APL Materials

NanoMine schema: An extensible data representation for polymer nanocomposites

Cite as: APL Mater. 6, 111108 (2018); https://doi.org/10.1063/1.5046839
Submitted: 02 July 2018 . Accepted: 07 November 2018 . Published Online: 30 November 2018

He Zhao, Yixing Wang, Anqi Lin, Bingyin Hu, Rui Yan, James McCusker, Wei Chen, Deborah L. McGuinness, Linda Schadler, and L. Catherine Brinson

NanoMine schema: An extensible data representation for polymer nanocomposites

He Zhao,[1,a] Yixing Wang,[1,a] Anqi Lin,[2] Bingyin Hu,[2] Rui Yan,[3] James McCusker,[3] Wei Chen,[1] Deborah L. McGuinness,[3] Linda Schadler,[4] and L. Catherine Brinson[2,b]

[1]*Department of Mechanical Engineering, Northwestern University, Evanston, Illinois 60208, USA*
[2]*Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina 27708, USA*
[3]*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180, USA*
[4]*Department of Material Science and Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, USA*

(Received 2 July 2018; accepted 7 November 2018; published online 30 November 2018)

Polymer nanocomposites consist of a polymer matrix and fillers with at least one dimension below 100 nanometers (nm) [L. Schadler *et al.*, Jom **59**(3), 53–60 (2007)]. A key challenge in constructing an effective data resource for polymer nanocomposites is building a consistent, coherent, and clear data representation of all relevant parameters and their interrelationships. The data resource must address (1) data representation for representing, saving, and accessing the data (e.g., a data schema used in a data resource such as a database management system), (2) data contribution and uploading (e.g., an MS Excel template file that users can use to input data), (3) concept and knowledge modeling in a computationally accessible form (e.g., generation of a knowledge graph and ontology), and (4) ultimately data analytics and mining for new materials discovery. This paper addresses the first three issues, paving the way for rich, nuanced data analysis. We present the NanoMine polymer nanocomposite schema as an XML-based data schema designed for nanocomposite materials data representation and distribution and discuss its relationship to a higher level polymer data core consistent with other centralized materials data efforts. We also demonstrate aspects of data entry in an accessible manner consistent with the XML schema and discuss our mapping and augmentation approach to provide a more comprehensive representation in the form of an ontology and an ontology-enabled knowledge graph framework for nanopolymer systems. The schema and ontology and their easy accessibility and compatibility with parallel material standards provide a platform for data storage and search, customized visualization, and machine learning tools for material discovery and design. https://doi.org/10.1063/1.5046839

Polymer nanocomposites consist of a polymer matrix and fillers with at least one dimension below 100 nm.[1] With tailored composition and chemical processing, nanocomposites have been demonstrated as a class of designer materials that typically exhibit a suite of material properties superior to both the macroscopic polymer matrix and the nano-scale fillers. Recent research efforts demonstrate promising potential for next generation functional materials for structural, mechanical, dielectric/electrical, and other applications.[2,3] The properties of nanocomposites can be influenced by many factors, such as constituent properties, processing methods, dispersion of the nanofiller, interfacial chemistry, and filler geometry. The state of the art for material design in this domain,

[a]H. Zhao and Y. Wang contributed equally to this work.
[b]Author to whom correspondence should be addressed: cate.brinson@duke.edu

however, remains trail and error experiements using few rigorous design reproaches to optimize properties, which limit wide application and commercialization of nanocomposite materials.[4–6]

Data-driven methods have been a recent thrust in many sub-disciplines of materials science research.[7,8] The goal is to combine the best practices of materials science, computer science, and informatics to utilize the vast array of past research data for new materials discovery. While large amounts of data have been accumulated and published in the scientific literature, handbooks, and commercial product catalogs, the vast majority of the data underpinning the publications is not readily accessible. A few efforts have begun to create online versions of material databases, with curated materials entries available for look-up.[8–10] For example, in the metals community, several mature materials databases are widely used, such as the CALPHAD database for alloy phase diagrams which is over three decades old.[10] Other efforts, such as the Materials Project,[9] leverages high performance computing (HPC) to accelerate materials discovery, particularly in the field of next generation batteries using simulated datasets of inorganic compounds.[8] Open quantum materials database (OQMD) has enabled researchers to quickly obtain density functional theory (DFT) calculated thermodynamic properties and apply tools to visualize phase diagrams and crystal structures for many complex alloys and compounds.[11,12] Data types and templates, as well as data sharing protocols, have become mature in these fields with active user involvement, constant feedback, and knowledge exchange.

Unlike the inorganic and metallic alloy fields, polymers and polymer nanocomposites are in the nascent stages in terms of data-driven and systematic informatics protocols. This status has largely been attributed to the complexity and high dimensionality of the polymer and polymer nanocomposites data space, which requires accounting for details in all aspects of composition, chemistry, processing, structure, and property spaces.[1] Additionally, with diverse types of data scattered across numerous sources, it is challenging to implement a universal gold standard that can contain all possible nanocomposite data. There exist several examples of online polymer data resources with reasonable datasets and a production grade user interface. These include the CRC POLYMERSnetBASE by the Taylor and Francis Group,[13] the Polymer Property Predictor and Database (PPPDB) by the University of Chicago and the PolyInfo database from NIMS of Japan.[14] All of these data resources distribute curated polymer data from publications and polymer handbooks, with detailed annotations of chemical properties and characteristics. However, there are limitations for those data resources: for PolyInfo, the lack of application program interface (API) access prevents the application of user-defined search and exploring of data; the CRC POLYMERSnetBASE requires paid access; and PPPDB covers only a few properties (notably the chi parameter and glass transition temperature) and is focused on polymer blends. Additionally, for all those data resources, very few records are related to composites or nanocomposites, and those data resources rarely contain the complete information needed to describe the nanocomposite processing, microstructure and properties. Therefore, a dedicated solution for nanocomposites is still required in order to apply data-driven approaches to nanocomposite materials.

As an exclusive data resource for the nanocomposite research community, we developed NanoMine as a prototype framework of an online, open resource system for nanocomposite materials data sharing, analysis, and material design.[15] It includes a nanocomposite database with online access and REST API, a suite of web-based tools for statistical analysis of microstructure images and material properties, and physics-based modeling and simulation of nanocomposite properties. Currently, the NanoMine database contains 182 papers, over 1000 unique samples, covering over 100 types of polymers/particles with over 20 000 data points and is continuously expanding.

While our previous work outlined a vision for NanoMine as a resource with data, analytical tools and simulation packages, and an initial nascent framework for a nanocomposite schema,[15] here we present the full development of and open access to the NanoMine Nanocomposite Schema and demonstrate its use toward effective and systematic nanocomposite data curation and archiving. We begin with an XML schema to ensure hierarchical relations between the parameters and to maintain a well-formed structure so that the major categories of factors and variables can be systematically recorded and queried. We also show that a more accessible representation can be obtained by a direct mapping of the data storage in XML to display in a spreadsheet-like table, making manual data curation more approachable for the general user. We present an initial translation and expansion of this schema into an ontology-enabled framework with semantic interpretation and discuss its impact on

new materials discovery, as well as coordination with other materials standards, such as the Polymer Data Core initiative by NIST.

