COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Li, H, et al. 2019. A Method for Extending Ontologies with Application to the Materials Science Domain. *Data Science Journal,* 18: 50, pp. 1–21. DOI: https://doi.org/10.5334/dsj-2019-050

RESEARCH PAPER

A Method for Extending Ontologies with Application to the Materials Science Domain

Huanyu Li[1,3], Rickard Armiento[2,3] and Patrick Lambrix[1,3]

[1] Department of Computer and Information Science, Linköping University, Linköping, SE

[2] Department of Physics, Chemistry and Biology, Linköping University, Linköping, SE

[3] The Swedish e-Science Research Centre, Linköping University, Linköping, SE

Corresponding author: Patrick Lambrix (patrick.lambrix@liu.se)

In the materials science domain the data-driven science paradigm has become the focus since the beginning of the 2000s. A large number of research groups and communities are building and developing data-driven workflows. However, much of the data and knowledge is stored in different heterogeneous data sources maintained by different groups. This leads to a reduced availability of the data and poor interoperability between systems in this domain. Ontology-based techniques are an important way to reduce these problems and a number of efforts have started. In this paper we investigate efforts in the materials science, and in particular in the nanotechnology domain, and show how such ontologies developed by domain experts, can be improved. We use a phrase-based topic model approach and formal topical concept analysis on unstructured text in this domain to suggest additional concepts and axioms for the ontology that should be validated by a domain expert. We describe the techniques and show the usefulness of the approach through an experiment where we extend two nanotechnology ontologies using approximately 600 titles and abstracts.

Keywords: Ontology; Materials Science; Nanotechnology; Nanoparticle; Topic Model; Formal Topical Concept Analysis

1 Introduction

From the beginning of the 2000s materials science has shifted towards its fourth paradigm, (big) data-driven science (Agrawal & Choudhary 2016). More and more researchers in materials science have realized that data-driven techniques could accelerate the discovery and design of materials. Therefore, a large number of research groups and communities have developed data-driven workflows including data repositories (for an overview see (Lambrix et al. 2019)) and data analytics tools for particular purposes. As data-driven techniques become widely used, big data challenges regarding volume, variety, variability and veracity (Lambrix et al. 2019) and challenges in reproducing, sharing, and integrating data (Kalidindi & De Graef 2015, Agrawal & Choudhary 2016, Tropsha et al. 2017, Karcher et al. 2018, Rumble et al. 2019) are growing at the same time.

These challenges also occurred in other fields. For instance, in (Lambrix 2005) the problems of locating, retrieving and integrating data in the biomedical field were addressed. These problems relate to the more recently introduced FAIR principles that aim to support machines to automatically find and use data, and individuals to reuse the data (Wilkinson et al. 2016). The FAIR principles state that data should be Findable, Accessible, Interoperable, and Reusable, respectively. In different areas research is on the way to conform data management to these principles, including in the materials science domain (Draxl & Scheffler 2018). One of the recognized enablers for the principles are ontologies and ontology-based techniques. Ontologies provide a shared standardized representation of knowledge of a domain. By describing data using ontologies, the data will be more findable. By using ontologies for representing the metadata, the level of accessibility can be raised. By using the same terminology as defined by ontologies, interoperability is enabled. Finally, as ontologies are shared and standardized, reusability is supported.

Taking nanotechnology as an example, in (Tropsha et al. 2017) it is stated that there exists a gap between data generation and shared data access. The domain lacks standards for collecting and systematically representing nanomaterial properties. In (Karcher et al. 2018) stakeholder-identified technical and operational challenges for the integration of data in the nanotechnology domain are presented. The technical challenges mainly refer to (i) the use of different data formats, (ii) the use of different vocabularies, (iii) the lack of unique identifiers, and (iv) the use of different data conceptualization methods. In terms of operational challenges, they refer to (i) the fact that organizations have different levels of data quality and completeness, and (ii) the lack of understandable documentation. To solve these challenges, it is proposed that ontologies and ontology-based techniques can play a significant role in the data-driven materials science and enable reproduction, sharing and integration of data. This was, for instance, the main outcome of a workshop on interoperability in materials modelling organized by the European Materials Modelling Council (European Materials Modelling Council 2017).

Although in its infancy, some organizations and research groups have started to develop ontologies and standards for the materials domain (Section 2.2), including in the nanotechnology domain. However, developing ontologies is not an easy task and often the resulting ontologies are not complete. In addition to being problematic for the correct modelling of a domain, such incomplete ontologies also influence the quality of semantically-enabled applications such as ontology-based search and data integration. Incomplete ontologies when used in semantically-enabled applications can lead to valid conclusions being missed. For instance, in ontology-based search, queries are refined and expanded by moving up and down the hierarchy of concepts. Incomplete structure in ontologies influences the quality of the search results. In experiments in the biomedical field, an example was given where a search in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), a large database with abstracts of research articles in the biomedical field, using the MeSH (Medical Subject Headings) (http://www.nlm.nih.gov/mesh/) ontology would miss 55% of the documents if the relation between the concepts *Scleral Disease* and *Scleritis* is missing (Liu & Lambrix 2010).

In this paper, we present a novel method for extending existing ontologies by detecting new concepts and relations in the concept hierarchy that should be included in the ontologies. We do this by presenting a new approach, formal topical concept analysis, that integrates a variant of topic modeling and formal concept analysis. Further, we apply our method to two ontologies (NanoParticle Ontology and eNanoMapper) in the materials science domain. The choice of the use of ontologies in the nanotechnology domain is motivated by the fact that, as we have shown before, there is an awareness of the need for ontologies to deal with interoperability and reusability issues. Further, there are not so many ontologies in materials science yet (see Section 2.2) and the chosen ontologies are among the more mature ontologies in the field. Therefore, they represent the most difficult case for extending ontologies.

The remainder of the paper is organized as follows. In Section 2 we describe what ontologies are, efforts on ontologies in the materials domain as well as work on extending ontologies. Section 3 describes our approach while Section 4 shows and discusses the results of the application of our approach in the nanotechnology domain. We show how NanoParticle Ontology and eNanoMapper were extended and evaluate the usefulness of the approach. We also compare our results to the results of an experiment with another popular system on the same data. Finally, the paper concludes in Section 5.

2 Background

2.1 Ontologies

Intuitively, ontologies can be seen as defining the basic terms and relations of a domain of interest, as well as the rules for combining these terms and relations. Ontologies are used for communication between people and organizations by providing a common terminology over a domain. They provide the basis for interoperability between systems, and can be used as an index to a repository of information as well as a query model and a navigation model for data sources. They are often used as a basis for integration of data sources, thereby alleviating the variety and variability problems. The benefits of using ontologies include reuse, sharing and portability of knowledge across platforms, and improved maintainability, documentation, maintenance, and reliability. Overall, ontologies lead to a better understanding of a field and to more effective and efficient handling of information in that field (e.g., (Stevens et al. 2000).

From a knowledge representation point of view, ontologies may contain four components: (i) concepts that represent sets or classes of entities in a domain, (ii) instances that represent the actual entities, (iii) relations, and (iv) axioms that represent facts that are always true in the topic area of the ontology. Axioms can represent such things as domain restrictions, cardinality restrictions, or disjointness restrictions. Ontologies can be classified according to which components and the information regarding the components they contain. As an example, **Figure 1** represents a small piece of the NanoParticle Ontology (Thomas et al. 2011)

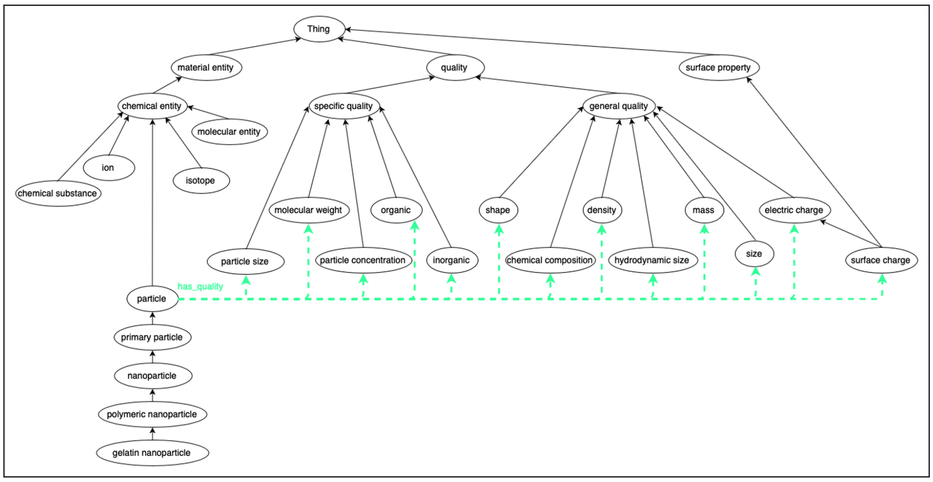
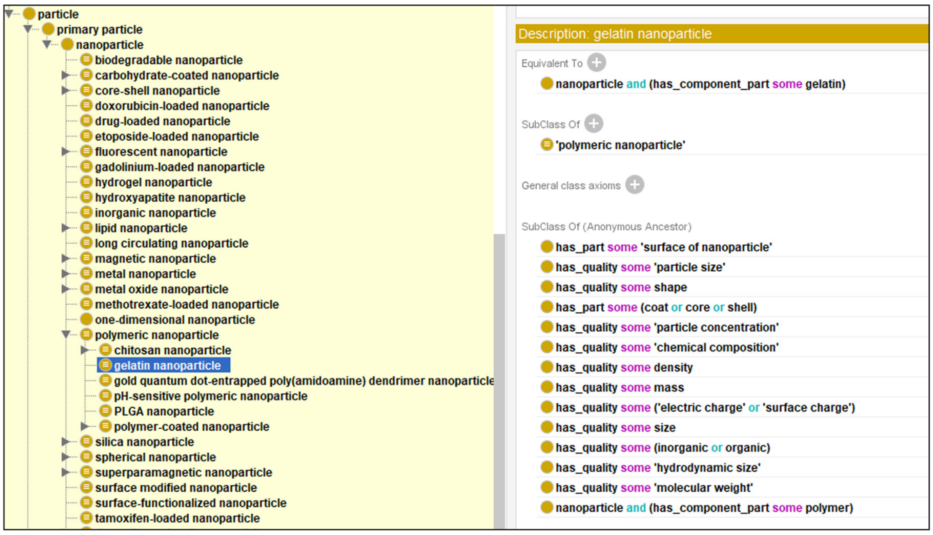


