A Controlled Vocabulary and Metadata Schema for Materials Science Data Discovery

PRACTICE PAPER

ANDREA MEDINA-SMITH
CHANDLER A. BECKER
RAYMOND L. PLANTE
LAURA M. BARTOLO
ALDEN DIMA
JAMES A. WARREN
ROBERT J. HANISCH

*Author affiliations can be found in the back matter of this article

ABSTRACT

The International Materials Resource Registries (IMRR) working group of the Research Data Alliance (RDA) was created to spur initial development of a federated registry system to allow for easier discovery and access to materials data. As part of this effort, a controlled vocabulary and metadata schema were developed with contributions from members of the working group and other experts. Here we describe the process, the resulting vocabulary and XML schema, and lessons learned in the development and use of the schema.

CORRESPONDING AUTHOR:
Andrea Medina-Smith
National Institute of Standards and Technology, Information Services Office, Gaithersburg, MD, United States of America
andrea.medina-smith@nist.gov

KEYWORDS:
controlled vocabulary; metadata schema; materials science; RDA; federated registry system

TO CITE THIS ARTICLE:
Medina-Smith, A, Becker, CA, Plante, RL, Bartolo, LM, Dima, A, Warren, JA and Hanisch, RJ. 2021. A Controlled Vocabulary and Metadata Schema for Materials Science Data Discovery. *Data Science Journal*, 20: 18, pp. 1–10. DOI: *https://doi.org/10.5334/dsj-2021-018*

INTRODUCTION

In early 2016, the Research Data Alliance (RDA) Working Group for International Materials Resource Registries (IMRR) was established to bring together experts in materials science and information technology to address the problem material science researchers face of finding and accessing data related to their work. The aim was to initiate development of an international federation of registries that can be used for global discovery of data resources for materials science. At a basic level, a resource registry makes available high-level metadata descriptions of resources such as data repositories, archives, websites, and services that are useful for data-driven research, not unlike a library's catalog. By making the collection searchable, it aids scientists across the discipline to discover data relevant to their research and work interests. With supporting infrastructure, the data can then be obtained and used as part of a larger ecosystem (Dima et al. 2016).

This paper presents part of this successful pilot of a registry federation for materials science data discovery. In particular, we cover how the eXtensible Markup Language (XML) defines our schema, which incorporates both generic and materials science-specific metadata. The domain-specific metadata are based on a high-level Materials Science Vocabulary developed as part of this effort. Finally, we outline an approach to schema definition based on extensions that enable the schema to evolve over time in a tractable way.

Developing a successful international materials science resource registry requires a combination of technical and social processes. The latter are important for establishing consensus around standards. The RDA Working Group was especially helpful in collecting input on a common Materials Science Vocabulary and getting contributions of resource descriptions from the global community. The pilot registry federation currently holds more than 350 resource description records distributed across two registry instances located at NIST (*https://materials.registry.nist.gov*) and the Materials Data Facility (*https://mrr.materialsdatafacility.org*). Some of these records, and an initial vocabulary focused on software resources, were created for the MGI Code Catalog (MGI Code Catalog 2015) and migrated for this effort. The software deployed to implement our pilot federation is a product called the Materials Resource Registry (MRR; Brady et al. 2019), illustrated in ***Figures 1*** and ***2***. More information on the federated registries architecture and implementation is available in a companion paper (Plante et al. *in preparation*).