Data representation is critical to the effectiveness of data management in data-driven materials science research.[18] The Polymer Property Predictor and Database (PPPDB) and PolyInfo are both examples of existing online polymer material databases,[14,19] where public domain data of pure polymers are cataloged and organized in structured ways. Figure 1 shows an example for polystyrene from the PPPDB with each field shown in a tabular format. The PPPDB contains the Flory-Huggins parameter (χ) and the glass transition temperature (T_g) for various polymers. It can be observed that the simplest element is just one value (string) for each attribute, such as a molecular formula, while other attributes have embedded structure, such as CAS numbers, with multiple rows and/or columns. Most of the higher-order relations in this database can be handled by linking multiple tables since the relationships between the tables is simple and the degree of complexity relatively low.

Similarly, PolyInfo describes polymer property data as tables of descriptive text, with the ability for simple visualization as shown in Fig. 2. PolyInfo resembles PPDB in terms of data layout (see Fig. 1), but also has more sections for different groups of properties. PolyInfo contains thermal [glass transition temperature (T_g), melting temperature (T_m), heat of fusion, thermal decomposition temperature], physical (density and water absorption), electrical (dielectric constant, dielectric loss tangent, and electric conductivity), and mechanical (tensile modulus, tensile stress (strength) at break, and elongation at break) properties for polymer systems and some of the data contains polymer composite samples.

On the other hand, the processing, structure, and property (PSP) data for nanocomposites contain more hierarchy, where the matrix, filler, and surface chemistry all add additional dimensions. Furthermore, the processing conditions associated with any given sample are critical to the document, as they dramatically impact the structure and properties,[20,21] and the associated images and text data adds another level of complexity to the relations. In terms of material property data, multiple XY plots in tabular form with or without image data cannot be easily represented by a single set of relational tables. Therefore a more rigorous and flexible format for the nanocomposite data platform is needed.

While SQL (Structured Query Language) databases (relational databases, which store the data in series of related data tables) were the primary data storage mechanisms over the past decades, NoSQL

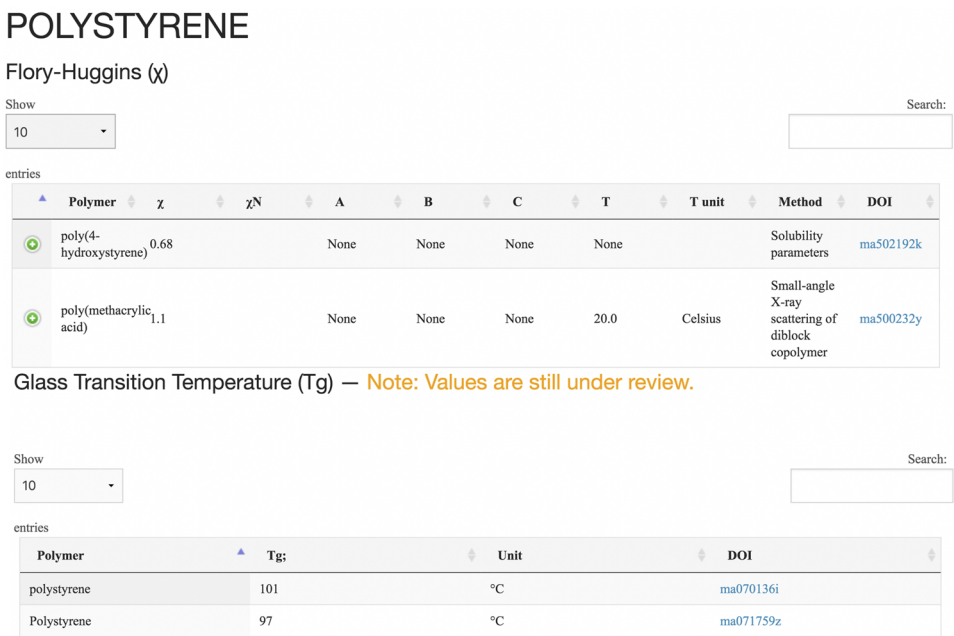

FIG. 1. Screen shot showing Flory-Huggins and glass transition temperature for polystyrene with each field shown in a tabular format in PPPDB.[19]

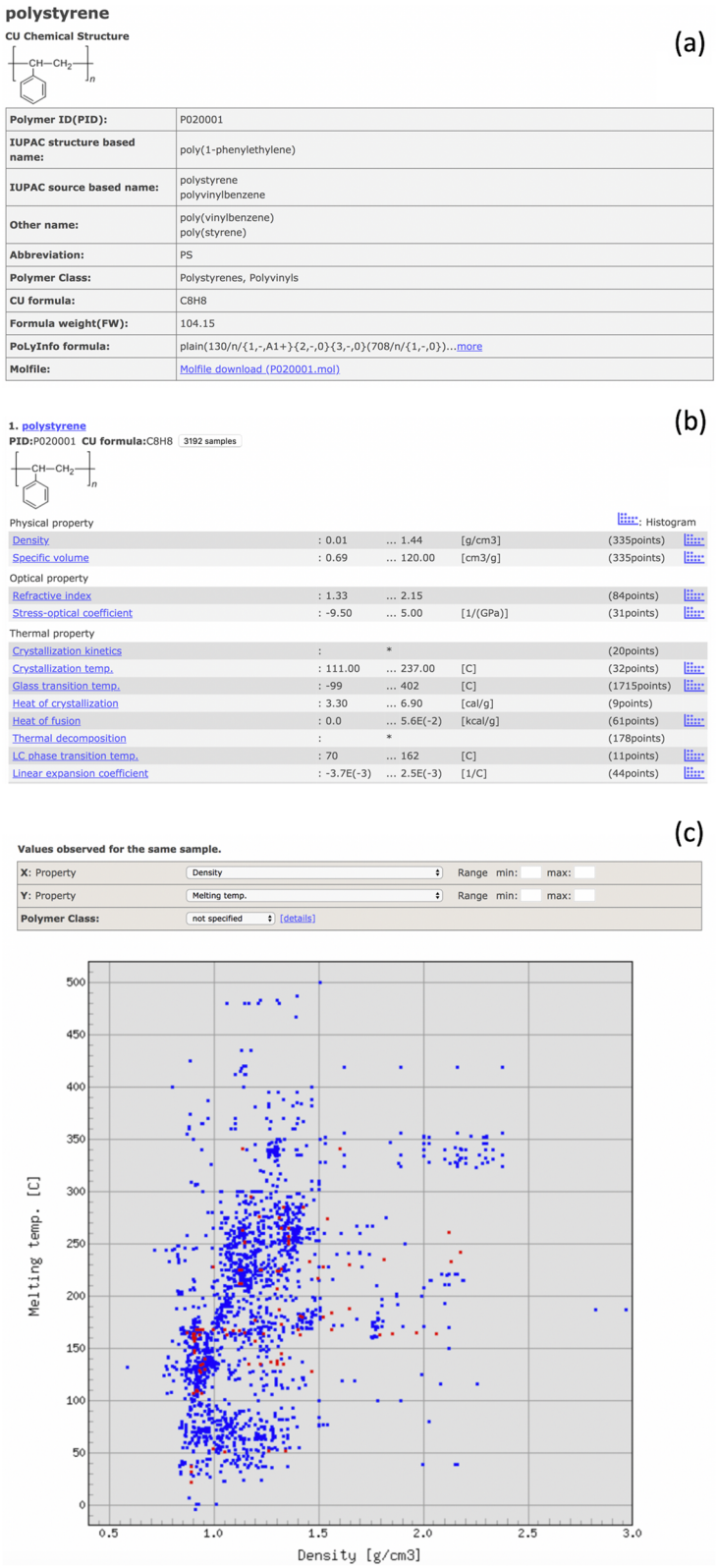

FIG. 2. Screenshots from PolyInfo showing different views of polymer data (polystyrene as an example).[14] (a) A table containing material information including the IUPAC name and abbreviation. (b) A summary table containing list of available material properties including physical, mechanical, electrical, and optical properties. (c) A sample visualization plotting the polymer density and glass transition temperature for both pure polymer and composites.