Figure 1: Example from NanoParticle Ontology.



particle
primary particle
nanoparticle
biodegradable nanoparticle
carbohydrate-coated nanoparticle
core-shell nanoparticle
doxorubicin-loaded nanoparticle
drug-loaded nanoparticle
etoposide-loaded nanoparticle
fluorescent nanoparticle
gadolinium-loaded nanoparticle
hydrogel nanoparticle
hydroxyapatite nanoparticle
inorganic nanoparticle
lipid nanoparticle
long circulating nanoparticle
magnetic nanoparticle
metal nanoparticle
metal oxide nanoparticle
methotrexate-loaded nanoparticle
one-dimensional nanoparticle
polymeric nanoparticle
chitosan nanoparticle
gelatin nanoparticle
gold quantum dot-entrapped poly(amidoamine) dendrimer nanoparticle
pH-sensitive polymeric nanoparticle
PLGA nanoparticle
polymer-coated nanoparticle
silica nanoparticle
spherical nanoparticle
superparamagnetic nanoparticle
surface modified nanoparticle
surface-functionalized nanoparticle
tamoxifen-loaded nanoparticle

Description: gelatin nanoparticle

Equivalent To
nanoparticle and (has_component_part some gelatin)

SubClass Of
'polymeric nanoparticle'

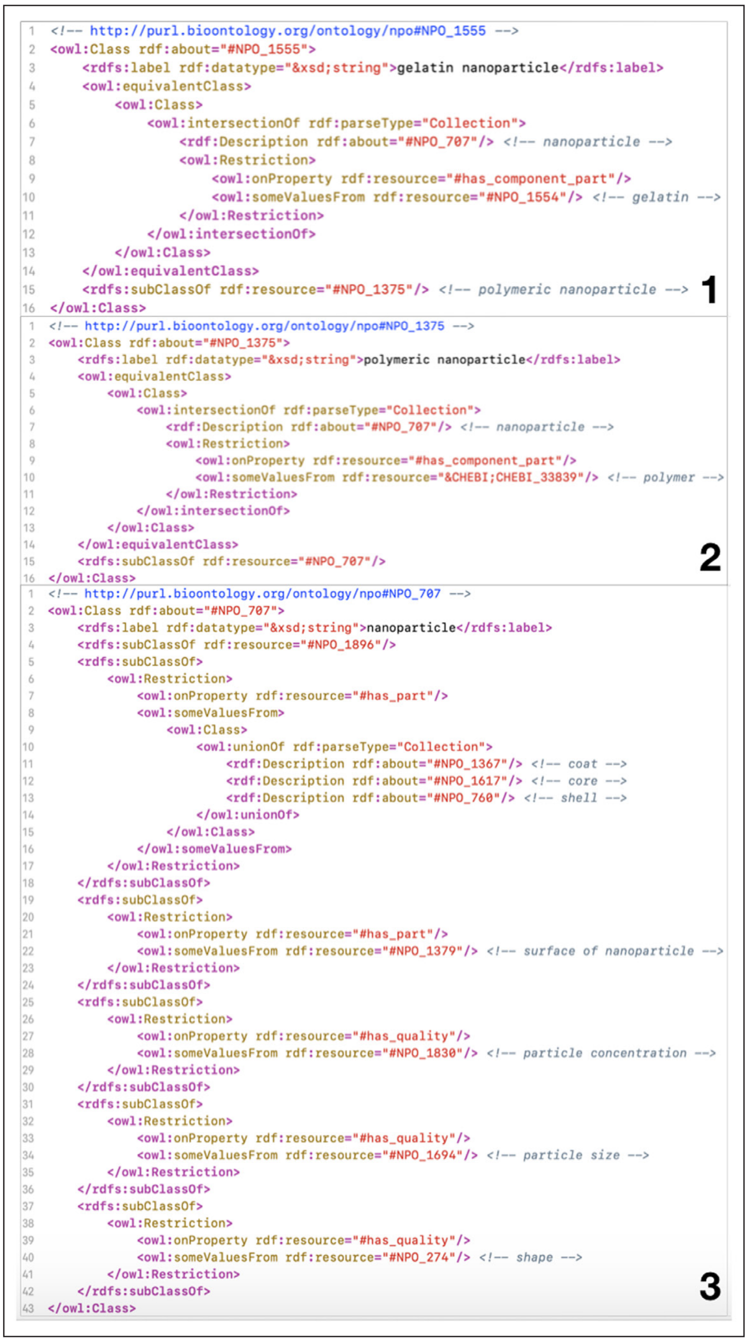
General class axioms

SubClass Of (Anonymous Ancestor)
has_part some 'surface of nanoparticle'
has_quality some 'particle size'
has_quality some shape
has_part some (coat or core or shell)
has_quality some 'particle concentration'
has_quality some 'chemical composition'
has_quality some density
has_quality some mass
has_quality some ('electric charge' or 'surface charge')
has_quality some size
has_quality some (inorganic or organic)
has_quality some 'hydrodynamic size'
has_quality some 'molecular weight'
nanoparticle and (has_component_part some polymer)

Figure 2: Example from NanoParticle Ontology opened in Protégé.

regarding 'chemical entity' and 'quality'. Regarding chemical entities NanoParticle Ontology contains, for instance, the concepts *chemical entity, chemical substance, ion, particle, isotope* and *molecular entity*. The black full arrows represent axioms representing is-a relations, i.e. if A is a B, then all entities that belong to concept A also belong to concept B. We also say then that A is a sub-concept of B. In this example we have that *chemical substance, particle, ion, isotope* and *molecular entity* are sub-concepts of *chemical entity*. Therefore, all chemical substances, particles, ions, isotopes, and molecular entities are also chemical entities. Further, all primary particles are particles, all nanoparticles are primary particles, all polymeric nanoparticles are nanoparticles and all gelatin nanoparticles are polymeric nanoparticles. The is-a relation is transitive such that, for instance, a gelatin nanoparticle is also a particle. Regarding different kinds of qualities NanoParticle Ontology contains, for instance, the concepts *particle size, molecular weight, particle concentration, organic, inorganic, shape, chemical composition, density, hydrodynamic size, mass, size*, and *electric charge*. Further, particles have qualities; this is represented by an axiom that states that concepts *particle* and *quality* are connected to each other by the relation *has quality* (green dashed arrows in **Figure 1**). Properties represented by relations are inherited via the is-a hierarchy. Therefore, also the subconcepts of particles are related to qualities.

