RDA Working Group Report

International Materials Resource Registries Working Group: Final Report and Recommendations

Version: 9 April 2021

Chandler A Becker[1], Raymond L Plante[1], Andrea Medina-Smith[1], Sharief Youssef[1], Alden Dima[1], Laura M Bartolo[2], James A Warren[1], Robert J Hanisch[1], Brian Matthews[3], Asahiko Matsuda[4], Raphael Ritz[5]

[1]National Institute of Standards and Technology, USA
[2]Northwestern University, USA
[3]Rutherford Appleton Laboratory, STFC, UK
[4]National Institute for Materials Science, Japan
[5]Max-Plank-Gesellschaft, Germany

Executive Summary

A number of national initiatives are aimed at accelerating the development of new materials, and a key component of the strategy is greater access to experimental and simulation data. Under the drive of such initiatives, we are seeing rapid growth in the data that is available over the web. Consequently, it is becoming increasingly difficult for researchers to learn what data is available and how to access it. To address this problem, the RDA Working Group for International Materials Science Registries (IMRR) was established to bring together materials science and information technology experts. The aim is to establish an international federation of registries that can be used for global discovery of data and information resources for materials science. A resource registry collects high-level metadata descriptions of resources such as data repositories, archives, websites, and services that are useful for data-driven research. By making the collection searchable, it aids scientists in industry, universities, and government labs in discovering data relevant to their research and work interests. It can also serve as the basis for a variety of distributed data discovery systems.

In this final report of the IMRR WG, we present the results of our successful piloting of a registry federation for materials science data discovery. In particular, we lay out a blueprint for creating such a federation that is capable of amassing a global view of all available materials science data, and we enumerate the requirements on the standards that make the registries interoperable within the federation. These standards include a protocol for exchanging resource descriptions and a standard metadata schema for encoding those descriptions. We summarize how we leveraged the existing Open Archives Initiative Protocol for Metadata

Harvesting (OAI-PMH) for metadata exchange and Extensible Markup Language (XML) to define our schema, incorporating both generic and materials science-specific metadata. The domain-specific metadata is based on a modest but general Materials Science Vocabulary. We outline an approach to schema definition based on extensions that enable the schema to evolve over time in a tractable way. Finally, we review the registry software developed to realize the federation and describe the user experience.

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 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 360 resource description records distributed across two registry instances.

Introduction: Targeted Problem Space and Goals

The Materials Science and Engineering (MSE) research domain is exceptionally broad and interdisciplinary with its origins most directly from metallurgy, ceramics, and polymer science, but also with important ties to other disciplines such as physics, chemistry, chemical engineering, geology, electronics, optics, and biology. As a global community, MSE is expanding rapidly worldwide through the establishment of large, multi-institutional academic research centers, government labs, industrial consortia, and computing facilities. MSE researchers often need to answer complex questions such as “What structural properties and processing methods are required to develop new lightweight materials that significantly improve fuel efficiency yet meet safety standards satisfied by traditional materials in use today?” To this end we have seen the creation of programs such as the Materials Genome Initiative (MGI) in the US and comparable international materials-focused initiatives in China, Europe, and Japan. These initiatives share a consistent aim -- to decrease the cost and time to develop new materials by a factor of two through more effective discovery, access, and interoperability of experimental and simulation data.

It is under the drive of these initiative that we have seen increased efforts to make materials science data accessible via the web. New data projects like the Materials Project[1] and Materials Commons[2] from academia and industrial initiatives join a legacy of existing data resources that in some cases pre-date the web. In this rapidly evolving climate for data, materials scientists and engineers who might want to make use of the growing digital wealth of information have lacked a comprehensive mechanism for learning what data even exist. At best, we might have consulted manually curated web pages that simply listed the most well-known projects (in the eyes of their curators). Such a web site is not likely to achieve comprehensive coverage of the

[1] https://materialsproject.org/

[2] https://materialscommons.org/

available data in the face of a growing data activity, particularly for the “long-tail” of published data.

In response to the challenge of finding data, the RDA/CODATA Materials Data, Infrastructure, and Interoperability Interest Group (MDII IG; co-chairs James Warren, NIST, and Laura Bartolo, Northwestern University) in collaboration with materials science professional societies, created the International Materials Resource Registries (IMRR) Working Group. The focus of this working group has been to pilot an international federation of data resource registries to enable data discovery. The IMRR WG assembled MSE domain experts representing different regions and sectors, including Asia, Europe, and North America. As part of establishing this federation, we must identify and define the standards and profiles necessary to operate in an open and scalable way.

Our approach to data discovery in this Working Group starts with a piece of infrastructure that has been applied successfully in other domains: a *resource registry*. We define a registry to be a searchable collection of data resource descriptions. A *data resource* is typically a data collection of some kind, like a database, a data publication, or a data repository; however, in general, it can refer to anything that is useful for doing data-enabled science, including software, services, web portals, informational web sites, and even the organizations that provide tools and data. We, of course, are primarily interested in web-accessible data resources. Data resource descriptions come in the form of digital metadata records, which include, most importantly, a URL for getting access and more information about the resource. By making these descriptions searchable, researchers have a way to discover resources related to a particular scientific topic.