databases (non-relational databases) have gained popularity. In a NoSQL format,[22,23] the data is stored more flexibly and dynamically, such as in paired values or as a graph. Data can be brought in or out of a NoSQL database, in many formats, such as JSON or Extensible Markup Language (XML).[24] Widely seen as a semantic web building block, XML has its roots as standards of Internet protocols and often is the initial encoding language on the way to more detailed representations that are captured in ontologies. An *XML Schema*, in the form of an XML Schema Definition (XSD) file, describes the structure for data intended for a specific NoSQL database and ensures the integrity of the XML documents.[25]

XML schemas have been applied throughout the physical sciences and engineering fields. For example, ThermoML has been developed by NIST to archive and exchange thermophysical and thermomechanical data with an extensive list of parameters that cover all compounds, mixtures, and chemical reaction information observed in experiments.[26,27] NIST also developed MatML as a high level XML schema for management and sharing of materials data.[24] The first version of MatML schema was launched in December 2002 with several subsequent updates leading to the present version of the MatML schema created in 2004, with no further updates. Applications of MatML include the development of the NSF Materials Digital Library (MatDL), a data exchange project by WMTR, Battelle, and Boeing, as well as collecting contaminant emissions data.[28] MatML was also extended to design an ontology to model the Heat Treating of Materials[29] and to develop CastML for material casting research, development, production, or design.[30]

To support the rapid growth in materials data and the need for shared platforms and exchange,[31,32] NIST has recently developed a flexible platform for material data infrastructure, the Materials Data Curation System (MDCS).[33] MDCS is open source, powered by a MongoDB database engine, provides a flexible API, and organizes the data using user-defined XML schemas. A template composer in MDCS allows the user to create a new schema, or build off an existing one, for any material system of interest. The MDCS platform also functions to make the schema XSD file actionable such that a regular user can make records that validate against the schema. In this work, we utilize MDCS platform and the top level categories of MatML to develop the NanoMine schema specifically for nanocomposite materials with the aim to link processing-structure-property relationship and nanocomposite material design.

In the remainder of this paper, we first illustrate the basic structure of the NanoMine schema and the underlying principles in the design of the schema, as well as how it can be related to a general material vocabulary for polymer data. Then, to provide a more accessible interface for materials researchers, a "flat" relational representation of the same template is illustrated as an MS Excel template along with modular conversion scripts to translate between the tree-like and relational structures. The ontology section applies the schema to construct an initial nanocomposite ontology for use in knowledge graph generation, term mappings, and conceptual understanding. The ontology is encoded using the World Wide Web Consortium's recommended language for ontologies on the web (OWL) and is compatible with the XML schema and we present it using the RDF/XML exchange syntax.

To appropriately capture the full suite of possible data for nanocomposites, there are six major sections in the NanoMine schema. In each category, many hierarchically related fields have been defined to store sufficient detail on nanocomposite data to enable subsequent data analytics and design tools. The structure and fields have been designed using a test set of published papers on nanocomposites and have been further refined with additional curation of data into the data resource. The top level categories in NanoMine and the description of the subfields organized underneath them are as follows:

- **Data provenance**: Metadata of the source of the literature guided by Dublin core standards.[34] The essential metadata includes the DOI of the cited source, author, title, keyword, time and source of publication, etc. This section of the schema supports both published and unpublished datasets. For published datasets, we have developed a robust automatic DOI information retrieving tool[35] that fills out the data source information for users and enhances the data quality by eliminating entry errors. NanoMine currently stores data from 12 publishers and the tool supports all of them.

- **Materials Composition**: Characteristics of the constituent materials, including the polymer matrix, the filler particle, and surface treatments for polymer and/or particle constituents. Bulk matrix and filler properties can be entered, along with compositions (volume/weight fraction) and pre-treatment (for example, grafting and other surface treatment methods).
- **Processing**: Extracted sequential description of chemical synthesis and experimental procedures. Currently, three major sub-categories are included: solution processing, melt mixing, and *in situ* polymerization. Detailed information at each processing step, such as the temperature, pressure, and time, can be entered.
- **Characterization**: specification of the material characterization equipment and methods and conditions used. This information includes details on common microscopic imaging (SEM and TEM), thermal and electrical measurement, mechanical property measurements, and nano-scale spectroscopy.
- **Properties**: measured data of materials performance and response. Properties include mechanical, viscoelastic, electrical, thermal, and volumetric properties. Format of data can be scalar or higher dimensional, such as in 2D plots or 3D maps.
- **Microstructure**: contains nanophase spatial dispersion and topological information from micro and nano-scale imaging. Multiple images can be archived to document microstructure in nanocomposites. Unlike traditional journal articles, where typically one "representative" image is recorded, we provide the ability to archive sets of images, which will increase robustness of statistical descriptors of microstructure[36,37] and the correlations[38] that can be gleaned between structure and processing or properties. Geometric descriptors can also be entered into this section to describe the statistical distribution of the microstructure.

The full representation of the current schema in the XML tree can be found on GitHub[39] and on the NanoMine website.[40] We utilize the NIST Materials Data Curation System (MDCS) platform[33] as the backbone of the NanoMine database. Experimental nanocomposite samples have been effectively populated into the NanoMine database using this schema. Figure 3 shows an example of a populated XML document of a graphite-PMMA nanocomposite sample. The hierarchical structure indicates some basic relationships across the parameters. For example, the "Particle Size" can be

FIG. 3. Populated XML tree for a given sample in NanoMine. As highlighted in two red boxes, "Particle Size," with two sub-elements "value" and "unit," is a child element of "Filler," which is in the "Materials Composition" upper level category.

defined specifically as a child-element under "Filler," with the value and unit as its subsequent child elements. We will see later that mapping this schema to an ontology will provide an even more flexible framework to handle both simple and complex relationships far beyond the parent-child concept.

The non-relational format of the schema enables efficient editing and versioning. The schema is a living system and will continue to evolve with curation. For example, solid state processing is a processing method that is not currently included and will be added to an upcoming version of the schema as data from papers utilizing it are curated. Editing the schema can be conveniently performed by retaining the original hierarchical relation and appending any new elements as child or parent to corresponding fields. For example, a new processing condition type (e.g., solid state processing) can be inserted along with associated new physical quantities (e.g., residence time) into the existing schema, retaining existing representation of all other previous terms to ensure backward compatibility.

Figure 4 illustrates a parallel coordinate plot showing an overview of Nanomine samples for some selected dimensions/parameters. The color of each line indicates the number of distinct properties recorded for each sample. From the plot, it can be seen that NanoMine has a wide range of interconnected data on each nanocomposite sample by using our designed NanoMine schema. The rich information set will enable robust discovery and development of search and visualization tools.

The NanoMine XML schema defines the parameters and relationships necessary to describe nanocomposite processing-structure-property datasets. We discuss the mapping of a superset of the NanoMine schema into a higher level centralized polymer data vocabulary, a methodology to convert between XML and a flattened excel format to enable accessibility of the data structure to domain materials research scientists, and the initial stages to transform the schema and our overall infrastructure into a robust ontology-enabled framework.

