HIVE-4-MAT: Advancing the Ontology Infrastructure for Materials Science

Jane Greenberg\(^{(1)}\), Xintong Zhao\(^{(1)}\), Joseph Adair\(^{(1)}\), Joan Boone\(^{(1)}\), and Xiaohua Tony Hu\(^{(1)}\)

Metadata Research Center, Drexel University, Philadelphia, PA 19104, USA
\{jg3243,xz485,jda57,jpb357,xh29\}@drexel.edu

Abstract. This paper introduces Helping Interdisciplinary Vocabulary Engineering for Materials Science (HIVE-4-MAT), an automatic linked data ontology application. The paper provides contextual background for materials science, shared ontology infrastructures, and knowledge extraction applications. HIVE-4-MAT’s three key features are reviewed: 1) Vocabulary browsing, 2) Term search and selection, and 3) Knowledge Extraction/Indexing, as well as the basics of named entity recognition (NER). The discussion elaborates on the importance of ontology infrastructures and steps taken to enhance knowledge extraction. The conclusion highlights next steps surveying the ontology landscape, including NER work as a step toward relation extraction (RE), and support for better ontologies.

Keywords: Materials science · Ontology · Ontology infrastructure · Helping interdisciplinary vocabulary engineering · Named entity recognition · Knowledge extraction

1 Introduction

A major challenge in materials science research today is that the artifactual embodiment is primarily textual, even if it is in digital form. Researchers analyze materials through experiments and record their findings in textual documents such as academic literature and patents. The most common way to extract knowledge from these artifacts is to read all the relevant documents, and manually extract knowledge. However, reading is time-consuming, and it is generally unfeasible to read and mentally synthesize all the relevant knowledge [26,28]. Hence, effectively extracting knowledge and data becomes a problem. One way to address this challenge is through knowledge extraction using domain-specific ontologies [18]. Unfortunately, materials science work in this area is currently hindered by limited access to and use of relevant ontologies. This situation underscores the need to improve the state of ontology access and use for materials science research, which is the key goal of the work presented here.

Supported by NSF Office of Advanced Cyberinfrastructure (OAC): #1940239.
© Springer Nature Switzerland AG 2021
E. Garoufallou and M.-A. Ovalle-Perandones (Eds.): MTSR 2020, CCIS 1355, pp. 297–307, 2021.
https://doi.org/10.1007/978-3-030-71903-6_28
This paper introduces Helping Interdisciplinary Vocabulary Engineering for Materials Science (HIVE-4-MAT), an automatic linked data ontology application. The contextual background covers materials science, shared ontology infrastructural knowledge extraction applications. HIVE-4-MAT's basic features are reviewed, followed by a brief discussion and conclusion identifying next steps.

2 Background

2.1 Materials Science

Materials science is an interdisciplinary field that draws upon chemistry, physics, engineering and interconnected disciplines. The broad aim is to advance the application of materials for scientific and technical endeavors. Accordingly, materials science researchers seek to discover new materials or alter existing ones; with the overall aim of offering more robust, less costly, and/or less environmentally harmful materials.

Materials science researchers primarily target solid matter, which retains its shape and character compared to liquid or gas. There are four key classes of solid materials: metals, polymers, ceramics, and composites. Researchers essentially process (mix, melt, etc.) elements in a controlled way, and measure performance by examining a set of properties. Table 1 provides two high-level examples of materials classes, types, processes, and properties.

