [15, 25, 29, 33, 43] later advancing to techniques based upon Formal Concept Analysis (FCA) [12, 19, 47, 48] and Natural Language Processing (NLP) [2, 8, 16, 41].

Both manually constructed and automatically-constructed ontologies require evaluation before they can be reused for a new task. Throughout this paper, we will use the term *candidate* ontology to refer to the ontology being evaluated. Raad et al. [39] reviewed ontology evaluation methods and identified seven evaluation criteria defined as follows.

- **Accuracy** refers to concept definition correctness.
- **Completeness** determines an ontology's coverage of the domain.
- **Conciseness** identifies irrelevant concepts in the ontology.
- **Adaptability** measures how well an ontology is suitable for its intended task.
- **Clarity** assesses how well the intended meaning of the ontology is being projected, i.e., concepts should be independent of the context.
- **Computational efficiency** measures the usage cost of the ontology in terms of performance.
- **Consistency** serves as a measure of contradictions within the ontology.

These criteria can be used in different evaluation methods that were classified into the following four categories based on the artifact used as a basis of comparison to evaluate the candidate ontology.

(1) **Gold standard-based** methods compare an ontology with a previously (typically manually) created ontology.
(2) **Corpus-based** methods extract terms from a corpus and use them to determine the evaluated ontology's fit to the domain represented by the corpus. These methods focus on the accuracy, completeness, and conciseness criteria.
(3) **Task-based** methods evaluate the ontology by its fit to solve a specific set of tasks that the ontology is designed for, focusing on the adaptability criterion.
(4) **Criteria-based** methods evaluate the ontology by computing scores based on a set of rules and constraints. This evaluation is centered upon the structure of the ontology and often addresses the clarity criterion.

In this work, we propose a novel corpus-based method that covers four evaluation criteria, namely, accuracy, completeness, conciseness, and consistency.

3 RELATED WORK

Brewster et al. [6] first proposed to extract terms from a document corpus to assess an ontology's completeness. They further suggested using WordNet [34] to expand the list of extracted concepts, although it remains limited to single-word concepts. Additionally, they perform a vector-space similarity comparison between the text corpus and the ontology to assess accuracy. We extend this approach by supporting phrasal (more than one word) concepts, utilizing a concept extraction method tuned to the domain at hand,

and addressing additional coverage-based criteria such as conciseness. In contrast with our work, their method does not address consistency as in our work, where we evaluate the correctness of relations within the candidate ontology. Furthermore, the accuracy and utility of the extracted concept set is suspect, as the authors employ a general-purpose WordNet thesaurus and PoS tagger for this purpose. In this work, we create an accurate representation of the domain by utilizing a large language model extensively trained over a large representative set of documents from the domain.

DiGiuseppe et al. [14] proposed another corpus-based approach in which an ontology is generated from the corpus and compared to the candidate ontology. In their approach, concepts are extracted using PoS tagging and mapped via vocabularies to determine their synonyms and synonym symmetry. The synonym information is used to derive the concepts' hierarchy. The process results in a corpus-based ontology. The generated ontology is then compared to the candidate ontology. The coverage analysis outputs scores for classes, class equivalence, hierarchies, and breadth. The approach is both a corpus-based and criteria-based method. Again, only single-word nouns are considered, which is a limitation of PoS. Furthermore, the external dictionaries used to determine the synonyms are general-purpose English dictionaries that do not reflect the true relations in the domain. In this work, we support multi-word concepts and utilize a domain-tuned language model to evaluate the candidate ontology's relations.

OOPS [38] is a web-based evaluation tool for OWL ontologies. Its evaluation is mostly based on lexical and structural patterns highlighting ontology pitfalls. This evaluation can be categorized as criteria-based since it employs rules and patterns. Although it can determine if an ontology is aligned with common standards, it cannot assess the fitness of the ontology to a particular domain as our work does.

Ontologies are sometimes mentioned in relation to linked data [5]. However, while ontologies focus on the conceptual description of a domain, linked data refers to large sets of related entities representing instances of these concepts and relation types. Work around the evaluation of Linked-Data (LD) has been proposed [18, 28, 40, 51], in which a rule-based approach is taken to find inconsistencies among data instances within an LD data source. In this work, we focus on evaluating ontologies rather than instances and records as in LD evaluation.

In a rare example of using large language models (LLM) in the context of ontologies, Liu et al. [32] present an approach for placing a set of new concepts within an existing ontology. In their paper, the authors utilize the BERT language model [13], specifically its next sentence prediction capabilities, to determine if a hierarchical relationship between two concepts exists. They do so by pre-training BERT on corpus text from the domain, then fine-tuning it using a set of pairs of concepts that exhibit a taxonomic relationship (i.e., "IS-A"), taken from the SNOMED biomedical ontology. They then test the model over concepts from the latest version of the ontology that were not present in the previous version that was used as training data. The results of the trained model yield an average of 95% recall and 85% precision. This suggests that a language model, such as BERT, can learn the semantic meaning of the concepts and provide accurate relationship predictions even of unseen concepts. However, Liu et al. [32] do not attempt to evaluate an ontology but only demonstrate the ability of an LLM to learn the semantics of the domain and the relations between its concepts. In this work, we utilize this ability to evaluate the completeness, accuracy, conciseness, and consistency of a domain ontology.

4 AUTOMATED ONTOLOGY EVALUATION

In this section, we describe our automated approach for evaluating an ontology with respect to a domain of interest. Our method allows the evaluation of completeness (coverage) and correctness (semantic relation coherence). Furthermore, we can use the evaluation's results to identify specific concepts missing from the ontology as well as misaligned semantic relations between its existing concepts.