The Polymer Data Core is a high level vocabulary for polymers based on both the NanoMine schema and the NIST "materials vocabulary" initially being used in its Material Resource Registry (MRR).[41–43] The MRR aims to help researchers build, find, and share large material science resources, such as collections and repositories. Many key terms from the NanoMine schema have been incorporated into the Polymer Data Core and the higher level terms from this selection have also been added to the general NIST MRR vocabulary to provide high level descriptions for polymers and their

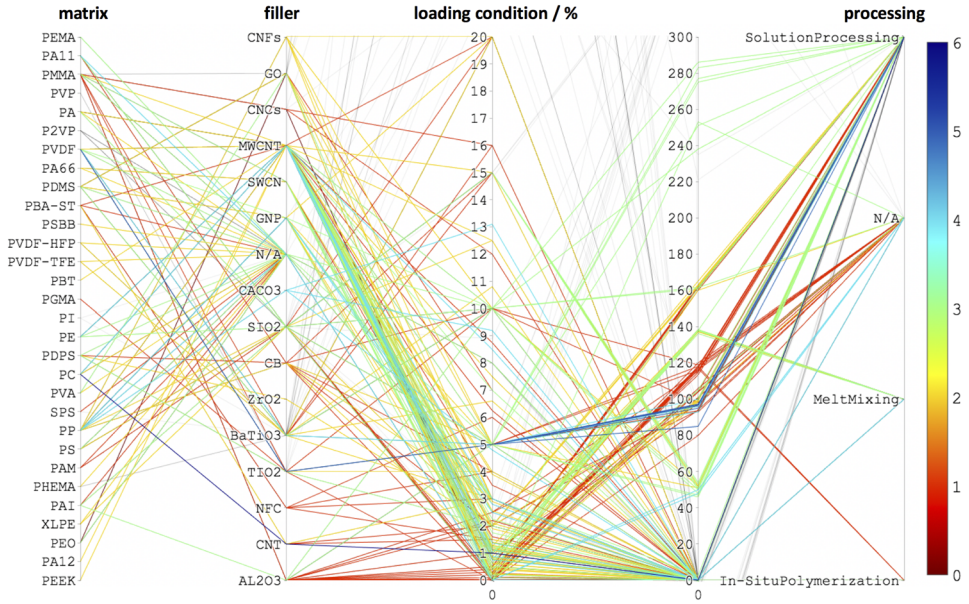

FIG. 4. Parallel coordinate plot of Nanomine samples for selected parameters (polymer, filler, loading, Tg, and characterization method).

composites that were not contained in the original MRR vocabulary. The use of common terms in describing materials will enhance the ability to cross reference and readily access and share information across current and future data repositories. Currently the general MRR materials vocabulary is a list of over 290 terms, divided into the categories of material types, structural features, properties addressed, characterization methods, computational methods, and synthesis and processing. The lack of formal definitions of the terms in this vocabulary list opens the door for more formal considerations, which are discussed in the ontology section of this paper. As an example, a small snippet of the MRR vocabulary is shown in Fig. 5 and the 2 highlighted terms were added specifically to reflect polymer nanocomposites. Overall, about a quarter of the terms in MRR are shared in common with NanoMine.

While the XML format of NanoMine schema is the official, robust, and rigorous data framework for the high dimensional data of nanocomposites, it is still helpful to be able to convert between the treelike structure of the XML and a traditional spreadsheet like MS Excel. This interoperability with excel is introduced in order to make the data entry more accessible for the typical researcher, who is not familiar with XML programing tools. The data structure captured in the full XML tree (six main branches and currently 350 distinct parameters) has been captured in an excel file where each major XML branch is a tab, and a structure has been provided to enable entry of all the individual descriptors in an easy, logical manner. While this flattened spreadsheet is convenient for data entry, the relationships have been lost. Therefore, to convert between the XML and excel tables, a parser script in Python has been written to associate keywords in each tab with parameters in the XML schema and accordingly extract key information from the excel template to XML files.[44] This parser is also used as a translator for the data uploader tools where the populated Excel spreadsheet and all associated files and images can be uploaded to the NanoMine website and curated into the repository.[40]

Figure 6 shows the electrical properties tab of the data uploader template. The first row serves as the header of data types for input cells. The first column shows the names of the attributes. The color shading denotes different rules of data input. For example, the atomic value (string or numerical value), filename (for files attached in the uploading), or drop-down menu are all denoted in separate colors. The master excel template can be downloaded on the NanoMine website[40] or in GitHub.[45]

While the master template contains all the possible entries in the schema, a typical data set will utilize only a fraction of the possible 350 terms. To further enhance accessibility of the data template, it can be customized to a more compact and accessible form on demand. To design a customized template, the papers in a specific sub-area or research group are used to identify the focus of the methods and properties and the master template is truncated accordingly. For example, for a series of papers focused on viscoelastic data from solvent processed samples, the electrical properties, melt mixing parameters, and other unrelated fields are removed resulting in a reduction of ~200 terms.[46] Such customizations provide a dramatically simpler and streamlined data template and an easier entry point for data for research groups, enhancing the likelihood of data curation. Self-curation is being currently piloted with a number of labs with positive feedback and compliant data, indicating the power of this approach to encourage data curation. Potential challenges in terms of the customization include the complication of translating the data from multiple truncated excel sheets to the single full XML schema, the method for which is currently under development.

Structural features	microstructures	nanocrystalline
Structural features	microstructures	particle distribution
Structural features	microstructures	particle shape
Structural features	microstructures	polycrystalline
Structural features	microstructures	polydispersity
Structural features	microstructures	porosity
Structural features	microstructures	precipitates

FIG. 5. A small section of the "structural features" terms in the MRR general materials vocabulary list.[43] Highlighted in red are terms added to this list to intersect with important terms for polymer nanocomposites.

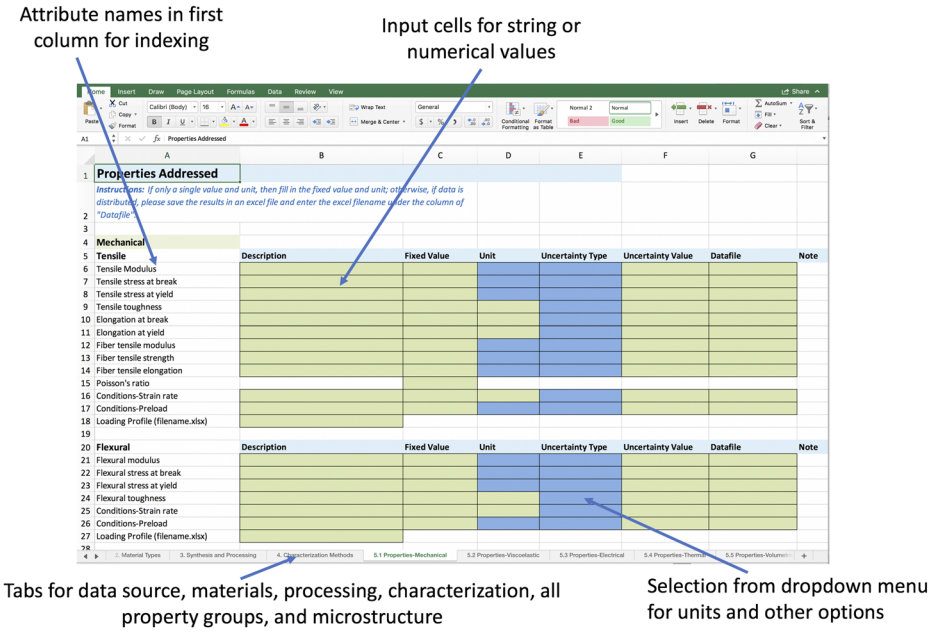

FIG. 6. Example screenshot of excel relational template for data uploading. This template is then processed by a script to convert into the XML format.