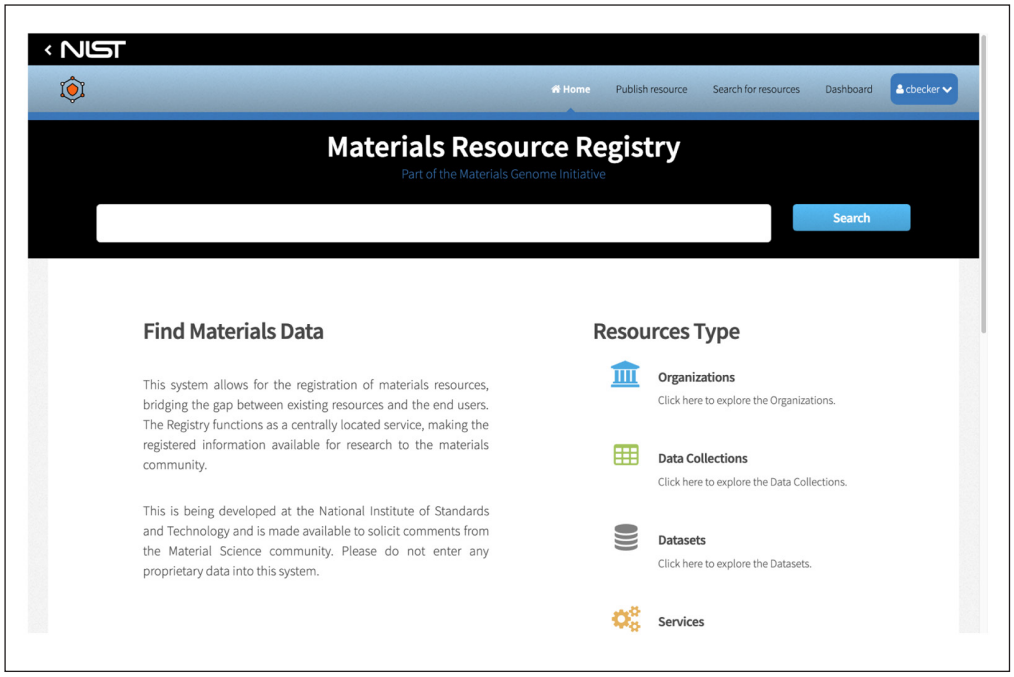

Figure 1 The main page for the NIST Materials Resource Registry, with options for publishing resources, searching for resources, and record management.

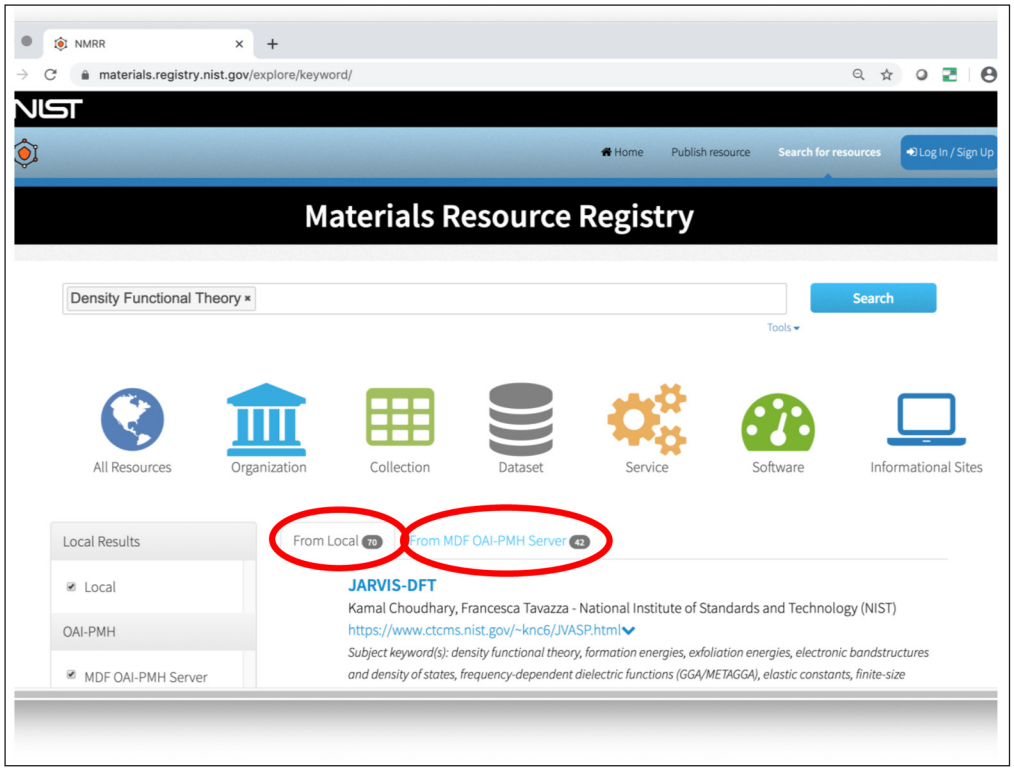

Figure 2 A search for "Density Functional Theory" returned 70 results from the NIST Materials Resource Registry instance, as well as 42 from the CHiMaD Materials Data Facility (MDF) instance.

THE RESOURCE METADATA

The uses a metadata schema encoded using XML Schema (Fallside and Walmsley 2004). XML as a metadata format satisfies the key format requirements of the registry system:

- XML Schema provides a means to define the schema in a formal way,
- XML namespaces provide a means to identify a schema via a URL and avoid collisions that may arise when the same terms are used in different contexts, and
- Open software is available to validate resource description documents against the XML Schema definition.

We note if we had used a different metadata format—namely, JSON—these features would still be critical.

The schema we assembled drew on existing schemas and vocabularies, most notably Dublin Core (DCMI Usage Board, 2012), DataCite (DataCite Metadata Working Group, 2017), and the Virtual Observatory's Resource Metadata for the generic, domain-unspecific concepts (Plante et al. 2008). We also reviewed the state of materials science-related vocabulary and ontology activities at that time in the hopes of adopting an existing set of terms for compatibility; we found that, while there was value in existing work, it did not satisfy the requirements for this system (i.e., high-level, general, and broad coverage of materials science concepts). This is described in detail in the Materials Science Vocabulary section.