Fig. 1 illustrates our proposed evaluation method in chronological order. In order to evaluate the candidate ontology, we must generate an accurate representation of the domain. This representation takes two forms. The first is a domain-trained language model (Domain BERT), used to judge relations between concepts using the semantics of the domain rather than their general-purpose use in English. The second is a collection of phrasal concepts (Domain concepts) extracted from the domain text corpus using a specialized named entity recognition (NER) model. The candidate ontology's concepts can then be compared to this concept collection. In the following sections, we detail each step in the proposed evaluation pipeline. We start with describing how a collection of documents or text corpus (Section 4.1) is created, followed by our method for training a specialized NER model (Section 4.2) and pre-training a language model (Section 4.5). Using the NER model, domain concepts are extracted from the text (Section 4.3) and are then matched with the concepts in the candidate ontology (Section 4.4), from which a sub-set of this ontology (Sub-ontology) is derived. Next, using the pre-trained language model, an evaluation (Section 4.6) takes place, generating a set of scores that reflect the correctness and completeness of the candidate ontology with respect to the domain.

4.1 Document Collection

Document collection is a crucial step since it serves as the core of this pipeline. It is assumed that the corpus encompasses the knowledge that is required to represent the domain since both the NER and language model are trained on it. The use of large collections of text to represent a domain is not new. Large text collections are routinely used in a variety of tasks, from training domain-specific language models [22] to guiding automated literature surveys [3]. We consider the use of peer-reviewed representations of domain knowledge created by thousands of experts to be a robust representation of the domain. Moreover, such ontologies often form the basis for ontology-based data access and data integration system, e.g., [10], and since scientific datasets are often described by research papers, we expect the same concepts to describe these datasets. In this work 10,000 papers from oceanographic journals were used, collected using a web crawler and the Crossref API based on previous research [44]. The papers were converted to raw text with ScienceParse[1], including only the title, abstract, and content.

[1] https://github.com/allenai/science-parse

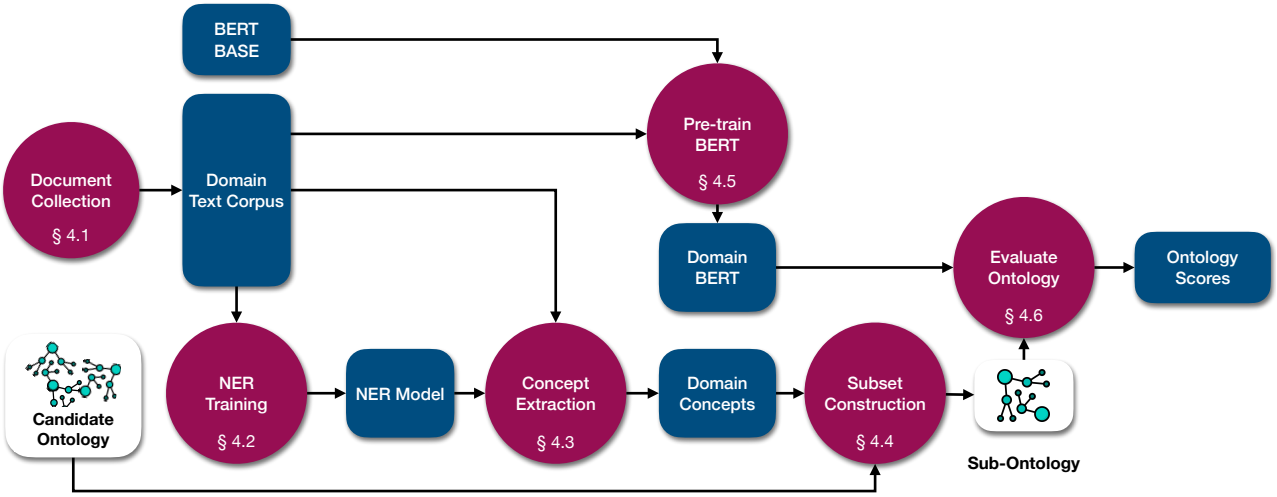

Figure (1) Ontology evaluation pipeline. Details of the steps can be found in the corresponding sections.

4.2 Domain Specific Named Entity Recognition

A typical NER model is capable of identifying phrases representing *named entities* in text, such as people (Marie Curie), places (Warsaw), or organizations (United Nations). But in order to be able to extract phrases representing (not necessarily named) *concepts* relevant to the domain such as *temperature*, one must train a custom NER model [31]. Using the collected text corpus (Section 4.1), a domain-specific NER model is trained. NER models are often created through supervised or semi-supervised approaches, requiring manual annotation of a sample of the corpus by domain experts. Then, this annotated sample can used by existing NER architectures (e.g., [1] that is used here) to train a domain specific model.

4.3 Concept Extraction

Here we use the previously described NER model to extract a set of concepts (hereafter, domain concepts). The NER model, can detect multi-word phrases as well as semantically label them into classes, such as *Organization* or *Measured Variable*. After extracting the concepts, a threshold is applied to remove concepts with a small number of occurrences, assuming these are not representative of the entire but perhaps only a small subset of it. Finally, the remaining concepts are considered to be the *domain concepts* (gray and red dots in Figure 2). We filtered and kept only concepts that have appeared in at least ten different papers.

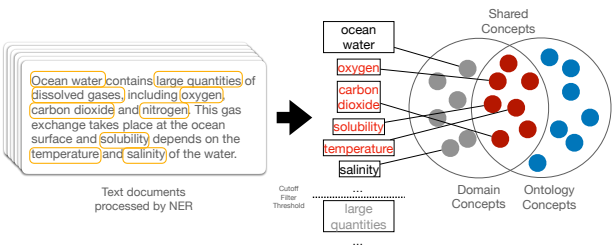

Figure (2) Concept extraction from text and the determination of shared concepts between the domain concepts and the ontology concepts.

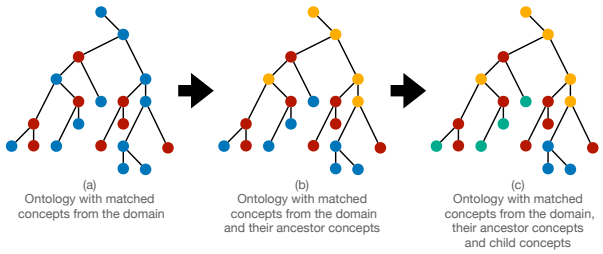

Figure (3) Ontology subset derivation phase in which shared concepts (red) are first identified among the candidate ontology's concept. Next, the ontology's is-a hierarchy is used to add their ancestors (yellow). Finally, the children (green) of the shared concepts are added and the remaining unconsidered concepts (blue) are removed.

