A First Step towards Extending the Materials Design Ontology

Mina Abd Nikooie Pour[1], Huanyu Li[1,3], Rickard Armiento[2,3], and Patrick Lambrix[1,3]

[1] Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden
[2] Department of Physics, Chemistry and Biology, Linköping University, 581 83 Linköping, Sweden
[3] The Swedish e-Science Research Centre, Linköping University, 581 83 Linköping, Sweden
firstname.lastname@liu.se

Abstract. Ontologies have been proposed as a means towards making data FAIR (Findable, Accessible, Interoperable, Reusable) and has recently attracted much interest in the materials science community. Ontologies for this domain are being developed and one such effort is the Materials Design Ontology. However, to obtain good results when using ontologies in semantically-enabled applications, the ontologies need to be of high quality. One of the quality aspects is that the ontologies should be as complete as possible. In this paper we show preliminary results regarding extending the Materials Design Ontology using a phrase-based topic model.

Keywords: ontology, ontology extension, materials design, topic model

1 Introduction

In many areas there is a recent interest in making data FAIR, i.e., Findable, Accessible, Interoperable, and Reusable [16]. Findable refers to the fact that data and metadata should be easy to find, accessible to the fact that it should be clear how to access the data, interoperable to the fact that the data needs to be integrated with other data and be usable by applications and workflows, and reusable to the fact that data and metadata are well described such that the data can be replicated or combined in different settings. Ontologies have been proposed as a means towards making data FAIR. Also in the materials science domain there is an awareness regarding the importance of the FAIR principles [4] and efforts are on the way to develop upper ontologies such as EMMO (European Materials & Modelling Ontology), and domain ontologies regarding different sub-domains of materials science such as Mat-Onto [2], Materials Ontology [1], NanoParticle Ontology [14], eNanoMapper ontology [6], ontologies related to computational molecular engineering [7], Materials Design Ontology (MDO) [11], and Materials Graph Ontology [15].

However, to obtain good results when using ontologies for semantically-enabled applications, the ontologies need to be of high quality. One of the quality aspects is that the ontologies should be as complete as possible which relates to the requirement of domain coverage in [12].[1] Many techniques exist for finding missing information in ontologies (see overview in [8]) and extending them. In this paper we show preliminary results of using a variant of the method for extending ontologies that we developed in [10] on MDO.

The remainder of the paper is organized as follows. In section 2 we describe MDO, while in section 3 we describe the method for extending ontologies. In section 4 we show preliminary results of applying the method to MDO. The paper concludes in section 5.

2 The Materials Design Ontology (MDO)

MDO [11] was developed using the NeOn ontology engineering methodology [13], as an answer to the need for an ontology to represent concepts which are the basis for materials design, such as structures of materials, properties of materials, materials calculations and relationships among them. The development was guided by the schemas of the Open Databases Integration for Materials Design (OPTIMADE[2]) project which aims at making materials databases interoperable by developing a common API. The OPTIMADE schemas are based on a consensus reached by several of the materials database providers in the field.

The current version of MDO is publicly available at w3id.org[3] and consists of four modules (Figure 1) [11]. The *Core* module consists of the top-level concepts and relationships of MDO that are reused in other modules. The *Structure* module represents the structural information of materials. The *Calculation* module represents a classification of different computational methods. The *Provenance* module represents provenance information of materials data and calculations. The OWL2 DL representation of the ontology contains 37 classes, 32 object properties, and 32 data properties.

3 Method for extending ontologies

In [10] we presented a general approach for extending ontologies, shown in Figure 2, and showed its use by extending two ontologies in the nanotechnology field. In this paper we use a variant of the approach. We mention the changes from the approach in [10] while describing how we extend MDO in section 4.

Our approach contains two steps. In the first step a phrase-based topic model is created using the ToPMine system [5]. Given a corpus of documents related

[1] In practice, it is difficult to know when an ontology is complete according to the domain, but it is possible to define an 'is more complete than' relation between ontologies which can be used for comparing completeness [8].