The importance of supporting metadata extensibility and evolution was an essential consideration based on experience with the Virtual Astronomical Observatory and reinforced here. The metadata standard will need to be updated over time, not just to correct mistakes but to add more concepts to support new functionality. Because metadata validation is built right into the application, all participating registries share a common basis for validation. For our pilot, this centers on ensuring that the registries have the same XML Schema definition document against which to validate records. Updating the schema can be disruptive as it involves not only redistributing and installing the new schema document to all the participating registries, but also updating existing records to the new standard and possibly updating the software. Thus, in the existing system, updates should be done with consideration and deliberation.

The Virtual Astronomical Observatory developed techniques for defining XML schemas that greatly mitigate the disruption caused by schema evolution. These techniques are based on a common core metadata schema and evolution accomplished through pluggable extensions to

that core (Plante et al. 2008). Likewise, our metadata schema is based on the Virtual Observatory approach with adjustments made to accommodate the current state of the software. We are further developing the registry software to take advantage of metadata extensions and make it more robust to an evolving metadata schema. This is described in more detail in the related Materials Resource Registry architecture paper (Plante et al. *in preparation*).

The GitHub[1] repository, mgi-resmd, captures the development of the metadata schema developed for use by our pilot. Because our general metadata model is designed to be extensible, our ideal schema would be organized as one schema file representing the core schema and additional schema files defining extensions (Plante et al. 2018). For integration with our registry software, we combined all definitions into a single schema document, *https://github.com/usnistgov/mgi-resmd* (Plante et al. 2018). The XML Schema file includes full documentation; in particular, each element that can accept a value has a definition spelling out the semantic meaning of the element.

THE METADATA MODEL

In this section, we summarize the overall high-level metadata design, as illustrated in ***Figure 3***. Readers can consult the schema file itself for precise definitions of individual metadata terms.

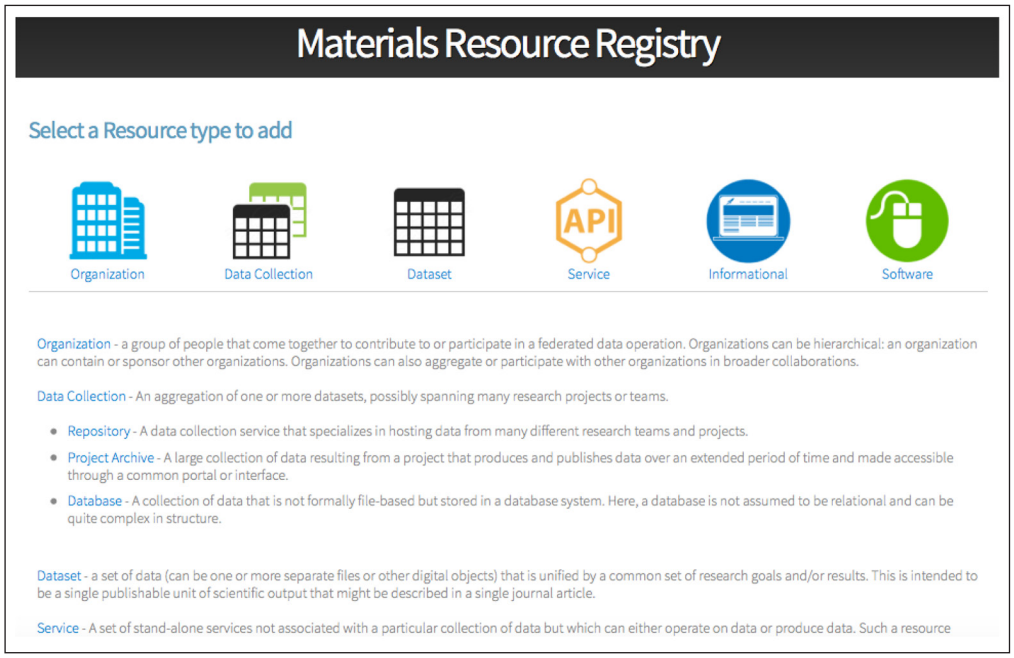

Figure 3 Resource types to add include Organization, Data Collection, Dataset, Service, Informational sites, and Software, with descriptions for each.

Our data resource metadata model reflects a few core principles:

- Our model separates the generic metadata from the domain-specific metadata.
- There are different *types* of resources — e.g., repositories, databases, web sites, and software — and while some metadata apply to all (or most) types of resources, we will also need to employ type-specific metadata to describe them. A resource may also belong to multiple types simultaneously.
- Because materials science overlaps heavily with other areas of science (physics, chemistry, biology, etc.), it is necessary to leverage metadata from different domains simultaneously within the resource description.
- We must identify multiple points for extensibility: in the future, we want to support new types of resources or plug in new domain-specific metadata.