4.4 Ontology Subset Construction

Here, we construct a domain-relevant subset of the candidate ontology. It is assumed that concepts excluded from this subset are not relevant to the domain, an assumption we evaluate in Section 6. Fig. 3 depicts the process. We first match all concepts from the candidate ontology to the previously extracted domain concepts. We begin by standardizing the textual representation of both by lower-casing and lemmatization. For example, concepts containing words like *Raining*, *Temperatures*, and *Solids* become *rain*, *temperature*, and *solid*, respectively. After applying this process to both the domain concepts and the candidate ontology's concepts, we perform an exact match search for overlapping concepts. We refer to these overlapping concepts as *Shared Concepts* (marked red in Figs. 2 and 3). Using the shared concepts we traverse the candidate ontology's hierarchy such that every ancestor in the hierarchy of a shared concept (yellow) is included, as well as every direct child of that concept (green).

4.5 Language Model Pre-training

Using the previously mentioned (Section 4.1 text corpus, we pretrain a BERT [13] language model such that it adjusts to the domain.

This process entails feeding the model with the text corpus using pairs of sentences where some follow each other and some not, letting the model learn to predict the most probable next sentence while at the same time masking some of the words to let the model predict the masked words. Previous research [23] has shown pre-training to increase the performance of downstream tasks utilizing such models. The final output of this phase is a pre-trained language model that is adapted to the domain. In the following section we utilize this model's embedding layer to encode the concepts into a vector space for similarity evaluation.

4.6 Evaluation Measures

To determine the completeness of the candidate ontology , we compute three metrics: *Ontology Relevance*, *Sub-Ontology Relevance*, and *Domain Coverage* based upon three quantifications of the concept overlap between the candidate ontology and the concept set extracted from the text corpus, representing the domain (see Venn diagram in Figure 2). $\mathcal{O}$ represents the number of concepts in the candidate ontology. $\mathcal{D}$ *Domain concepts* counts the number of concepts extracted from the domain corpus after pruning (Section 4.3), $\mathcal{S}$ counts the number of shared concepts found between the candidate ontology and the domain concepts extracted from the corpus, and $\mathcal{H}$ the number of concepts in the subset of the candidate ontology constructed by taking the shared concepts and expanding them using the ontology's hierarchical relations (Section 4.4). The measures are defined as follows.

$$\text{Domain Coverage} = \frac{\mathcal{S}}{\mathcal{D}} \tag{1}$$

$$\text{Ontology Relevance} = \frac{\mathcal{S}}{\mathcal{O}} \tag{2}$$

$$\text{Sub-Ontology Relevance} = \frac{\mathcal{S}}{\mathcal{H}} \tag{3}$$

Revisiting the terminology introduced by Raad and Cruz [39], Eq. 1 represents a *completeness* measure, evaluating the completeness of the candidate ontology concept set with respect to the domain. Equations 2 and 3 measure *conciseness*, or the extent to which the candidate ontology (or its subset) is relevant to the domain.

Semantically similar concepts are expected to share properties [30]. Thus, defining measures that estimate this similarity is important. Therefore, to measure the correctness of the semantic relationships between concepts within the ontology, we define the following measures. All of the proposed measures rely on the measured cosine similarity between concept pairs using a vector space where a high-dimensional vector represents each concept. This representation is done by encoding the concepts using the domain-adapted BERT language model. Since the model is domain-adapted, the similarity of the concept vectors is derived from the similarity of their contextual environment in the document corpus. Thus, terms used in the same grammatical role in similar sentences will be similar in the vector space.

We now define three measures intended to be used to evaluate a single concept family (Hereafter CF, Definition 2.2). The first (CSS) represents an *accuracy* measure as it evaluates the correctness of the CF as constructed. The final two measure consistency, as they measure the extent to which the same relations (is-A) within a CF agree with each other.

(1) **Child Similarity Score** - CSS is the mean cosine similarity between every pair of siblings in a CF. We define this function as follows where M is the number of CF child concepts.

$$CSS(CF) = \frac{1}{M} \sum_{i=1, j=i+1}^{M-1} similarity(C_i, C_j) \tag{4}$$

(2) **Parent Similarity Score** - PSS is the mean cosine similarity between the parent and each of its direct child concepts.

$$PSS(CF) = \frac{1}{M} \sum_{i=1}^{M} similarity(C_i, C_p) \tag{5}$$

Where C_p is the parent concept and M is the number of child concepts.

(3) **Parent Difference Agreement** - PDA makes use of the standard deviation of the similarity between the parent concept and its direct children. We can interpret this value as the amount of agreement between the siblings towards the parent with respect to similarity. It is defined as:

$$PDA(CF) = 1 - \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} [similarity(C_i, C_p) - PSS(CF)]^2} \tag{6}$$

Using the defined measures, we iterate over all concept families within the ontology with two or more child concepts and compute the mean of CSS, PSS, and PDA. All of the values are within the range of [0-1]. Thus, having computed the measures, we determine the accuracy, completeness, conciseness, and semantic consistency using *CSS*, *domain coverage*, *ontology relevance*, and *PDA*, respectively.

5 EVALUATION

Here, we demonstrate our approach by performing an automated evaluation on three ontologies with respect to the oceanographic domain. We begin by describing the domain and candidate ontologies (Section 5.1), followed by the results and a comparison with previous work (Section 5.2). We then demonstrate how the measures can be used to improve an ontology (Section 5.3) and conclude with a discussion of the results (Section 5.4).