Table 1. Examples of Materials classes and types, processes, and properties

| Material class & type | Manufacturing process | Properties (examples) |
|-----------------------|-----------------------|-----------------------|
| Class: Polymer Type: Polyethylene [21] | Polymerization (distillation of ethane into fractions, some of which are combined with catalysts) | Melt temperature, Tensile strength, Flexural strength, Shrink Rate |
| Class: Metal Type: Steel | Iron ore is heated and forged in blast furnaces, where the impurities are altered and carbon is added | Yield strength, Tensile strength, Thermal conductivity, Resistance to wear/corrosion, Formability |

The terms in Table 1 have multiple levels (sub-types or classes) and variants. For example, there is stainless steel and surgical steel. Moreover, the universe of properties, which is large, extends even further when considering nano and kinetic materials. This table illustrates the language, hence the ontological underpinnings, of materials science, which is invaluable for knowledge extraction. Unfortunately, the availability of computationally ready ontologies applicable to materials science is severely limited, particularly compared to biomedicine and biology.
2.2 Ontologies: Shared Infrastructure and Knowledge Extraction Applications

Ontologies have provided a philosophical foundation and motivation for scientific inquiry since ancient times [15]. Today, computationally ready ontologies conforming to linked data standards [9] offer a new potential for data driven discovery [14]. Here, the biomedical and biology communities have taken the lead in developing a shared infrastructure, through developments such as the National Center for Biological Ontologies (NCBO) Bioportal [4,29] and the OBO foundry [6,25]. Another effort is the FAIRsharing portal [1,23], providing access to a myriad of standards, databases, and other resources [31].

Shared ontology infrastructures help standardize language and support data interoperability across communities. Additionally, the ontological resources can aid knowledge extraction and discovery. Among one of the best known applications in this area is Aronson’s [8] MetaMap, introduced in 2001. This application extracts key information from textual documents, and maps the indexing to the metathesaurus ontology. The MetaMap application is widely-used for extraction of biomedical information. The HIVE application [16], developed by the Metadata Research Center, Drexel University, also supports knowledge extraction in a same way, although results are limited by the depth of the ontologies applied. For example, biomedicine ontologies, which often have a rich and deep network of terms, will produce better results compared to more simplistic ontologies targeting materials science [32,33].

Overall, existing ontology infrastructure and knowledge extraction approaches are applicable to materials science. In fact, biology and biomedical ontologies are useful for materials science research, and researchers have been inspired by these developments to develop materials science ontologies [7,11,17]. Related are nascent efforts developing shared metadata and ontology infrastructures for materials science. Examples include the NIST Materials Registry [5] and the Industrial Ontology Foundry [2]. These developments and the potential to leverage ontologies for materials science knowledge extraction motivate our work to advance HIVE-4-MAT. They have also had a direct impact on exploring the use of NER to assist in the development richer ontologies for materials science [33].

2.3 Named Entity Recognition

The expanse and depth of materials science ontologies is drastically limited, pointing to a need for richer ontologies; however, ontology development via manual processes is a costly undertaking. One way to address this challenge is to through relation extraction and using computational approaches to develop ontologies. To this end, named entity recognition (NER) can serve as an invaluable first step, as explained here.

The goal of Named Entity Recognition (NER) is to recognize key information that are related to predefined semantic types from input textual documents [20]. As an important component of information extraction (IE), it is widely applied
in tasks such as information retrieval, text summarization, question answering and knowledge extraction.

The semantic types can vary depending on specific task types. For example, when extracting general information, the predefined semantic types can be location, person, or organization. NER approaches have been also proven effective to biomedical information extraction; an example from SemEval2013 task 9 [24] about NER for drug-drug interaction is shown in Fig. 1 below.

![Fig. 1. An NER example from a SemEval task](image)

As shown in the Fig. 1, the NER pharmaceutical model receives the textual input (e.g. sentences), and returns whether there are important information entities that belong to any predefined labels, such as brand name and drug name.

A similar undertaking has been pursued by Weston et al. [28], with their NER model designed for inorganic materials information extraction. Their model includes seven entity labels and testing has resulted in an overall f1-score of 0.87 [28]. This work has inspired the HIVE team to use NER, as a step toward relation extraction, and the development of richer ontologies for materials science.

### 3 Purpose and Goals

Goals of this paper are to:

1. Introduce HIVE
2. Demonstrate HIVE’s three key features Vocabulary browsing, term search and selection, and knowledge extraction/indexing
3. Provide an example of our NER work, as a foundation for relation extraction.

