Materials Registry Working Group Meeting

Chandler Becker and Ray Plante

Sharief Youssef, Alden Dima, Zachary Trautt,
Kimberly Tryka, Robert Hanisch, Jim Warren, Mary Brady
National Institute of Standards and Technology

Laura Bartolo
Northwestern Univ.

RDA Plenary, 15 Sept 2016

Current WG members (9/13/16)

Say a few words about your project.

- Brian Matthews
  - Science and Technology Facilities Council
- Chandler Becker
  - National Institute of Standards and Technology
- Clare Paul
  - Air Force Research Laboratory
- Deborah Mies
  - Granta Design, Ltd.
- Haiqing Yin
  - Beijing Univ. of Science and Tech.
- James Warren
  - National Institute of Standards and Technology
- Kathleen Fontaine
  - Rochester Polytechnic Institute (RDA)
- Laura Bartolo
  - Northwestern Univ.
- Raphael Ritz
  - Max Planck Society, Garching
- Raymond Plante
  - National Institute of Standards and Technology
- Robert Hanisch
  - National Institute of Standards and Technology
- Sharief Youssef
  - National Institute of Standards and Technology
- Tobias Weigel
  - German Climate Computing Center (DKRZ)
- Vasily Bunakov
  - Science and Technology Facilities Council
- Zachary Trautt
  - National Institute of Standards and Technology

Any new people who haven’t joined yet?

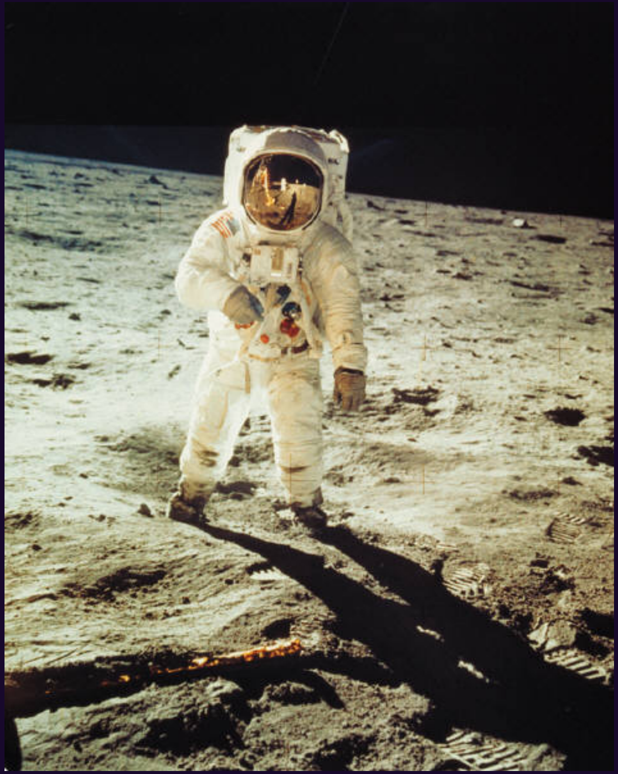
Overview

- Motivation for WG and related efforts
- Summary of timeline from case statement
- Proposed activities and topics to address in WG
- NIST pilot effort
- Thoughts on how to proceed, followed by discussion

Motivation for the working group

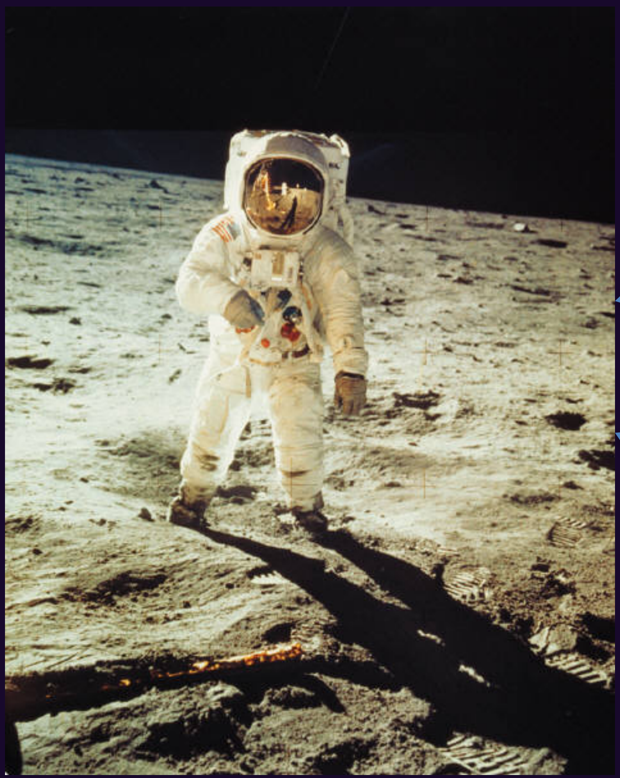
- Many materials resources exist (datasets, websites, repositories, registries, etc.), and the number is growing.
- How can we link them in a way that makes it easier to find and share relevant information and data?