For the scope of this working group, we have restricted ourselves to what can be referred to as high-level resources like repositories, databases, and web portals. We have not focused on individual datasets, data records, or measurements. The aim here is to direct users to the home pages for data that are supported by the people and organizations who have made the data available; there, the users can leverage the collection-specific tools to delve deeper to find the individual datasets or records that they need (rather than circumventing those tools). Another reason to keep our registry at a high-level view of what data are available is that additional challenges arise if the registry must scale up to support potentially millions of records. One of those challenges is keeping the registry continuously up to date. Active data resources could be continuously adding new data, making it more likely that the registry goes out of date; on the other hand, the metadata describing web sites and large data aggregations will change more slowly. Most importantly, we feel that the metadata that distinguishes fine-grain resources like individual datasets or records will be much more diverse, more challenging to integrate, and is best curated by the providers of that data in the systems they built to handle that metadata.

We note that while our registry contains only a coarse-grained view of the materials science data that is out there, this does not prevent us from using it to discover and download individual datasets. As an example from the astronomical community, the virtual observatory framework [1] features a registry as the first step in an automated data discovery process: the registry is used to find the repositories and portals where data is served. Users can be directed to those portals; however, if the sites have also registered search services, those sites can be queried automatically to find individual datasets or measurements. Since the registry itself features its own application programming interface (API) for searching, third-party tools can carry out the searches on behalf of users without them realizing that a registry is involved. With a layered approach to discovery like this, building a community registry represents a practical, tractable first step that is still useful on its own.

Finally, a key aim of our working group is not simply to build a single registry but to establish a *federated registry framework*. In this approach (also borrowed from the virtual observatory federation), there is no primary or central registry but rather a federation of multiple registries distributed around the world. Each registry can collect descriptions from different sub-communities and then share those descriptions with other registries through a standard API. Any registry can then *harvest* resource descriptions from all of its peers to get a globally comprehensive view of all available data. A federated approach that's not anchored by a single registry makes for a system that is more robust against failures or interruptions of any individual registry (including funding failures). Just as important, however, is how a federated approach can serve to distribute the curation of the records: each organization running a registry can curate the records they create. Thus, records can be curated by those who understand the records best. To make this federation effective and scalable, we need not just a common API for harvesting; we also need a common metadata schema for encoding the resource descriptions.

In summary, our Working Group set out to prototype and demonstrate a model for data discovery built around federated registries. In particular, our goals were to:

- Create and build consensus around a common resource metadata schema that can include material science concepts important for data discovery,
- Build a working registry system capable of creating, collecting, and searching data resource descriptions based on the common metadata schema,
- Populate a working registry with a useful number of material science resource descriptions, and
- Demonstrate a registry federation with multiple registries capable of exchanging resource descriptions.

An RDA Working Group is an appropriate forum for prototyping and demonstrating such a model for data discovery. In particular, it has allowed us to engage a global community of domain researchers, understand their needs, and collect their input. (This has been most valuable with developing the registry metadata schema.) It also provides a channel for translating this approach to data discovery to other research domains.

Summary of Working Group Deliverables

In this final report, we present the following as deliverables of the International Materials Resource Registries Working Group:

- A model for a federated registry framework that can support the distributed creation, sharing, and searching of data resource descriptions. We describe that model in this report.
- A model for extendable resource metadata that combines domain-independent concepts with domain or community-specific extensions. We summarize that model in this report.
- An open source implementation of a federatable registry for XML-based resource descriptions developed by the US National Institute of Standards and Technology (NIST) Information Technology Laboratory.
- A vocabulary describing Materials Science data (documented at https://www.rd-alliance.org/materials-vocabulary-draft-21-mar-2017).
- An XML-based resource metadata schema that includes:
  - Domain-neutral metadata including Dublin Core concepts,
  - An extension for metadata related to Materials Science that leverages the Materials Science vocabulary.
- A pilot demonstration of our federated registry framework applied to Materials Science featuring the NIST registry implementation deployed by two different institutions—NIST and the Center for Hierarchical Materials Design (CHiMaD)[3]—which share records.

The Federated Registry Framework

Our model for a registry-mediated data search process is based on a *federated* architecture (adapted from the model used by the Virtual Observatory in the astronomy domain). Specifically, this means that there is no one master registry; rather, we have a network of registries working together. Any registry can pull in the resource descriptions from each of the other registries using a common metadata exchange protocol in order to build a comprehensive collection of resource descriptions of all known resources in the network. It may then make this collection searchable.