In **Figure 2** we show the part of NanoParticle Ontology that represents particles using the ontology development system Protégé (https://protege.stanford.edu/). On the left hand side the concepts and the is-a hierarchy are shown. The is-a relations are represented by indentation. For instance, *gelatin nanoparticle* (highlighted in **Figure 2**) is a sub-concept of *polymeric nanoparticle* which in its turn is a sub-concept of *nanoparticle*. On the right-hand side of **Figure 2** information related to the axioms are shown using a special notation reflecting constructs in the representation language OWL (http://www.w3.org/TR/owl-features/, http://www.w3.org/TR/owl2-overview/), a knowledge representation language that is often used for representing ontologies and that is based on description logics (Baader et al. 2010). For instance, we note that the concept *gelatin nanoparticle* was defined to be equivalent to *nanoparticle* **and** (*has_component_part* **some** *gelatin*). This means that every gelatin nanoparticle is a nanoparticle that has a component that is gelatin, and vice versa, whenever a nanoparticle has a component that is gelatin, then it is a gelatin nanoparticle. Further, there is information about the types of qualities that gelatin nanoparticles have (inherited from the *particle* concept). An advantage of using a description logics-based representation is that it allows for reasoning. In the ontology it was defined that *gelatin nanoparticle* is equivalent to *nanoparticle* **and** (*has_component_part* **some** *gelatin*) (as we just noted), that *polymeric nanoparticle* is equivalent to *nanoparticle* **and** (*has_component_part* **some** *polymer*), and that *gelatin* is a subconcept of *protein* which is a subconcept of *biopolymer* which is in its turn a subconcept of *polymer*. Based on these axioms the system can derive the additional information that a *gelatin nanoparticle* is a *polymeric nanoparticle*, which is also shown on the right-hand side of **Figure 2** (under 'SubClass Of'). **Figure 3** shows the actual OWL representation for the concepts *gelatin nanoparticle*, *polymeric nanoparticle* and *nanoparticle*.



```
<!-- http://purl.bioontology.org/ontology/npo#NPO_1555 -->
<owl:Class rdf:about="#NPO_1555">
    <rdfs:label rdf:datatype="&xsd;string">gelatin nanoparticle</rdfs:label>
    <owl:equivalentClass>
        <owl:Class>
            <owl:intersectionOf rdf:parseType="Collection">
                <rdf:Description rdf:about="#NPO_707"/> <!-- nanoparticle -->
                <owl:Restriction>
                    <owl:onProperty rdf:resource="#has_component_part"/>
                    <owl:someValuesFrom rdf:resource="#NPO_1554"/> <!-- gelatin -->
                </owl:Restriction>
            </owl:intersectionOf>
        </owl:Class>
    </owl:equivalentClass>
    <rdfs:subClassOf rdf:resource="#NPO_1375"/> <!-- polymeric nanoparticle -->
</owl:Class>
```

```
<!-- http://purl.bioontology.org/ontology/npo#NPO_1375 -->
<owl:Class rdf:about="#NPO_1375">
    <rdfs:label rdf:datatype="&xsd;string">polymeric nanoparticle</rdfs:label>
    <owl:equivalentClass>
        <owl:Class>
            <owl:intersectionOf rdf:parseType="Collection">
                <rdf:Description rdf:about="#NPO_707"/> <!-- nanoparticle -->
                <owl:Restriction>
                    <owl:onProperty rdf:resource="#has_component_part"/>
                    <owl:someValuesFrom rdf:resource="&CHEBI;CHEBI_33839"/> <!-- polymer -->
                </owl:Restriction>
            </owl:intersectionOf>
        </owl:Class>
    </owl:equivalentClass>
    <rdfs:subClassOf rdf:resource="#NPO_707"/>
</owl:Class>
```

```
<!-- http://purl.bioontology.org/ontology/npo#NPO_707 -->
<owl:Class rdf:about="#NPO_707">
    <rdfs:label rdf:datatype="&xsd;string">nanoparticle</rdfs:label>
    <rdfs:subClassOf rdf:resource="#NPO_1896"/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#has_part"/>
            <owl:someValuesFrom>
                <owl:Class>
                    <owl:unionOf rdf:parseType="Collection">
                        <rdf:Description rdf:about="#NPO_1367"/> <!-- coat -->
                        <rdf:Description rdf:about="#NPO_1617"/> <!-- core -->
                        <rdf:Description rdf:about="#NPO_760"/> <!-- shell -->
                    </owl:unionOf>
                </owl:Class>
            </owl:someValuesFrom>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#has_part"/>
            <owl:someValuesFrom rdf:resource="#NPO_1379"/> <!-- surface of nanoparticle -->
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#has_quality"/>
            <owl:someValuesFrom rdf:resource="#NPO_1830"/> <!-- particle concentration -->
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#has_quality"/>
            <owl:someValuesFrom rdf:resource="#NPO_1694"/> <!-- particle size -->
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#has_quality"/>
            <owl:someValuesFrom rdf:resource="#NPO_274"/> <!-- shape -->
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

Figure 3: Example from NanoParticle Ontology – OWL/XML Syntax Format.

2.2 Ontologies in materials domain

Within the materials domain the use of semantic technologies is in its infancy with the development of ontologies and standards. According to (Zhang, Zhao & Wang 2015) domain ontologies have been used to organize materials knowledge in a formal language, as a global conceptualization for materials information integration (e.g. (Cheng et al. 2014)), for linked materials data publishing, for inference support for discovering new materials and for semantic query support (e.g., (Zhang, Luo, Zhao & Zhang 2015, Zhang et al. 2017)). Most ontologies focus on specific sub-domains of the materials field (e.g., metals, ceramics, thermal properties, nanotechnology) and have been developed with a specific use in mind (e.g., search, data integration, discovery). Some examples of ontologies are the Materials Ontology (Ashino 2010) for data exchange among thermal property databases, PREMΛP ontology (Bhat et al. 2013) for steel mill products, MatOnto ontology (Cheung et al. 2008) for oxygen ion conducting materials in the fuel cell domain, and the FreeClassOWL ontology (Radinger et al. 2013) for the construction and building materials domain. An ontology design pattern regarding material transformations was proposed in (Vardeman II et al. 2017). Since recently, the European Materials Modelling Council is developing the European Materials Modelling Ontology (European Materials Modelling Council 2017).

In the sub-field of nanotechnology, the NanoParticle Ontology (Thomas et al. 2011) was created for understanding biological properties of nanomaterials, searching for nanoparticle relevant data and designing nanoparticles. It builds on the Basic Formal Ontology (BFO, http://basic-formal-ontology.org/) (Arp et al. 2015) and Chemical Entities of Biological Interest Ontology (ChEBI) (de Matos et al. 2010) to represent basic knowledge regarding physical, chemical and functional features of nanotechnology used in cancer diagnosis and therapy. The eNanoMapper ontology (Hastings et al. 2015) aims to integrate a number of ontologies such as the NanoParticle Ontology for assessing risks related to the use of nanomaterials.

Furthermore, standards for exporting data from databases and between tools are being developed. These standards provide a way to exchange data between databases and tools, even if the internal representations of the data in the databases and tools are different. They are a prerequisite for efficient materials data infrastructures that allow for the discovery of new materials (Austin 2016).

In several cases the standards formalize the description of materials knowledge and thereby create ontological knowledge. For instance, one effort is by the European Committee for Standardization which organized workshops on standards for materials engineering data of which the results are documented in (European Committee for Standardization 2010). Another recent effort is connected to the European Centre of Excellence NOMAD (Ghiringhelli et al. 2016).

2.3 Extending ontologies from unstructured text

The ontology extension problem that we tackle deals mainly with concept discovery and concept hierarchy derivation. These are also two of the tasks in the problem of ontology learning (Buitelaar et al. 2005). Therefore, most of the related work comes from that area. For instance, a recent survey (Asim et al. 2018) discusses 140 research papers. Different techniques can be used for concept and relationship extraction. In this setting, new ontology elements are derived from text using knowledge acquisition techniques.

Linguistic techniques use part-of-speech tagged corpora for extracting syntactic structures that are analyzed regarding the words and the modifiers contained in the structure. One kind of linguistic approach is based on linguistics using lexico-syntactic patterns. The pioneering research conducted in this line is in (Hearst 1992), which defines a set of patterns indicating is-a relationships between words in the text. Other linguistic approaches may make use of, for instance, compounding, the use of background and itemization, term co-occurrence analysis or superstring prediction (e.g. (Wächter et al. 2006, Arnold & Rahm 2013)).

Another paradigm is based on machine learning and statistical methods which use the statistics of the underlying corpora, such as k-nearest neighbors approach (Maedche et al. 2003), association rules (Maedche & Staab 2000), bottom-up hierarchical clustering techniques (Zavitsanos et al. 2007), supervised classification (Spiliopoulos et al. 2010) and formal concept analysis (Cimiano et al. 2005). There are also some approaches that use topic models (Schaal et al. 2005, Lin et al. 2012, Rani et al. 2017) but they focus on concept names that are words, rather than phrases as in our approach.

Ontology evolution approaches (Hartung et al. 2011, Dos Reis et al. 2013) allow for the study of changes in ontologies and using the change management mechanisms to detect candidate missing relations. An approach that allows for detection and user-guided completion of the is-a structure is given in (Ivanova & Lambrix 2013, Lambrix et al. 2015) where completion is formalized as an abduction problem and the RepOSE tool is presented.

3 Approach