1 *Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.*

A resource description using our schema is divided into sections (where each section is potentially extendable). The sections containing generic metadata include:

- **Identity** – how the resource is named and referenced
- **Providers** – who is responsible for the resource
- **Role** – what type of resource it is (e.g., database, web portal, data collection, software, etc.)
- **Content** – what the resource is about and what it contains
- **Access** – how one can access the resource
- **Related** – other related resources

A resource description can have more than one **Role** section, each describing its role as a different type of resource. The types (and subtypes) of resources we currently support are:

- Organization
  - Institution
  - Project
- Data Collection
  - Repository
  - Archive
- Dataset
  - Database
- Service
  - Application Programming Interface (API)
- Software

Where appropriate, a **Role** section can have additional type-specific metadata included with it.

We note that wherever the schema can refer to another resource, it is the best practice to do so via a global identifier. The Identity section supports associating a resource with multiple identifiers including a DOI and the identifier assigned by the registry.

In addition to the generic metadata sections, an additional section, **Applicability**, is defined in order to capture domain-specific metadata. Specifically, an Applicability section captures metadata that describe how the resource *applies* or relates to a particular domain. A resource description can have multiple Applicability sections, each leveraging domain metadata from a different domain. The intent is that consumers of the metadata document will interpret the Applicability sections for domains it understands and ignore those that it does not. For this reason, it is acceptable if the different domains include metadata that overlap in their semantics. XML namespaces are the technology used to avoid collisions between the schema.

For our pilot, we defined an Applicability section for materials science that leverages in large part the materials science vocabulary discussed in the next section.

THE MATERIALS SCIENCE VOCABULARY

The materials science vocabulary defines controlled terms that identify attributes of materials and material research. Using a controlled vocabulary provides a number of advantages that simplifies creating records and searching for records. This vocabulary was not meant to exhaustively cover all domains of materials science at all levels; rather, it was intended to assist with discovering high level data resources described in the registry (Plante et al. in preparation); consequently, it focuses on attributes of data and data service collections rather than individual datasets.

The process of developing the vocabulary for this application began in 2015 and involved examining then-existing work in the area (Zhang, 2015; *Matml.org*. 2005; *Trc.nist.gov*. 2006; Cheung et al. 2009; Bhat, 2015; Ashino, 2010; MIF Schema. [online]; Bercaru, 2009; *Wiki.knoesis.org*. 2015), iterating with experts (including members from the RDA-IMRR Working Group) and making use of the terms in the MRR pilot application to refine the vocabulary. This

refinement took the form of discussions at Working Group meetings, emails, other discussions, participation in a VoCamp workshop (November 2016), and feedback from users who were registering their own resources. This general process is shown in ***Figure 4*** with the inner loop representing changes within a single version and the outer loop representing major revisions. We also attempted to be consistent with the draft Polymers Core vocabulary being developed at NanoMine (*Materialsmine.org*., 2014; Zhao et al. 2018; *Rd-alliance.org*. 2017). These are not the only efforts in the discipline; we note that another materials polymer data repository, Polymer Property Predictor and Database (PPPDb, n.d.) is incorporating the summary description format of PolyInfo (*Polymer.nims.go.jp*, n.d.).

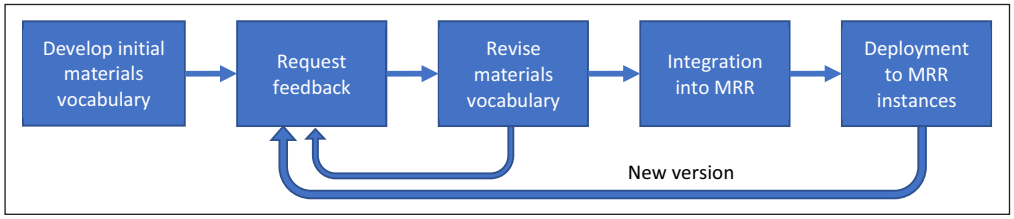

Figure 4 Process for developing, deploying, and revising materials science vocabulary for the Materials Resource Registry.

While the vocabulary originally had two levels of hierarchy, more specificity was needed and a third level was added. This structure, combined with free text fields available in the subject keyword sections, balances the need for minimal burden when entering metadata and information specific to a particular effort.

From that point the terms were normalized, and some terms deprecated in favor of those more commonly used. At each point in this process the draft versions of the vocabulary were sent out to the Working Group and comments were requested. In the end there were nearly 500 distinct terms across the hierarchy.

The vocabulary developed is a simple type of thesaurus. A thesaurus is defined as a "a specialized authority list (usually restricted to a particular subject area) of controlled vocabulary terms... terms represent single concepts together with any references, scope notes, and subdivisions... and are organized so that the relationships between concepts are made explicit." (Taylor 2006, pg 546) The Materials Science Vocabulary is hierarchical, but relationships beyond the Broader Than (BT) and Narrower Than (NT) are not currently noted, nor are there scope notes defining the terms themselves. Preferred terms were also not discussed, though these richer concepts would be useful in future versions.