5.1 Domain and Candidate Ontologies

Following the previously described method (Fig. 1), we use a pre-existing corpus of 10,000 academic papers collected in the oceanographic domain (Section 4.1) and a NER model that was trained on it [4] (Section 4.2). Using the NER model and a general-purpose PoS tagger [36], we extract the domain concepts from the text corpus (Section 4.3). This phase generated 455,051 unique concepts. After applying additional constraints and filters such as frequency and term length, 17,516 concepts remained. We then pre-trained the BERT[2] [13] language model on the corpus (Section 4.5), resulting in an oceanography-domain BERT model.

[2]BERT Base, Transformers 4.17.0, https://huggingface.co/, accessed June 6th, 2022

Table (2) Automated evaluation results of three ontologies

Ontology	Original Size	Reduced Size	Ontology Relevance	Sub-Ontology Relevance	Domain Coverage	CSS Mean	PSS Mean	PDA Mean
ENVO	6,566	2,585	0.11	0.28	0.05	0.72	0.65	0.90
OMIT	87,816	5,379	0.01	0.26	0.10	0.71	0.69	0.92
SWEET	4,533	3,241	0.34	0.48	0.11	0.71	0.68	0.89

Table (1) Evaluated Ontologies

Ontology	Description	Concepts
ENVO	Ontology of environmental features and habitats	6,566
OMIT	Ontology to establish data exchange standards and common data elements in the microRNA (miR) domain	87,816
SWEET	Semantic Web for Earth and Environment Technology Ontology	4,533

Our candidate ontologies (Table 1) are ENVO [7], OMIT [26], and SWEET [42]. While both ENVO and SWEET are environmental ontologies, OMIT is considered a microRNA ontology. However, due to its relatively large size, it substantially overlaps the oceanographic domain. We match each candidate ontology to the set of domain concepts extracted from the document corpus (Section 4.4) allowing us to perform the evaluation (Section 4.6).

5.2 Evaluation Results

The results are presented in Table 2. In terms of relevance (conciseness measures, Eqs. 2 and 3) and domain coverage (Eq. 1), SWEET achieved the best results, with 34%, 48%, and 11% respectively. In terms of consistency, OMIT achieved the highest PDA score of 92% indicating a high level of similarity agreement among the children and the parent concepts (Eq. 6). Lastly, all ontologies received a CSS (Eq. 4) value between 71-72% indicating an average level of accuracy. The source code and datasets used are available online[3].

Table (3) Coverage and relevance comparison between our method and LSA [6]

		Ours	LSA
ENVO	Coverage	0.05	0.01
	Relevance	0.11	0.09
OMIT	Coverage	0.10	0.02
	Relevance	0.01	0.01
SWEET	Coverage	0.11	0.02
	Relevance	0.34	0.25

We now compare our method to the following recreation of Brewster et al. [6] (Table 3). The text corpus was fed into the LSA (Latent Semantic Analysis) algorithm with 20 clusters. From each cluster, the 15,000 most dominant words were fetched, resulting in a set of 33,754 unique words. Next, the WordNet expansion was applied in which two levels of hypernyms were fetched for each word. This expansion resulted in a larger set of 42,603 unique terms. From here, coverage and relevance (recall and precision in the original paper's terminology) were measured for each candidate ontology. Results show that LSA consistently assigns lower coverage and relevance figures than our method. We discuss this and the previous results in Section 5.4.

[3]https://github.com/Minitour/ontology-evaluation

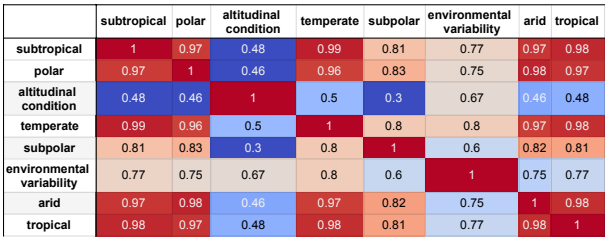

Figure (4) Similarity matrix of child concepts of *Environmental Condition* from ENVO

5.3 Improving an Ontology using CSS and PDA

Here, we demonstrate how one can utilize our method to improve an Ontology. CSS and PDA are defined (Eqs. 4, 6) for a single concept family (Def. 2.2). Thus, to gain better insight into the type of problems the model has identified or specific relations that may be inconsistent for future repair, one can use detailed similarity matrices (Fig. 4) for a family that received low scores. The matrix presents the cosine similarity between every pair of child concepts in the family. CSS is defined as the sum of this matrix.

Fig. 4 presents such a matrix of all child concepts in the concept family of *Environmental Condition*. As highlighted by the colors, most child concepts are highly similar (>0.97 cosine similarity, colored red). However, the concepts *environmental variability* and *altitudinal condition* received a relatively low similarity score. Indeed (in this domain), these two concepts have a different relationship with the parent concept.

To demonstrate how one can use these results to improve the ontology, we create an interim concept separating the set of concepts (*temperate, tropical, subtropical, subpolar, polar, arid*) from their original parent concept *Environmental Condition*. We consulted with domain experts who suggested a few possible candidates. Out of the proposed candidates, the *Climate Model* concept achieved the highest value of PDA when introduced into the concept family (Fig. 5). As can be seen in the figure, introducing the new concept markedly increases the CSS scores.

5.4 Discussion

Discussing the results obtained by our evaluation method over the different ontologies with domain experts yielded some interesting observations. The low overall relevance and domain coverage of OMIT, an mRNA ontology, was expected. However, the fact that we could extract a large and relevant sub-ontology from it using our

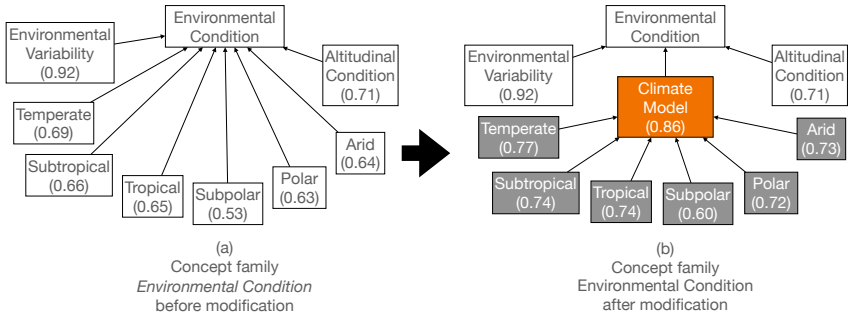

Figure (5) Manual refinement of the *Environmental Condition* concept family. A new intermediate concept, *Climate Model*, is selected from a set of expert-suggested replacements using its PDA score, grouping similar concepts together. The numerical values represent the similarity to the direct parent (CSS).