Our approach for extending ontologies, shown in **Figure 4**, contains the following steps. In the first step, *creation of a phrase-based topic model*, documents related to the domain of interest are used to create topics. The phrases as well as the topics are suggestions that a domain expert should validate or interpret and relate to concepts in the ontology. In the second step the (possibly validated and updated) topics are used in a *formal topical concept analysis* which returns suggestions to the domain expert regarding relations between topics and thus concepts in the ontology. Both steps lead to the addition of new concepts and (subsumption) axioms to the ontology. In the following subsections we describe these steps.

3.1 Phrase-based Topic Model

In our first step we use the phrases-based topic model in the ToPMine system (El-Kishky et al. 2014). Given a corpus of documents and the number of requested topics, representations of latent topics in the documents are computed. Essentially, topics can be seen as a probability distribution over words or phrases. The ToPMine approach is purely data-driven, i.e., it does not require domain knowledge or specific linguistic rule sets. This is important for our application domain as there is a lack of annotated background knowledge. An important property of the system is that it works on bags-of-phrases, rather than the traditional bag-of-words. This means that words occurring closer together have more weight than words far away. Further, as we assume existing ontologies, it is very likely that concepts with one-word names are already in the ontology and we therefore focus on phrases.

The approach consists of two parts: phrase mining and topic modelling. In the first part frequent contiguous phrases are mined, which consists of collecting aggregate counts for all contiguous words satisfying a minimum support threshold. Then the documents are segmented based on the frequent phrases. Further, an agglomerative phrase construction algorithm merges the frequent phrases guided by a significance score. In the second part topics are generated using a variant of Latent Dirichlet Allocation, called PhraseLDA, that deals with phrases, rather than words.

3.2 Formal Topical Concept Analysis

In the second step we define a new variant of Formal Concept Analysis (e.g., (Ganter & Wille 2012)) and use this new variant on topics. These topics can come directly from the previous step or can be a modified version of the topics of the previous step, where non-relevant topics or phrases are removed.

We first define the notions of formal topical context, formal topical concept and topical concept lattice. (Note that formal topical concepts should not be confused with concepts in the ontologies.)

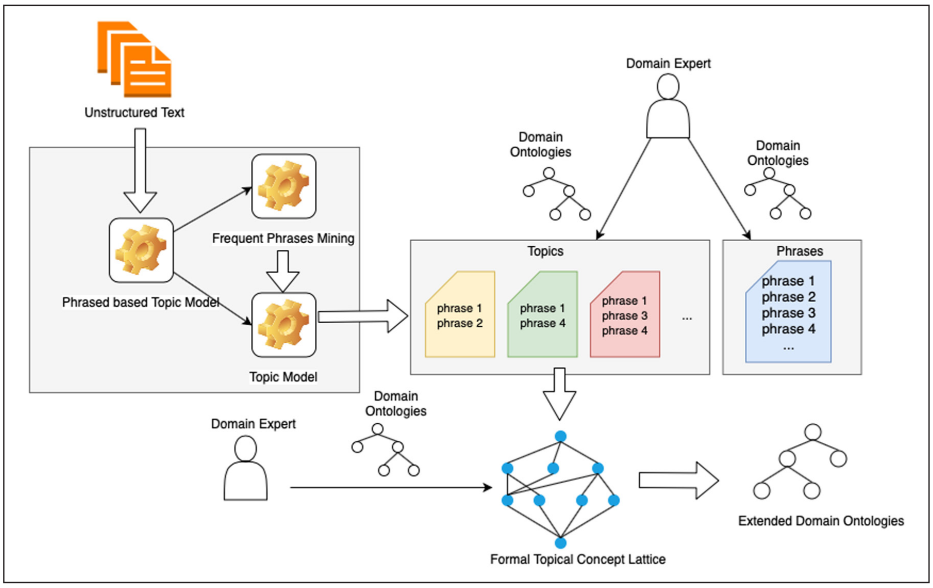


Figure 4: Approach: The upper part of the Figure shows the creation of a phrase-based topic model with as input unstructured text and as output phrases and topics. The lower part shows the formal topical concept analysis with as input topics and as output a topical concept lattice. In both parts a domain expert validates and interprets the results.

Definition 1. (Formal Topical Context) *A formal topical context is a triple (P, T, I) where P is a set phrases, T is a set topics, and I is a binary relation between P and T ($I \subseteq P \times T$).*

Definition 2. (Formal Topical Concept) (*A*, *B*) *is a formal topical concept of (P, T, I) iff* $A \subseteq P$, $B \subseteq T$, $A' = B$, $B' = A$ *where* $A' := \{t \in T \mid \forall p \in A : <p, t> \in I\}$ *and* $B' := \{p \in P \mid \forall t \in B : <p, t> \in I\}$. *A is the extent and B is the intent of* (*A*, *B*).

Definition 3. (Topical Concept Lattice) *Topical formal concepts can be ordered. We say that* $(A_1, B_1) \leq (A_2, B_2)$ *iff* $A_1 \subseteq A_2$. *The set* $\Phi(P, T, I)$ *of all formal topical concepts of (P, T, I), with this order, is called the topical concept lattice of (P, T, I).*

As an example, in **Figure 5(a)** we show a matrix representing the occurrence of phrases in topics in a topic model, the resulting formal topical concepts in **Figure 5(c)** and the topical concept lattice in **Figure 5(b)**. In the lattice a node represents a formal topical concept (same numbering as in **Figure 5(a)**). For a formal topical concept (*A*, *B*), its extent (phrases) is found by collecting all phrases in its node as well as its descendants. The intent (topics) is found by collecting all topics in its node as well as its ancestors.

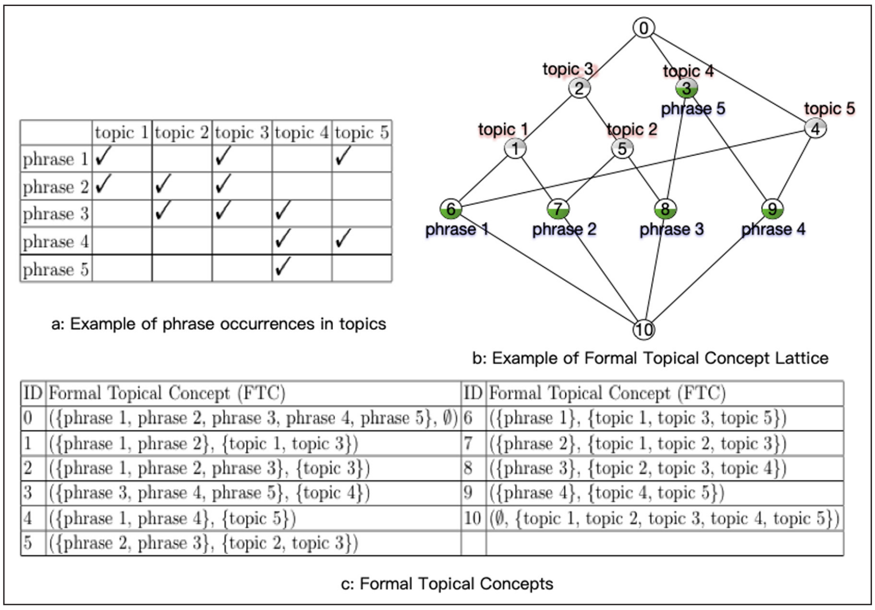
3.3 Domain Expert Validation

As shown in **Figure 4**, a domain expert is involved in the different steps in our approach to validate and interpret the results of the phrase-based topic model and the formal topical concept analysis.

The domain expert validates or **interprets all phrases** that appear in all topics. The outcome can be one of the following:

(i) The phrase is a meaningful representation of a concept in the specific domain and it is already in the ontology. For example, *gold nanoparticle* is a specific concept within the nanotechnology domain and it is already in the NanoParticle Ontology. We distinguish two cases: (1) a concept with the same name or a name that is a synonym of the original form of the phrase already exists in the ontology (EXIST) or (2) a concept with a name that is a modified form of the phrase already exists in the ontology (EXIST-m).

(ii) The phrase is a meaningful representation of a concept in the specific domain but it is not in the ontology. For example, *microcrystalline silicon* is a meaningful representation of a concept but such concept does not exist in the ontology. We distinguish two cases: (1) a concept with the same name as the original form of the phrase should be added into the ontology (ADD) or (2) a concept with as name a modified form of the phrase should be added into the ontology (ADD-m).



	topic 1	topic 2	topic 3	topic 4	topic 5
phrase 1	✓		✓		✓
phrase 2	✓	✓	✓		
phrase 3		✓	✓	✓	
phrase 4				✓	✓
phrase 5				✓	

a: Example of phrase occurrences in topics

b: Example of Formal Topical Concept Lattice

ID	Formal Topical Concept (FTC)	ID	Formal Topical Concept (FTC)
0	({phrase 1, phrase 2, phrase 3, phrase 4, phrase 5}, ∅)	6	({phrase 1}, {topic 1, topic 3, topic 5})
1	({phrase 1, phrase 2}, {topic 1, topic 3})	7	({phrase 2}, {topic 1, topic 2, topic 3})
2	({phrase 1, phrase 2, phrase 3}, {topic 3})	8	({phrase 3}, {topic 2, topic 3, topic 4})
3	({phrase 3, phrase 4, phrase 5}, {topic 4})	9	({phrase 4}, {topic 4, topic 5})
4	({phrase 1, phrase 4}, {topic 5})	10	(∅, {topic 1, topic 2, topic 3, topic 4, topic 5})
5	({phrase 2, phrase 3}, {topic 2, topic 3})		

c: Formal Topical Concepts

Figure 5: Examples of **(a)** phrase occurrences in topics, **(b)** Formal Topical Concept Lattice and **(c)** Formal Topical Concepts.

(iii) No concept related to the phrase should be added to the ontology. This can happen because the phrase does not make sense in the domain (No), but also because it is a meaningful representation of a concept in a more general domain (No-g). For example, *electron transfer* is a general concept within the perspective in materials science, but should not necessarily be in a nanotechnology ontology.

A second interaction with the domain expert occurs in the **interpretation of topics**. The outcome can be one of the following:

(i) Using the representative phrases in a topic, the domain expert labels the topic. Using this label as a phrase, we have the outcomes EXIST, EXIST-m, ADD, ADD-m, No-g and No, as above. Furthermore, we add an outcome Q (for query) when the label for the topic is too specific for adding to the ontology, but could be defined using concepts in the ontologies and OWL constructs.
(ii) Using a subset of representative phrases in a topic, the domain expert labels the subset. Using this label as a phrase, we have the outcomes EXIST, EXIST-m, ADD, ADD-m, No-g, No, and Q as above. This can be done for different subsets.

Finally, the domain expert **interprets the lattice**.

(i) Given the relationships in the lattice, as well as the connections of the topics and phrases to concepts in the ontology, new relationships between ontology concepts can be identified.

4 Extending NanoParticle and eNanoMapper Ontologies

In the following subsections, we show the usefulness of our approach by extending two ontologies in the nanotechnology domain.

4.1 Corpus and ontologies

The corpus that we use is based on reports on nanoparticles from the Nanoparticle Information Library (http://nanoparticlelibrary.net). For each nanoparticle report, we take the text in 'Research Abstract' as well as the abstracts (or only the titles if there is no abstract) from the publications in 'Related Publications'. The final corpus contains 117 abstracts from the 'Research Abstract' field in the reports and 510 abstracts (or titles) from publications. We have chosen to only retrieve titles and abstracts rather than full texts. The title and abstract cover the basic content of an article. For a research article in the materials science domain they will generally contain a summary of the problem, experiments, simulations and computations. As the ontologies aim to represent basic knowledge in the domain, these parts of a research article often contain enough information for extraction of concepts. When using the full text, more proposals for concepts may be generated, but many of those will not be relevant. In related fields, it has been shown that the use of titles (and abstracts) may be a reasonable approach (e.g., (Galke et al. 2017)).

The ontologies that we extend are the NanoParticle Ontology (Thomas et al. 2011) (1904 concepts and 81 relations) and the eNanoMapper ontology (Hastings et al. 2015) (12,531 concepts and 4 relations). Both ontologies are available via BioPortal (https://bioportal.bioontology.org/).

4.2 Experiments Setup

In our experiments, we configure the phrases mining threshold with two values (high and low), and the PhraseLDA with different numbers of requested topics (20, 30 and 40). The other parameters of PhraseLDA are set as follows: the total number of Gibbs sampling iterations over the entire data is 1000, the hyperparameters are $\alpha = 50/T$ and $\beta = 0.01$ where T is the number of topics. These initial values for the hyper-parameters are justified in (Steyvers & Griffiths 2007). Thus we have six experiments over the data.