Although we ultimately encoded the vocabulary into our resource description schema, we developed it originally independently of XML Schema. This was done because we expected that this vocabulary could be useful beyond the application of the registry. The Materials Vocabulary descriptions document captures the terms in a human-readable format. We created a SKOS definition of the vocabulary as well (Medina-Smith et al. 2017).

As mentioned above, the vocabulary is organized into three tiers of increasing detail. The first tier identifies attributes of materials science data, its origins, and its context. These are:

- **Data origin** (i.e., experiments, simulations, or informatic analysis)
- **Material types**
- **Structural features**
- **Properties addressed**
- **Characterization methods**
- **Computational methods**
- **Synthesis and processing**

The second and third tiers define categories and sub-categories in each of these attributes, as shown in ***Figure 5***. For example, categories of **Material types** include **ceramics**, **metals and alloys**, and **polymers** (among others). Sub-categories of polymers include **elastomers**, **liquid crystals**, and **thermoplastics**. Using a controlled vocabulary means that a data provider, when describing a dataset, can quickly check off all of the different material types the dataset explores. As a tiered vocabulary, a provider can refer to all polymers generally or specific types of polymers.

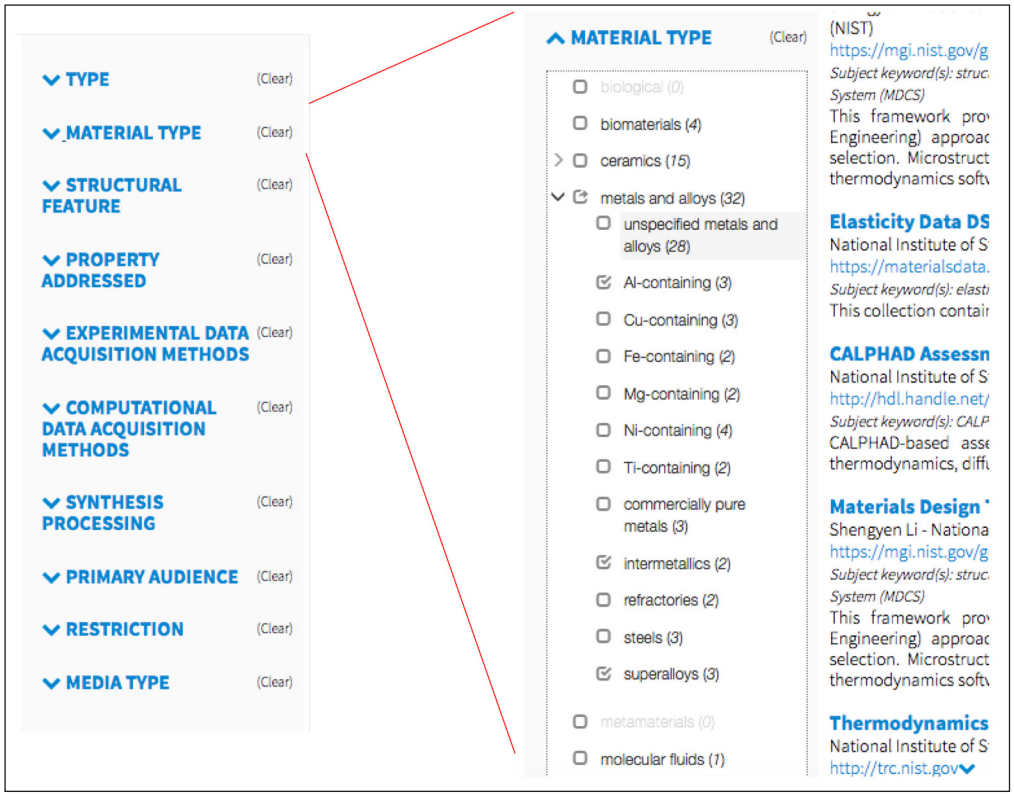

Figure 5 Details of the materials vocabulary used for tagging and filtering results in the Materials Resource Registry user interface.

The detail captured in the vocabulary was intentionally limited to the three tiers in an attempt to balance the advantages of a rich vocabulary with the increasing difficulty and overhead it incurs in making use of the detail (e.g. measured in the time it takes to interpret and select the appropriate terms). It should be noted that the MRR application allows the free text entry of keywords and descriptions. In conjunction with the controlled vocabulary, these unstructured terms allow for both high-level compatibility across MSE and the specificity necessary for materials practitioners to assess the usefulness of particular resources.