[2] `https://www.optimade.org/`

[3] `https://w3id.org/mdo/full/1.0/`

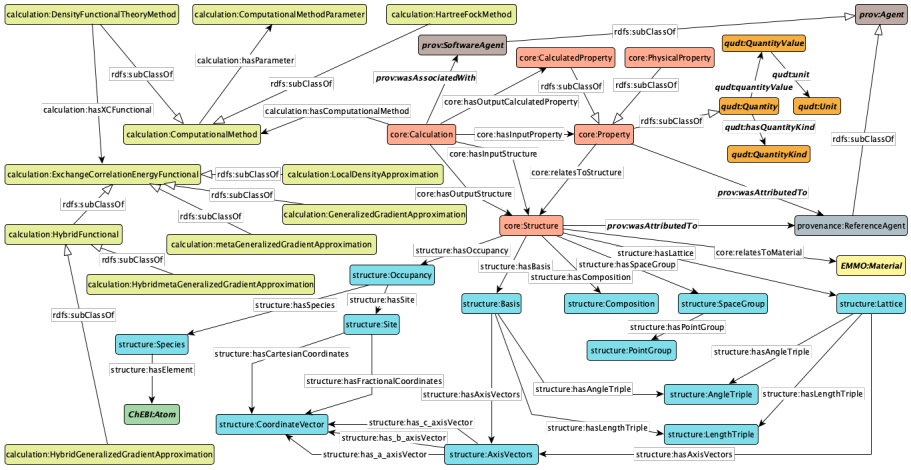


Fig. 1. The Materials Design Ontology [11].

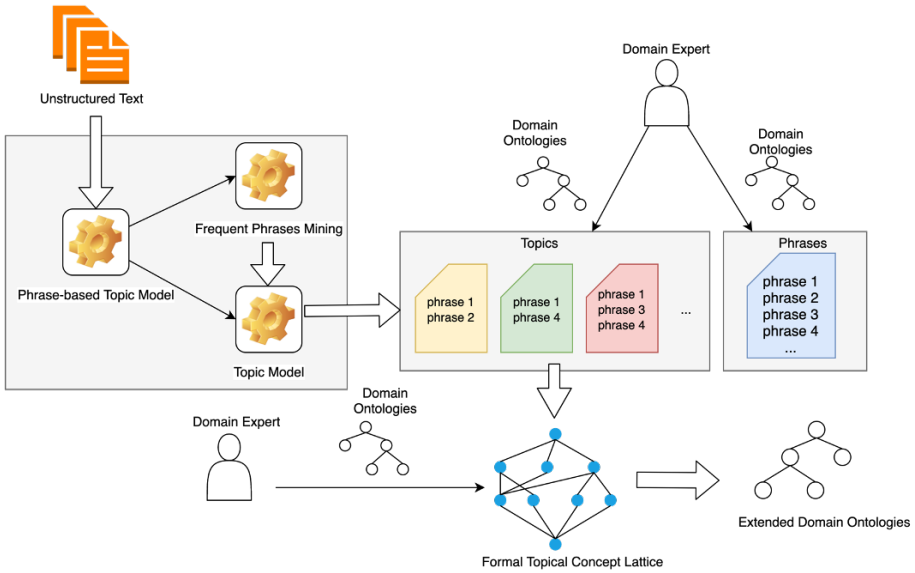


Fig. 2. Approach: The upper part of the figure shows the creation of a phrase-based topic model with unstructured text as input and phrases and topics as output. The lower part shows the formal topical concept analysis with as input topics and as output a topical concept lattice. In both parts a domain expert validates and interprets the results. [10]

to the domain of interest and the number of requested topics, representations of latent topics in the documents are computed. The phrases as well as the topics

are suggestions that a domain expert should validate or interpret and relate to concepts in the ontology.

The second step generates suggestions to the domain expert regarding relations between topics based on formal topical concept analysis [10].

Based on the validations and interpretations of the domain expert, concepts and axioms are added to the ontology.