From a technical perspective, the federation is an open one: any organization can host a registry as means of advertising their own resources to the world. We refer to a registry whose primary function is to export resource descriptions out to the federation as a *publishing registry*. In this role, the registry providers take responsibility for creating and curating the description records for a specific subset of resources. By “curating,” we mean keeping the records accurate in their content, up-to-date, and compliant with the metadata standards in use. If the registry is operated by a data center that provides a variety of resources—

[3] https://chimad.northwestern.edu

databases, data collections, services, and perhaps a portal to navigate them all—the registry would then curate the records describing the resources it provides. A registry might be run by a particular sub-community in a domain, curating records on the behalf of its constituents; in our pilot, the CHiMaD registry manages records for resources provided by the Center's member organizations. When a data provider only shares a few resources (say, a single database or a few datasets), it may not make sense for them to also operate a registry; instead, there can be publishing registries that host records on community providers' behalf. In our pilot, the NIST registry plays that role: the registry web portal allows a provider to login, create resource descriptions, publish them out, and update them over time.

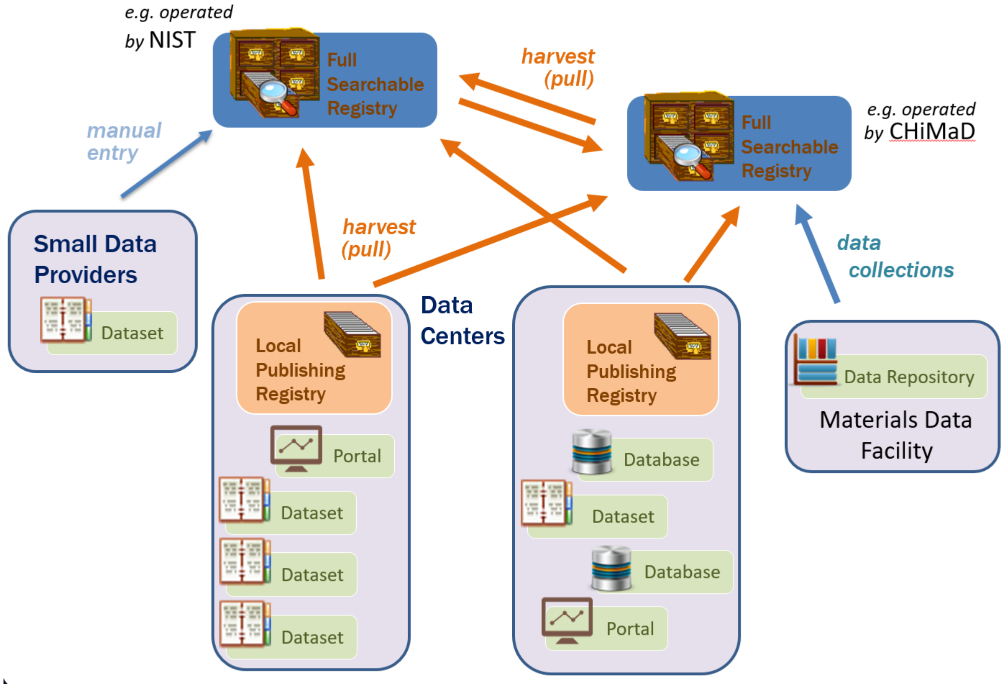


Figure 1 Record interchange within the registry federation. *Searchable registries harvest resource descriptions from all publishing registries via a standard protocol. Some registries, like the one operated by CHiMaD, can manage records for a particular sub-community, possibly pulling in records through customized APIs. Other registries, like the one operated at NIST, can serve the at-large community where providers can create and maintain records through a GUI interface. (Figure was adapted from the architecture figure presented in [1].)*

The other important role of a registry, of course, is making the resource descriptions searchable. Not every registry in the federation needs to do this, so we distinguish this role by referring to *searchable registries*. In our model, registries are *not* obligated to provide search capabilities in the same way, particularly through their user-oriented web sites; rather, a searchable registry can tailor its search services to the primary audience they serve. (There is great value in providing a common search API for powering remote clients; however, this is beyond the scope of the goals of the Working Group.)

Why Federate?

We recommend a federated model as part of an approach to sustainable infrastructure. In particular, federation brings these key features:

1. Distributed metadata curation

   With a single, centralized registry, there is a danger that registry records become effectively divorced from the people that care about the things being described, and so it is common under such a model for records to be inaccurate and become out of date. In the federated model, the curation of the registry records can be distributed across the community and kept closer to the experts responsible for providing the resources the records describe. While record curation will still be a sociological challenge, it is made more tractable when more of the community can be involved.

2. No single point of failure

   When there are multiple registries that have complete collections of resource descriptions, discovery services need not completely shut down when one registry goes off-line: users and search clients have alternate registries they can connect to. Robustness to registry failure can include funding failure when the federation spans the globe.

3. Allows for innovation

   The registry federation need not present a one-size-fits-all solution for data discovery. That is, searchable registries can specialize their capabilities to a particular sub-community who is their mission to serve, whether it’s in the search interface that is presented to the user or the way the records are indexed.