As an example of how the system can be used, a search for "interatomic potentials" is illustrated in ***Figures 6*** and ***7***. Nine records associated with that term are returned (***Figure 6***), each with a summary and link to the complete record, as well as a direct link to the resource itself. A complete record is shown in ***Figure 7***, with free-text keywords and description fields in addition to any terms selected from the controlled materials vocabulary. This listing of relevant resources is part of a growing list of resources relevant to materials scientists seeking data about this field.

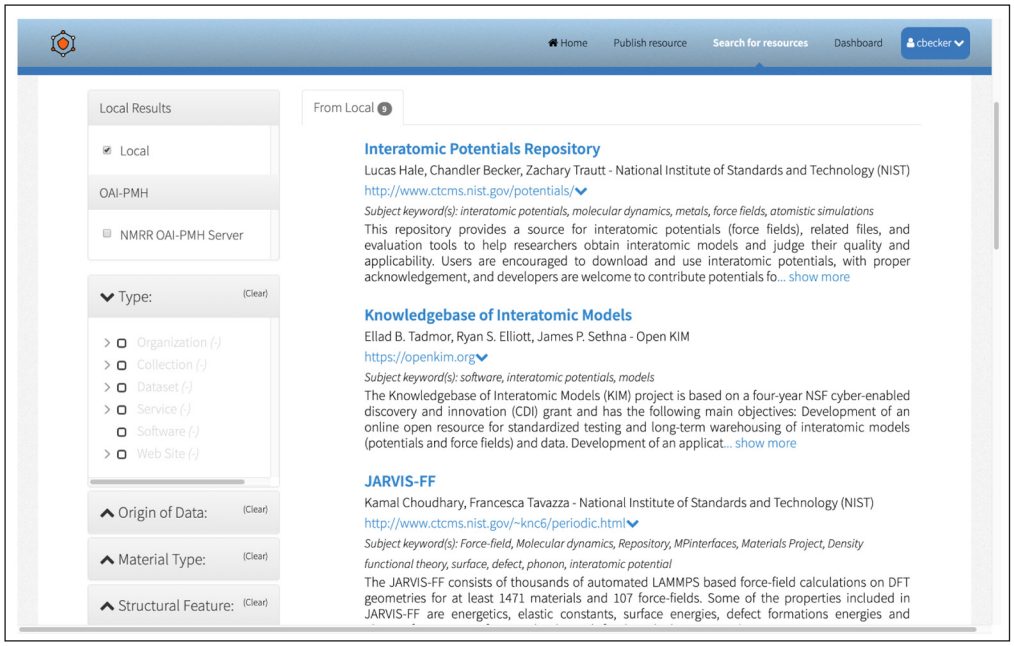

Figure 6 A search for "interatomic potentials" returns nine resources from the NIST Materials Resource Registry.

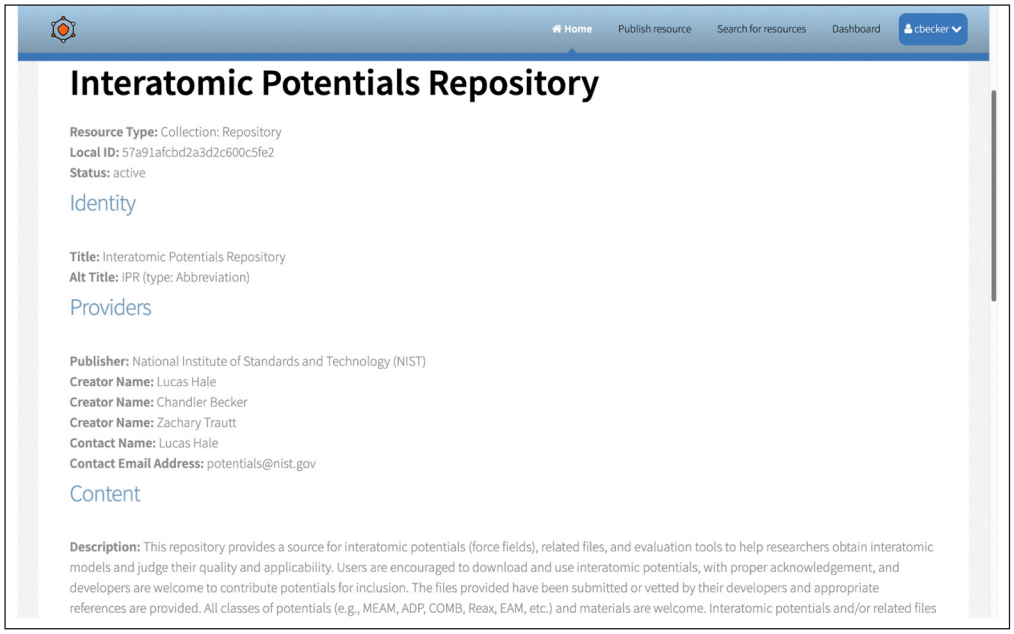

Figure 7 The record for the NIST Interatomic Potentials Repository contains information about contributors, keywords, a free-text description, and links to the project.