4 Extending the Materials Design Ontology

4.1 Data

A first step is to collect the corpus that is used as input. The approach in [10] does not specify how the corpus should be collected. In that paper we used an existing library of documents related to the field. In this paper we use MDO as a seed for querying journal databases. We use two journals in the field of materials design: NPJ Computational Materials[4] and Computational Materials Science[5]. We use the 37 concepts of MDO as search phrases in the two journals to find relevant articles and retrieve titles and abstracts of the returned articles. The corpus contains titles and abstracts from 403 articles of NPJ Computational Materials and 8,193 from Computational Materials Science.

In the preprocessing step characters are set to lower case and punctuations are removed. Further, we remove words of length one or two. After preprocessing there are 21,548 distinct words which together occur 808,862 times. An overview of the frequency of the words is presented in Table 1. Most of the words (72.27%) occur less than 10 times, while there are 17 words that occur more than 3000 times. These are 'based', 'properties', 'method', 'calculations', 'phase', 'materials', 'study', 'structure', 'temperature', 'density', 'results', 'energy', 'electronic', 'model', 'molecular', 'simulations', 'surface'.

Table 1. The distribution of word frequency after preprocessing.

Frequency	Percentage of words
less than 10	72.27
10-30	13.25
31-100	7.76
101-500	5.25
501-1000	0.83
1001-2000	0.44
2001-3000	0.12
More than 3000	0.08

[4] https://www.sciencedirect.com/journal/computational-materials-science
[5] https://www.nature.com/npjcompumats/

4.2 Frequent phrases

Given a minimum support threshold *min_support*, we say that phrases that occur at least *min_support* times are *frequent phrases*. ToPMine generates frequent phrases of a length up to a maximum length that is given as an input parameter. In our experiments this was set to 10. Further, ToPMine does not generate all frequent phrases but uses a method based on partitioning documents and using a significance score for deciding which words likely belong together, to produce high-quality frequent phrases [5].

The second column of Table 2 shows the number of frequent phrases that ToPMine generates for different values of *min_support*. The higher the *min_support*, the fewer frequent phrases are generated.

Table 2. Number of frequent phrases for *min_support* 10, 15, 20, 25 and 30 respectively, and three different versions of the ToPMine algorithm.

min_support	original TopMine	New ToPMine without stemming	New ToPMine with stemming
10	6901	6,478	5,452
15	3826	3,578	3,022
20	2542	2,402	2,046
25	1816	1,722	1,477
30	1375	1,298	1,119

In addition, in this paper we also define a maximum support threshold *max_support_word*. Words that occur more than *max_support_word* times are removed. These words are usually very general terms that are not interesting for an ontology or that would not be interesting for a domain ontology, but possibly for an upper ontology. We do note, however, that some of these words could be useful such as 'method', 'electronic', 'model', and 'molecular'. In the remainder we call 'New ToPMine' the algorithm that adds *max_support_word* as well as the preprocessing step. The second column in Table 3 shows how *max_support_word* influences the number of generated frequent phrases with a constant *min_support* of 10. The higher *max_support_word*, the more frequent phrases are generated. Note that no word occurs more than 8000 times in our corpus, so setting *max_support_word* to 8000 allows all words (or, in other words, *max_support_word* is not used).

Another way to look at the influence of *min_support* and *max_support_word* is to compare how many of the frequent phrases are the same and different for different settings. In Figure 3 we show this comparison of different settings to the base setting where *min_support* is 10 and *max_support_word* is 8000 (i.e., *max_support_word* is not used) which is shown in the middle of the figure. The 'Same' bars show how many generated phrases occur both in the base setting and the compared setting. The 'Removed' bars show how many frequent phrases occur in the base setting, but not in the compared setting. For the cases where

Table 3. Number of frequent phrases for *min_support* to 10 and for *max_support_word* 500, 1000, 3000, 5000, and 8000, respectively for two different versions of the ToPMine algorithm.

max_support_word	New ToPMine without stemming	New ToPMine with stemming
8,000	6,478	5,452
5,000	5,947	5,023
3,000	4,692	4,090
1,000	1,878	1,692
500	932	866