Requirements for Interoperable Registries

A functioning registry federation requires some interoperability between registries in how records are traded. In summary, registries must have (1) a common metadata exchange protocol and (2) a common metadata schema and format for passing records within that protocol. Multiple open standards exist today which can be adopted to define a registry federation; to the extent that the standards are general and not community- or application-specific, additional requirements are needed to define the profile of those standards for a particular community:

1. The profile of the common metadata exchange protocol should:
    a. Provide a means for identifying the record format(s) and schema(s) that can be used to encode resource descriptions. (Often the metadata schema and format are coupled together as a single standard.)
    b. Set a distinction between the records that have been created and curated by the registry sharing their records, and the records that it has harvested from other registries; the protocol should allow (or require) delivery of only the former. This ensures that harvesters only receive one copy of a record from its definitive source.
    c. Be able to communicate that a resource (and its resource description) is no longer available.
    d. Require minimal validation of records before they are made available to users and clients (for searching, harvesting, etc.). In a distributed system, when something goes wrong, it is often unclear to users who is responsible. By requiring validation of resource descriptions before exporting them from their registry of origin, problems in the resource records can be detected close to where they are best fixed.

2. The common metadata format should:
    a. Be openly defined,
    b. Have a unique identifier associated with it, and
    c. Be validatable.

When a community standardizes the requirements for participating in a registry federation, these are the minimum features the community must define. We note that other best practices regarding the definition and use of a metadata standard should be applied, including connecting the metadata schema to community-recognized vocabulary and semantics; these could also be part of the standard.

We note that the need for a common metadata schema is specifically for unifying the high-level discovery process. Within our framework, we want to enable the creation of rich and varied applications for indexing and searching through the resource records; thus, developers need to know how to extract particular kinds of information across all records. Consequently, a single, common schema makes this possible in the simplest way. Nevertheless, throughout the community that produces this metadata, many schemas and formats are in use. Some schemas are adopted locally because they capture information important to a particular sub-community (say, geographic locations or astronomical positions). It is not the intention of this framework to cut off access to this richer information; rather, this richer information can be more easily leveraged if the metadata exchange protocol is capable of sharing records in multiple formats.

OAI-PMH as a Standard Exchange Protocol

For our pilot for the Materials Science community, we chose the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) [2]. This was chosen for three reasons. First, as an XML-based protocol, it is well suited for transmitting our XML-formatted metadata records. Second, it is broadly used across many communities and has an established track record for successfully enabling interoperability between data centers, registries, and end-user tools. Finally, the protocol not only meets the requirements set out in the previous section, but provides additional features that make it an efficient means to exchange metadata, including incremental harvesting (i.e., harvesting new or changed records since some given date) and record paging.

OAI-PMH is a “pull” protocol: a registry (or other consumer) wanting records—the *harvester*—asks for new or changed records from another registry and pulls them in to its own collection. In this protocol model, the harvester gets to choose which other registries it will collect records from and when. This is in contrast to a “push” protocol, where the registry sends its records out to other interested registries; this requires that the registry know in advance who wants its records (its *subscribers*).

The Resource Metadata

In our demonstration for materials science, we developed an XML-encoded metadata schema using XML Schema [3]. XML as a metadata format satisfies the key format requirements described above:

- 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
- Open software is available to validate resource description documents against the XML Schema definition.

The schema we assembled drew on existing schemas and vocabularies, most notably, Dublin Core [4], DataCite [5], and the virtual observatory’s Resource Metadata [6] for the generic, domain-unspecific concepts. We also reviewed the current state of materials science-related vocabulary and ontology activities; we found that much of the work in this area was either still in development (or dormant) or highly specific to particular applications.

A lesson we took from the virtual observatory experience and which we experienced ourselves during our pilot is the importance of supporting metadata extensibility and evolution. It is inevitable that we will have to update our metadata standard 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, it is important that all participating registries share a common basis for validation, and for our pilot, this centers on making sure all registries have the same XML Schema definition document to validate records. Updating the schema can be a highly disruptive endeavor 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.

The virtual observatory developed some techniques for defining XML schemas that greatly mitigates the disruption caused by schema evolution. These techniques are based on the idea that there is a common core metadata schema and that evolution is accomplished through pluggable extensions to that core [6]. We note, however, that we have not yet completed support for all of these techniques in our registry software (as we have, in this phase, put more emphasis on the user experience). Nevertheless, we designed our metadata schema based on the virtual observatory approach but then made adjustments to the schema to accommodate the current state of the software. Despite this less-than-ideal approach, we are further developing the registry software to take advantage of metadata extensions and make it more robust to an evolving metadata schema.

The GitHub[4] repository, mgi-resmd [7], 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. For integration with our registry software, we combined all definitions into a single schema document, mgi-resmd.xsd [8]. 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 design of the metadata from a high level. Readers can consult the schema file itself for precise definitions of individual metadata terms.