Start by creating catalogs of resources



Hosted in many different locations with diverse content

Then connect them

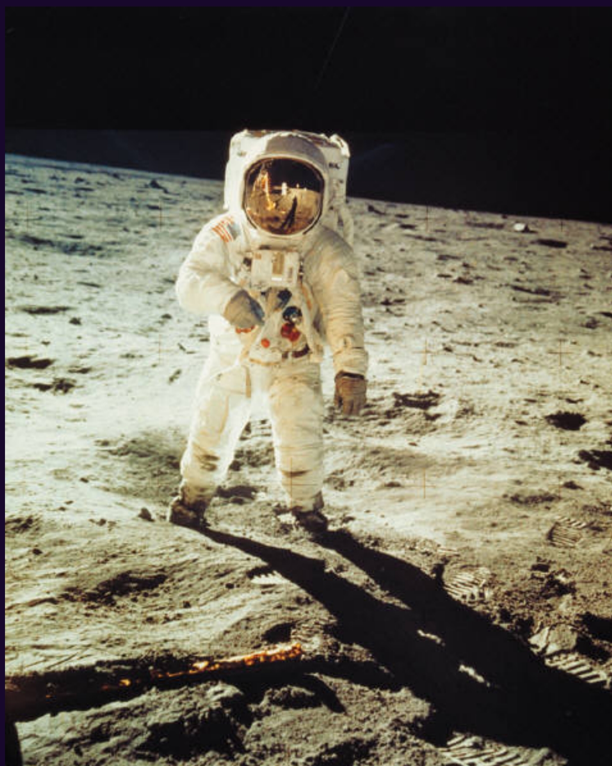


Via data- and information-sharing protocols

What is a registry?

- Registry is a catalog containing descriptions of *resources** that are useful for (materials science) data-driven research
  - * Mainly datasets, databases, and data services
  - * Can also be portals, software, organizations, …
- A starting point for *discovering* useful data and tools
  - – By making the metadata descriptions searchable
  - – Can direct users to the web sites that host the data

Connected catalogs

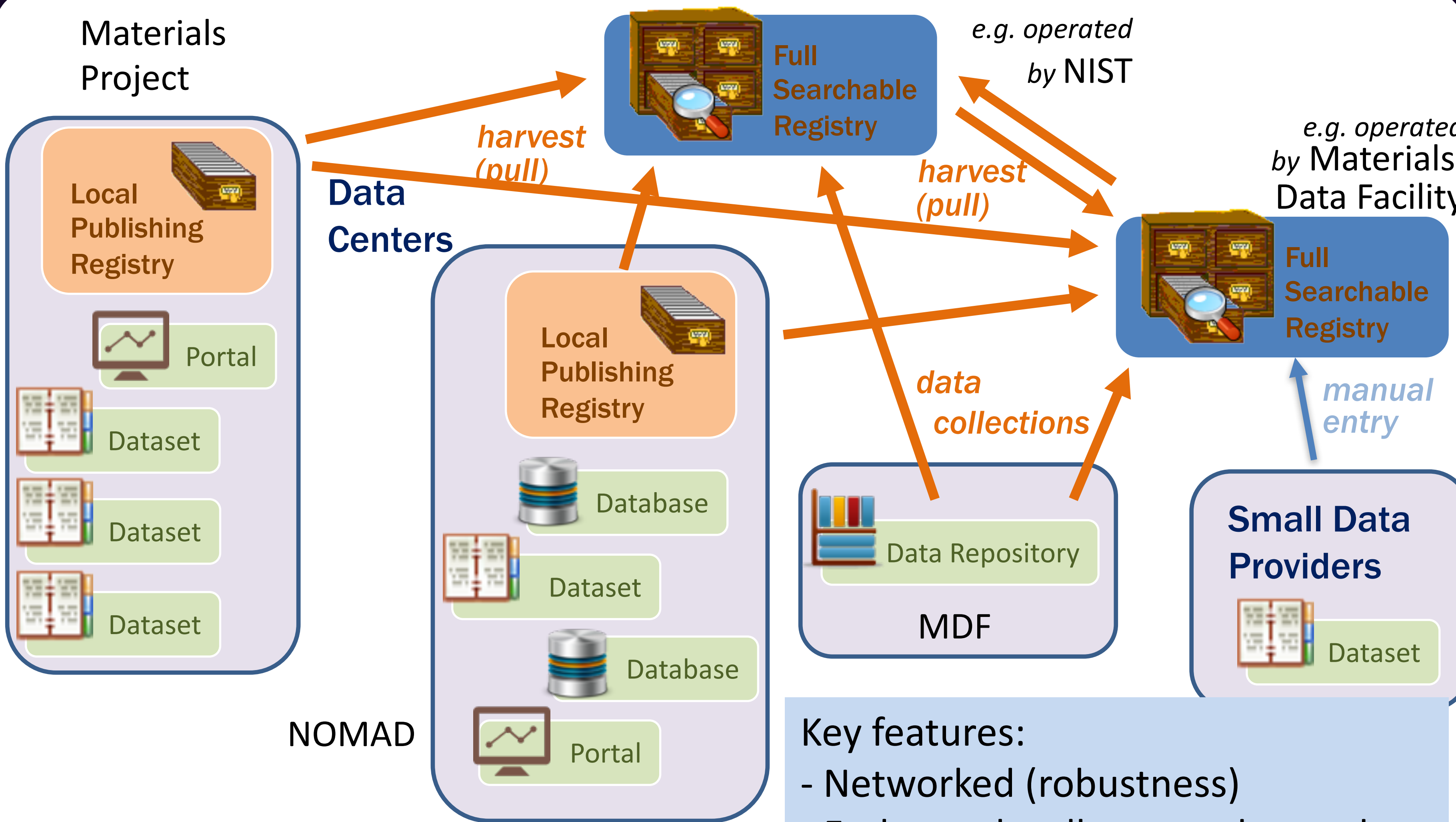


Turn into…

Building a Registry Federation

- What does federation mean?
  - Comprised of a network of registries; there is no single Registry
    - Any registry can collect a globally-comprehensive collection of resource descriptions and make it searchable
  - Resource metadata exchange
    - There a common mechanism(s) for sharing descriptions of available data resources
  - Allow local metadata curation
    - Any organization can run registry of their own data resources and share it with the world
- Why federate?
  - Distribute metadata curation
    - Allow experts who provide/operate data resources to manage how they are described, update descriptions as they evolve
  - No single point of failure (including funding failure)
  - Allow innovation in providing search capabilities
- How do we federate?
  - Common metadata exchange mechanism
    - We propose starting with OAI-PMH
  - Common metadata schema