method can form the basis for an automated ontology construction method in the future that can obtain significant portions of partially relevant ontologies to piece together a comprehensive domain ontology. The fact that the SWEET ontology, which purports to cover the entire earth science domain (including oceanography), scored so low on coverage was surprising. It prompted us to perform a meta-evaluation of our method to ensure we were looking for actual domain concepts and not irrelevant concepts indiscriminately collected from the text. The results of this meta-evaluation are presented in the following section. When comparing to the current state of the art [6], we get a better coverage and relevance score. This was expected as the limitations of the LSA method cause it to miss phrasal concepts and many domain-specific concepts. We validate the assumption that, indeed, the method misses more of the domain concepts in the following section as well.

Introducing new intermediate concepts in an ontology is normally manual and time-consuming. However, as demonstrated here, using measures such as CSS and PDA, one can automate this process by finding the most suitable candidate concepts and testing which best maximizes the measures.

6 META-EVALUATION

Evaluating an evaluation method requires special care as it must be based upon sound assumptions of what is considered a *good result*. Here, we present a meta-evaluation that evaluates our proposed pipeline in two aspects. We begin by measuring the external agreement of our method with our intended target audience, oceanographic researchers. we then perform a statistical analysis to see how the different evaluation measures agree with each other and provide different perspectives on the candidate ontologies.

6.1 External Agreement - Coverage

To validate the coverage values obtained for the ontologies, we collect domain-specific concepts from two oceanographic researchers, one from the marine biology sub-domain and the other from computational oceanography. We received 43 concepts the experts had suggested upon reviewing their latest publications. We then compared these to the domain concepts collected as described in Section 4.3 and to the three ontologies - ENVO, OMIT, and SWEET. If, indeed, our evaluation method is sound, the results should reflect a high agreement between the domain concepts and the experts' concept list and coverage values close to those found for the candidate ontologies by our method (Table 2). We found 88% of the experts' concepts in the domain concepts that were extracted from the text by our method, which is as expected, representing a good domain coverage. ENVO, OMIT, and SWEET covered 23.2%, 13.9%, and 34.8% of our experts' concepts. The SWEET result is perfectly in line with our coverage score. The ENVO and OMIT results are higher, but this reflects an inherent bias in this meta-evaluation that over-represents marine biology concepts which are more prevalent in these two ontologies than in the domain at large. To validate our assumption that the lower LSA method scores in Table 3 are due to its poor coverage of the domain concepts, we tested it here as well and found it to be low, as expected, at 28%.

Table (4) Example of real and fake concept pairs.

Child Concept	Parent Concept	Real or Fake
organic acid	environmental material	Fake
natural plastic	organic molecule	Fake
leather dye	dye	Real
mesenchyme	tissue	Real

6.2 External Agreement - Accuracy and Consistency

Here we evaluate whether our CSS and PDA measures indeed measure the accuracy and consistency of the ontology. We compare the ruling of two domain experts over parent-child concept pairs to the effect on our measures of including these pairs in their concept family. For pairs that our experts believe have a parent-child relationship between them, we expect their inclusion in the same concept family to increase the CSS and PDA scores. The reverse is also true. We randomly sampled 300 concept pairs with a hierarchical (IS-A) relationship between them from the ENVO ontology alongside 300 auto-generated pairs that do not have a hierarchical relationship.A few examples are displayed in Table 4.

Of the 600 pairs, only 326 were used due to the lack of familiarity of the experts with the others. Out of the 326 labeled entries, 144 were labeled true, and the remaining 182 were labeled false. We measured a Kappa agreement score of 0.75 between the two experts over their overlapping pairs which can be interpreted as a substantial level of agreement. We iterate over each of the concept pairs and compute the following score before and after their inclusion in

a concept family.

$$score(CF) = CSS(CF) \times 0.9 + PDA(CF) \times 0.1 \quad (7)$$

In consultation with the domain experts, CSS was given a higher weight than PDA due to the nature of the task, which is to determine if a concept belongs to the concept family or not. This decision, according to our domain experts, is substantially more impacted by the similarity to existing *siblings* than by the extent of difference from the parent.

Table (5) Confusion Matrix Definition

	Labeled Positive	**Labeled Negative**
Predicted Positive	The relation is *labeled true* and including the concept *increases* the score **(True Positive)**	The relation is *labeled false* and including the concept *increases* the score **(False Positive)**
Predicted Negative	The relation is *labeled true* and including the concept *decreases* the score **(False Negative)**	The relation is *labeled false* and including the concept *decreases* the score **(True Negative)**

Table (6) Examples of concept pairs and how they were labeled by experts and predicted by the model

Child Concept	Parent Concept	Expert Label	Model Prediction	Classification
hill	mount	True	True	TP
leaf	isoprenoid	False	False	TN
organ	multicellular anatomical structure	True	False	FP
ore	cellular organisms	False	True	FN

The final results of the model evaluation are as follows. Of 326 concept pairs, 127 are true positives, 139 are true negatives, 43 are false positives, and 17 are false negatives. Some of the example concept pairs are presented in Table 6. Thus, the model achieved an accuracy of 81%, a precision of 74%, a recall of 88%, and an F1 score of 81% (F1 Score), which strengthens the claim that the model's ability to correctly identify inconsistencies and inaccuracies is on par with that of domain experts.