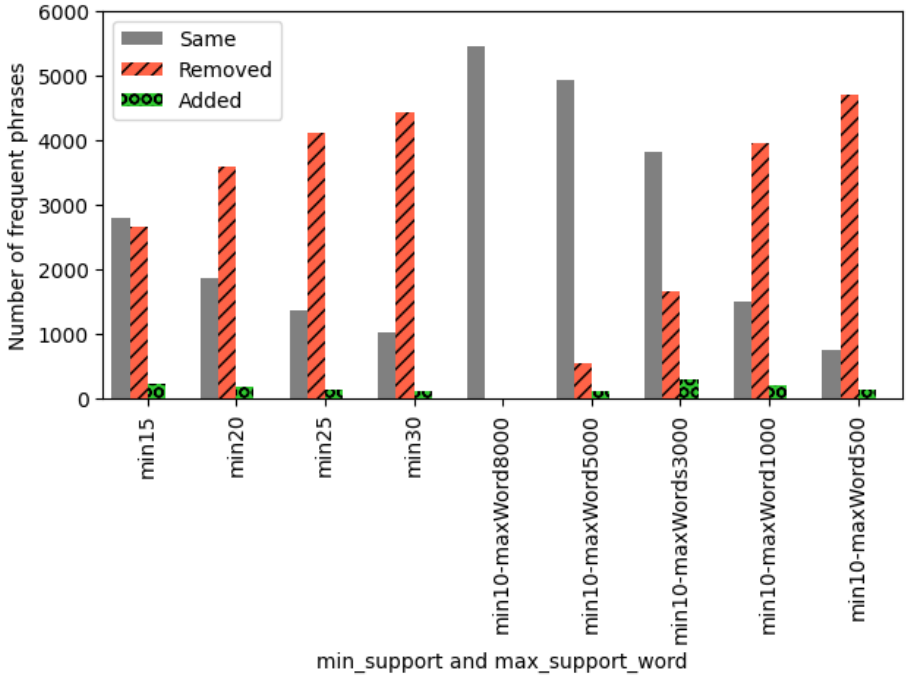


Fig. 3. Comparison of the frequent phrases of new ToPMine algorithm with *min_support* 10 (and *max_support_word* 8000) to settings with *min_support* in 15, 20, 25 and 30, respectively, and settings with *min_support* 10 and *max_support_word* 500, 1000, 3000, 5000, respectively.

we change *min_support*, these would be phrases that are frequent phrases for *min_support* 10, but not for the higher *min_support* in the compared setting. For example 'computational screening' is removed for *min_support* 15. For the cases where we change the *max_support_word*, these would be phrases with words that occur more often than the *max_support_word* in the compared setting. For instance, 'sheet metal forming' contains the word 'metal' with frequency 3457

and would be removed for *max_support_word* 1000. The 'Added' bars show which frequent phrases occur newly in the compared settings. This happens, as stated before, because ToPMine does not generate all frequent phrases, but focuses on high-quality frequent phrases. As an example, 'exchange correlation potential' appears at least 10 times and less than 30 times and 'exchange correlation' appears at least 30 times. Both are frequent phrases for *min_support* 10. However, ToPMine does not generate 'exchange correlation' for *min_support* 10, but it does generate 'exchange correlation potential'. For *min_support* 30 'exchange correlation potential' is not a frequent phrase, while 'exchange correlation' is, and ToPMine does generate 'exchange correlation' as a frequent phrase.

Further, in this paper we also investigate using stemming on the frequent phrases. As an example, the phrases 'molecular dynamics simulations', 'molecular dynamics simulation', 'molecular dynamic simulations' and 'molecular dynamic simulation' have the same stem 'molecular dynam simul'. Stemming allows for removing redundant phrases and thus reduces the work of the domain expert. The influence on the number of generated phrases can be seen by comparing the last two columns in Tables 2 and 3. A disadvantage is that in some cases possible concept candidates may be removed. To alleviate this problem we show the domain expert for each of the stemmed frequent phrases the list of corresponding original phrases. This also helps the domain expert to choose terms to be added to the ontology.