A Registry Federation



Key features:
- Networked (robustness)
- Exchange locally curated metadata
- Support search and discovery

Want This…

Common protocols

Mappings between content and approaches of different projects

Not This…

Lots of incompatible resources and catalogs

Confusion, frustration, data loss, missed opportunity

Words, words, words

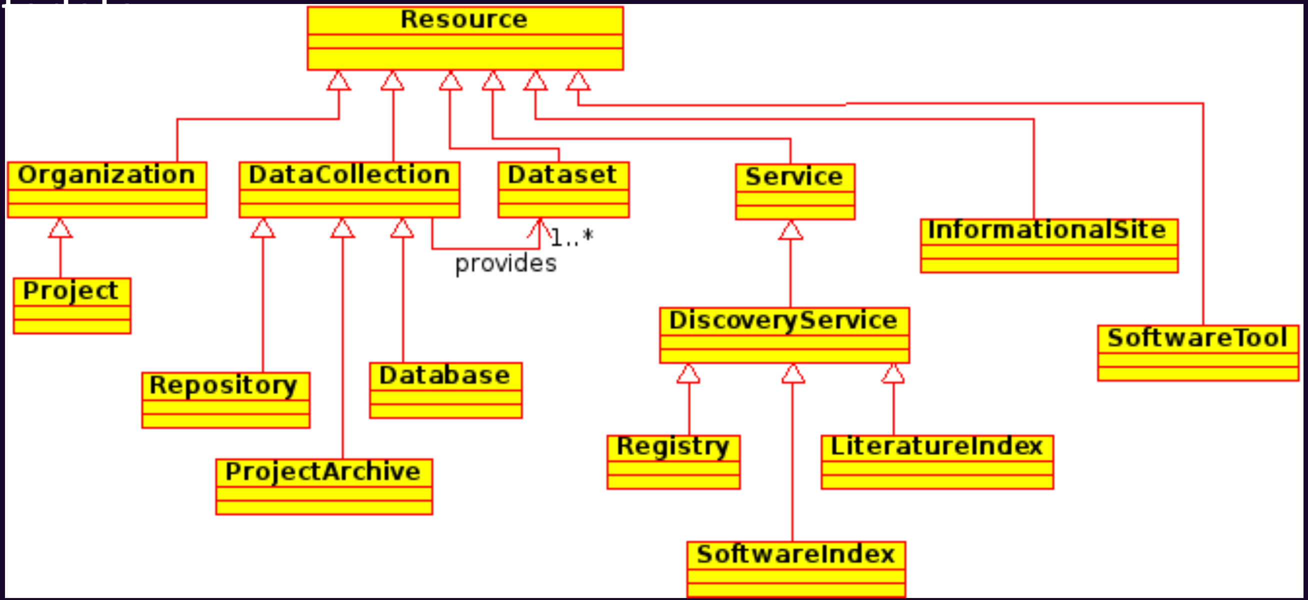
- For this to work, we need words **that describe the resources** being registered
- Some terms are generic (based on Dublin Core (dublincore.org)):
  - Organization
  - Contact information
  - Access methods and locations
- But others have to be domain- (i.e., materials-) specific
- Not the complete metadata required to fully document the data in the resource
- Want to be user-friendly, which currently means selecting from a relatively limited list of high-level terms and using searchable free text

Resource Concept Model

- A *Resource* is
  - A thing we want to describe and discover
  - *An identified, described, and discoverable component of the distributed data environment*
  - Different types of resources

    (some can be of multiple types simultaneously)
  - Model implies some common metadata, each subtype can add additional me[...]

What kinds of Resources do we want to share and discover?



Categories of Resource Metadata

- **Identity** -- how we recognize it
- **Role** – what type of Resource is it
- **Publication** -- who is responsible
- **Content** -- what it is about
- **Access** -- how to get at it
- **Applicability** -- how it applies to different domains
  - Can have multiple entries, each containing metadata specific to a different domain
  - Include a section for Materials Science metadata

Metadata Exchange: Formats

- Format: How to encode metadata
- Common encoding mechanisms currently in use:
  - XML (as defined by XML Schema)
    - recommended for WG deliverable
  - JSON (as defined by JSON Schema)
  - JSON-LD
- Work at NIST: interoperability between these formats
  - Best practices for define format schemas
  - Provides technical mechanism for supporting extensibility
  - Enable well-defined mechanisms to convert between
  - https://github.com/usnistgov/mgi-resmd
- Collaboration on schema welcome
  - General Resource metadata, formatting (via Schema)
  - Materials Science-specific metadata

Technical Collaboration

- "Entry-level" involvement
  - Describe your resources at one of the community registries
- Contribute to metadata schema development
- Operate a registry for your organization
  - Can run an instance of the NIST MRR application
  - Good if you have a larger number of records to share
  - Can connect to your local metadata infrastructure
- Create your own registry application
  - Support exchange format
  - Support OAI-PMH; help set profile
  - Prototype alternate exchange mechanisms

"Do I have to give you my data?"