### 4 HIVE-4-MAT: Prototype Development and Features

Hive is a linked data automatic metadata generator tool developed initially as a demonstration for the Dryad repository [16,30], and incorporated into the DataNet Federation Consortium’s iRODS system [12]. Ontologies encoded in the Simple Knowledge Organization System (SKOS) format are shared through a HIVE-server. Currently, HIVE 2.0 uses Rapid Automatic Keyword Extraction (RAKE), an unsupervised algorithm that processes and parses text into a set of
HIVE-4-MAT: Advancing the Ontology Infrastructure for Materials Science

(candidate keywords based on co-occurrence [22]. Once the list of candidate keywords is selected from the SKOS encoded ontologies, the HIVE system matches candidate keywords to terms in the selected ontologies. Figure 2 provides an overview of the HIVE model.

HIVE-4-MAT builds on the HIVE foundation, and available ontologies have been selected for either broad or targeted applicability to materials science. The prototype includes the following ten ontologies: 1) Bio-Assay Ontology (BioAssay), 2) Chemical Information Ontology (CHEMINF), 3) Chemical Process Ontology (prochemical), (4) Library of Congress Subject Headings (LCSH), 5) Metals Ontology, 6) National Cancer Institute Thesaurus (NCIT), 7) Physico-Chemical Institute and Properties (FIX), 8) Physico-chemical process (REX), 9) Smart Appliances REFerence Ontology (SAREF), and 10) US Geological Survey (USGS).

[Diagram of HIVE structure]

Currently, HIVE-4-MAT has three main features:

- Vocabulary browsing (Fig. 3 and Fig. 4)
- Term search and selection (Fig. 5)
- Knowledge Extraction/Indexing (Fig. 6)

The vocabulary browsing feature allows a user to view and explore the ontologies registered in HIVE-4-MAT. Figure 3 presents the full list of currently available ontologies, and Fig. 4 provides an example navigating through the hierarchy of the Metals ontology. The left-hand column (Fig. 4) displays the hierarchical levels of this ontology; the definition, and the right-hand side displays the alternative name, broader concepts and narrow concepts.
4.1 Mapping Input Text to Ontologies

The term search and selection feature in Fig. 5 allows a user to select a set of ontologies and enter a search term. In this example, eight of the 10 ontologies are selected, and the term thermoelectric is entered as a search concept. Thermoelectrics is an area of research that focuses on materials conductivity of temperature (heat or cooling) for energy production. In this example, the term was only found in the LCSH, which is a general domain ontology. The lower-half of Fig. 5 shows the term relationships. There are other tabs accessible to see the JSON-LD, SKOS-RDF/XML and other encoding. This feature also allows a user to select an encoded term for a structure database system, such as a catalog, or for inclusion in a knowledge graph.

Figure 6 illustrates the Knowledge Extraction/Indexing Feature. To reiterate, reading research literature is time-consuming. Moreover, it is impossible for a researcher to fully examine and synthesize all of the knowledge from existing work. HIVE-4-MAT’s indexing functionality allows a researcher or a digital content curator to upload a batch of textual resources, or simply input a uniform resource locator (URL) for a web resource, and automatically index the textual content using the selected ontologies. Figure 6 provides an example using the
Fig. 4. Vocabularies/ontologies structure

Fig. 5. Term search
Wikipedia content for Wikipedia page on Metal [3]. The visualization of the HIVE-4-MAT’s results helps a user to gain an understanding of the knowledge contained within the resource, and they can further navigate the hypertext to confirm the meaning of a term within the larger ontological structure.

4.2 Building NER for Information Extraction

Inspired by the work of Weston et al. [28], the HIVE team is also exploring the performance and applications of NER as part of knowledge extraction in materials science. Research in this area may also serve to enhance HIVE. Weston et al. [28] focus on inorganic materials, and appear to be one of the only advanced initiative’s in this area. Our current effort focuses on building a test dataset for organic materials discovery, with the larger aim of expanding research across materials science.

To build our corpus, we used Scopus API [27] to collect a sample of abstracts from a set of journals published by Elsevier that cover organic materials. The research team has identified and defined a set of seven key entities to assist with the next step of training our model. These entities have the following semantic labels: (1) Molecules/fragments, (2) Polymers/organic materials, (3) Descriptors, (4) Property, (5) Application, (6) Reaction and (7) Characterization method. Members of our larger research team are actively annotating the abstracts using these semantic labels as shown in Fig. 7. The development a test dataset is an important research step, and will help our team move forward testing our NER model and advancing knowledge extraction options for materials science in our future work.
5 Discussion

The demonstration of HIVE and reporting of initial work with NER is motivated by the significant challenge materials science researchers face gleaning knowledge from textual artifacts. Although this challenge pervades all areas of scientific research, disciplines such as biology, biomedicine, astronomy, and other earth sciences have a much longer history of open data and ontology development, which drives knowledge discovery. Materials science has been slow to embrace these developments, most likely due to the disciplines connection with competitive industries. Regardless of the reasons impacting timing, there is clearly increased interest and acceptance of a more open ethos across materials science, as demonstrated by initiatives outlined by Himanen et al. in 2019 [19]. Two key examples include NOMADCoE [13] and the Materials Data Facility [10], which are inspired by the FAIR principles [13,31]. These developments provide access to structured data, although, still the majority of materials science knowledge remains hidden in textually dense artifacts. More importantly, these efforts recognize the value of access to robust and disciplinary relevant ontologies. HIVE-4-MAT complements these developments and enables materials science researchers not only to gather, register, and browse ontologies; but, also the ability to automatically apply both general and targeted ontologies for knowledge extraction. Finally, the HIVE-4-MAT output provides researchers with a structured display of knowledge that was previously hidden within unstructured text.

6 Conclusion

This paper introduced the HIVE-4-MAT application, demonstrated HIVE’s three key features, and reported on innovative work underway exploring NER. The progress has been encouraging, and plans are underway to further assess the strengths and limitation of existing ontologies for materials science. Research here will help our team target areas where richer ontological structures are needed. Another goal is to test additional algorithms with the HIVE-4-MAT application, as reported by White, et al. [30]. Finally, as the team moves forward, it is critical to recognize that ontologies, alone, are not sufficient for extracting knowledge, and it is important to consider other approaches for knowledge
extraction, such as Named Entity Recognition (NER) and Relation Extraction (RE) can complement and enrich current approaches. As reported above, the HIVE team is also pursuing research in this area as reported by Zhao [33], which we plan to integrate with the overall HIVE-4-MAT.

Acknowledgment. The research reported on in this paper is supported, in part, by the U.S. National Science Foundation, Office of Advanced Cyberinfrastructure (OAC): Grant: #1940239. Thank you also to researchers in Professor Steven Lopez’s lab, Northeastern University, and Semion Saiki, Kebotix for assistance in developing the entity set for organic materials.

References

1. Fairsharing. https://fairsharing.org/
2. Industrial ontology foundry. https://www.industrialontologies.org/
3. Metal-wikipedia entry. https://en.wikipedia.org/wiki/Metal/
4. Ncbo bioportal. http://bioportal.bioontology.org/
5. Nist materials registry. https://materials.registry.nist.gov/
6. Obo foundry. http://www.obofoundry.org/
7. Anikin, A., Litovkin, D., Sarkisova, E., Petrova, T., Kultsova, M.: Ontology-based approach to decision-making support of conceptual domain models creating and using in learning and scientific research. In: IOP Conference Series: Materials Science and Engineering, vol. 483, page 012074 (2019)
8. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: The metamap program (2001)
9. Bizer, C.: The emerging web of linked data. IEEE Intell. Syst. 24(5), 87–92 (2009)
10. Blaiszik, B., Chard, K., Pruyne, J., Ananthakrishnan, R., Tuecke, S., Foster, I.: The materials data facility: data services to advance materials science research. JOM 68(8), 2045–2052 (2016)
11. Cheung, K., Hunter, J., Drennan, J.: MatSeek: an ontology-based federated search interface for materials scientists. IEEE Intell. Syst. 24(1), 47–56 (2009)
12. Conway, M.C., Greenberg, J., Moore, R., Whitton, M., Zhang, L.: Advancing the DFC semantic technology platform via HIVE innovation. In: Garoufallou, E., Greenberg, J. (eds.) MTSR 2013. CCIS, vol. 390, pp. 14–21. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-03437-9_3
13. Draxl, C., Scheffler, M.: Nomad: the fair concept for big data-driven materials science. MRS Bull. 43(9), 676–682 (2018)
14. Eisenberg, I.W., et al.: Uncovering the structure of self-regulation through data-driven ontology discovery. Nat. Commun. 10(1), 1–13 (2019)
15. Greenberg, J.: Philosophical foundations and motivation via scientific inquiry. In: Lee, H.-L., Smiraglia, R. (eds.) Ontology in Knowledge Organization, pp. 5–12. Würzburg : Ergon (2015)
16. Greenberg, J., Losee, R., Agüera, J.R.P., Scherle, R., White, H., Willis, C.: Hive: Helping interdisciplinary vocabulary engineering. Bull. Am. Soc. Inf. Sci. Technol. 37(4), 23–26 (2011)
17. Greenberg, J., Zhang, Y., Ogletree, A., Tucker, G.J., Foley, D.: Threshold determination and engaging materials scientists in ontology design. In: Garoufallou, E., Hartley, R.J., Gaitanou, P. (eds.) MTSR 2015. CCIS, vol. 544, pp. 39–50. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24129-6_4
18. Haendel, M.A., Chute, C.G., Robinson, P.N.: Classification, ontology, and precision medicine. N. Engl. J. Med. 379(15), 1452–1462 (2018)
19. Himanen, L., Geurts, A., Foster, A.S., Rinke, P.: Data-driven materials science: status, challenges, and perspectives. Adv. Sci. 6(21), 1900808 (2019)
20. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition (2020)
21. Rogers, T.: Everything you need to know about polyethylene (pe), creative mechanisms (2015)
22. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Min. Appl. theory 1, 1–20 (2010)
23. Sansone, S.-A., et al.: Fairsharing as a community approach to standards, repositories and policies. Nat. Biotechnol. 37(4), 358–367 (2019)
24. Segura-Bedmar, I., Martínez, P., Herrero-Zazo, M.: SemEval-2013 task 9: extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, June 2013, pp. 341–350. Association for Computational Linguistics (2013)
25. Smith, B., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25(11), 1251–1255 (2007)
26. Tshitoyan, V., et al.: Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571(7763), 95–98 (2019)
27. Wan, K.: What are Scopus APIs and how are these used? (2019). Accessed 17 Oct 2020
28. Weston, L., et al.: Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, June 2019
29. Whetzel, P.L., et al.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res. 39(suppl.2), W541–W545 (2011)
30. White, H., Willis, C., Greenberg, J.: The hive impact: contributing to consistency via automatic indexing. In: Proceedings of the 2012 iConference, pp. 582–584 (2012)
31. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data 3(1), 1–9 (2016)
32. Zhang, X., Zhao, C., Wang, X.: A survey on knowledge representation in materials science and engineering: an ontological perspective. Comput. Ind. 73, 8–22 (2015)
33. Zhao, X., Greenberg, J., Menske, V., Toberer, E., Hu, X.: Scholarly big data: computational approaches to semantic labeling in materials science. In: Proceedings of the Workshop on Organizing Big Data, Information, and Knowledge at JCDL 2020 (2020)