6.3 Statistical Analysis

Here, we perform a correlation analysis between our consistency measures (Eqs. 4, 5, 6) using Spearman's correlation (the measures are not normally distributed). The measures are defined over a concept family (Cf, Def. 2.2), and we wish to ensure that they measure different aspects of the CF. Results (Table 7) over the SWEET ontology (657 concept families) show a low to moderate correlation between the measures, confirming our assumption that they capture somewhat different aspects of the CF. An analysis of the other ontologies returned similar results that are omitted for brevity.

Table (7) Correlation analysis of consistency measures using Spearman's correlation coefficient.

SWEET N=657 concept families		**CSS**	**PSS**	**PDA**
	CSS	1	0.213	0.484
	PSS	0.213	1	0.145
	PDA	0.484	0.145	1

7 CONCLUSIONS AND FUTURE WORK

In this work, we showcase a novel approach for the automated evaluation of ontologies with respect to a domain. We do so by pre-training a bi-directional transformer-based language model in an unsupervised fashion on a text corpus from the domain. We define measures that make use of the language model to assess the accuracy and consistency of the ontology. Additionally, we use a NER model and PoS tagger to extract key concepts from the corpus, with which we create a concept set to evaluate the completeness and conciseness of an ontology. We validate the applicability of our approach by comparing the output of the model to that of domain experts. The results further strengthen the notion that language models such as BERT can adapt and encapsulate domain knowledge that can be utilized for a variety of tasks. Additionally, we showcase the potential applicability of our tools in both detecting a problem in the ontology and solving it. In this work, only hierarchical relations were considered due to the limitations of publicly available ontologies as well as computer-generated ontologies which are simple and lack other kinds of relations. However, the method can be expanded to work with other kinds of relationships as well as we intend to do in our future work.

ACKNOWLEDGMENTS

This work was partially supported by the Data Science Research Center at the University of Haifa through the Israel PBC grant *Advancing Data Science to Serve Humanity and Protect the Global Environment* (grant no. 100009443), the Danish Council for Independent Research (DFF) under grant agreement no. DFF-8048-00051B, and the Poul Due Jensen Foundation.

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1638–1649. https://aclanthology.org/C18-1139

[2] Mazen Alobaidi, Khalid Mahmood Malik, and Susan Sabra. 2018. Linked open data-based framework for automatic biomedical ontology generation. *BMC bioinformatics* 19, 1 (2018), 1–13.

[3] Claus Boye Asmussen and Charles Møller. 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data* 6, 1 (2019), 1–18.

[4] Koby Bar. 2020. *Oceanic NER Project*. University of Haifa. https://osf.io/my2nk/

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data-The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)* 5, 3 (2009), 1–22.

[6] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. 2004. Data Driven Ontology Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association, Lisbon, Portugal, 641–644. http://www.lrec-conf.org/proceedings/lrec2004/summaries/737.htm

[7] Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, and Suzanna E Lewis. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics* 4, 1 (2013), 1–9.

[8] Denis Eka Cahyani and Ito Wasito. 2017. Automatic ontology construction using text corpora and ontology design patterns (ODPs) in Alzheimer's disease. *Jurnal Ilmu Komputer dan Informasi* 10, 2 (2017), 59–66.

[9] Enrico G Caldarola and Antonio M Rinaldi. 2016. An approach to ontology integration for ontology reuse. In *2016 IEEE 17th international conference on information reuse and integration (IRI)*. IEEE, IEEE, Pittsburgh, PA, USA, 384–393.

[10] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. 2011. The MASTRO system for ontology-based data access. *Semantic Web* 2, 1 (2011), 43–53.

[11] Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, and Riccardo Rosati. 2018. *Using Ontologies for Semantic Data Integration*. Springer

International Publishing, Cham, 187–202. https://doi.org/10.1007/978-3-319-61893-7_11

[12] Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Sabrina Senatore. 2009. Towards an automatic fuzzy ontology generation. In *2009 IEEE International Conference on Fuzzy Systems*. IEEE, Jeju, Korea (South), 1044–1049.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://aclanthology.org/N19-1423

[14] Nicholas DiGiuseppe, Line C Pouchard, and Natalya F Noy. 2014. SWEET ontology coverage for earth system sciences. *Earth Science Informatics* 7, 4 (2014), 249–264.

[15] J Ding, D Berleant, D Nettleton, and E Wurtele. 2002. Mining MEDLINE: abstracts, sentences, or phrases?. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. World Scientific, Kauai, Hawaii, USA, 326–337.

[16] Kristina Doing-Harris, Yarden Livnat, and Stephane Meystre. 2015. Automated concept and relationship extraction for the semi-automated ontology management (SEAM) system. *Journal of biomedical semantics* 6, 1 (2015), 1–15.

[17] Damion M Dooley, Emma J Griffiths, Gurinder S Gosal, Pier L Buttigieg, Robert Hoehndorf, Matthew C Lange, Lynn M Schriml, Fiona SL Brinkman, and William WL Hsiao. 2018. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food* 2, 1 (2018), 1–10.

[18] Bouchra El Idrissi, Salah Baïna, and Karim Baïna. 2013. Automatic generation of ontology from data models: a practical evaluation of existing approaches. In *IEEE 7th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, Paris, France, 1–12.

[19] Gaihua Fu. 2016. FCA based ontology development for data integration. *Information processing & management* 52, 5 (2016), 765–782.

[20] Daniel Libonati Gomes and Thiago Henrique Bragato Barros. 2020. The bias in ontologies: An analysis of the foaf ontology. In *Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference*. Ergon-Verlag, Aalborg, Denmark, 236–244.

[21] Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 2 (1993), 199–220. https://doi.org/10.1006/knac.1993.1008

[22] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[23] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[24] Junheng Hao, Muhao Chen, Wenchao Yu, Yizhou Sun, and Wei Wang. 2019. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA). Association for Computing Machinery, New York, NY, USA, 1709–1719.

[25] Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*. ACL, Nantes, France, 539–545.

[26] Jingshan Huang, Jiangbo Dang, Glen M Borchert, Karen Eilbeck, He Zhang, Min Xiong, Weijian Jiang, Hao Wu, Judith A Blake, Darren A Natale, et al. 2014. OMIT: dynamic, semi-automated ontology development for the microRNA domain. *PLoS One* 9, 7 (2014), e100855.