After the interpretation of the phrases by the domain expert, for each setting, all (rows regarding) phrases interpreted with No are removed from the phrase occurrence matrix. The updated matrix (with all EXIST(-m), ADD(-m) and No-g phrases) are used as input for the formal topical concept analysis and a formal topical concept lattice is generated.

For the interpretation of the phrases, topics and lattice results a domain expert (second author) worked together with two ontology engineering experts (first and third author). In a first 2 hour session the three experts went through the phrases of all topics for one of the settings (low mining threshold, 40 topics) of the topic model approach. Each phrase was discussed regarding whether it was relevant for a nanotechnology ontology, checked whether concepts with the same or similar names existed in the NanoParticle Ontology, and a decision was made regarding EXIST(-m)/ADD(-m)/No(-g) as well as which axioms may be needed to add to the ontology. In addition to investigating the ontologies, in some cases terms were checked via wikipedia or research articles. As a preparation for the second session, the knowledge engineers prepared suggestions for the phrases for the other settings, based on the interpretation results of the first session and search in the two ontologies. During the second session (4 hours) the phrases for all settings were interpreted and related to both ontologies. Further, the topics for one setting were interpreted. In the third (2 hour) session the remaining topics as well as the lattice results were interpreted.

4.3 Results and discussion of results

In **Table 1** we show the results regarding the interpretation of the phrases. In addition to the number of concepts in the EXISTS(-m), ADD(-m), and No(-g) categories, we also show the precision. The precision of the system is the ratio of the number of relevant proposed concepts to the number of proposed concepts. We decided to define a relevant proposed concept as a proposed concept that the domain expert recognizes as a relevant concept, whether it be in the ontology, or more specific than concepts in the ontology, or could belong to a more general ontology. Therefore, the relevant proposed concepts are the ones that do not belong to the 'No' category. This conforms to what is relevant in the ontology learning setting.

We note that some phrases may contribute to the addition of multiple concepts and axioms. Furthermore, the low mining threshold settings generate the most number of phrases (in total and per topic). Except for one 'No' phrase, all phrases generated by any of the high mining threshold settings are also generated by at least one (and usually all) low mining threshold settings. For the low mining threshold settings there are only small differences regarding the phrases that occur in topics. There are 29 phrases that are generated by all settings. Of these do 13 exist in the ontologies and relate, among others, to kinds of nanotubes, microscopy, spectroscopy, and various properties of nanoparticles. Furthermore, 7 exist in a modified form, e.g., temperature for low/high/room temperature and core-shell nanoparticle for the phrase core shell. The remaining 9 should be added to the ontologies in the same or modified form. These relate to properties (resolution, pore size, band gap, electrical conductivity, crystallinity), a technique (vapor deposition) and nano-objects (mesoporous silica nanoparticle, thin film). Reverse micelle-synthesized quantum dot leads to the creation of a specific kind of quantum dots as well as a specific synthesis technique. Regarding the phrases that are only found by low mining threshold settings, they relate to different kinds of silicons, nanoparticles, properties and techniques, of which many should be added to the ontologies. There are, however, also several phrases that relate to more general concepts in the materials domain that should not necessarily be added to an ontology in the nanotechnology domain. In all settings, we find most EXIST(-m) cases, which shows that the phrases are relevant with respect to the existing ontologies. Furthermore, we found many ADD(-m) cases which lead to new concepts and axioms. There are also some phrases that relate to more general concepts and some phrases that do not lead to anything meaningful in the context of extending

Table 1: Result of interpreting phrases. The first column defines the case using the number of topics, low or high mining threshold, and ontology. The precision is truncated.

	ADD	ADD-m	EXIST	EXIST-m	No-g	No	precision
20, low, NanoParticle	32	4	26	19	16	9	0.91
20, low, eNanoMapper	29	3	24	25	14	12	0.88
30, low, NanoParticle	30	4	26	18	16	9	0.91
30, low, eNanoMapper	28	3	24	26	12	11	0.89
40, low, NanoParticle	32	4	26	15	16	10	0.90
40, low, eNanoMapper	29	3	24	22	14	12	0.88
20, high, NanoParticle	9	1	14	7	4	0	1.00
20, high, eNanoMapper	8	2	12	10	3	0	1.00
30, high, NanoParticle	8	2	14	8	0	1	0.96
30, high, eNanoMapper	7	1	12	10	0	1	0.96
40, high, NanoParticle	9	2	14	12	4	4	0.91
40, high, eNanoMapper	9	2	12	14	2	4	0.90

For the meanings of ADD(-m), EXIST(-m) and No(-g), see Section 3.3.
For ADD and ADD-m, a new concept is defined in the ontology and one or more subsumption axioms are added.

the ontology. From **Table 2** we note that the more topics the system generates, the lower the percentage of topics that contribute to EXIST(-m) and ADD(-m) categories.

In **Table 3** we show the results regarding the interpretation of the topics. We note that the high mining threshold settings generate the most concepts to add to the ontologies. In each setting there are one or two concepts that were not found during the interpretation of the phrases (e.g., high resolution experiment, water soluble reverse micelle systems, core-shell semiconductors). All EXIST(-m) concepts were also found during the interpretation of the phrases. The No-g category consists of earlier found phrases or specializations of those. Furthermore, many of the topics are very specific and it was decided they should not be added to the ontology, but queries (or complex concepts) using concepts in the ontologies and OWL constructs can be constructed. We also observe that the results for the two ontologies are almost the same, which may be because the topic labels are (much) more specific than the phrase labels and the ontologies do not model concepts at the lowest levels of specificity.

In the final step we generated lattices for all settings. As an example, a part of the lattice for the case of 40 requested topics with a low mining threshold is shown in **Figure 6**. Nodes that contain one topic/one

Table 2: The number (and truncated percentage in parentheses) of topics that contribute to extending the ontologies. The first column defines the case using the number of topics, low or high mining threshold, and ontology.

	Contribute to ADD and ADD-m	**Contribute to EXIST and EXIST-m**	**Contribute to No-g**
20, low, NanoParticle	18 (90.0%)	16 (80.0%)	6 (30.0%)
20, low, eNanoMapper	18 (90.0%)	16 (80.0%)	5 (40.0%)
20, high, NanoParticle	11 (55.0%)	13 (65.0%)	3 (15.0%)
20, high, eNanoMapper	11 (55.0%)	13 (65.0%)	2 (10.0%)
30, low, NanoParticle	19 (63.0%)	19 (63.0%)	11 (36.6%)
30, low, eNanoMapper	18 (60.0%)	20 (66.6%)	11 (36.6%)
30, high, NanoParticle	10 (33.3%)	19 (63.3%)	3 (10.0%)
30, high, eNanoMapper	9 (30.0%)	20 (66.6%)	2 (6.6%)
40, low, NanoParticle	22 (55.0%)	21 (52.5%)	12 (30.0%)
40, low, eNanoMapper	21 (52.5%)	23 (57.5%)	9 (22.5%)
40, high, NanoParticle	13 (32.5%)	16 (40.0%)	4 (10.0%)
40, high, eNanoMapper	12 (30.0%)	18 (45.0%)	3 (7.5%)

Table 3: Result of interpreting topics. The first column defines the case using the number of topics, low or high mining threshold, and ontology. Note that some topics may be empty and some topics may require several concepts. The values in parentheses show the number of added concepts that were not found in the phrase interpretation phase.

	ADD	**ADD-m**	**EXIST**	**EXIST-m**	**No-g**	**Q**	**No**	**precision**
20, low, both	3(1)	0	2	0	1	13	0	1.00
30, low, both	8(2)	0	4	0	1	13	0	1.00
40, low, both	16(1)	0	11	1	2	10	5	0.88
20, high, both	8(1)	0	3	2	0	7	0	1.00
30, high, both	3(2)	0	10	2	0	7	0	1.00
40, high, NanoParticle	10(2)	0	10	3	2	3	2	0.93
40, high, eNanoMapper	10(2)	0	9	4	2	3	2	0.93

For the meanings of ADD(-m), EXIST(-m), No(-g) and Q, see Section 3.3.
For ADD and ADD-m, a new concept is defined in the ontology and one or more subsumption axioms are added.

phrase and have as child the bottom node and as parent the top node are not shown. These have been dealt with in the phrase interpretation step and as there are no connections to other nodes (except top and bottom), no additional information can be gained for those nodes.

The lattices were used in the following ways. First, the domain expert labeled the nodes based on the phrases connected to the nodes. These may be the extents or subsets of the extents of topics. The results are given in **Table 4**. Some new concepts were found that are more general than concepts related to topics (e.g., core-shell cdse nanoparticles), but in general, few additional information was found.

Secondly, the domain expert labeled the nodes based on the phrases connected to the nodes and their descendants. As a node contains less phrases than all its ancestors, a labeling may lead to the definition of a new concept that is a super-concept of the concepts related to the ancestor topics (and relevant axioms). As, according to the topic interpretation step, many topics are very specific, this approach may give a way to decide on the appropriate level of specificity for concepts to add to the ontology. In our experiments, however, the lattices were very flat and the nodes with empty intent contained only one phrase and thus did not lead to additional concepts.

Thirdly, the domain expert used the lattice as a visualization tool to check the original topic interpretation. According to the domain expert, the use of the lattice provides significant help in interpreting the topics. As it groups phrases that are in common between different topics and distinguishes phrases that are specific for certain topics, the structure of complex concepts (based on other concepts) is clarified. It results in a better organization and visualization of the topics and their underlying notions. For instance, for a topic with phrases ‘particle size’, ‘quantum dot’, and ‘gold nanoparticle’, the phrase ‘particle size’ was in common with another topic. By removing ‘particle size’ from the phrase list of the topic, it was easier to see that the topic was a combination of ‘particle size’ and a notion of ‘quantum dots of gold nanoparticles’.