- **NO.**
- The data can be hosted somewhere and an entry added to the NMRR (or another instance) to point to where the data is and how to access it.
- Companies, universities, other agencies, professional societies, etc., are all welcome to participate, **maintaining control over how their data is stored and accessed.**
- Great value to smaller projects and targeted collections
- We are ready to start testing metadata exchange via OAI-PMH

Intent for NIST registry instances

- Work with others to improve data sharing and discovery through a federated system
- Possible “registry of registries” to facilitate access across multiple registries and institutions
- Eventually primarily have records for NIST-specific resources (projects, data, software, etc.)
- Host focused registry instances for particular applications in which NIST works or has an interest

Working group overview

- Case statement submitted Jan. 2016
- Proposed timeline of 12-18 months for a pilot materials resource registry system
- Approved July 2016
  - thus dates are now shifted back six months from the original proposal

Full timeline

- Month 1 (Jul ’16)
  - recruit domain specialists to participate in WG
- Month 2 (Aug/Sep ’16)
  - initiate discussions about conducting a survey of existing materials science data providers
  - develop 20 typical data discovery queries to inform metadata discussions
- Month 3 (Sep/Oct ’16)
  - hold meeting to draft 1st version of metadata extensions to Dublin Core
- Months 4-8 (Oct ‘16-Feb ’17)
  - disseminate draft to the materials science community, both within and external to RDA, and solicit feedback
- Month 8 (Feb ’17)
  - hold second two-day meeting to refine **metadata extensions and establish implementation pilot program**
  - E.g., NMRR, MDF, others TBD within WG
- Months 9-12 (Mar – Jun ’17)
  - implement pilot federated registry and recruit testers/evaluators
  - evaluate granularity issues
  - ***write best practices guidelines document***
- Months 13-15 (Jul – Sep ’17)
  - fine tune metadata definitions and document metadata development process: what worked well, what didn’t
  - expand content of pilot registry
- Months 16-18 (Oct – Dec ’17)
  - Prepare final document for delivery to RDA

Deliverables

- Two main deliverables for WG:
  1. Report containing materials metadata extensions to Dublin Core
  2. Pilot with connected registries to demonstrate harvesting

- Plus smaller items along the way (meetings, drafts, etc.)

Where are we?

- WG has been created with an initial roster of members
- At this meeting, we are identifying known efforts and discussing materials science queries
- Need to determine mechanism and date of next meeting. Telecon? Part of an existing meeting (e.g., CHiMaD)?
- Need to plan meeting for approx. March 2017

Identification of existing efforts

- Registries and projects with data sharing enabled
  - E.g., nanoHUB, Materials Data Facility, NoMaD, NIMS, Citrine, + ?
- Ontologies, vocabularies, etc.
  - Collect items on WG wiki page for this effort?
  - XML-based schema repository under development

Previous word~~play~~ work

- Some schemas, vocabularies, and ontologies
  - MatML, ThermoML, Plinius ontology, Ashino ontology, MatOnto, PREMΛP, ONTORULE (steels), SLACKS, MatOWL, matvocab
  - Nice review article:
    - X. Zhang, C. Zhao, and X. Wang, Computers in Industry, 73 (2015) 8-22.
- Cover various areas but not everything
- Some are being developed (at all levels), others are dormant
- Others are proprietary or haven't been publicly released