XML has been widely applied throughout the physical sciences and engineering fields to archive material data in a non-relational data structure. However, there is more than one way to map a given dataset to an XML tree, and more than one way to interpret such a mapping. As shown in Fig. 7, the volume or mass fraction of a filler in a composite could be represented accurately in each of the hierarchies shown in Figs. 7(a)–7(c). Using an ontology, it is possible to generate a standardized high-level graph that represents the material properties and all the relationships between those properties. Here we describe using our XML schema as the seed to create an ontology for nanocomposites. This ontology formalizes relationships implicit and informal in our XML schema and can act as a translator to accept multiple XML formats, enhancing the ability to share across different data resources.

Figure 7(d) shows a section of the Resource Description Framework (RDF) graph of the same nanocomposite shown in Fig. 7(a), using terms from the NanoMine Ontology. RDF is an abstract model for data and knowledge representation that provides a formal semantic interpretation of data. This interpretation is possible because RDF uses Uniform Resource Identifiers (URIs), a superset of Uniform Resource Locators (URLs), as unambiguous identifiers—URIs can only refer to one thing, and every data producer who uses that URI agrees that they are talking about the same thing. Web Ontology Language (OWL) ontologies use RDF as a representational foundation, giving it the same clear semantics, and allows data producers to provide metadata, such as definitions of terms, in the same representation as the data itself. RDF-based query languages can provide an integrated view of data and its definitions, giving the software a better ability to "understand" the data it is processing.[47,48] OWL also has additional expressive power over XML, so it can also be used to capture more detailed relationships between terms.

Originally, ontology was defined as "a branch of metaphysics concerned with the nature and relations of being" which comes from philosophy.[49] One of the most quoted definitions of ontology for computer science is as follows: an ontology is "a specification of a conceptualization."[50] A more detailed definition from Noy and McGuinness, 2001,[49] is as follows: "an ontology is a formal, explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions))."

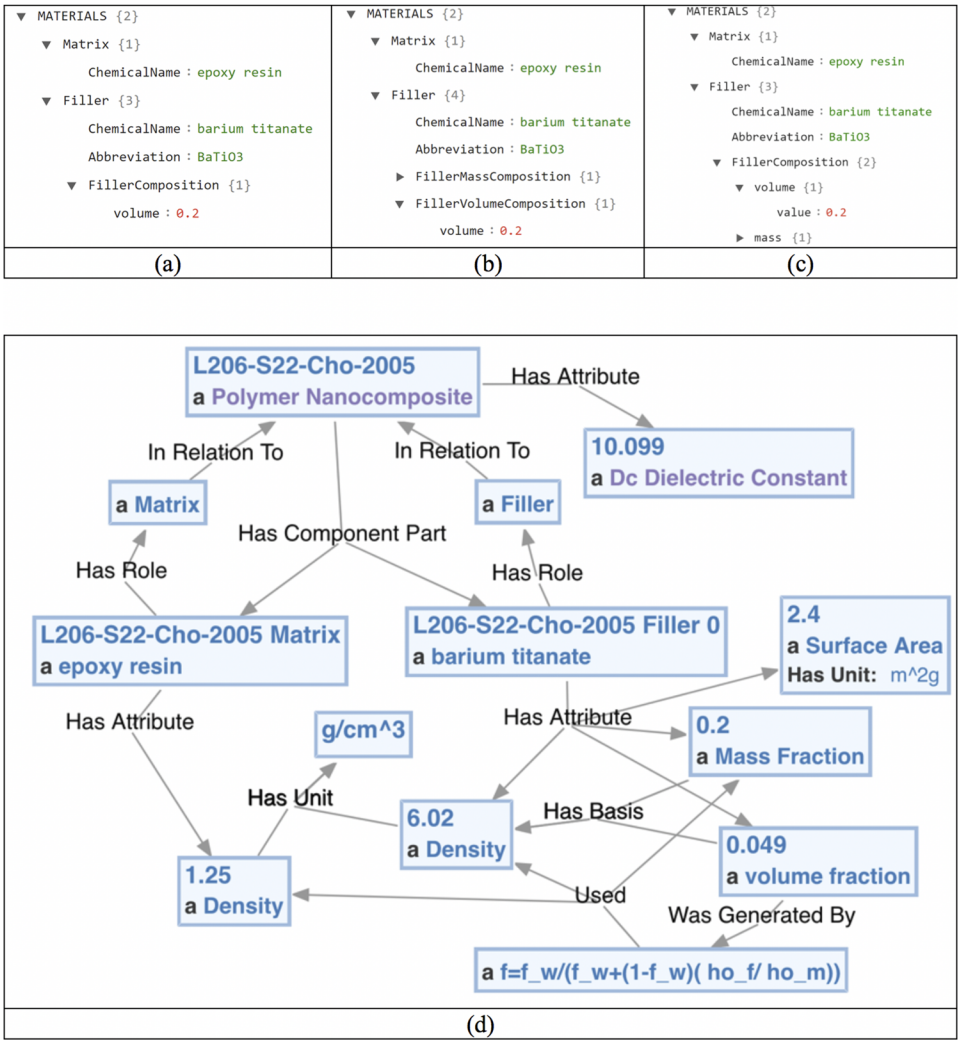

FIG. 7. (a)–(c) are several possible XML tree representations that can each adequately represent this small portion of the XML tree, where they corresponds to our implementation. (d) shows the OWL-based representation of the knowledge initially represented using the XML. Note that in RDF, we are able to express that both the volume and mass fraction are based on the density of the filler material and are thus inherently related to one another.

Informally, an ontology can be viewed as a taxonomy of *classes* of objects that are defined by the attributes and relationships between them. The ontology contains a rigorous set of definitions of its objects, including possible attributes for each object, constraints, and allowed values. In the NanoMine ontology framework, each kind of material descriptor is a "class" and the specific value of that descriptor for a given sample is an "instance" (or, equivalently, an "object"). For example, each field prefixed with an "a" in Fig. 7(d) is a *class*, with the specific *instance* for this sample provided above it. Table I illustrates the definitions underlying a class and an instance of that class in the ontology. The left column provides the human readable definition of the class "Polymer Nanocomposite" and the right column contains the machine-readable definition. "Nanocomposite" and "Polymer" each have their own formal definitions as well (not shown). Volume fraction, another class shown in Fig. 7(d), also has the definition "*The volume of a constituent divided by the volume of all constituents of the mixture prior to mixing*" ("The NanoMine ontology is available at https://github.com/tetherless-world/nanomine-ontology").

The relationships of the objects in the ontology are defined in a directed labeled graph, represented in an RDF graph, as sets of *subject–predicate-object* statements. Figure 7(d) shows a small portion

TABLE I. Example of definitions from the ontology for polymer nanocomposites and their usage to represent a specific sample (Sample L206-S22-Cho-2005).

Class: "Polymer nanocomposite"	
Human readable definition	Machine readable definition in OWL
"*Polymer nanocomposites (PNC) consist of a polymer or copolymer having nanoparticles or nanofillers dispersed in the polymer matrix. These may be of different shape (e.g., platelets, fibers, spheroids), but at least one dimension must be in the range of* 1–50 *nm*"	"*Polymer Nanocomposite: Nanomaterial and hasPart some (Polymer that hasRole some Matrix) and hasPart some (Nano-particle that hasRole some Filler)*"
Example (**instance**) of the definitions using L206-S22-Cho-2005	
Human readable description	Machine readable description in RDF turtle
"*Polymer nanocomposite* L206-S22-Cho-2005 *consists of a PMMA matrix containing Silica nanofillers*"	*"*L206-S22-Cho-2005 *hasPart [* *a PMMA;* *hasRole [* *a Polymer;* *inRelationTo Sample 1c* *]* *], [* *a Silica;* *hasRole [* *a Filler;* *inRelationTo Sample 1c* *]* *]."*

of the ontology including the relationships of the terms in Figs. 7(a)–7(c), based on the semantic interpretation we developed for the XML files. For example, Fig. 7(d) states in *subject–predicate-object* form: the *L*206-*S*22-*Cho*-2005 *Filler* 0 (subject), which is of class "barium titanate"—*has attribute* (predicate)—of value 0.2 (object) in the class "mass fraction." Note that Fig. 7(d) is rendered from subject-predicate-object statements using instances. The instances of each class for the specific sample are provided in the box above the class identifier. Specifically, L206-S22-Cho-2005 Filler 0 is the instance of the class barium titanate in this sample and 0.2 is the instance of the mass fraction of barium titanate in this sample.[62]

By using a well-structured API and knowledge representation, algorithms (and sophisticated humans) can parse and consume the data found in our knowledge graph without a detailed knowledge of the data or the API. While "the knowledge graph" has been defined in a variety of manners in the literature, we define a knowledge graph as a graph (or network), composed of a set of meaningful labeled links between entities (vertices) using unambiguous identifiers, justification for why the knowledge is true, and a limited set of link labels.[51] Here we use the RDF framework coupled with the NanoMine ontology to create the NanoMine knowledge graph. From a machine learning standpoint, exposing the data through a semantic API enables discovery and analytics. For example, data stored in XML as shown in Figs. 7(a)–7(c) can be fully understood only if the consumer of the data already understands what a filler is and that the filler is part of a nanocomposite material. Without that knowledge, the consumer (whether a human or a machine) can only infer from the XML tree that "filler" is a child of "materials." By using an ontology language to represent the relationships between elements, we are able to encode more precise information in our knowledge graph using those well-defined and unambiguous terms. This allows the human and machine, for example, to understand that a nanocomposite is a material that is composed of a matrix and a filler and that this filler has a filler composition defined by the volume or mass fraction, which are in turn related by known expressions. The meta-knowledge comes from the definitions of classes and relationships in the ontology.

The expressiveness of the ontology is very useful for validation of data within the knowledge graph that is submitted to the database, specifying relationships and constraints, through a set of

validation queries against the knowledge graph. These explicit relationships between elements of the ontology enable validation through configuration, where the ontology defines the behavior of the validators and the way that the data in NanoMine can be viewed, parsed, and analyzed. For example, a validator for NanoMine checks that a filler has a stated volume or mass fraction and that the value falls between the bounds of 0 and 1. A data set that violates these expectations is flagged for manual review.

Additionally, the structure of the ontology also allows straightforward additions of relevant physics relationships between the objects. For example, there is a mathematical relationship between the volume fraction and mass fraction by way of density, as show in Fig. 7(d). The density of the filler and matrix can be used to infer the volume fraction from the mass fraction and vice-versa, but only if we encode that knowledge. With a well-managed knowledge graph and an ontology that encodes those relationships, it is possible to automatically add relevant conversions to the graph as they are available, simply by specifying an equation and providing bindings of the equation variables to specific properties. Equation-solving software provides the missing values by solving for the unknown variable. Similarly, some metadata about materials (like density) are available from public databases such as PubChem, even if the original nanocomposite paper did not include a density value. Utilizing this resource together with coded inferences, existing knowledge (like density) and inferred knowledge (like computed the volume fraction from mass fraction) can be automatically incorporated rather than manually curated into the knowledge graph. The records can therefore be the minimal set of knowledge needed to bootstrap a fully-realized set of knowledge about nanocomposite samples.

Figure 8 shows a larger section of the XML tree and the associated knowledge graph for a specific sample in NanoMine. While the structure of the XML tree [Fig. 8(a)] is useful for data curation, it does not encode everything that might be known about material or property interrelationships. Figure 8(b) shows the same sample information encoded in our knowledge graph. We have developed a Semantic Extract, Transform, and Load (SETL) script[52] that provides a semantic interpretation to the XML files for inclusion in the knowledge graph. This knowledge is backed by the ontology, which enriches the graph with definitions of entities in the graph, clear indication of the provenance of the knowledge, and indication of where the knowledge was curated from. The journal article metadata aligns with existing standards, making it easy to integrate metadata published through the DOI system. The material types in turn can link to entities in PubChem, which provide additional knowledge about materials. The ontology is able to also provide knowledge about relationships between properties at a higher level—we can support questions about, for instance, which electrical properties co-vary, and how much semantic similarity they have.

This work demonstrates a preliminary version of our use of an ontology to provide semantics to the NanoMine schema. A well-defined XML schema assisted in the initial construction of our ontology which builds off of the Semanticscience Integrated Ontology (SIO)[53] and the W3C Provenance Ontology (PROV-O).[54] We developed the ontology by collecting data-bearing elements from the XML schema into a spreadsheet, then mapped those elements to classes, reusing existing classes from our core ontologies. We then added human-readable labels and definitions, along with machine-readable metadata about the default units of measure. The spreadsheet was then transformed into OWL/RDF using SETLr, a powerful tool for creating RDF from tabular sources.[31] The resulting ontology was loaded into our knowledge graph. Further, SETLr[52] has allowed us to generate the knowledge graph from the XML described by that schema. Further refinement of the XML and conversion process has allowed us to begin connecting entities in the knowledge graph to external knowledge sources like PubChem[55,56] and additional ontologies like the Units Ontology.[57]

This approach, along with the use of a general upper ontology, allows us to easily expand the NanoMine ontology into other materials science domains by expanding the modeling of properties, processes, and characterization methods. In our ongoing work, the knowledge graph is being expanded to answer semantic queries and fill in the gaps between existing information and desired solutions. It will also be expanded to incorporate data analytics and mining for new materials discovery and optimization built on the techniques we developed in this area.[6,58–60] We use the nanopublication framework,[61] Fig. 8(b), to track provenance from particular data sources or computational processes.