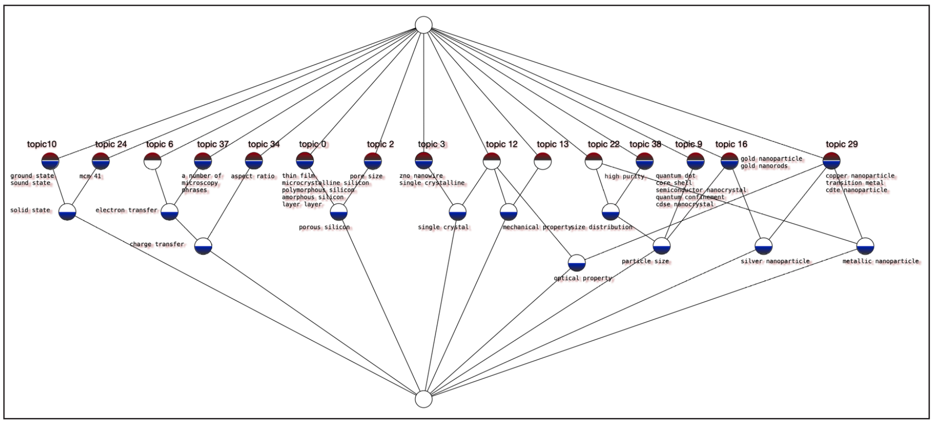


Figure 6: Part of the lattice for the 40 topics and low mining threshold setting. Nodes that contain one topic/one phrase and have as child the bottom node and as parent the top node are not shown.

Table 4: Result of interpreting lattice nodes. The first column defines the case using the number of topics, low or high mining threshold, and ontology. The values in parentheses show the number of added concepts that were not found in the phrase or topic interpretation phases.

	ADD	ADD-m	EXIST	EXIST-m	No-g	Q	No	precision
20, low, both	1(0)	0	1	0	2	0	0	1.00
30, low, NanoParticle	4(2)	0	3	0	1	0	0	1.00
30, low, eNanoMapper	3(2)	0	4	0	1	0	0	1.00
40, low, both	3(0)	0	1	0	0	0	0	1.00
20, high, both	0(0)	0	1	0	1	1	0	1.00
30, high, both	1(1)	0	1	0	0	0	0	1.00
40, high, both	0(0)	0	0	0	0	0	0	1.00

For the meanings of ADD(-m), EXIST(-m), No(-g) and Q, see Section 3.3.
For ADD a new concept is defined in the ontology and one or more subsumption axioms are added.

4.4 General discussion

For the experiments we have currently used few resources, i.e. circa 600 abstracts and less than 10 hours for each of the three experts. Even with these limited resources our approach finds 35 and 32 new concepts for the NanoParticle Ontology and the eNanoMapper ontology, respectively as shown in **Table 5**, as well as 42 and 37 new axioms, respectively, as shown in **Table 6**. In addition to the new concepts and new axioms, also other concepts are influenced. Indeed, for a new axiom A is-a B, the sub-concepts of A receive B and all its super-concepts as its super-concepts (and thus inherit their properties), and all super-concepts of B receive A and its sub-concepts as sub-concepts (and thus all instances of these concepts are also instances of B and

Table 5: New concepts for the NanoParticle and eNanoMapper ontologies.

Concepts	NanoParticle	eNanoMapper
amorphous silicon	✓	
band gap	✓	
Barium Titanate	✓	✓
block copolymer	✓	✓
copolymer	✓	✓
polymer		✓
CdSe nanocrystal	✓	✓
CdTe nanoparticle	✓	✓
copper nanoparticle	✓	
conductivity	✓	✓
electrical	✓	✓
gold nanorod	✓	✓
growth mechanism	✓	✓
resolution	✓	✓
layer by layer growth	✓	✓
liquid solid	✓	
pressure	✓	
MCM 41	✓	✓
mechanical property	✓	✓
viscosity		✓
melt spin	✓	✓
mesoporous silica nanoparticle	✓	✓
mesoporous silica nanosphere	✓	✓
microcrystalline silicon	✓	✓
optical property		✓
polymorphous silicon	✓	✓
pore size	✓	
porous silicon	✓	✓
quantum confinement	✓	✓
reverse micelle-type quantum dot	✓	✓
semiconductor nanocrystal	✓	✓
nanocrystal	✓	✓
silicon thin film	✓	✓
thin film	✓	✓
crystallinity	✓	✓
thermal conductivity	✓	✓
tunnel spectroscopy	✓	✓
ZnO nanowire	✓	✓
	35	32

Table 6: New axioms for the NanoParticle and eNanoMapper ontologies.

Axioms	NanoParticle	eNanoMapper
amorphous silicon is a silicon	✓	
band gap is a quality	✓	
Barium Titanate is an inorganic compound or molecule	✓	
Barium Titanate is a chemical substance		✓
block copolymer is a copolymer	✓	✓
copolymer is a polymer	✓	✓
polymer is an organic material		✓
CdSe nanocrystal is a nanocrystal	✓	✓
CdTe nanoparticle is a nanoparticle	✓	✓
copper nanoparticle is a metal nanoparticle	✓	
conductivity is an independent general individual quality	✓	
conductivity is a quality		✓
electrical conductivity is a conductivity	✓	✓
gold nanorod is a nanorod	✓	✓
growth mechanism is a process	✓	✓
resolution is an independent general individual quality	✓	
resolution is a quality		✓
layer by layer growth is a mechanism process	✓	✓
liquid solid is a liquid solid interface	✓	
pressure is an independent general individual quality	✓	
MCM 41 is a mesoporous silica nanoparticle	✓	✓
mechanical property is a realizable entity	✓	
mechanical property is a quality		✓
viscosity is a mechanical property	✓	✓
melt spin is a technique	✓	✓
mesoporous silica nanoparticle is a nanoparticle	✓	✓
mesoporous silica nanosphere is a nanosphere	✓	✓
microcrystalline silicon is a silicon	✓	
microcrystalline silicon is a chemical substance		✓
nanotube array has part nanotube	✓	✓
optical property is a property		✓
polymorphous silicon is a silicon	✓	
polymorphous silicon is a chemical substance		✓
pore size is a nanoparticle property	✓	
porous silicon is a silicon	✓	
porous silicon is a chemical substance		✓
raman scatter is a synonym of raman spectroscopy	✓	✓
quantum confinement	✓	✓
reverse micelle-type quantum dot is a quantum dot	✓	✓
semiconductor nanocrystal is a semiconductor and is a nanocrystal	✓	✓
nanocrystal is a nano-object and is a crystal	✓	✓
silicon thin film is a thin film	✓	✓
thin film is a fiat material part and one-dimensional nano-object	✓	✓
crystallinity is an independent general individual quality	✓	
crystallinity is a quality		✓
transition metal is a synonym of transition element	✓	

(Contd.)

Axioms	NanoParticle	eNanoMapper
thermal conductivity is a conductivity	✓	✓
tunnel spectroscopy is a spectroscopy	✓	✓
scanning tunneling spectroscopy is same as tunnel spectroscopy	✓	✓
chemical vapor disposition is a vapor disposition	✓	✓
physical vapor disposition is a vapor disposition	✓	✓
ZnO nanowire is a nanowire	✓	✓
	42	37

its super-concepts). In this experiment, 72 concepts from NanoParticle Ontology are influenced by the new axioms. Therefore, the quality of semantically-enabled applications is improved whenever one of the 35 new or 72 influenced concepts is used. For the eNanoMapper ontology the number of influenced existing concepts by adding new axioms is 37. In general, if domain and range are used for the definition of relations in the ontologies, even more concepts would be influenced. Thus, adding these axioms improves the quality of the ontologies and the semantically-enabled applications that use these ontologies. It is clear that the effort for extending the ontologies is worth-while.

The current corpus is mainly related to the themes of *Chemical synthesis*, *Engine Emissions*, *Flame Combustion*, and *Furnace Emissions*. A larger corpus would allow us to find more concepts and axioms as well as extend the coverage, i.e., larger parts of the ontologies could be extended.

Our results show that the approach generates many EXIST(-m) cases. This provides a sanity check for our approach as it shows that existing concepts can be found. In a future system we may want to filter out suggestions by checking the existence of the term or a similar term in the ontologies before showing the domain expert. This may lead to less unnecessary validation work for the domain expert as EXIST(-m) cases would be removed. However, this may also lead to missing some new concepts as the terms used in different ontologies may not always mean the same. For instance, in (Ivanova et al. 2012) it was shown that 'metabolism' in MeSH has a different meaning than 'metabolism' in ToxOntology. Therefore, only using (approximate) string matching and using synonyms may not be enough to filter out EXIST(-m) cases.

For the domain expert it was easier to interpret and label the topics for the settings with high mining thresholds. As mentioned, the number of phrases for topics for the low mining threshold settings is larger than for the high mining threshold settings. Often the topics for the low mining thresholds contained too many phrases to easily interpret the topic. In an extreme case, the domain expert thought that a topic "looked like the subject of a particular research article".

One issue that the domain expert noted was that it was not always easy to decide which level of granularity to use during the interpretation. The question is how specific or how general the interpretation could be and still make sense for the ontology. Although our approach gives much flexibility in this sense, it does give much responsibility to the domain expert and some way to automate recommendations would be helpful. Another related issue is the fact that we found several concepts that were too general for the nanotechnology domain, but that are still relevant. In this case we did not add these to the ontology, but one may reflect on how to deal with this issue, e.g., by importing or linking to other ontologies.

In this experiment we did not find cases where the lattice was in conflict with the ontologies. In our method the domain expert is involved in interpreting the lattice. Therefore, if there would be a conflict between the domain expert's validation and the ontologies, there are two possibilities. First, it is possible that the domain expert made a mistake, and by observing the conflict could rectify the mistake. Second, there may be a mistake in the ontologies. By observing the conflict, we now have an opportunity for debugging the ontology using specialized tools (e.g., (Lambrix 2019).

4.5 Comparison to Other Approaches

Literature As mentioned before, we are mainly dealing with concept discovery and concept hierarchy derivations. As these are also two tasks in ontology learning, we find most related work in that area. While we addressed different methods in Section 2.3, in this section we address systems. A number of ontology learning systems generate concepts. Examples are ASIUM (Faure & Poibeau 2000), CRCTOL (Jiang & Tan 2010), OntoGain (Drymonas et al. 2010), OntoLearn (Navigli et al. 2004) and Text2Onto (Cimiano & Völker 2005). ASIUM applies linguistics-based sentence parsing, syntactic structure analysis, and sub-categorization frames to return concepts. CRCTOL implements both linguistics-based methods and relevance analysis. OntoGain extracts concepts by using linguistics-based part-of-speech tagging, shallow parsing, and relevance analysis. OntoLearn generates concepts based on the concepts and glossary from WordNet. Finally, Text2Onto uses statistics-based co-occurrence analysis. We show the performance of these five systems in **Table 7** according to (Wong et al. 2012).