In Table 4, we show the candidate concepts based on the validation of a domain expert on the frequent phrases from the experiment with *min_support* 30 and *max_support_word* 500. In total, 88 candidate concepts are suggested based on 81 out of 131 frequent phrases generated by the experiment. Some candidate concepts can be added into MDO as sub-concepts of existing concepts. For instance, 'Linearized Augmented Plane Wave Method' is a sub-concept of 'Density Functional Theory Method'. Some candidate concepts are relevant to materials design domain but may be not interesting for data access or data integration over materials design databases. For instance, 'Covalent Bond' is a bonding type that can be used to describe materials structures.

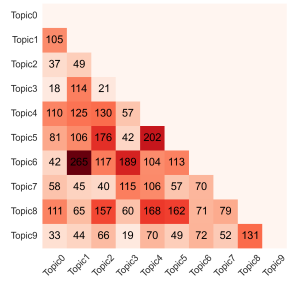
4.3 Topics

Using the frequent phrases, PhraseLDA, a variant of Latent Dirichlet Allocation, is used to generate topics. The number of topics (*num_topic*) is an input parameter to ToPMine. Each topic contains a set of phrases and these sets do not have to be disjoint. For instance, Figure 4 shows the overlap of phrases between topics for different settings of input parameters. In general, when we increase the number of topics, the number of frequent phrases in each topic decreases and the overlap between topics decreases as well.

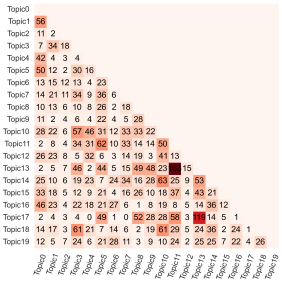
The domain expert validates these topics and if possible, labels them to generate concepts for the ontology. In Table 5, we show the domain expert validation on 10 topics generated by the New ToPMine with stemming, *min_support* 30 and *max_support_word* 500. Among these topics, there are two topics (topics 0 and 9) that are interpreted with multiples labels, i.e., the domain expert divided the

Table 4. Candidate concepts based on domain expert validation on the experiment with *min_support* 30 and *max_support_word* 500.

Stacking Fault	Stone-wales Defect	Cement Paste
Van der Waals Force	Covalent Bond	Perdew-Burke-Ernzerhof (PBE) Exchange-Correlation Functional
Functionally Graded Material	Symmetric Tilt Grain Boundary Structure	Fatigue Limit
Linearized Augmented Plane Wave Method	Asymmetric Tilt Grain Boundary Structure	Edurance Limit
Face Centered Cubic	Rock Salt Structure	Porous Media
Boron Nitride	Rock Salt	Microstructural Features
Nearest Neighbor	Projector Augmented Wave Method	Hall-Petch Relation
Body Centered Cubic	Iron	Conduction Band
Coarse Grained Model	Cahn–Hilliard Equation	Slip Plane
Fiber Reinforced	Cauchy-Born Rule	Vapor Deposition
Zinc Blende	Domain Wall	Spinodal Decomposition
Core Shell	Armchair	Spontaneous Polarization
Rare Earth	Zigzag	Absorption Spectrum
Refractive Index	Double Walled Nanotube	Charpy Impact Test
Half metallicity	Power Factor	Alkaline Earth Metal
X-ray diffration	Carbon Nanotube (cnt)	Contact Angle
Modified Embedded Atom Method	Mixed Mode Fracture	Vickers Hardness
Unit Cell	Homo-lumo Energy Gap	Rutile Titanium Dioxide (TiO_2)
Absorption Spectra	Stainless Steels	Kinematic Hardening
Glass Formation	Vibrational Modes	Hexagonal Close Packed (hcp)
Brillouin Zone	Domain Switching	Anomalous Hall Effect
Lennard Jones	Sound Velocity	Valence Band
Dispersion Curves	Anatase (TiO_2)	Voight Model
Cohesive Zone Model	Austenitic Stainless Steel	Reuss Model
Quasi-harmonic Debye Model	Crystallographic Orientation	Solute Segregation
Additive Manufacturing	Brittle Transition	Directional Solidification
Real Space Methods	Ductile Transition	Muffin-tin Orbital method
Quasi-harmonic Model	Brittle-Ductile Transition	Muffin-tin Orbital Approximation
Quantum Dot	Modified Becke-Johnson Exchange-Correlation Functional	
Hexagonal Boron Nitride	Kohn-Sham	