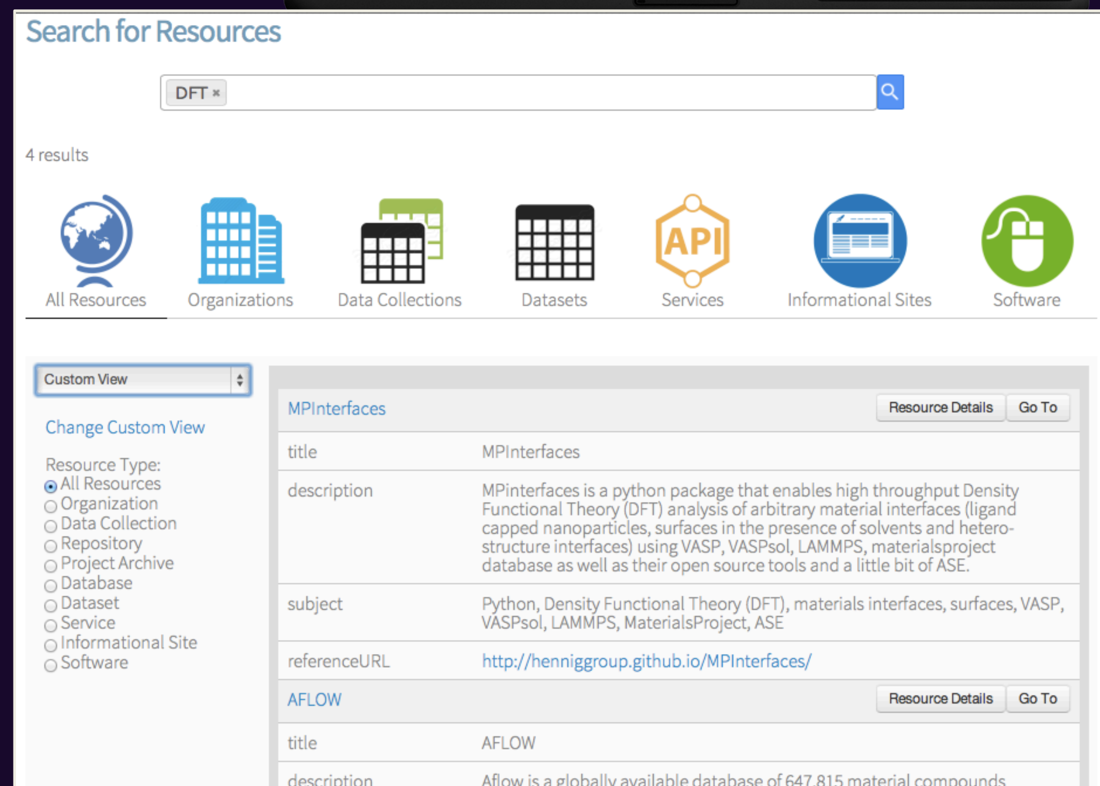
Example effort: NIST pilot

NIST Materials Resource Registry

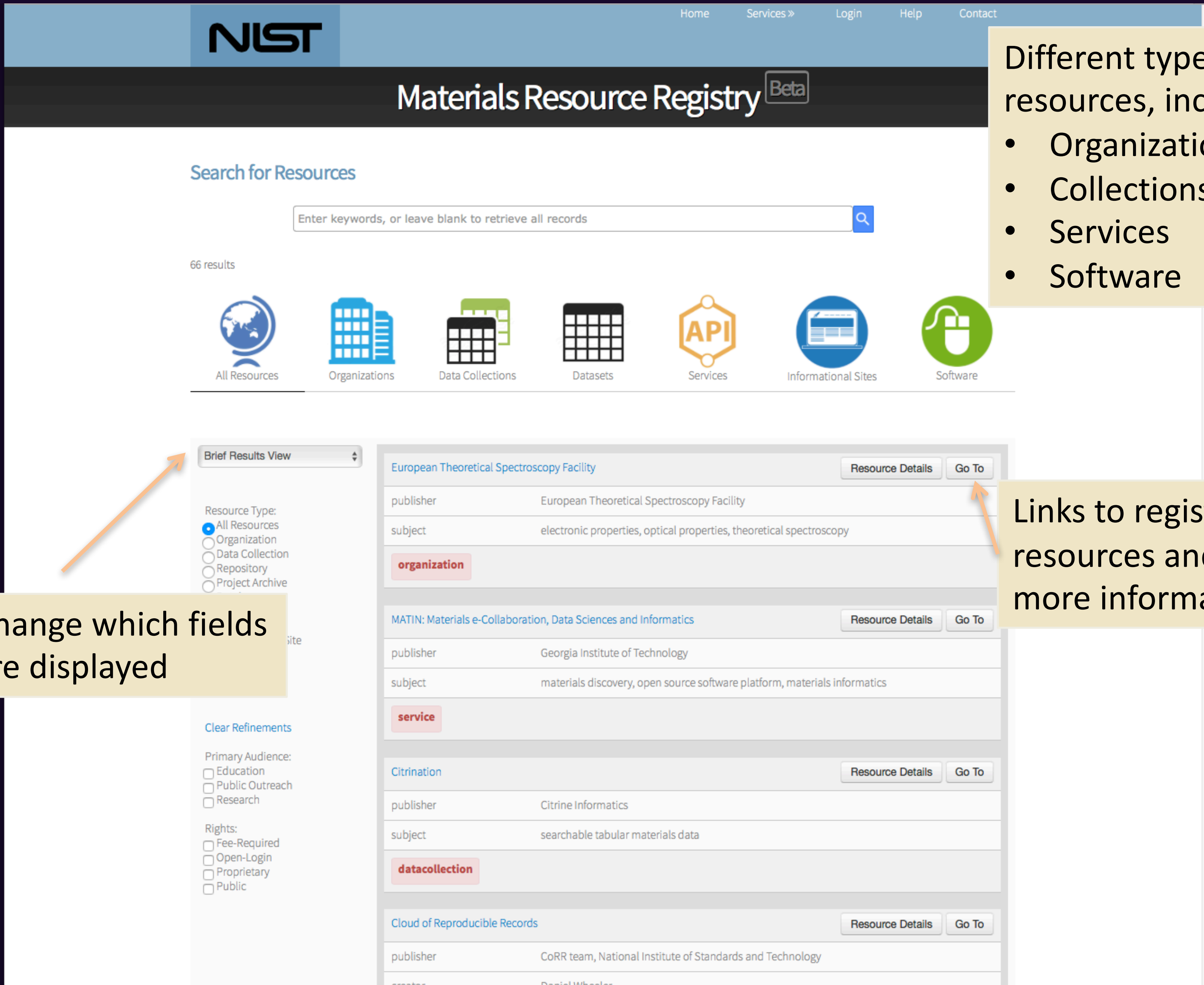
- General materials science resources
  - ~ 70 resources at the moment; working to migrate others from the MGI code catalog

- Intended to interact with other registries that are more focused and/or housed at other institutions

- OAI-PMH protocol enabled, built on the Materials Data Curation System platform
  - Code on GitHub
  - But don't require others to use the same software!



Browse Registered Resources



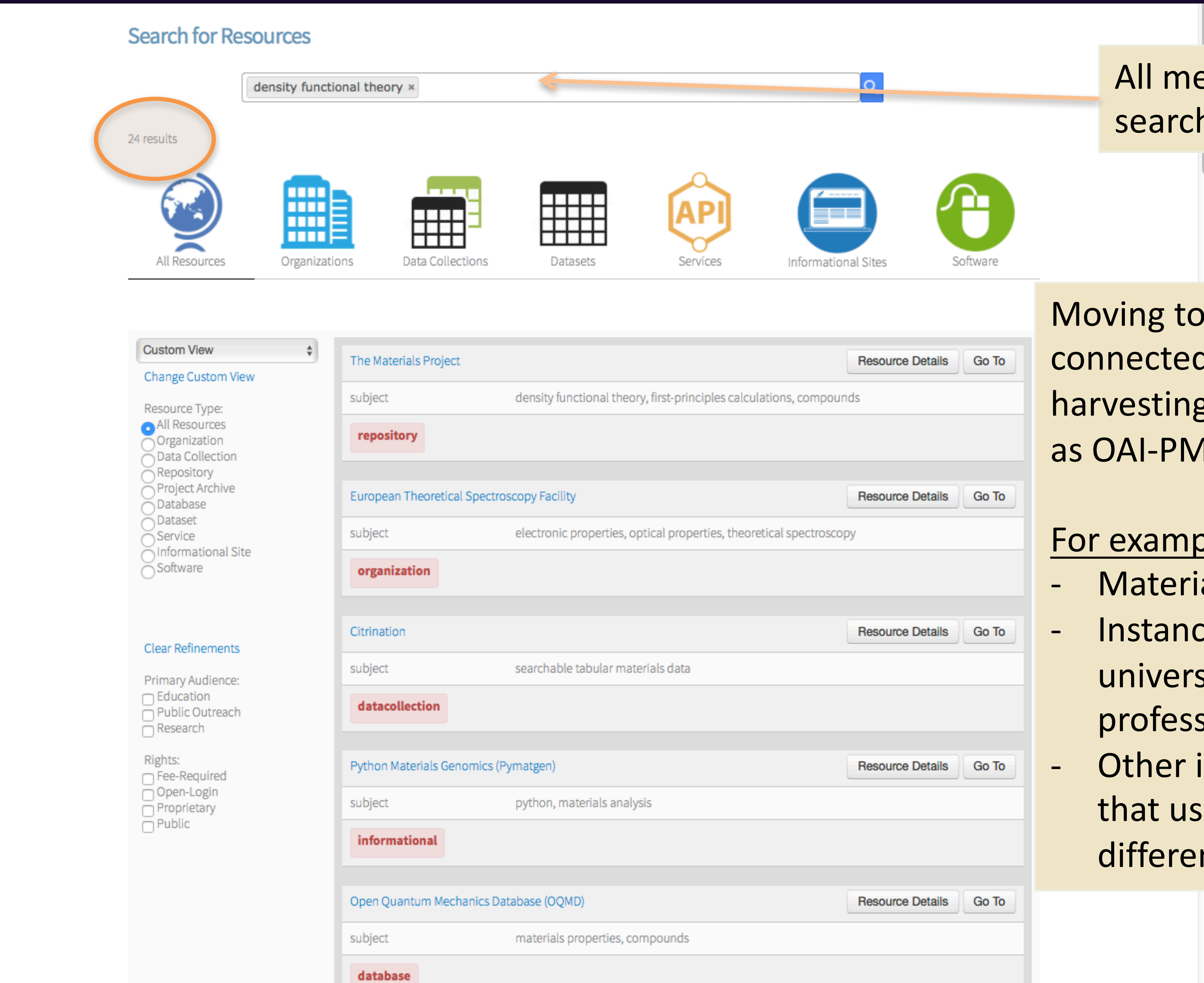
Different types of resources, including:

- Organizations
- Collections
- Services
- Software

Links to registered resources and more information

Change which fields are displayed

Search for resources



All metadata text is searchable

Moving toward resources connected by metadata harvesting protocols such as OAI-PMH

For example:

- Materials data facility
- Instances hosted by universities or professional societies
- Other implementations that use OAI-PMH but different code

Get More Information

Project REST API

Resource Details

National Renewable Energy Laboratory Materials Database

localid	DYJM46H37MU0FHYBE8NE
status	active
title	National Renewable Energy Laboratory Materials Database
shortName	NRELMatDB
publisher	National Renewable Energy Laboratory
contributor	Ann Deml, Stephan Lany, Haowei Peng, Vladan Stevanovic, Jun Yan, Pawel Zawadzki
name	
description	NRELMatDB is a computational materials database with the specific focus on materials for renewable energy applications including, but not limited to, photovoltaic materials, materials for photo-electrochemical water splitting, thermoelectrics, etc. The main goal of NRELMatDB is to enable and facilitate the access and exchange of computational data between different research groups following the guidelines outlined in the Presidential Materials Genome Initiative (http://www.whitehouse.gov/mgi).
subject	renewable energy, enthalpies of formation, enthalpies of decomposition, band gaps, dielectric functions
referenceURL	http://materials.nrel.gov/
referenceCitation	"Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies". V. Stevanovic, S. Lany, X. Zhang, A. Zunger, Physical Review B 85, 115104 (2012), http://dx.doi.org/10.1103/PhysRevB.85.115104, "Band-structure calculations for the 3d transition metal oxides in GW". S. Lany, Physical Review B 87, 085112 (2013), http://dx.doi.org/10.1103/PhysRevB.87.085112, "Semiconducting transition metal oxides". S. Lany, J. Phys.: Cond. Matter 27, 283203 (2015), http://dx.doi.org/10.1088/0953-8984/27/28/283203
primaryAudience	research
materialType	metal, semiconductor
structuralMorphology	bulk, crystalline
propertyClass	thermodynamic, optical, simulated
computationalDataAcquisitionMethod	density functional theory calculation
rights	public

database

Ok

Detailed information about resources, including

- who created them
- who maintains them
- what they contain
- how to access them

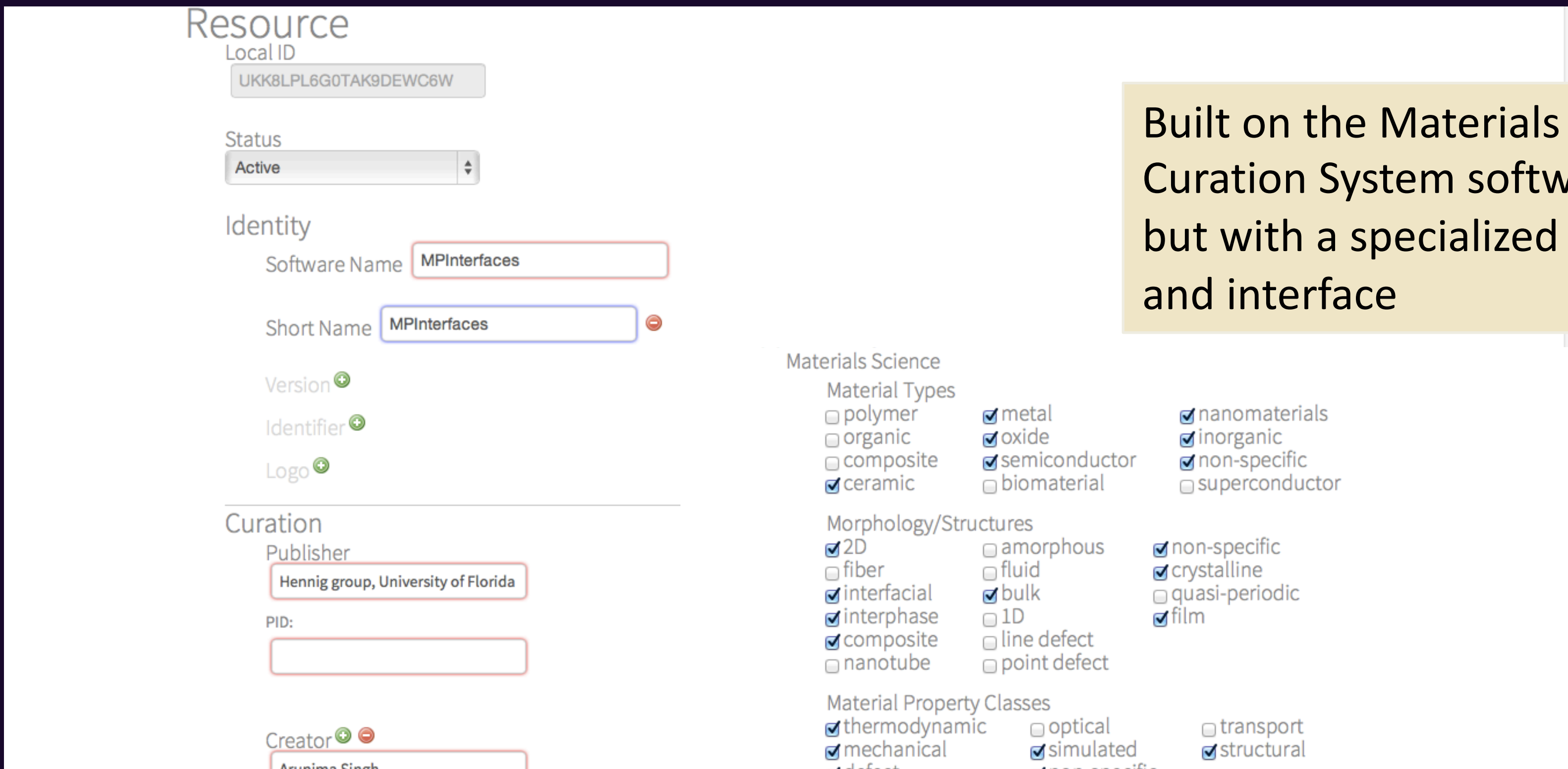
Plus links to the resources themselves

Experimental & computational

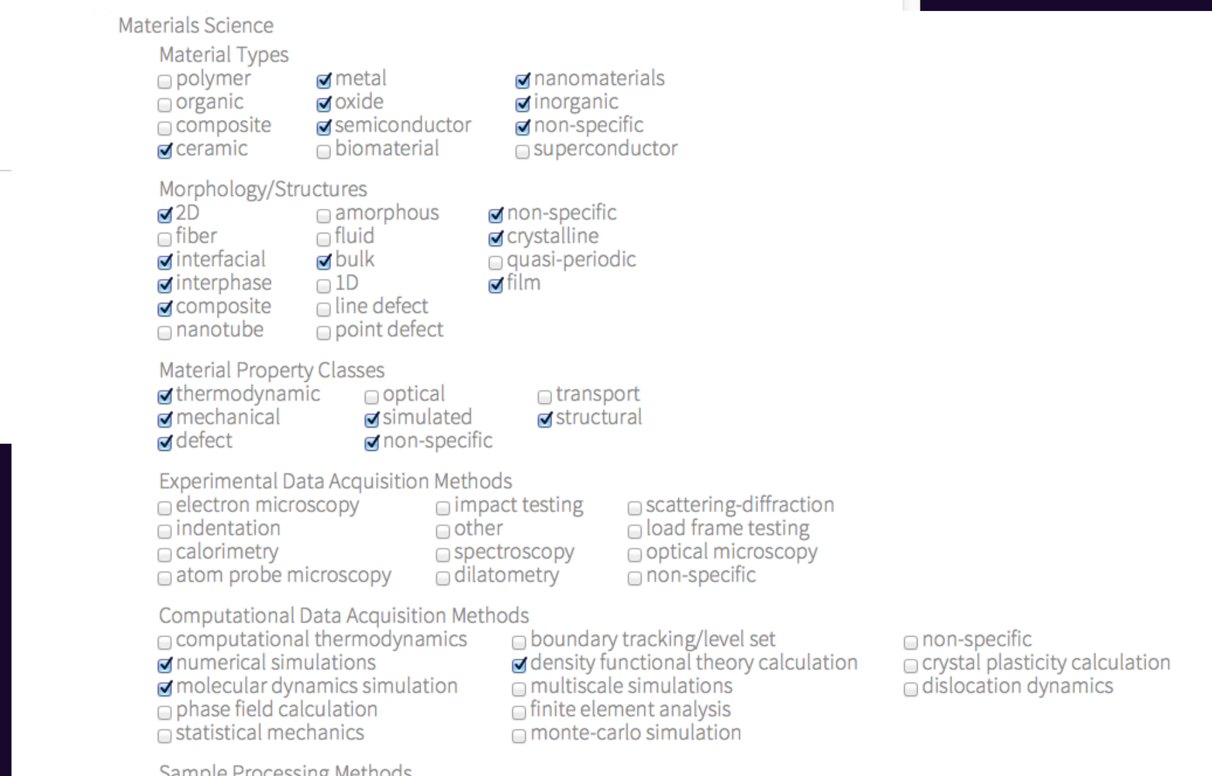
High-Throughput (Combinatorial) "Foundry" for Inorganic Materials: "Data on Demand"	
localid	8AWGGH6T0XYUAELMF98P
status	active
title	High-Throughput (Combinatorial) "Foundry" for Inorganic Materials: "Data on Demand"
publisher	National Institute of Standards and Technology (NIST)
creator	Martin Green
name	Nam Nguyen
emailAddress	htems@nist.gov
phoneNumber	(301) 975-5160
description	The high-throughput "foundry" for inorganic materials develops combinatorial measurement methods and metrologies for the rapid generation of comprehensive and consistent datasets. Manufacturers of devices based on functional inorganic materials can leverage these methodologies and/or datasets to screen and select new materials more rapidly and intelligently.
subject	combinatorial materials science
referenceURL	https://mgi.nist.gov/high-throughput-combinatorial-foundry-inorganic-materials-data-demand
primaryAudience	research
materialType	inorganic
structuralMorphology	film
propertyClass	optical, non-specific
experimentalDataAcquisitionMethod	spectroscopy, scattering-diffraction
sampleProcessing	vapor deposition
rights	public
projectarchive	

Ok

Add a Resource



Built on the Materials Data Curation System software, but with a specialized schema and interface



Seed NMRR metadata fields

Version 1. These will change based on WG efforts!

Material Types	Metal Semiconductor Ceramic Polymer Biomaterial	Organic Inorganic Oxide Composite Nanomaterials	Superconductor Non-Specific Other	?	(recommended)
Morphology/Structures	Crystalline Amorphous Fluid Quasi-periodic Bulk 2-Dimentional	1-Dimentional Film Nanotube Fiber Composite Interfacial	Interphase Line Defect Point Defect Non-Specific Other	?	(recommended)
Material Property Classes	Optical Mechanical Thermodynamic	Structural Simulated Diffusion	Defect Non-Specific Other	?	(recommended)
Experimental Data Aquisition Methods	Electron Microscopy Scattering/Diffraction Calorimetry Load Frame Testing	Atom Probe Microscopy Spectroscopy Optical Microscopy Impact Testing	Indentation Dilatometry Other	?	(recommended)
Computational Data Aquisition Methods	Density Functional Theory Molecular Dynamics Simulation Numerical Simulations Multiscale Finite Element Analysis Computational Thermodynamics	Statistical Mechanics Dislocation Dynamics Phase Field Crystal Plasticity Other		?	(recommended)
Sample Processing Methods	Casting Annealing Vapor Deposition Milling	Extrusion Pressing Exfoliation Melt Blending	Polymerization Curing Evaporation Other	?	(recommended)

Working group activities

WG items for discussion

- Common location for our work on RDA WG website
- Identification of existing projects and resources. What other efforts are represented here?
- Identification of vocabs/ontologies/etc.
- Identification of technical issues
- Planning for follow-up meetings and activities
- Identification of volunteers and interested people

What do you want to share or be able to find?
What data sharing efforts are under way?

Sample queries

- Al6065 mechanical properties
- Environmental degradation data for PE in humidity
- Finite element models of turbine blades
- Optical micrographs of gamma phases in Ni_3Al
- Compound formation energies for B2-NiAl
- Sintering temperatures for zirconia powders
- Dielectric properties for GaAs
- Calphad models of InGaAs and related materials
- Data for x alloy processed with y method and analyzed with z equipment
- …

Additional Queries?

RDA websites

- Interest group
  - https://rd-alliance.org/groups/rdacodata-materials-data-infrastructure-interoperability-ig.html
- Working group
  - https://rd-alliance.org/groups/working-group-international-materials-resource-registries.html
- Case statement
  - https://rd-alliance.org/group/international-materials-resource-registries-wg/case-statement/case-statement-rda-working-group