Experiment with Text2Onto To compare our approach with another system, we have chosen to experiment with Text2Onto (Cimiano & Völker 2005). It was the only system that we found that we could download and install. However, it is one of the most popular and well-known ontology learning systems and therefore a good choice. Text2Onto is an ontology learning system based on mining textual resources. For extracting concepts from the textual resource, Text2Onto implements four algorithms which are entropy-based, C-value/NC-value-based, relative term frequency-based, and term frequency-based and inverted document frequency (TF-IDF)-based respectively. As shown above, it performed well in different domains.

In this experiment, we use Text2Onto on the same corpus as in the experiment for our approach. We split the corpus into segments as Text2Onto uses too much memory when applied on the whole corpus. We apply Text2Onto with default settings for its four algorithms on our corpus. For each of the settings, Text2Onto returns thousands of candidates ranked based on relevance. We apply the same domain expert validation as in our method in terms of interpreting phrases presented in Section 3.3. Instead of using the complete ranked lists of thousands of proposed concepts, we decided to investigate the results of the sub-lists containing the 100, 200, 300 and 400 top elements in the lists, respectively. The results are shown in **Table 8**. The entropy-based and C-Value/NC-Value-based methods return exactly the same results. For the relative

Table 7: Performance of ontology learning systems in different domains (Wong et al. 2012). (Precision is truncated).

System	Domain	Precision
ASIUM	French journal Le Monde	0.86
CRCTOL	Patterns of Global Terrorism	0.92
OntoGain	Computer Science corpus	0.86
	Medical corpus	0.89
OntoLearn	Tourism	0.85
Text2Onto	Text from the paper (Navigli & Velardi 2004)	0.61
	Patterns of Global Terrorism	0.74

Table 8: The results of Text2Onto with different algorithms and different number of returned candidates. (Precision is truncated).

# of elements	Algorithm	ADD	ADD-m	EXIST	EXIST-m	No-g	No	precision
100	Entropy	5	0	39	19	4	33	0.67
	C-value/NC-value	5	0	39	19	4	33	0.67
	Relative term frequency	5	0	39	20	4	32	0.68
	TF-IDF	17	0	22	12	6	43	0.57
200	Entropy	7	1	63	43	8	79	0.60
	C-value/NC-value	7	1	63	43	7	79	0.60
	Relative term frequency	7	1	63	42	8	79	0.60
	TF-IDF	24	1	38	19	19	99	0.50
300	Entropy	12	1	80	52	16	139	0.53
	C-value/NC-value	12	1	80	52	16	139	0.53
	Relative term frequency	13	1	78	52	16	140	0.53
	TF-IDF	28	1	58	36	29	148	0.50
400	Entropy	18	1	98	62	20	199	0.50
	C-value/NC-value	18	1	98	62	20	199	0.50
	Relative term frequency	19	1	100	61	20	199	0.50
	TF-IDF	36	1	70	44	38	211	0.47

term frequency-based method the 160 highest ranked proposed concepts are the same as the 160 highest ranked proposed concepts for the entropy-based and C-Value/NC-Value-based methods. The precision for the entropy-based and C-Value/NC-Value-based methods is the highest for each fixed number of proposed concepts, closely followed by the relative term frequency-based method. The TF-IDF-based method has the lowest precision. However, the TF-IDF-based method finds the largest number of relevant new concepts (ADD(-m)). Further, the precision decreases and the number of relevant new concepts increases for all algorithms, when we take larger sub-lists of top elements.

In **Table 9**, we show the results for Text2Onto when all algorithms are used together for the different sub-lists of top elements and compare it to our method. In **Table 10** we show all the new concepts found by our method and Text2Onto for NanoParticle Ontology. 14 concepts were found by both methods. Further, our method found 21 new concepts that were not found by Text2Onto, while Text2Onto found 28 new concepts that were not found by our method. The two methods seem therefore to be complementary.

Table 9: Results for Text2Onto using all algorithms per setting and our method for extending NanoParticle Ontology. (Precision is truncated).

	ADD	ADD-m	EXIST	EXIST-m	No-g	No	precision
Text2Onto-100	20	0	51	27	11	71	0.60
Text2Onto-200	29	1	84	55	26	164	0.54
Text2Onto-300	39	1	118	78	44	266	0.51
Text2Onto-400	41	1	120	73	47	313	0.47
Our Method	32	3	25	18	14	22	0.80

Table 10: New concepts found by our method and Text2Onto for the NanoParticle Ontology.

Concepts	Our method	Text2Onto
acid group		✓
activation energy		✓
amorphous silicon	✓	
band gap	✓	✓
Barium Titanate	✓	✓
Barium Titante nanowire		✓
block copolymer	✓	✓
boron nanowire		✓
catalyst		✓
cluster		✓
copolymer	✓	✓
crystallite		✓
crystallinity	✓	
CdSe nanocrystal	✓	
CdTe nanoparticle	✓	
copper nanoparticle	✓	✓
conductivity	✓	✓
diblock copolymer		✓
electrical conductivity	✓	
esterification		✓
ethylene oxide		✓
gold nanorod	✓	✓

(Contd.)

Concepts	Our method	Text2Onto
growth mechanism	✓	✓
intensity		✓
resolution	✓	
layer by layer growth	✓	
liquid solid	✓	
pressure		✓
MCM 41	✓	
mechanical property	✓	
melting		✓
melt spin	✓	
mesoporous silica nanoparticle	✓	
mesoporous silica nanosphere	✓	
microcrystalline silicon	✓	✓
nano colloid		✓
nano composite		✓
nanocrystal	✓	✓
nano crystalline silicon particle		✓
nanogrid		✓
nano ribbon		✓
nanotube array	✓	✓
nanowire array		✓
oxidation		✓
photo activity		✓
polyelectrolyte		✓
polymorphous silicon	✓	
pore size	✓	✓
porous silicon	✓	
pressure		P
quantum confinement	✓	✓
reverse micelle-type quantum dot	✓	
semiconductor nanocrystal	✓	✓
silicon thin film	✓	
silica nanosphere		✓
silicon nanowire		✓
silicon nanowire array		✓
superlattice nanowire		✓
thin film	✓	
titanium nanotube		✓
thermal conductivity	✓	
tunnel spectroscopy	✓	
ZnO nanowire	✓	
	35	42

5 Conclusions and Future Work

In this paper we have used a phrase-based topic model approach and introduced a formal topical concept analysis for extending ontologies. A domain expert interprets the results which are phrases, topics and a lattice. This leads to the confirmation of ontological concepts (EXIST(-m)) or to the addition of new concepts and axioms (ADD(-m)). The latter is the actual extension of the ontologies. Also, concepts from more general or other domains may be found, as well as very specific concepts in the domain that need not be added to the ontology. We have shown the usefulness of the approach by extending two ontologies in the nanotechnology domain using approximately 600 abstracts.

In the future we will investigate how to help the domain expert dealing with the granularity issue. In particular, the topical concept lattice explored in this work appears to help refining topics into classifiers of content that are more general and meaningful in the domain. This may be a useful step forward towards a higher level of automation in the process of extracting ontology information out of unstructured text. Furthermore, we will investigate the scalability of our approach by experimenting with more documents. Another possible direction is to investigate synergy possibilities between the topics and the ontology concepts, e.g., by using the ontologies to generate the corpora, or by iterating between topic generation and interpretation.

Acknowledgements

This work has been financially supported by the Swedish e-Science Research Centre (SeRC), the Swedish Research Council (Vetenskapsrådet, dnr 2018-04147), and the Swedish National Graduate School in Computer Science (CUGS).

Competing Interests

The authors have no competing interests to declare.

Author Contributions

All authors were involved in the definition of the project and writing of the paper. HL implementend the methods. All authors participated in the evaluation as domain expert (RA) or as knowledge engineer (HL, PL).

References

Agrawal, A and **Choudhary, A.** 2016. Perspective: materials informatics and big data: realization of the Fourth paradigm of science in materials science. *APL Materials*, 4: 053208, 1–10. DOI: https://doi.org/10.1063/1.4946894

Arnold, P and **Rahm, E.** 2013. Semantic enrichment of ontology mappings: A linguisticbased approach. In: Catania, B, Guerrini, G and Pokorny, J (eds.), *17th East European Conference on Advances in Databases and Information Systems*, 42–55. DOI: https://doi.org/10.1007/978-3-642-40683-6

Arp, R, Smith, B and **Spear, AD.** 2015. *Building Ontologies with Basic Formal Ontology*. The MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262527811.001.0001

Ashino, T. 2010. Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. *Data Science Journal*, 9: 54–61. DOI: https://doi.org/10.2481/dsj.008-041

Asim, MN, Wasim, M, Khan, MUG, Mahmood, W and **Abbasi, HM.** 2018. A survey of ontology learning techniques and applications. *Database*, 2018: bay101, 1–24. DOI: https://doi.org/10.1093/database/bay101

Austin, T. 2016. Towards a digital infrastructure for engineering materials data. *Materials Discovery*, 3: 1–12. DOI: https://doi.org/10.1016/j.md.2015.12.003

Baader, F, Calvanese, D, McGuinness, DL, Nardi, D and **Patel-Schneider, PF.** 2010. *The Description Logic Handbook: Theory, Implementation and Applications*. 2nd edn. Cambridge University Press.

Bhat, M, Shah, S, Das, P, Kumar, P, Kulkarni, N, Ghaisas, SS and **Reddy, SS.** 2013. Premlp: knowledge driven design of materials and engineering process. *ICoRD'13*, 1315–1329. Springer. DOI: https://doi.org/10.1007/978-81-322-1050-4_105

Buitelaar, P, Cimiano, P and **Magnini, B.** 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press.

Cheng, X, Hu, C and **Li, Y.** 2014. A semantic-driven knowledge representation model for the materials engineering application. *Data Science Journal*, 13: 26–44. DOI: https://doi.org/10.2481/dsj.13-061

Cheung, K, Drennan, J and **Hunter, J.** 2008. Towards an Ontology for Data-driven Discovery of New Materials. *Semantic Scientific Knowledge Integration AAAI/SSS Workshop*, 9–14.

Cimiano, P, Hotho, A and **Staab, S.** 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24: 305–339. DOI: https://doi.org/10.1613/jair.1648

Cimiano, P and **Völker, J.** 2005. Text2Onto. In: Montoyo, A, Muñoz, R and Métais, E (eds.), *Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005,* Alicante, Spain, June 15–17, 2005, Proceedings, 227–238.

de Matos, P, Dekker, A, Ennis, M, Hastings, J, Haug, K, Turner, S and **Steinbeck, C.** 2010. ChEBI: a chemistry ontology and database. *Journal of cheminformatics*, 2(S1): P6, 1. DOI: https://doi.org/10.1186/1758-2946-2-S1-P6

Dos Reis, J, Dinh, D, Pruski, C, Da Silveira, M and **Reynaud-Delaitre, C.** 2013. Mapping adaptation actions for the automatic reconciliation of dynamic ontologies. *22nd ACM International Conference on Information and Knowledge Management*, 599–608. DOI: https://doi.org/10.1145/2505515.2505564