IMPACT AND OUTLOOK

We used the challenges of materials science research—specifically, the problem of finding materials science data—as a vehicle for exploring the more general problem of data discovery within and across all domains. It was hoped that by looking at the problem through the lens of a specific community with some well-defined needs, we could stay focused on deliverables with practical value. Nevertheless, we have kept the more general problem in view and attempted to structure our deliverables to allow for broader application in other fields. We have had success in this effort with the deployment of registries serving the metrology and greenhouse-gas research communities based on the same software and model.

While the larger Resource Registry project used the challenges of materials science research to view a complex project, the specifics of developing a controlled vocabulary for describing data resources in the material science domain was a drill-down exercise. The effort to build a community of experts in both taxonomies/ontologies and the materials science domain and then translating that expertise into a useable vocabulary was a valuable addition to the federated resource registry.

We note that that our collaboration with the Center for Hierarchical Materials Design (CHiMaD) has been important for reaching out to the materials science community because of that Center's leadership of the NSF-sponsored Midwest Big Data Spoke on Integrative Materials Design which features member institutions including the University of Chicago, Northwestern University, and the Universities of Illinois and Michigan. Each member institution of the Spoke leads significant government-funded Materials Genome Initiative programs and also incorporates a wide network of academic and industrial partners located across the Midwestern United States. With the MRR software and workflow functionality, the Materials Data Facility finds and prepopulates metadata records for the CHiMaD MRR instance; sends prepublished records to the Spoke member institutions for their expertise; and results in robust linkages of Midwest materials resources harvested and available throughout the federation of MRRs.

The metadata-specific deliverables of the RDA-IMRR Working Group can be transferred to other communities in these ways:

- We have laid out an approach to defining metadata schemas that combines generic and domain-specific metadata in an orderly way. This approach, which features a generic core with extensions for both different types of resources and metadata from different domains, allows for the schema to evolve in a tractable manner.
- We have presented a specific metadata schema based on the above principles that can be easily extended and adapted for other domains.

- We have built a community of practice around the controlled vocabulary that can be replicated by other knowledge communities.
- We have produced a specific controlled vocabulary for materials science that can be extended and used in other systems.

While the working group has finished, there are plans for moving the registry forward. Maintaining and improving metadata schema and vocabulary are important corollaries to the work on the Resource Registry software. This improvement will be most visible through the extension of the vocabulary into new and niche disciplines of materials science or into fields not yet covered by the current registry. Another way forward with the vocabulary is to formalize it into a taxonomy with preferred terms, and relationships specified between terms outside of the hierarchy. Adding scope notes will increase its usefulness. The work of updating and maintaining will be a collaborative effort involving materials scientists and information scientists. Specifically, the schema and vocabulary will be revisited through RDA working groups, particularly the RDA/CODATA Materials Data, Infrastructure & Interoperability Interest Group. It will also be discussed and revisited in other Materials Science meetings as appropriate to get additional subject-matter expert input. To support this effort, NIST will continue to maintain GitHub versions that will facilitate adoption and revisions. Development of the platform also continues through development of the Materials Resource Registry application that encodes these schema and terms.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Andrea Medina-Smith *orcid.org/0000-0002-1217-701X*
National Institute of Standards and Technology, Information Services Office, Gaithersburg, MD, United States of America

Chandler A. Becker *orcid.org/0000-0002-3653-0199*
National Institute of Standards and Technology, Material Measurement Laboratory, Office of Data and Informatics, Gaithersburg, MD, United States of America

Raymond L. Plante *orcid.org/0000-0002-9279-4877*
National Institute of Standards and Technology, Material Measurement Laboratory, Office of Data and Informatics, Gaithersburg, MD, United States of America

Laura M. Bartolo *orcid.org/0000-0002-2093-2302*
Northwestern University, Center for Hierarchical Materials Design, Evanston, IL, United States of America

Alden Dima *orcid.org/0000-0003-0547-3117*
National Institute of Standards and Technology, Information Technology Laboratory, Gaithersburg, MD, United States of America

James A. Warren *orcid.org/0000-0001-6887-1206*
National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, MD, United States of America

Robert J. Hanisch *orcid.org/0000-0002-6853-4602*
National Institute of Standards and Technology, Material Measurement Laboratory, Office of Data and Informatics, Gaithersburg, MD, United States of America

REFERENCES

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

Bercaru, C. 2009. *Ei Thesaurus*, 6th edn. New York: Elsevier Engineering Education.