[27] Madhura Jayaratne, Dinithi Nallaperuma, Daswin De Silva, Damminda Alahakoon, Brian Devitt, Kate E Webster, and Naveen Chilamkurti. 2019. A data integration platform for patient-centered e-healthcare and clinical decision support. *Future Generation Computer Systems* 92 (2019), 996–1008.

[28] Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen, and Amrapali Zaveri. 2014. Test-driven evaluation of linked data quality. In *Proceedings of the 23rd international conference on World Wide Web*. ACM, New York, NY, USA, 747–758.

[29] Naresh Kumar, Minakshi Kumar, and Manjeet Singh. 2016. Automated ontology generation from a plain text using statistical and NLP techniques. *International Journal of System Assurance Engineering and Management* 7, 1 (2016), 282–293.

[30] Juan J Lastra-Diaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana Garcia-Serrano, Mohamed Ben Aouicha, and Eneko Agirre. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence* 85 (2019), 645–665.

[31] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (2020), 50–70.

[32] Hao Liu, Yehoshua Perl, and James Geller. 2020. Concept placement using BERT trained by transforming and summarizing biomedical ontology structure. *Journal of Biomedical Informatics* 112 (2020), 103607.

[33] Diana Maynard, Adam Funk, and Wim Peters. 2009. Using lexico-syntactic ontology design patterns for ontology creation and population. In *Proc. of the Workshop on Ontology Patterns*. CEUR-WS.org, Aachen, DEU, 39–52.

[34] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[35] Giovanni Modica, Avigdor Gal, and Hasan M Jamil. 2001. The use of machine-generated ontologies in dynamic information seeking. In *International Conference on Cooperative Information Systems*. Springer-Verlag, Berlin Heidelberg, Germany, 433–447.

[36] Ines Montani, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, et al. 2023. *explosion/spaCy: v3.5.0: New CLI commands, language updates, bug fixes and much more*. Explosion. https://doi.org/10.5281/zenodo.7553910

[37] Kamran Munir and M Sheraz Anjum. 2018. The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics* 14, 2 (2018), 116–126.

[38] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. 2014. Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)* 10, 2 (2014), 7–34.

[39] Joe Raad and Christophe Cruz. 2015. A survey on ontology evaluation methods. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. SciTePress, Lisbonne, Portugal, 179–186.

[40] Frédéric Raimbault, Gildas Ménier, Pierre-François Marteau, et al. 2011. *On the detection of inconsistencies in RDF data sets and their correction at ontological level*. Technical Report. VALORIA - Laboratoire de Recherche en Informatique et ses Applications de Vannes et Lorient.

[41] Desi Ramayanti, Vina Ayumi, Handrie Noprisson, Anita Ratnasari, Inge Handriani, Marissa Utami, and Erwin Dwika Putra. 2020. Tuberculosis Ontology Generation and Enrichment Based Text Mining. In *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, Bandung, Indonesia, 429–434.

[42] R. G. Raskin. 2010. SWEET 2.1 Ontologies. In *AGU Fall Meeting Abstracts*, Vol. 2010. SAO/NASA Astrophysics Data System, Washington, DC, USA, Article IN44B-06, IN44B-06 pages.

[43] Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, and Shivakumar Vaithyanathan. 2008. An Algebraic Approach to Rule-Based Information Extraction. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008*, Gustavo Alonso, José A. Blakeley, and Arbee L. P. Chen (Eds.). IEEE Computer Society, Cancún, Mexico, 933–942. https://doi.org/10.1109/ICDE.2008.4497502

[44] Tomer Sagi, Yoav Lehahn, and Koby Bar. 2020. Artificial intelligence for ocean science data integration: current state, gaps, and way forward. *Elementa: Science of the Anthropocene* 8 (05 2020), 20 pages. https://doi.org/10.1525/elementa.418 arXiv:https://online.ucpress.edu/elementa/article-pdf/doi/10.1525/elementa.418/434891/418-7170-1-pb.pdf

[45] Stefan Schulz. 2018. The Role of Foundational Ontologies for Preventing Bad Ontology Design. In *Proceedings of the Joint Ontology Workshops 2018 Episode IV: The South African Spring co-located with the 10th International Conference on Formal Ontology in Information Systems (FOIS 2018), September 17-18, 2018 (CEUR Workshop Proceedings, Vol. 2205)*, Ludger Jansen, Daniele Paolo Radicioni, and Dagmar Gromann (Eds.). CEUR-WS.org, Cape Town, South Africa, 8 pages. http://ceur-ws.org/Vol-2205/paper22_bog1.pdf

[46] B Selvalakshmi and M Subramaniam. 2019. Intelligent ontology based semantic information retrieval using feature selection and classification. *Cluster Computing* 22, 5 (2019), 12871–12881.

[47] Quan Thanh Tho, Siu Cheung Hui, Alvis Cheuk M Fong, and Tru Hoang Cao. 2006. Automatic fuzzy ontology generation for semantic web. *IEEE transactions on knowledge and data engineering* 18, 6 (2006), 842–856.

[48] Amel Grissa Touzi, Hela Ben Massoud, and Alaya Ayadi. 2013. Automatic ontology generation for data mining using fca and clustering. *CoRR* abs/1311.1764 (2013), 10 pages. arXiv:1311.1764 http://arxiv.org/abs/1311.1764

[49] Guohui Xiao, Diego Calvanese, Roman Kontchakov, Domenico Lembo, Antonella Poggi, Riccardo Rosati, and Michael Zakharyaschev. 2018. Ontology-Based Data Access: A Survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018*, Jérôme Lang (Ed.). ijcai.org, Stockholm, Sweden, 5511–5519. https://doi.org/10.24963/ijcai.2018/777

[50] Binbin Yu. 2019. Research on information retrieval model based on ontology. *EURASIP Journal on Wireless Communications and Networking* 2019, 1 (2019), 1–8.