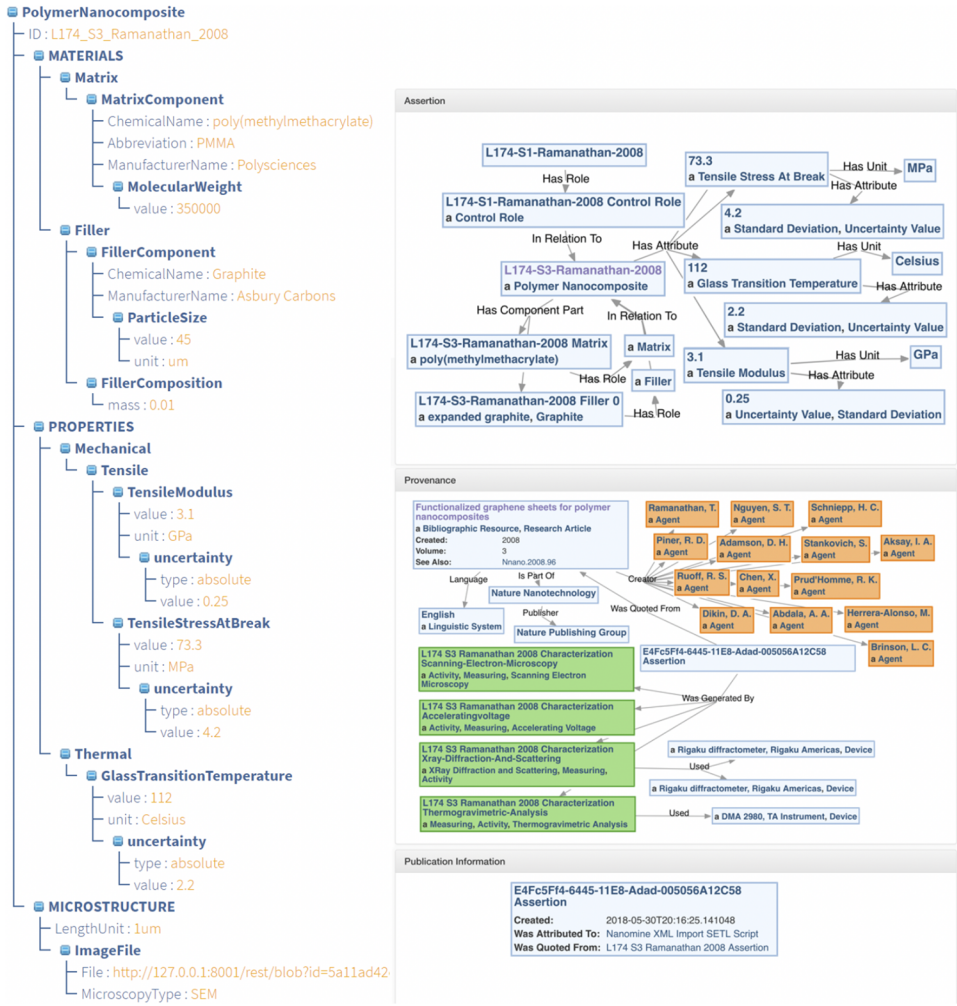

FIG. 8. (a) Part of an XML tree served the function for data curation and storage. (b) Comparable knowledge stored in a knowledge graph fragment, called a nanopublication, showing clear definitions of entities in the graph, indication of the provenance of the knowledge, and indication of where the knowledge was curated from.

In this work, we describe an XML-based materials data schema for polymer nanocomposites, representation, and ingest methods for data, and the use of a schema as a seed for a rigorous ontology and associated knowledge graph framework. The XML schema provides a robust, hierarchical relationship between provenance, composition, processing, structure, and property parameters in a nanocomposite sample and has been deployed in the NanoMine data system. The developed schema for nanocomposites has been deliberately integrated with higher level materials data vocabularies to enable increased ongoing interoperability of platforms. While non-relational in nature, the NanoMine schema has also been translated into a relational representation for facile data ingest from users, and additionally focused to create shorter, custom templates to encourage self-curation of data. From the NanoMine XML schema with over 350 parameters, we developed an initial NanoMine ontology, including rigorous definitions for each parameter, semantic data representation, and encoding relationships to enable knowledge graph development. We implemented a NanoMine ontology-enabled framework that includes the NanoMine Knowledge graph and can be used to support versatile data exchange and distribution and to assist in data exploration and play a crucial role in new nanocomposite materials development.

As a class of materials, nanocomposites contain a broad range of material constituents with vastly different processing techniques, reporting a wide array of physical material properties. The

NanoMine schema and ontology capture the key features and have enabled curation of 182 papers (over 1000 unique samples) into the system. Search, visualization, and analytic tools, which will enable new knowledge to be extracted from the organized curated data, are part of the platform and ongoing development.

Challenges and opportunities include implementation of machine learning algorithms on sparse datasets, as each individual paper reports only a fraction of possible parameters. In addition to algorithms developed to handle sparse data, the ontological framework enables development of inference agents to provide robust estimates of missing data, leveraging both internal and external resources. Another opportunity is developing intelligent agents to couple related data and enable physical understanding—for example, using microstructure characterization tools on provided images to obtain structural parameters which can be quantitatively linked to processing parameters or physical properties. Curation of the literature is an ongoing significant challenge. The papers to date have been manually curated into the system, which is time consuming and slow. Significant efforts are underway, however, to make the current curation process scalable and tractable. For example, natural language processing (NLP) techniques have been demonstrated as a potential solution to automatically extract targeted data and information from the scientific literature[16,17] and these methods are being adapted into the nanocomposite domain. Pilot projects are exploring author self-curation, using some of the accessible data transfer methods described in this paper. These and other developments will greatly accelerate the rate of data incorporation into this and other data resources.

As new subareas of nanocomposite research are curated, the flexible and robust format of the schema and ontology will enable continued expansion to incorporate new terms and relationships as needed. We also anticipate expanding from the current experimentally focused schema to add a branch to account for computational modeling, with parameters describing conditions used in simulations of nanocomposite behavior. Overall, the development of this schema and ontology will enable researchers to develop and test broad-reaching hypotheses about how inter-relationships between different materials processing methods and composition result in specific changes in material properties.

The authors gratefully acknowledge support of NSF (DMR-1310292), NSF DIBBS A12761, 1640840, NSF DMREF 1818574, 1729743, CMMI – 1729452, ACI - 1640840, NIST (70NANB14H012 Amd 5) and the CHiMaD center based at the Northwestern University.

[1] L. Schadler, L. Brinson, and W. Sawyer, "Polymer nanocomposites: A small part of the story," Jom **59**(3), 53–60 (2007).
[2] J. K. Nelson, "Overview of nanodielectrics: Insulating materials of the future," in *Electrical Insulation Conference and Electrical Manufacturing Expo, 2007* (IEEE, 2007).
[3] T. Ramanathan *et al.*, "Functionalized graphene sheets for polymer nanocomposites," Nat. Nanotechnol. **3**(6), 327 (2008).
[4] P. M. Ajayan, L. S. Schadler, and P. V. Braun, *Nanocomposite Science and Technology* (John Wiley & Sons, 2006).
[5] J. Jordan *et al.*, "Experimental trends in polymer nanocomposites—A review," Mater. Sci. Eng.: A **393**(1-2), 1–11 (2005).
[6] Y. Wang *et al.*, "Identifying interphase properties in polymer nanocomposites using adaptive optimization," Compos. Sci. Technol. **162**, 146–155 (2018).
[7] C. E. Campbell, U. R. Kattner, and Z.-K. Liu, "The development of phase-based property data using the CALPHAD method and infrastructure needs," Integr. Mater. Manuf. Innovation **3**(1), 12 (2014).
[8] 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, "The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).
[9] G. Ceder and K. Persson, The Materials Project: A Materials Genome Approach, 2010.
[10] H. L. Lukas, S. G. Fries, and B. Sundman, *Computational Thermodynamics: The Calphad Method* (Cambridge University Press, Cambridge, 2007), Vol. 131.
[11] S. Kirklin *et al.*, "The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies," npj Comput. Mater. **1**, 15010 (2015).
[12] J. E. Saal *et al.*, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," Jom **65**(11), 1501–1509 (2013).
[13] B. Ellis and R. Smith, *Polymers: A Property Database* (CRC Press, 2008).
[14] S. Otsuka *et al.*, "PoLyInfo: Polymer database for polymeric materials design," in *International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2011* (IEEE, 2011).
[15] H. Zhao *et al.*, "Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design," APL Mater. **4**(5), 053204 (2016).
[16] E. Kim *et al.*, "Materials synthesis insights from scientific literature via text extraction and machine learning," Chem. Mater. **29**(21), 9436–9444 (2017).
[17] T. Onishi, T. Kadohira, and I. Watanabe, "Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity," Sci. Technol. Adv. Mater. **19**(1), 649–659 (2018).