Our metadata model for describing data resources reflects a few core principles:

- Some of the metadata we need to collect is generic and some of it is specific to our domain; these should be kept separate in our model.
- There are different *types* of resources—namely, repositories, databases, web sites, software, etc.—and while some metadata may apply to all (or most) types of resources, we may need to employ type-specific metadata to describe them. Further, a resource may reasonably qualify as being of multiple types simultaneously.
- Reflecting how materials science overlaps heavily with other areas of science (i.e., physics, chemistry, biology, etc.), it may be necessary to leverage metadata from different domains simultaneously within the resource description.
- We should identify multiple points for future extensibility: in the future, we may want to support new types of resources or plug in new domain-specific metadata.

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

[4] GitHub is a commercially-operated, cloud-based software repository; its use by this working group follows the practice of the RDA community. Its mention in this document does not imply recommendation or endorsement by the US National Institute of Standard and Technology (NIST).

- **Identity** – how the resource is named and referred to
- **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** – what other resources it is related to

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

- Organization
    - Institution
    - Project
- DataCollection
    - Repository
    - Archive
- Dataset
    - Database
- Service
    - 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 can 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 describes how the resource *applies* or relates to a particular domain. A resource description can have multiple Applicability sections, each one 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 doesn't. For this reason, it is acceptable if the different domains include metadata that overlap in their semantics.

For our pilot, we defined an Applicability for materials science. This section 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 make both creating records and searching for them easier.

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 [9] captures the terms in a human-readable format. We also created a SKOS definition of the vocabulary as well [10].

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. 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.

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 leveraging the detail (e.g., measured in the time it takes to interpret and select the appropriate terms).

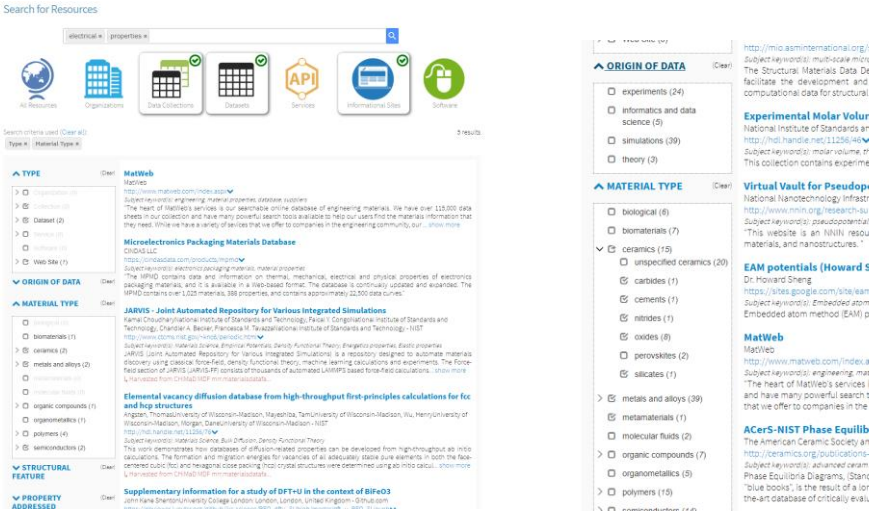
It is worth noting that other semantic developments—namely, an ontology for materials design and research--have been undertaken since work on this vocabulary which we discuss more under the “Future Work” section. An ontology, depending on its exact scope, could also provide terms (that could either extend this vocabulary or replace it) that could be used effectively to describe and enable discovery of resource within a registry.

The Registry Application

Through the Information Systems Group, Software and Systems Division, of the NIST Information Technology Laboratory, we developed a registry application based on their existing software, the NIST Materials Data Curation System [11]. The registry software (which is still under active development) is available from the GitHub repository, nmrr [12]. The software runs as a web application, and in our pilot, there are two instances in operation:

- http://materials.registry.nist.gov/
- http://mrr.materialsdatafacility.org/

Users looking for data resources can visit the website, access the search page, submit simple keyword-based search queries, and receive back a listing of matching resources. The search results page provides faceted browsing of the results which leverages the controlled vocabulary. The user can drill down into subsets of results by clicking on different resource attributes and their categories. For example, the user can select particular resource types (like databases or software) that provide experimental data related to stress corrosion of metals. Each search result includes a link to the resource's native landing page as provided by the data provider; thus, the user can now visit the resource's web site directly, download data or use the native tools provided by the data provider.



MatWeb
MatWeb
http://www.matweb.com/index.aspx
Subject keyword(s): engineering, material properties, database, suppliers
"The heart of MatWeb's services is our searchable online database of engineering materials. We have over 115,000 data sheets in our collection and have many powerful search tools available to help our users find the materials information that they need. While we have a variety of sevices that we offer to companies in the engineering community, our ... show more

Microelectronics Packaging Materials Database
CINDAS LLC
https://cindasdata.com/products/mpmd
Subject keyword(s): electronics packaging materials, material properties
"The MPMD contains data and information on thermal, mechanical, electrical and physical properties of electronics packaging materials, and it is available in a Web-based format. The database is continually updated and expanded. The MPMD contains over 1,025 materials, 388 properties, and contains approximately 22,500 data curves."