(a) *min_support* 10, *num_topic* 10



(b) *min_support* 10, *num_topic* 20

Fig. 4. Number of common phrases between pairs of topics.

topic in different parts. The other topics received one label. Further, representative phrases are given for each topic. The labels and the representative phrases can all lead to new concepts.

Table 5. Topic labelling based on domain expert validation on the experiment with *min_support* 30 and *max_support_word* 500 (Up to five representative phrases are selected for each label).

Topic NO.	Topic Labels	Representative Phrases
0	Computational Method Categories	Linearized Augmented Plane Wave Method
		Hartree-Fock Method
		Perdew-Burke-Ernzerhof (PBE) Exchange-Correlation Functional
		Modified Becke-Johnson Exchange Correlation Functional
		Kohn-Sham
	Materials Properties and Features	Absorption Spectrum
		Refractive Index
		Homo-lumo Energy Gap
		Alkaline Earth Metal
		Dispersion curves
	Electronic Structure Features	Conduction Band
		Valence Band
	Materials Categorizations	Half metallicity
		Rare Earth
	Experimental Method Categories	X-ray Diffraction
	Specific Materials	Zinc Blende
	Applications	Optoelectronic Devices
1	Hardness-related Materials Concepts	Quasi-harmonic Debye Model
		Quasi-harmonic Model
		Rock Salt
		Sound Velocity
		Zinc Blende
2	Materials Strength-related Concepts	Stacking Fault
		Van der Waals Force
		Tension Compression
		Uniaxial Tension
		Symmetric Tilt Grain Boundary Structure
3	Materials Fatigue/Fracture-related Concepts	Functionally Graded Material
		Fiber Reinforced
		Cohesive Zone Model
		Unit Cell
		Cement Paste
4	Materials Synthesis Concepts	Additive Manufacturing
		Vapor Deposition
		Directional Solidification
		Microstructural Features
		Crystallographic Orientations
5	Battery-related Materials Concepts	Ion Batteries
		Anatase (TiO_2)
		Lithium Ion Batteries
		Rutile Titanium Dioxide (TiO_2)
		Boron Nitride
6	Materials Structural Categorizations	Face Centered Cubic
		Body Centered Cubic
		Coarse Grained Model
		Hexagonal Close Packed (hcp)
		Iron
7	Nanotube-related Concepts	Armchair
		Boron Nitride
		Hexagonal Boron Nitride
		Carbon Nanotube (cnt)
		Cross Section
8	Artificial Intelligence-Methods (NO)	Artificial Neural
		Neural Networks
		Open Source
		Degrees Freedom
		Artificial Neural Networks
9	Materials Concepts for Solar-cells	Solar Cells
		Quantum Dots
		Domain Wall
		Power Factor
		Electric Fields
	Materials Magnetism Concepts	Domain Switching
		Anomalous Hall Effect
	Materials Polarization Concepts	Spontaneous Polarization

5 Conclusion

In this paper we started our work on extending MDO using a topic model-based approach that relies on domain experts to validate whether candidate concepts should be added to the ontology. We investigated the influence of different settings on the number of frequent phrases that are generated. This is important as it influences the amount of work for the domain expert. Further, we have shown preliminary results on candidate concepts that are generated in the frequent phrases phase and the topics generation phase.

For future work we continue to validate the results of the different variants and settings of the approach for generating frequent phrases and topics. We will also decide which of the candidate concepts should be added to MDO. Then, we will perform formal concept analysis to produce relations between the added concepts. Further, we will use complementary approaches such as Text2Onto [3] and RepOSE [9] to find more concepts and relations.