[18] S. R. Kalidindi, "Data science and cyberinfrastructure: Critical enablers for accelerated development of hierarchical materials," Int. Mater. Rev. **60**(3), 150–168 (2015).
[19] See http://pppdb.uchicago.edu/ for Polymer Property Predictor and Database. [cited 2018 05/26].
[20] C. Calebrese *et al.*, "A review on the importance of nanocomposite processing to enhance electrical insulation," IEEE Trans. Dielectr. Electr. Insul. **18**(4), 938 (2011).
[21] K. Wakabayashi *et al.*, "Polymer–graphite nanocomposites: Effective dispersion and major property enhancement via solid-state shear pulverization," Macromolecules **41**(6), 1905–1908 (2008).
[22] L. Perkins, E. Redmond, and J. Wilson, *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement* (Pragmatic Bookshelf, 2018).
[23] G. Harrison, *Next Generation Databases: NoSQL and Big Data* (Apress, 2015).
[24] E. Begley and C. P. Sturrock, "MatML: An XML for standardizing web-based materials property data," JOM **52**(7), 56 (2000).
[25] H. S. Thompson *et al.*, "W3C XML schema definition language (XSD) 1.1 Part 1: Structures," The World Wide Web Consortium (W3C), W3C Working Draft Dec, 2009, Vol. 3.
[26] R. D. Chirico *et al.*, "ThermoML an XML-based approach for storage and exchange of experimental and critically evaluated thermophysical and thermochemical property data. 2. Uncertainties," J. Chem. Eng. Data **48**(5), 1344–1359 (2003).
[27] See https://www.nist.gov/mml/acmd/trc/thermoml for NIST. ThermoML. [cited 2018 May 29].
[28] A. S. Varde, E. F. Begley, and S. Fahrenholz-Mann, "MatML: XML for information exchange with materials property data," in *Proceedings of the 4th International Workshop on Data Mining Standards, Services and Platforms* (ACM, 2006).
[29] A. Varde *et al.*, "Semantic extensions to domain-specific markup languages," *Proceedings of IEEE's CCCT* (IEEE, 2004), pp. 55–60.
[30] A. Stawowy, R. Wrona, and A. Macioł, "CastML—A language for description of casting products and processes," Arch. Foundry Eng. **8**(4), 205–208 (2008).
[31] N. R. Council, *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security* (The National Academies Press, Washington, DC, 2008), p. 152.
[32] J. P. Holdren, *Materials Genome Initiative for Global Competitiveness* (National Science and Technology Council OSTP, Washington, USA, 2011).
[33] A. Dima *et al.*, "Informatics infrastructure for the materials genome initiative," Jom **68**(8), 2053–2064 (2016).
[34] See http://dublincore.org/documents/dces/ for *Dublin Core* Metadata Element Set. [cited 2018 05/29].
[35] See https://github.com/Duke-MatSci/doi-crawler for *DOI crawler*. [cited 2018 08/10].
[36] H. Xu *et al.*, "A descriptor-based design methodology for developing heterogeneous microstructural materials system," J. Mech. Des. **136**(5), 051007 (2014).
[37] H. Xu *et al.*, "Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials," Comput. Mater. Sci. **85**, 206–216 (2014).
[38] R. Bostanabad *et al.*, "Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques," Prog. Mater. Sci. (2018).
[39] See https://github.com/Duke-MatSci/nanomine-schema/tree/master/xml for Nanomine Schema. [cited 2018 06/24].
[40] See http://www.nanomine.org for *Nanomine Homepage*. [cited 2018 06/24].
[41] See https://www.rd-alliance.org/polymer-data-core-vocabulary-draft-september-2017 for Polymer Data Core Vocabulary. 2017 [cited 2018 05/29].
[42] See https://www.rd-alliance.org/plenary-10-evolving-materials-resource-registry-vocabulary-0 for Plenary 10 Evolving Materials Resource Registry Vocabulary. [cited 2018 06/25].
[43] See https://www.rd-alliance.org/materials-vocabulary-draft-21-mar-2017 for Materials Resource Registry Materials Science Vocabulary. [cited 2018 05/29].
[44] See https://github.com/bingyinh/nanomine_xlsx2xml for NanoMine Excel Parser. [cited 2018 08/10].
[45] *NanoMine master excel template*. [cited 2018 06/28].
[46] See https://github.com/Duke-MatSci/nanomine-schema/blob/master/xls-input-forms/master_template_visco_custom.xlsx for Customized template for viscoelastic properties of nanocomposites. [cited 2018 08/10].
[47] F. Manola, E. Miller, and B. McBride, "RDF primer," W3C Recommendation **10**(1-107), 6 (2004).
[48] P. Hitzler *et al.*, "OWL 2 web ontology language primer," W3C Recommendation **27**(1), 123 (2009).
[49] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, Stanford, CA, 2001.
[50] T. R. Gruber, "A translation approach to portable ontology specifications," Knowl. Acquis. **5**(2), 199–220 (1993).
[51] J. P. McCusker *et al.*, "What is a knowledge graph," Semantic Web J. (submitted).
[52] See https://github.com/tetherless-world/setlr/wiki/SETLr-Tutorial for *SETLr* [cited 2018 05/29].
[53] M. Dumontier *et al.*, "The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery," J. Biomed. Semantics **5**(1), 14 (2014).
[54] See https://www.w3.org/TR/prov-o/ for PROV-O: The PROV Ontology. 2013 [cited 2018 05/29].
[55] See https://pubchem.ncbi.nlm.nih.gov/ for *PubChem* [cited 2018 05/29].
[56] E. E. Bolton *et al.*, "PubChem: Integrated platform of small molecules and biological activities," in *Annual Reports in Computational Chemistry* (Elsevier, 2008), pp. 217–241.
[57] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, "The units ontology: A tool for integrating units of measurement in science," Database **2012**, bas033.
[58] H. Xu *et al.*, "A machine learning-based design representation method for designing heterogeneous microstructures," J. Mech. Des. **137**(5), 051403 (2015).
[59] I. Hassinger *et al.*, "Toward the development of a quantitative tool for predicting dispersion of nanocomposites under non-equilibrium processing conditions," J. Mater. Sci. **51**(9), 4238–4249 (2016).

[60] Y. Zhang *et al.*, "Microstructure reconstruction and structural equation modeling for computational design of nanodielectrics," Integr. Mater. Manuf. Innovation **4**(1), 14 (2015).
[61] J. McCusker *et al.*, "A nanopublication framework for biological networks using cytoscape. Js," in ICBO, 2014, CiteSeer.
[62] Strictly speaking it would not be correct to state that "barium titanate—has attribute—of mass fraction," the reason being that the class barium titanate may not always be used as a filler in a nanocomposite and therefore may not always have a mass fraction.