JARVIS - Joint Automated Repository for Various Integrated Simulations
Kamal ChoudharyNational Institute of Standards and Technology, Faical Y. CongoNational Institute of Standards and Technology, Chandler A. Becker, Francesca M. TavazzaNational Institute of Standards and Technology - NIST
http://www.ctcms.nist.gov/~knc6/periodic.html
Subject keyword(s): Materials Science, Empirical Potentials, Density Functional Theory, Energetics properties, Elastic properties
JARVIS (Joint Automated Repository for Various Integrated Simulations) is a repository designed to automate materials discovery using classical force-field, density functional theory, machine learning calculations and experiments. The Force-field section of JARVIS (JARVIS-FF) consists of thousands of automated LAMMPS based force-field calculations. ... show more
↳ Harvested from CHiMaD MDF mrr.materialsdatafa...

Elemental vacancy diffusion database from high-throughput first-principles calculations for fcc and hcp structures
Angsten, ThomasUniversity of Wisconsin-Madison, Mayeshiba, TamUniversity of Wisconsin-Madison, Wu, HenryUniversity of Wisconsin-Madison, Morgan, DaneUniversity of Wisconsin-Madison - NIST
http://hdl.handle.net/11256/76
Subject keyword(s): Materials Science, Bulk Diffusion, Density Functional Theory
This work demonstrates how databases of diffusion-related properties can be developed from high-throughput ab initio calculations. The formation and migration energies for vacancies of all adequately stable pure elements in both the face-centered cubic (fcc) and hexagonal close packing (hcp) crystal structures were determined using ab initio calcul... show more
↳ Harvested from CHiMaD MDF mrr.materialsdatafa...

Supplementary information for a study of DFT+U in the context of BiFeO3
John Kane ShentonUniversity College London: London, London, United Kingdom - Github.com

http://mio.asminternational.org/
Subject keyword(s): multi-scale micr
The Structural Materials Data De facilitate the development and computational data for structural

Experimental Molar Volur
National Institute of Standards ar
http://hdl.handle.net/11256/46
Subject keyword(s): molar volume, th
This collection contains experime

Virtual Vault for Pseudop
National Nanotechnology Infrastr
http://www.nnin.org/research-su
Subject keyword(s): pseudopotential
"This website is an NNIN resou materials, and nanostructures."

EAM potentials (Howard S
Dr. Howard Sheng
https://sites.google.com/site/eam
Subject keyword(s): Embedded atom
Embedded atom method (EAM) p

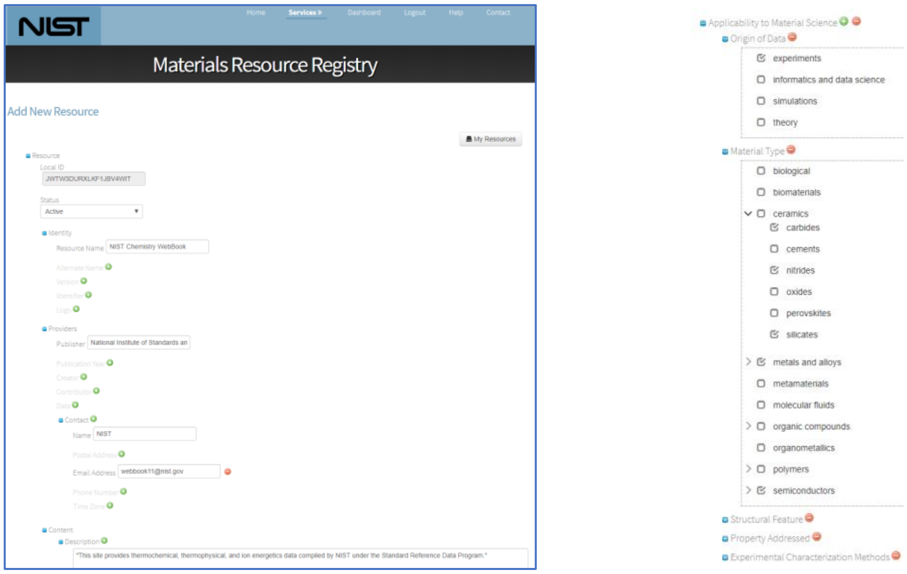
MatWeb
MatWeb
http://www.matweb.com/index.a
Subject keyword(s): engineering, mat
"The heart of MatWeb's services i and have many powerful search t that we offer to companies in the

ACerS-NIST Phase Equilib
The American Ceramic Society an
http://ceramics.org/publications-
Subject keyword(s): advanced ceram
Phase Equilibria Diagrams, (Stanc "blue books", is the result of a lor the-art database of critically evalu

Figure 2 Registry search results. *(left) Example of the search results page with faceted browsing filters to the left. (right) a zoomed view of the Material Type filters showing how general type, "ceramics", can be clarified with more specific types of ceramics.*

Data providers can visit the website to register the existence of their resources. When they create an account, they have a space where they can create descriptions of different kinds of resources. After selecting a resource type, they are presented with a form where they can enter the metadata. In particular, the materials science metadata is presented as checkboxes with expandable detail.