Draxl, C and **Scheffler, M.** 2018. Nomad: The fair concept for big data-driven materials science. *MRS Bulletin*, 43(9): 676–682. DOI: https://doi.org/10.1557/mrs.2018.208

Drymonas, E, Zervanou, K and **Petrakis, EG.** 2010. Unsupervised ontology acquisition from plain texts: the OntoGain system. *International Conference on Application of Natural Language to Information Systems*, 277–287. DOI: https://doi.org/10.1007/978-3-642-13881-2_29

El-Kishky, A, Song, Y, Wang, C, Voss, CR and **Han, J.** 2014. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3): 305–316. DOI: https://doi.org/10.14778/2735508.2735519

European Committee for Standardization. 2010. A guide to the development and use of standards compliant data formats for engineering materials test data.

European Materials Modelling Council. 2017. Report on workshop on interoperability in materials modelling.

Faure, D and **Poibeau, T.** 2000. First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. *ECAI-2000 Ontology Learning Workshop*, 7–12.

Galke, L, Mai, F, Schelten, A, Brunsch, D and **Scherp, A.** 2017. Using titles vs. full-text as source for automated semantic document annotation. In: Corcho, Ó, Janowicz, K, Rizzo, G, Tiddi, I and Garijo, D (eds.), *Proceedings of the Knowledge Capture Conference, K-CAP 2017,* Austin, TX, USA, December 4–6, 2017, 20: 1–4. DOI: https://doi.org/10.1145/3148011.3148039

Ganter, B and **Wille, R.** 2012. Formal concept analysis: mathematical foundations. Springer Science & Business Media.

Ghiringhelli, LM, Carbogno, C, Levchenko, S, Mohamed, F, Huhs, G, Lueders, M, Oliveira, M and **Scheffler, M.** 2016. Towards a Common Format for Computational Materials Science Data. *PSI-K Scientific Highlights*. July.

Hartung, M, Terwilliger, J and **Rahm, E.** 2011. Recent advances in schema and ontology evolution. In: Bellahsene, Z, Bonifati, A and Rahm, E (eds.), *Schema Matching and Mapping*, 149–190. DOI: https://doi.org/10.1007/978-3-642-16518-4

Hastings, J, Jeliazkova, N, Owen, G, Tsiliki, G, Munteanu, CR, Steinbeck, C and **Willighagen, E.** 2015. eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. *Journal of Biomedical Semantics*, 6: 10, 1–15. DOI: https://doi.org/10.1186/s13326-015-0005-5

Hearst, MA. 1992. Automatic acquisition of hyponyms from large text corpora. *14th International Conference on Computational Linguistics*, 539–545. DOI: https://doi.org/10.3115/992133.992154

Ivanova, V, Bergman, JL, Hammerling, U and **Lambrix, P.** 2012. Debugging Taxonomies and their Alignments: the ToxOntology – MeSH Use Case. In: Lambrix, P, Qi, G and Horridge, M (eds.), *Proceedings of the First International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2012,* Galway, Ireland, October 8, 2012, 25–36.

Ivanova, V and **Lambrix, P.** 2013. A unified approach for aligning taxonomies and debugging taxonomies and their alignments. In: Cimiano, P, Corcho, Ó, Presutti, V, Hollink, L and Rudolph, S (eds.), *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013*, Montpellier, France, May 26–30, 2013, Proceedings, 1–15. DOI: https://doi.org/10.1007/978-3-642-38288-8

Jiang, X and **Tan, AH.** 2010. CRCTOL: A semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology*, 61(1): 150–168. DOI: https://doi.org/10.1002/asi.21231

Kalidindi, SR and **De Graef, M.** 2015. Materials data science: current status and future outlook. *Annual Review of Materials Research*, 45: 171–193. DOI: https://doi.org/10.1146/annurev-matsci-070214-020844

Karcher, S, Willighagen, EL, Rumble, J, Ehrhart, F, Evelo, CT, Fritts, M, Gaheen, S, Harper, SL, Hoover, MD, Jeliazkova, N, Lewinski, N, Robinson, RLM, Mills, KC, Mustad, AP, Thomas, DG, Tsiliki, G and **Hendren, CO.** 2018. Integration among databases and data sets to support productive nanotechnology: Challenges and recommendations. *NanoImpact*, 9: 85–101. DOI: https://doi.org/10.1016/j.impact.2017.11.002

Lambrix, P. 2005. Towards a semantic web for bioinformatics using ontology-based annotation. *14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise (WETICE'05)*, 3–7. DOI: https://doi.org/10.1109/WETICE.2005.58

Lambrix, P. 2019. Completing and debugging ontologies: state of the art and challenges. arXiv: 1908.03171.

Lambrix, P, Armiento, R, Delin, A and **Li, H.** 2019. Big semantic data processing in the materials design domain. In: Sakr, S and Zomaya, AY (eds.), *Encyclopedia of Big Data Technologies*. DOI: https://doi.org/10.1007/978-3-319-63962-8

Lambrix, P, Wei-Kleiner, F and **Dragisic, Z.** 2015. Completing the is-a structure in light-weight ontologies. *Journal of Biomedical Semantics*, 6: 12, 1–26. DOI: https://doi.org/10.1186/s13326-015-0002-8

Lin, Z, Lu, R, Xiong, Y and **Zhu, Y.** (2012). Learning ontology automatically using topic model. *IEEE International Conference on Biomedical Engineering and Biotechnology*, 360–363. DOI: https://doi.org/10.1109/iCBEB.2012.263

Liu, Q and **Lambrix, P.** 2010. A System for Debugging Missing Is-a Structure in Networked Ontologies. In: Lambrix, P and Kemp, G (eds.), *Data Integration in the Life Sciences, 7th International Conference, DILS 2010*, Gothenburg, Sweden, August 25–27, 2010, Proceedings, 50–57. DOI: https://doi.org/10.1007/978-3-642-15120-0

Maedche, A, Pekar, V and **Staab, S.** 2003. Ontology learning part one – on discovering taxonomic relations from the web. In: Zhong, N, Liu, J and Yao, Y (eds.), *Web Intelligence*, 301–320. Springer. DOI: https://doi.org/10.1007/978-3-662-05320-1

Maedche, A and **Staab, S.** 2000. Discovering conceptual relations from text. *14th European Conference on Arti_cial Intelligence*, 321–325.

Navigli, R and **Velardi, P.** 2004. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2): 151–179. DOI: https://doi.org/10.1162/089120104323093276

Navigli, R, Velardi, P, Cucchiarelli, A, Neri, F and **Cucchiarelli, R.** 2004. Extending and enriching WordNet with OntoLearn. *Proc. 2nd Global WordNet Conf. (GWC)*, 279–284.

Radinger, A, Rodriguez-Castro, B, Stolz, A and **Hepp, M.** 2013. Baudataweb: the Austrian building and construction materials market as linked data. *9th International Conference on Semantic Systems*, 25–32. ACM. DOI: https://doi.org/10.1145/2506182.2506186

Rani, M, Dhar, AK and **Vyas, OP.** 2017. Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence*, 63: 108–125. DOI: https://doi.org/10.1016/j.engappai.2017.05.006

Rumble, J, Broome, J and **Hodson, S.** 2019. Building an international consensus on multi-disciplinary metadata standards: A codata case history in nanotechnology. *Data Science Journal*, 8: 12: 1–11. DOI: https://doi.org/10.5334/dsj-2019-012

Schaal, M, Müller, RM, Brunzel, M and **Spiliopoulou, M.** 2005. RELFIN – topic discovery for ontology enhancement and annotation. In: Gómez-Pérez, A and Euzenat, J (eds.), *The Semantic Web: Research and Applications, Second European Semantic Web Conference, ESWC 2005*, Heraklion, Crete, Greece, May 29 – June 1, 2005, Proceedings, 608–622. DOI: https://doi.org/10.1007/11431053_41

Spiliopoulos, V, Vouros, GA and **Karkaletsis, V.** 2010. On the discovery of subsumption relations for the alignment of ontologies. *Journal of Web Semantics*, 8: 69–88. DOI: https://doi.org/10.1016/j.websem.2010.01.001

Stevens, R, Goble, CA and **Bechhofer, S.** 2000. Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1(4): 398–414. DOI: https://doi.org/10.1093/bib/1.4.398

Steyvers, M and **Griffiths, T.** 2007. Probabilistic topic models. In: Landauer, TK, McNamara, DS, Dennis, S and Kintsch, W (eds.), *Latent semantic analysis: A road to meaning.*

Thomas, DG, Pappu, RV and **Baker, NA.** 2011. Nanoparticle ontology for cancer nanotechnology research. *Journal of Biomedical Informatics*, 44(1): 59–74. DOI: https://doi.org/10.1016/j.jbi.2010.03.001

Tropsha, A, Mills, KC and **Hickey, AJ.** 2017. Reproducibility, sharing and progress in nanomaterial databases. *Nature nanotechnology*, 12: 1111–1114. DOI: https://doi.org/10.1038/nnano.2017.233

Vardeman, C, II, Krisnadhi, A, Cheatham, M, Janowicz, K, Ferguson, H, Hitzler, P and **Buccellato, A.** 2017. An ontology design pattern and its use case for modeling material transformation. *Semantic Web*, 8(5): 719–731. DOI: https://doi.org/10.3233/SW-160231

Wächter, T, Tan, H, Wobst, A, Lambrix, P and **Schroeder, M.** 2006. A corpus-driven approach for design, evolution and alignment of ontologies. *Winter Simulation Conference*, 1595–1602. DOI: https://doi.org/10.1109/WSC.2006.322932

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, JW, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJ, Groth, P, Goble, C, Grethe, JS, Heringa, J, 't Hoen, PA, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, SA, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and **Mons, B.** 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3: 160018, 1–9. DOI: https://doi.org/10.1038/sdata.2016.18

Wong, W, Liu, W and **Bennamoun, M.** 2012. Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, 44(4): 20. DOI: https://doi.org/10.1145/2333112.2333115

Zavitsanos, E, Paliouras, G, Vouros, GA and **Petridis, S.** (2007). Discovering subsumption hierarchies of ontology concepts from text corpora*. IEEE/WIC/ACM International Conference on Web Intelligence*, 402–408. DOI: https://doi.org/10.1109/WI.2007.55

Zhang, X, Chen, H, Ruan, Y, Pan, D and **Zhao, C.** 2017. MATVIZ: a semantic query and visualization approach for metallic materials data. *International Journal of Web Information Systems*, 13: 260–280. DOI: https://doi.org/10.1108/IJWIS-11-2016-0065

Zhang, X, Zhao, C and **Wang, X.** 2015. A survey on knowledge representation in materials science and engineering: An ontological perspective. *Computers in Industry*, 73: 8–22. DOI: https://doi.org/10.1016/j.compind.2015.07.005

Zhang, Y, Luo, X, Zhao, Y and **Zhang, HC.** 2015. An ontology-based knowledge framework for engineering material selection. *Advanced Engineering Informatics*, 29: 985–1000. DOI: https://doi.org/10.1016/j.aei.2015.09.002

How to cite this article: Li, H, Armiento, R and Lambrix, P. 2019. A Method for Extending Ontologies with Application to the Materials Science Domain. *Data Science Journal,* 18: 50, pp. 1–21. DOI: https://doi.org/10.5334/dsj-2019-050

Submitted: 22 June 2019 **Accepted:** 23 September 2019 **Published:** 03 October 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS