Knowledge Graph-Empowered Materials Discovery

Xintong Zhao
Information Science
Drexel University
Philadelphia, USA
xz485@drexel.edu

Jane Greenberg
Information Science
Drexel University
Philadelphia, USA
jg3243@drexel.edu

Scott McClellan
Information Science
Drexel University
Philadelphia, USA
sm4522@drexel.edu

Yong-Jie Hu
Materials Science and Engineering
Drexel University
Philadelphia, USA
yh593@drexel.edu

Steven Lopez
Chemistry and Chemical Biology
Northeastern University
Boston, USA
s.lopez@northeastern.edu

Semion K. Saikin
Kebotix, Inc.
Cambridge, USA
semion@kebotix.com

Xiaohua Hu
Information Science
Drexel University
Philadelphia, USA
xh29@drexel.edu

Yuan An
Information Science
Drexel University
Philadelphia, USA
ya45@drexel.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671503

***Abstract*—In this position paper, we describe research on knowledge graph-empowered materials science prediction and discovery. The research consists of several key components including ontology mapping, materials data annotation, and information extraction from unstructured scholarly articles. We argue that although big data generated by simulations and experiments have motivated and accelerated the data-driven science, the distribution and heterogeneity of materials science-related big data hinders major advancements in the field. Knowledge graphs, as semantic hubs, integrate disparate data and provide a feasible solution to addressing this challenge. We design a knowledge-graph based approach for data discovery, extraction, and integration in materials science.**

***Index Terms*—Knowledge Graph, Materials Discovery, Information Extraction, Ontology, Natural Language Processing**

I. Introduction

Simulations and advanced experiments in materials science research have generated unprecedented big data [1], [2]. Recently, data-driven machine learning methods have shown great potential to accelerate materials discovery [30], [31]. However, significant amounts of the big data remain underutilized due to data isolation, distribution, and heterogeneity. Example data sources include curated structured databases, such as Inorganic Crystal Structure Database (ICSD) [3], the databases in Materials Project [4], Material Genome Initiative [8], and unstructured text containing much more extensive data, such as over millions of published peer-reviewed research articles and patent documents [5]. It is imperative to develop semantic approaches for unifying the distributed and disparate big data to empower data-driven materials science. In this paper, we present our position in applying knowledge graph techniques for semantic data discovery, extraction, and integration in materials science. The workflow of our ongoing research is presented in Figure 1.

the U.S. National Science Foundation, Office of Advanced Cyberinfrastructure (OAC): Grant# 1940239 and 1940307

II. Data Issues in Current Data-Driven Materials Science

Materials scientists aim to design and discover new materials. They combine chemistry, physics, mathematics, and engineering methods through experimental, theoretical, and computational approaches. Prior to the big data era, the conventional materials discovery process was time-consuming, and inefficient. With the advances in machine learning algorithms, data-driven approaches define the fourth paradigm of materials study, following earlier paradigms shaped by experiment, theory and computation [2], [6]. In data-driven materials science, researchers systematically extract new knowledge by analyzing large-scale materials datasets and predict properties of new materials by building sophisticated machine learning systems [7]. Rich data sources have brought out salient opportunities for data-driven materials research. However, how to efficiently use these data becomes a significant research gap. Many challenging issues are related to the FAIR principles (Findable, Accessible, Interoperable, Reusable) [9]. In particular, data *heterogeneity* and *volume* stand as two main barriers to data-driven materials research [10].

The first challenge is data *heterogeneity*. Materials data are heterogeneous in terms of source and format. As proposed by [34], although current materials informatics greatly benefit from existing data sharing infrastructures, these resources are disconnected with each other and hence can result in the loss of inherent interconnections between data. In addition, material data can appear in many different forms such as texts, numeric values, or coordinates. These heterogeneities hinder materials data analytics from reaching higher potential performance.

The second challenge is data *volume*. As mentioned by [2], researchers' ability to collect data has surpassed the capacity to analyze it. However, given this volume, manually searching for valuable data by reading possibly related articles seems no longer efficient enough.

Based on the above challenges, materials science researchers are overwhelmed by the large amount of data in

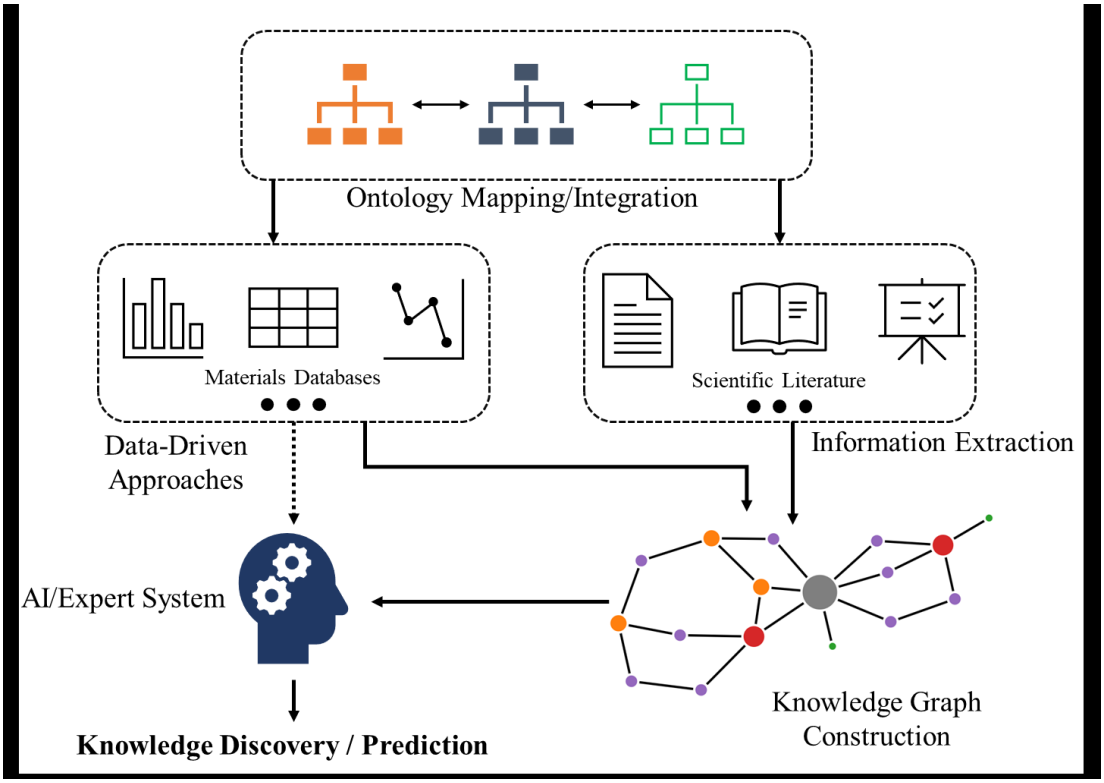

Fig. 1. An overview of the research on empowering materials science discovery with knowledge graphs: the top box describes the research on materials ontology mapping and integration. In the middle layer, there are two key tasks: annotating materials science data with the terms in the integrated ontology and extracting structured knowledge from unstructured text in materials science. On the right-bottom lies the knowledge graph constructed from the tasks on higher layers. Finally, the knowledge graph provides additional semantic data and explanations to the AI/ML models for data-driven materials prediction and discovery.

various forms. To address this research gap, we propose a framework design to integrate and structure heterogeneous data by domain-specific knowledge graph construction. We describe our in-progress research on constructing knowledge graph-empowered approach for semantic materials knowledge discovery, extraction and integration in this position paper.

III. Knowledge Graph-Empowered Approach

Knowledge graphs have emerged as a promising solution to addressing data heterogeneity problems. A knowledge graph consists of a set of triples called **(subject, predicate, object)** or **(head, relation, tail)** , which comprise a labeled graph. Both structured data and unstructured text are automatically converted to triples by data annotation and information extraction techniques. To represent the semantics of the triples, a knowledge graph links the elements in triples to the concepts and relationships defined in a domain ontology or several related ontologies. Thus, our approach involves three key components: materials science ontology mapping and integration, semantic annotation of materials science data, and information extraction from unstructured scholarly articles.

A. Materials Science Ontology Mapping and Integration

Several materials science ontologies have been developed including Ashino's Materials Ontology [12], ChEBI (Chemical Entities of Biological Interest) [13], European Materials and Modelling Ontology (EMMO), Materials Design Ontology (MDO) [14], and the NIST controlled vocabulary [15]. Figure 2 shows a visualization of several concepts defined in the MDO, in which materials Structure, Composition, Occupancy, and Calculation are linked through various relationships (Object Properties). With the proliferation of domain ontologies in chemistry, biology, and materials science, it is necessary to discover the linkages between semantically similar terms in different ontologies for interoperability. To this end, we have developed an ontology matching system, OTMapOnto [16], which applies Optimal Transport ontology embeddings. Our experimental results showed OTMapOnto could achieve mappings with higher recall compared to several state-of-the-art tools. We are improving the OTMapOnto on both precision and recall for the Material Sciences and Engineering (MSE) ontology matching Benchmark [17].

B. Semantic Annotation of Materials Science Data

Semantic interoperability refers to an agreement between the schemas and data items in multiple disparate data sources for data exchange. With the availability of ontologies and mappings between them, different data sources can annotate their data with concepts and relationships in ontologies which clarifies their meaning, and an exchange of information can

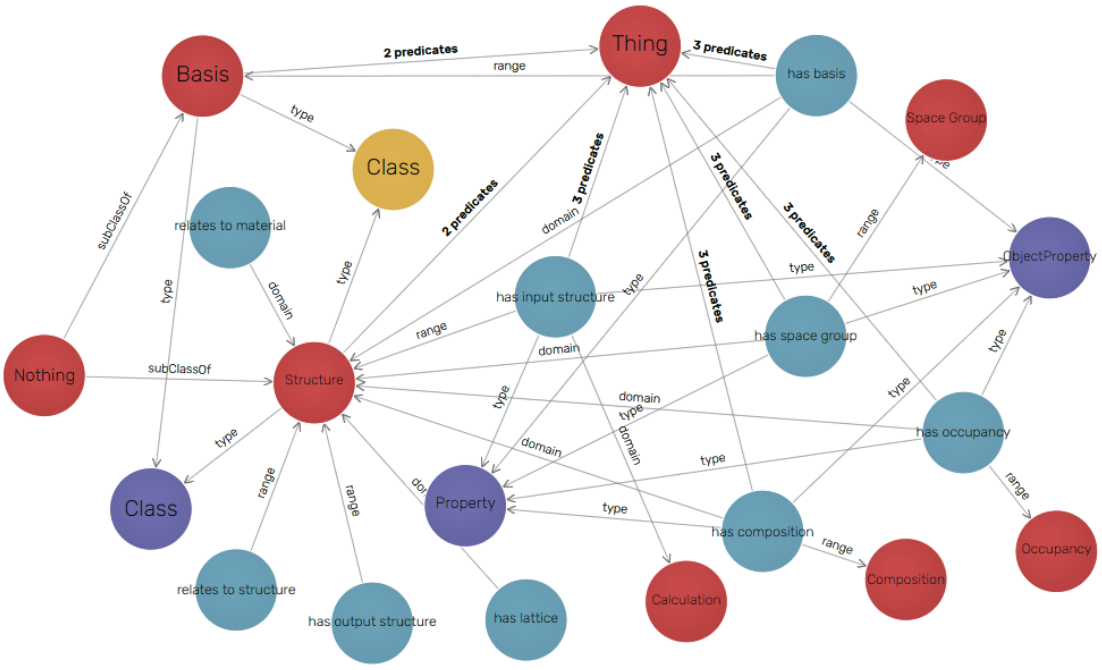

Fig. 2. A visualization of several concepts defined in the Materials Design Ontology (MDO).

rely on standardized semantics. Mapping databases to ontologies is a long-standing problem and various methods have been developed for relational and XML data [18], [19]. We extend the existing methods to structured data in a variety of formats including CSV, JSON, relational, and XML. An embedding-based method is under development where both database and ontology are converted to low-dimensional, dense numerical vectors and semantic mappings are discovered by aligning embedding spaces. For example, the Virtual Excited State Reference for the Discovery of Electronic Materials Database (VERDE materials DB) [20] is an open and searchable database containing ground and excited state properties of organic molecules. We can annotate the molecules in the VERDE DB with concepts such as "Organic Molecule" from the ChEBI ontology. If there are other databases annotated with the same concepts, a knowledge graph will be built by the triples representing data from different databases.

C. Information Extraction from Unstructured Scholarly Articles in Materials Science

Scholarly literature is a major knowledge source to learn from frontier research works. Materials science is no exception; materials researchers usually include valuable knowledge in their published articles, such as material structures, properties with their numerical measures, synthesis methods, and other features. However, unlike structured relational datasets, this important knowledge is often embedded in unstructured text data, which is neither machine-readable nor can it be used directly in machine learning approaches.

The large volume of literature greatly hinders researchers from manually discovering key knowledge. This motivates researchers to explore natural language processing (NLP) techniques to automatically extract important information from text [22]–[24], [26]–[29]. These studies have been conducted across various domains; most of them applied rule-based methods for text extraction, which have less adaptability since predesigned rules may not work on other corpora. [5] used a deep learning model with static word representations (Word2Vec) to address the above limitation. More recently, attention-based pre-trained language models [32], [33] have achieved superior performance in various NLP tasks. These pre-trained language models have great potential to push materials knowledge extraction studies further. In our study, we propose deep-learning enhanced information extraction as a sound direction toward structuring knowledge from unstructured data, and we formulate IE as a sequence-to-sequence labeling problem.

Information Extraction (IE) is one of the fundamental tasks in the field of NLP and can be divided into two main substeps: (1) Named entity recognition (NER), and (2) relation extraction (RE). Then we merge entity-relation-entity triples extracted from the IE process to form the material domain knowledge graph. A demonstration is presented in Figure 3 and described below.

1) Named Entity Recognition: As the first step in the information extraction process, NER aims to recognize important entity mentions in texts that fall into a predefined entity type. We formulate this problem as a sequence labeling task.

Formally, given a natural sentence S consisting of a sequence of N tokens $S =< t_1, t_2, t_3, ..., t_N >$ and we have the

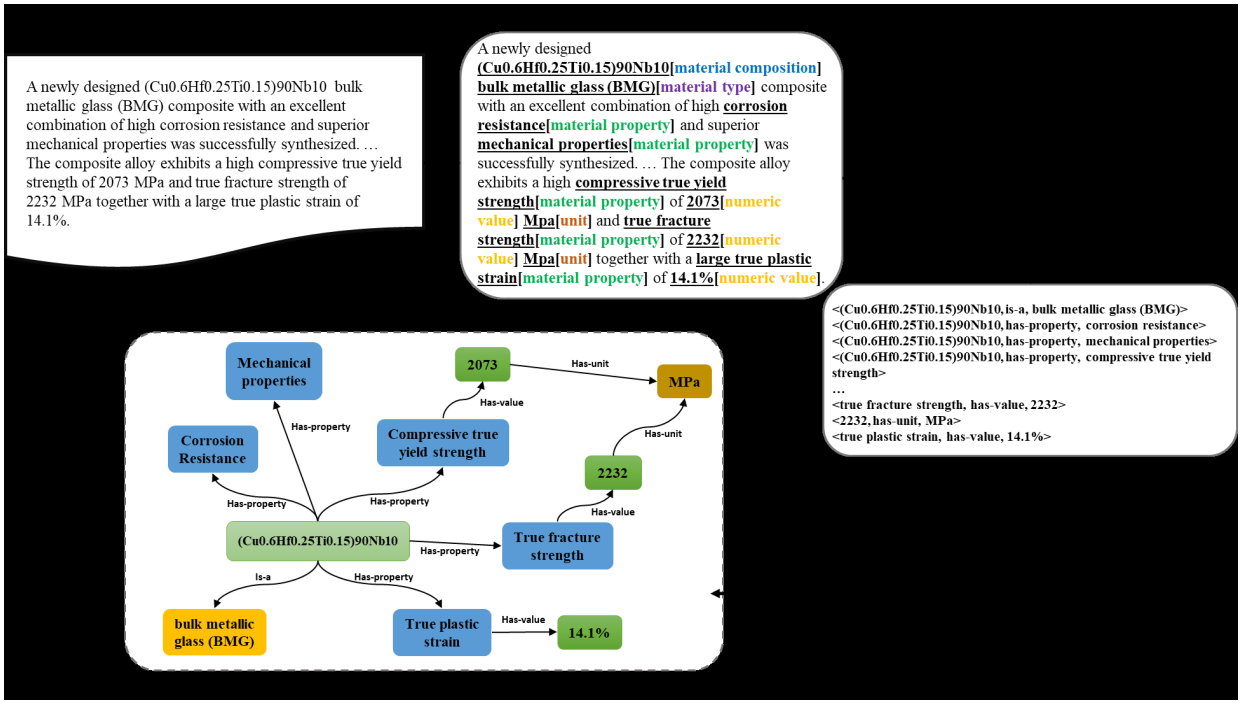

Fig. 3. A demonstration of constructing a knowledge graph from unstructured texts. There are four main components: (1) the overall process takes natural sentences as input, then (2) important entity mentions are extracted with their type, (3) their relations are predicted during the relation extraction step, and finally (4) we merge extracted triples together to obtain a materials domain knowledge graph. The piece of research article used in the example originally from [35].

corresponding labels of each token $L =< l_1, l_2, l_3, ..., l_N >$, our goal is to recognize the list of tuples (t_i, l_k), where t_i is a named entity mention and its label l_k falls into one of the predefined entity types.

2) Relation Extraction: Based on the extracted entities from the previous step, the goal of RE is to determine whether a relation exists between an entity pair and classify the relation type if it exists. Based on the problem formulation above, the goal of RE is to recognize any entity relation pair $< t_i, t_k, r_j >$ in S where t_i, t_k are extracted entity mentions and r_j is one of the designated relation types.

IV. Conclusion

We have developed preliminary prototypes for each of the three components for the knowledge graph-empowered materials discovery. We are integrating the components into the HIVE (Helping Interdisciplinary Vocabulary Engineering) [21] platform which will become a powerful assistant to researchers and practitioners in materials science. HIVE currently addresses interoperability and cost challenges associated with using multiple ontologies, as researchers often seek to work with multiple systems, but HIVE currently requires human interaction. Our knowledge graph-empowered design will transform this technology to a more machine-driven innovation. Finally, our IE component which aims to extract knowledge from unstructured scholarly literature will detect/structure an extensive amount of material knowledge and enrich the knowledge graph enabling expert system/knowledge discovery.

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 and 1940307. We also acknowledge the support of our domain side collaborators from Drexel University, Northeastern University and Kebotix,Inc.

References

[1] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, "Data-Driven materials science: Status, challenges, and perspectives," Advanced Science, vol. 6, no. 21, p. 1900808, 2019.

[2] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the 'fourth paradigm' of science in materials science," APL Materials, vol. 4, no. 5, p. 053208, 2016.

[3] "The world's largest database for completely identified inorganic crystal structures," ICSD. [Online]. Available: https://icsd.products.fiz-karlsruhe.de/. [Accessed: 04-Nov-2021].

[4] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Materials, vol. 1, no. 1, p. 011002, 2013.

[5] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, K. Persson, G. Ceder, and A. Jain, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," 2019.

[6] T. Hey, "The fourth paradigm – data-intensive scientific discovery," Communications in Computer and Information Science, pp. 1–1, 2012.

[7] J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton, and J.-C. Zhao, "New frontiers for the materials genome initiative," npj Computational Materials, vol. 5, no. 1, 2019.

[8] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The Fair Guiding Principles for Scientific Data Management and Stewardship," Scientific Data, vol. 3, no. 1, 2016.

[9] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, and B. Meredig, "Materials science with large-scale data and Informatics: Unlocking New Opportunities," MRS Bulletin, vol. 41, no. 5, pp. 399–409, 2016.

[10] N. Guarino, D. Oberle, and S. Staab, "What is an ontology?," Handbook on Ontologies, pp. 1–17, 2009.

[11] T. Ashino, "Materials ontology: An infrastructure for exchanging materials information and knowledge," Data Science Journal, vol. 9, pp. 54–61, 2010.

[12] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: A database and ontology for chemical entities of Biological Interest," Nucleic Acids Research, vol. 36, no. Database, 2007.

[13] H. Li, R. Armiento, and P. Lambrix, "An ontology for the Materials Design Domain," Lecture Notes in Computer Science, pp. 212–227, 2020.

[14] A. Medina-Smith, C. A. Becker, R. L. Plante, L. M. Bartolo, A. Dima, J. A. Warren, and R. J. Hanisch, "A controlled vocabulary and metadata schema for materials science data discovery," Data Science Journal, vol. 20, 2021.

[15] Y. An, A. Kalinowski, and J. Greenberg, "OTMapOnto: Optimal Transport-based Ontology Matching," Ontology Alignment Evaluation Initiative, 2021.

[16] EngyNasr, "EngyNasr/MSE-benchmark: MSE Track for Evaluating Ontology Matching Tools," GitHub. [Online]. Available: https://github.com/EngyNasr/MSE-Benchmark. [Accessed: 04-Nov-2021].

[17] Y. An, A. Borgida, and J. Mylopoulos, "Inferring complex semantic mappings between relational tables and ontologies from simple correspondences," Lecture Notes in Computer Science, pp. 1152–1169, 2005.

[18] Y. An, A. Borgida, and J. Mylopoulos, "Constructing complex semantic mappings between XML data and Ontologies," The Semantic Web – ISWC 2005, pp. 6–20, 2005.

[19] B. G. Abreha, S. Agarwal, I. Foster, B. Blaiszik, and S. A. Lopez, "Virtual excited state reference for the discovery of Electronic Materials Database: An open-access resource for ground and excited state properties of organic molecules," The Journal of Physical Chemistry Letters, vol. 10, no. 21, pp. 6835–6841, 2019.

[20] J. Greenberg, X. Zhao, J. Adair, J. Boone, and X. T. Hu, "HIVE-4-MAT: Advancing the ontology infrastructure for materials science," Metadata and Semantic Research, pp. 297–307, 2021.

[21] Greenberg, J., Zhao, X., Monselise, M., Grabus, and S., Boone, J. (In Press, 2021). "A Knowledge Network for AI with Helping Interdisciplinary Vocabulary Engineering (HIVE)," Cataloging & Classification Quarterly, 60 (7).

[22] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, "Materials synthesis insights from scientific literature via text extraction and machine learning," Chemistry of Materials, vol. 29, no. 21, pp. 9436-9444, 2017.

[23] S. Huang and J. M. Cole, "A database of battery materials auto-generated using ChemDataExtractor," Scientific Data, vol. 7, no. 1, pp. 1-13, 2020.

[24] C. J. Court and J. M. Cole, "Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction," Scientific data, vol. 5, no. 1, pp. 1-12, 2018.

[25] E. Kim, K. Huang, O. Kononova, G. Ceder, and E. Olivetti, "Distilling a materials synthesis ontology," Matter, vol. 1, no. 1, pp. 8-12, 2019.

[26] F. Ren et al., "Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments," Science advances, vol. 4, no. 4, p. eaaq1566, 2018.

[27] X. Zhao, J. Greenberg, V. Meschke, E. Toberer, and X. Hu, "An exploratory analysis: Extracting materials science knowledge from unstructured scholarly data," The Electronic Library, vol. ahead-of-print, no. ahead-of-print, 2021.

[28] Z. Jensen et al., "A machine learning approach to zeolite synthesis enabled by automatic literature data extraction," ACS central science, vol. 5, no. 5, pp. 892-899, 2019.

[29] X. Zhao, S. Lopez, S. Saikin, X. Hu, and J. Greenberg, "Text to insight: Accelerating organic materials knowledge extraction via deep learning," Proceedings of the Association for Information Science and Technology, vol. 58, no. 1, pp. 558–562, 2021.

[30] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Computational Materials, vol. 2, no. 1, pp. 1-7, 2016.

[31] M. H. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI," Nature, vol. 555, no. 7698, pp. 604-610, 2018.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[33] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," arXiv preprint arXiv:1903.10676, 2019.

[34] D. Mrdjenovich, M. K. Horton, J. H. Montoya, C. M. Legaspi, S. Dwaraknath, V. Tshitoyan, A. Jain, and K. A. Persson, "Propnet: A knowledge graph for materials science," Matter, vol. 2, no. 2, pp. 464–480, 2020.

[35] C. L. Qin, W. Zhang, K. Asami, H. Kimura, X. M. Wang, and A. Inoue, "A novel Cu-based BMG composite with high corrosion resistance and excellent mechanical properties," Acta Materialia, vol. 54, no. 14, pp. 3713–3719, 2006.