The application exploits the general capabilities of the underlying curation software. In particular, both the resource registration form (where resource descriptions are created) and the search results page (where hits can be filtered) are generated on the fly based directly from the XML Schema document. This means that the application can be easily adapted to other schemas and domains. In fact, NIST has re-used this software and its underlying model to set up registry federations for other domain communities, including metrology (http://imrr.bipm.org/) and greenhouse gas research.



Add New Resource

Resource

Local ID: JWTW3DURXLKF1JBV4WIT

Status: Active

Identity

Resource Name: NIST Chemistry WebBook

Alternate Name

Version

Identifier

Logo

Providers

Publisher: National Institute of Standards an

Publication Year

Creator

Contributor

Date

Contact

Name: NIST

Postal Address

Email Address: webbook11@nist.gov

Phone Number

Time Zone

Content

Description: "This site provides thermochemical, thermophysical, and ion energetics data compiled by NIST under the Standard Reference Data Program."

Applicability to Material Science

Origin of Data

- ☑ experiments
- □ informatics and data science
- □ simulations
- □ theory

Material Type

- □ biological
- □ biomaterials
- □ ceramics
  - ☑ carbides
  - □ cements
  - ☑ nitrides
  - □ oxides
  - □ perovskites
  - ☑ silicates
- ☑ metals and alloys
- □ metamaterials
- □ molecular fluids
- □ organic compounds
- □ organometallics
- □ polymers
- ☑ semiconductors

Structural Feature

Property Addressed

Experimental Characterization Methods

Figure 3 Resource Registration Form. (left) A portion of the form used to create a resource description and which is automatically generated from the schema. (right) a blow-up of the portion of the form where Material Science attributes can be select; these terms come from the Materials Vocabulary.

The registry application also features a rich API for both searching its contents and uploading new records. We envision the latter being important for integrating a registry with a data center's infrastructure: the API can be used to automatically push descriptions of resources based on metadata from the center's native systems. The registry application also supports an OAI-PMH harvesting service. Not only can it expose this service to other registries, the application can be configured to harvest from other OAI-PMH services at regular intervals.

As of this writing, the registry contains over 350 records describing resources from 52 organizations. (We note that many of these were registered by NIST WG members on behalf of those organizations to help seed the registry and attract their interest.) Most of the resources describe software; this reflects a collaboration with the Materials Genome Initiative (MGI) in which we absorbed records from the MGI Code Catalog. The registry also currently describes

33 databases, 23 repositories, 2 project archives, 4 services, and 22 portals and web sites. Further work is focused on expanding the contents of the registry system, particularly organizations and services.

Future Plans

Although this working group is finished, we hope that the community we have built around this work can continue to grow. In particular, the participants of this working group are interested in organizing a Task Group within the RDA/CODATA Materials Data Interest Group to explore expanding the impact of vocabularies and registry systems for material science and other communities. In particular, the work presented here in the area of vocabularies has begun to generate greater discussion of vocabularies, ontologies, and related semantic assets for fields related not only to material science but also for beam-line research and chemistry.

NIST, CHiMaD, and our partners in the NSF-funded Midwest Big Data Hub plan to continue development in the registry federation. We will seek to involve more providers in registering their resources as well as recruit centers that can host their own registries and integrate them into the registry federation for materials science.

With the expectation of continued growth, there are a number of improvements and new pilots that we envision for improving registry-based data discovery. As mentioned above, NIST continues to develop the registry software. The software's core (now called the Configurable Data Curation System) is being updated to make it more modular and, therefore, more adaptable to new domains and applications. As part of this transformation, NIST is moving more to an open source development process to leverage contributions from the growing external user community. In this new phase of development, we hope to address a number of critical and desired features of the software. In particular, we would like to:

- Provide improved XML Schema support that can enable community-based standardization of schemas.
- Improve support for various persistent identifiers and facilitate robust linking of related resources (such as linking services to the organizations that provide them).
- Add support for evolving schemas with robust validation.
- Explore support for other metadata format types such as JSON.

We would also like to explore new pilots in advanced data discovery. An important one would be connecting the registry (which describes only large or high-level resources) with more fine-grained discovery services provided by data centers. Data providers can already register web services they provide. We would like to enrich their descriptions so that the registry can recognize certain kinds of search services; in such a case, the registry—or any third-party tool using the registry—could automatically call the services on behalf of the user. In this way, a user might submit a query to the registry, and the registry can pass on that query to those services likely to have data. This tiered model for data discovery is used by the virtual observatory to drill down to individual datasets that can be downloaded or individual database records.