As a side effect of the validation work by the domain expert we found that in addition to a validation protocol, it would be valuable for the domain expert if there would be a system that helps the expert, e.g., by recommending validations, by allowing for easy search in the results and by clustering similar results together. Further, the system would allow for easy validation, notify when concepts with the same or similar names already exist in the ontology and generate OWL statements for the ontology extension. Developing such a system is one of our current priorities.

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

References

1. Ashino, T.: Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. Data Science Journal **9**, 54–61 (2010). https://doi.org/10.2481/dsj.008-041
2. Cheung, K., Drennan, J., Hunter, J.: Towards an Ontology for Data-driven Discovery of New Materials. In: Semantic Scientific Knowledge Integration AAAI/SSS Workshop. pp. 9–14 (2008)
3. Cimiano, P., Völker, J.: Text2Onto. In: Montoyo, A., Muñoz, R., 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. pp. 227–238 (2005). https://doi.org/10.1007/11428817_21
4. Draxl, C., Scheffler, M.: Nomad: The fair concept for big data-driven materials science. MRS Bulletin **43**(9), 676–682 (2018). https://doi.org/10.1557/mrs.2018.208
5. El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J.: Scalable topical phrase mining from text corpora. Proceedings of the VLDB Endowment **8**(3), 305–316 (2014). https://doi.org/10.14778/2735508.2735519

6. Hastings, J., Jeliazkova, N., Owen, G., Tsiliki, G., Munteanu, C.R., Steinbeck, C., Willighagen, E.: eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. Journal of Biomedical Semantics **6**, 10:1–15 (2015). https://doi.org/10.1186/s13326-015-0005-5
7. Horsch, M.T., Niethammer, C., Boccardo, G., Carbone, P., Chiacchiera, S., Chiricotto, M., Elliott, J.D., Lobaskin, V., Neumann, P., Schiffels, P., Seaton, M.A., Todorov, I.T., Vrabec, J., Cavalcanti, W.L.: Semantic interoperability and characterization of data provenance in computational molecular engineering. Journal of Chemical & Engineering Data **65**(3), 1313–1329 (2020). https://doi.org/10.1021/acs.jced.9b00739
8. Lambrix, P.: Completing and debugging ontologies: state of the art and challenges (2020), arXiv:1908.03171
9. Lambrix, P., Ivanova, V.: A unified approach for debugging is-a structure and mappings in networked taxonomies. Journal of Biomedical Semantics **4**, 10 (2013). https://doi.org/10.1186/2041-1480-4-10
10. Li, H., Armiento, R., Lambrix, P.: A method for extending ontologies with application to the materials science domain. Data Science Journal **18**(1) (2019). https://doi.org/10.5334/dsj-2019-050
11. Li, H., Armiento, R., Lambrix, P.: An ontology for the materials design domain. In: Pan, J.Z., Tamma, V.A.M., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12507, pp. 212–227. Springer (2020). https://doi.org/10.1007/978-3-030-62466-8_14
12. Sabou, M., Fernandez, M.: Ontology (network) evaluation. In: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology engineering in a networked world, pp. 193–212. Springer (2012). https://doi.org/10.1007/978-3-642-24794-1_9
13. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn methodology for ontology engineering. In: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology engineering in a networked world, pp. 9–34. Springer (2012). https://doi.org/10.1007/978-3-642-24794-1_2
14. Thomas, D.G., Pappu, R.V., Baker, N.A.: Nanoparticle ontology for cancer nanotechnology research. Journal of Biomedical Informatics **44**(1), 59–74 (2011). https://doi.org/10.1016/j.jbi.2010.03.001
15. Voigt, S., Kalidindi, S.: Materials Graph Ontology. Materials Letters (2021). https://doi.org/10.1016/j.matlet.2021.129836
16. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR guiding principles for scientific data management and stewardship. Scientific data **3**, 160018:1–9 (2016). https://doi.org/10.1038/sdata.2016.18