[51] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soeren Auer. 2016. Quality assessment for linked data: A survey. *Semantic Web* 7, 1 (2016), 63–93.

[52] Hansi Zhang, Yi Guo, Qian Li, Thomas J George, Elizabeth Shenkman, François Modave, and Jiang Bian. 2018. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC medical informatics and decision making* 18, 2 (2018), 129–147.

A SUPPLEMENTARY MATERIAL

In order to verify the applicability of the pre-trained model, we showcase a few examples of the fill-mask task, in which the model is given a sentence with one of the tokens being masked and the task is to fill it. The suggestions of both the original and pre-trained model are presented in Table 8.

Table 9 showcases a set of concepts and concept families with respect to the relevant metrics. The *Relevance* column presents concepts from the different ontologies where the concept does not belong to the domain of interest, whereas on the other hand, the *Coverage* column showcases concepts that did exist in our domain concept set but did not appear in the ontology.

Finally, Table 10 presents a comparison of expert-provided terms and their presence in our domain concepts dataset as well as the ontologies we examined.

Table (8) Comparison between BERT models before and after pre-training on the domain corpus on fill-mask task.

Sentence	Base BERT	Domain BERT
Upwelling water consists of [MASK].	groundwater, gravel	nutrients, diatoms
Arid, subpolar, and polar are all [MASK] models.	standard, dynamic	ocean, climate
phytoplankton use chlorophyll for [MASK].	growth, food, reproduction	photosynthesis, growth, carbon
A thermocline is the transition [MASK] between the warmer mixed water at the surface and the cooler deep water below.	state, point, metal	zone, layer, region
A coccolithophore is a unicellular, eukaryotic [MASK].	cell, organism	organism, phytoplankton

Table (9) Examples of different concepts with respect to different metrics

	Relevance	Coverage	CSS	PSS	PDA
	Example of concepts that are part of the ontology but have no relevance to the domain	Examples of concepts that are part of the domain but are not part of the ontology	Two Concept families with high and low child similarity score	Two Concept families with high and low parent similarity score	Two concept families with high and low parent difference agreement
ENVO	Sofa ENVO_01000588	Northern Hemisphere, Atlantic Ocean	Mining has a CSS score of 0.95 Sedimentary rock has a CSS score of 0.54	Particulate organic matter has a PSS of 0.95 Elevation has a PSS of 0.15	Coastal inlet has a PDA of 0.99 Liquid environmental material has a PDA of 0.68
OMIT	Paintings OMIT_0011154	Seafloor	Nucleic Acids has a CSS score of 0.98 Heterocyclic Compounds 1-Ring has a CSS score of 0.46	Chlorofluorocarbons has a PSS score of 0.94 Food has a PSS score of 0.25	Stramenopiles has a PDA of 0.99 Silicon Dioxide has a PDA of 0.75
SWEET	Civil aviation, management system, recrystallization	Resin	Ecosystem has a CSS of 0.98	Volcanic Activity has a PSS of 0.92	Bioprospecting has a PDA of 0.96

Table (10) The comparison of expert-provided concepts to extracted domain concepts and the three ontologies

Concept Name	Domain Concepts	ENVO	OMIT	SWEET
Lagrangian analysis	analysis, reanalysis, microanalysis, lagrangian	-	-	lagrangian
biogeography	biogeography	-	-	-
jellyfish	jellyfish	-	-	jellyfish
coccolithophores	coccolithophores	-	-	-
slicks	slicks	-	-	-
sea surface microlayer	sea surface, sea surface waters, microlayer	oceanic sea surface microlayer biome	-	-
mixed layer depth	mixed layer depth	-	-	-
stratification	stratification	stratification	-	-
water masses	water masses	-	-	-
ocean circulation	ocean circulation	-	-	ocean circulation
mesoscale	mesoscale	mesoscale marine eddy, marine mesoscale eddy field	-	mesoscale wind, mesoscale disturbance, mesoscale cellular convection, mesoscale convective complex, mesoscale eddy
thermohaline circulation	thermohaline circulation	-	-	thermohaline circulation
Ekman transport	ekman transport	-	-	-
upwelling	upwelling	upwelling	-	upwelling
sea surface height	sea surface height	-	-	-
geostrophic currents	currents	-	-	-
nitrate	nitrate	nitrate	-	-
thermocline	thermocline	thermocline	-	thermocline
altimetry	altimetry	-	-	altimetry
ocean colour	ocean, colour	-	-	-
carbon cycle	global carbon cycle	carbon cycle	Carbon Cycle	carbon cycle
phytoplankton	phytoplankton	-	Phytoplankton	phytoplankton
diatoms	diatoms	diatoms	Diatoms	-
sediment traps	sediment traps	-	-	-
transparent exopolymer particles	transparent exopolymer particles	-	-	-
chlorophyll	chlorophyll	chlorophyll	Chlorophyll	chlorophyll
remote sensing	remote sensing	-	Remote Sensing Technology	remote sensing
Phycosphere	-	-	-	-
Heterotrophic bacteria	heterotrophic bacteria	-	-	-
SAR11	SAR	-	-	-
Copiotroph	-	-	-	-
Succession	succession	-	-	succession
Phytoplankton bloom	phytoplankton bloom	-	-	-
Carbon use efficiency	carbon fluxes, carbon cycling, carbon biomass	-	-	-
Sinking flux	sinking, sinking rates	-	-	-
Choanoflagellate	-	-	-	-
Redfield ratio	redfield ratio	-	-	-
Biomass objective function	biomass values, biomass productivity, biomass accumulation	-	-	-
Monod equation	-	-	-	-
Plume	plume	plume	-	plume
Diffusion boundary layer	diffusion, layer, boundary	-	-	-
Eddy	eddy	-	-	eddy
Quorum sensing	-	-	Quorum Sensing	-
TOTAL: 43	**38**	**10**	**6**	**15**
Coverage	**88.37%**	**23.25%**	**13.95%**	**34.88%**