Finally, we note that a new task group within the Materials Science interest group has been begun work on ontologies for this discipline. This task group brings together regional ontology development efforts for coordination and possible collaboration. Most notably, this includes the European Materials and Modeling Ontology (EMMO) [14] as well as efforts within the US (include CHiMaD). As noted above, an ontology could provide technically well-defined terms that could be leveraged in a registry context and possibly enable more sophisticated interactions with resources. It is has been suggested in our WG discussions that there may be some incompatibilities between the registry vocabulary presented here and the EMMO ontology; thus, it may be necessary to consider whether the registry schema can be made better aligned with an emergent community ontology or should be replaced with something based directly on the ontology.

Impact for the RDA and the Broader Community

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 some success already in this effort with the deployment of registries serving the metrology and greenhouse-gas research communities based on the same software and model.

We note that that our collaboration with CHiMaD has been important for reaching out into 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, Michigan, and Illinois. Each member institution of the Spoke leads significant government funded Materials Genome Initiative programs which also incorporates a wide network of academic and industrial partners located across the Midwest. 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.

We summarize several ways in which the deliverables of this working group can be transferred to other communities:

- We have laid out a blueprint for a registry federation that supports global data discovery. In this report, we have attempted to call out the key features of such a federation in a manner that is independent of the specific technology choices. For example, these same features could be applied to a federation based on JSON-encoded metadata, using the ResourceSync protocol for metadata exchange [13], and some completely different software implementation.

- 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 demonstrated a working registry system based on open software that can be readily adapted for other domains.

Appendix A: Working Group Members and Contributors

Andrea Medina-Smith
Ann Racuya-Robbins
Brian Matthews
Chandler Becker
Charles Vardeman II
Clare Paul
Daniel Mietchen
Deborah Mies
Gerhard Goldbeck
Haiging Yin
Hilary Goodson
James Warren
Kathleen Fontaine
Laura Bartolo
Mark Leggott
Matthew Lange
Brian Matthews
Debbie Mies
Raphael Ritz
Raymond Plante
Robert Hanisch
Scott Henry
Sharief Youssef
Thomas Proffen
Tim Austin
Timea Biro
Tobias Weigel
Vasily Bunakov
Yibin Xu
Haiquing Yin
Zachary Trautt

References

1. Hanisch, R.J., Berriman, G.B., Lazio, T.J.W., Emery Bunn, S., Evans, J., McGlynn, T.A., Plante, R. (2015). *The Virtual Astronomical Observatory: Re-engineering access to astronomical data*, Astronomy and Computing, 11, pp. 190-209. https://doi.org/10.1016/j.future.2019.10.030
2. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S. (2002). *The Open Archive Initiative Protocol for Metadata Harvesting*, v2.0, Open Archives Initiative, https://www.openarchives.org/OAI/openarchivesprotocol.html
3. Fallside, D. C., Walmsley, P. (2004). *XML Schema Part 0: Primer Second Edition*, W3C Recommendation 28 October 2004, https://www.w3.org/TR/2004/REC-xmlschema-0-20041028.
4. DCMI Usage Board (2012). *DCMI Metadata Terms*, Dublin Core Metadata Initiative, http://dublincore.org/documents/2012/06/14/dcmi-terms/
5. DataCite Metadata Working Group. (2017). *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*. Version 4.1. DataCite e.V. doi:10.5438/0014
6. 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
7. mgi-resmd GitHub repository: https://github.com/usnistgov/mgi-resmd
8. Materials Resource Registry schema in use as of 2018-03-15:

9. Becker, C, et al. (2017), Materials Vocabulary Draft, v15, 2017-03-21, https://rd-alliance.org/system/files/documents/Materials_Registry_vocab_draft_170321.pdf
10. Medina-Smith, A. et al. (2017), the Materials Vocabulary in SKOS,
11. Dima, A., S. Bhaskarla, C. Becker, M. Brady, C. Campbell, P. Dessauw, R. Hanisch, et al. 2016. “Informatics Infrastructure for the Materials Genome Initiative.”JOM Journal of the Minerals Metals and Materials Society 68 (8). https://doi.org/10.1007/s11837-016-2000-4.doi:10.1007/s11837-016-2000-4
12. nmrr GitHub repository: https://github.com/usnistgov/nmrr
13. Klein, M., Van der Sompel, H., Warner, S. et al. (2017). *The ResourceSync Framework Specification*, Open Archives Initiative, http://www.openarchives.org/rs/1.1/resourcesync
14. European Committee for Standardization (CEN) (2018). *Material modelling – Terminology, classification and metadata*, https://www.cen.eu/news/workshops/pages/ws-2017-012.aspx