Bhat, T, Bartolo, L, Kattner, U, Campbell, C and **Elliott, J.** 2015. Strategy for Extensible, Evolving Terminology for the Materials Genome Initiative Efforts. *JOM*, 67(8): 1866–1875. DOI: *https://doi.org/10.1007/s11837-015-1487-4*

Cheung, K, Hunter, J and **Drennan, J.** 2009. MatSeek: An Ontology-Based Federated Search Interface for Materials Scientists. *IEEE Intelligent Systems*, 24(1): 47–56. DOI: *https://doi.org/10.1109/MIS.2009.13*

DataCite Metadata Working Group. 2017. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*. Version 4.1. DataCite e.V. DOI: *https://doi.org/10.1007/s11837-016-2000-4*

Dima, A, Bhaskarla, S, Becker, C, Brady, M, Campbell, C, Dessauw, P, Hanisch, R, Kattner, U, Kroenlein, K, Newrock, M, Peskin, A, Plante, R, Li, S, Rigodiat, P, Amaral, G, Trautt, Z, Schmitt, X, Warren, J and **Youssef, S.** 2016. Informatics Infrastructure for the Materials Genome Initiative. *JOM*, 68(8): 2053–2064. DOI: *https://doi.org/10.1007/s11837-016-2000-4*

DCMI Usage Board. 2012. *DCMI Metadata Terms*. Dublin Core Metadata Initiative. *http://dublincore.org/documents/2012/06/14/dcmi-terms/*

Fallside, DC and **Walmsley, P.** 2004. *XML Schema Part 0: Primer Second Edition W3C Recommendation 28 October 2004*. [online] Available at: *https://www.w3.org/TR/2004/REC-xmlschema-0-20041028*. [Accessed 30 Sep. 2019].

Materialsmine.org. 2014. *NanoMine Nanocomposites Data Resource*. [online] Available at: *https://materialsmine.org/nm#/* [Accessed 30 Sep. 2019].

Matml.org. 2005. *MatML Overview*. [online] Available at: *https://www.matml.org/* [Accessed 30 Sep. 2019].

Medina-Smith, A, Becker, C and **Tryka, KA.** 2017. *Simple Knowledge Organization System (SKOS) version of Materials Data Vocabulary*, National Institute of Standards and Technology. DOI: *https://doi.org/10.18434/T4/1435037* [Accessed 2019-10-7].

MIF Schema. [online] Available at: *https://citrineinformatics.github.io/mif-documentation/* [Accessed 30 Sep. 2019].

MGI Code Catalog. 2015 [online]. Available at *https://www.nist.gov/programs-projects/mgi-code-catalog* [Accessed 7 Oct 2019].

Plante, R, Becker, CA, Medina-Smith, A, Youssef, S, Dima, A, Bartolo, LM, Warren, JA and **Hanisch, RJ.** (in preparation). *Implementing a registry federation for materials science data discovery*.

Plante, R, Benson, K, Graham, M, et al. 2008. *VOResource: an XML Encoding Schema for Resource Metadata*, v1.03, IVOA Recommendation 22 Feb 2008, *http://adsabs.harvard.edu/abs/2008ivoa.spec.0222P*

Polymer.nims.go.jp. (n.d.). *Polymer Database (PoLyInfo)*. [online] Available at: *https://polymer.nims.go.jp/index_en.html* [Accessed 30 Sep. 2019].

Pppdb.uchicago.edu. (n.d.). *Polymer Property Predictor and Database*. [online] Available at: *http://pppdb.uchicago.edu/* [Accessed 30 Sep. 2019].

Rd-alliance.org. 2017. *Polymer Data Core*. [online] Available at: *https://www.rd-alliance.org/system/files/documents/Final-Draft_Polymer-Data-Core_High-Level-Description.pdf* [Accessed 30 Sep. 2019].

Taylor, A. 2006. *Introduction to Cataloging and Classification*, 10th ed., Libraries Unlimited, Westport, Conn.

Trc.nist.gov. 2006. *ThermoML Archive*. [online] Available at: *https://trc.nist.gov/ThermoML.html* [Accessed 30 Sep. 2019].

Wiki.knoesis.org. 2015. *MatVocab – Knoesis wiki*. [online] Available at: *http://wiki.knoesis.org/index.php/KnowledgeWiki* [Accessed 30 Sep. 2019].

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

Zhao, H, Wang, Y, Lin, A, Hu, B, Yan, R, McCusker, J, Chen, W, McGuinness, D, Schadler, L and **Brinson, L.** 2018. NanoMine schema: An extensible data representation for polymer nanocomposites. *APL Materials*, 6(11): 111108. DOI: *https://doi.org/10.1063/1.5046839*

TO CITE THIS ARTICLE:
Medina-Smith, A, Becker, CA, Plante, RL, Bartolo, LM, Dima, A, Warren, JA and Hanisch, RJ. 2021. A Controlled Vocabulary and Metadata Schema for Materials Science Data Discovery. *Data Science Journal*, 20: 18, pp. 1–10. DOI: *https://doi.org/10.5334/dsj-2021-018*

Submitted: 03 April 2020
Accepted: 15 January 2021
Published: 29 April 2021

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