Materials Data IG, IMMR WG & Ontologies TG

Co-Chairs James Warren (NIST),Laura Bartolo, (CHiMaD/Northwestern University) Takuya Kadohira (NIMS), Adham Hashibon (Fraunhofer IWM), Alysia Garmulewicz (Universidad de Santiago de Chile)

- General Agenda: Thursday 24th October 2019 Breakout 4 09:00 - 10:30

Undergraduate Center - U149 U6 (Konecranes)

- Brief Introduction: James Warren, National Institute of Standards & Technology
  - 5 minutes
- Panel Discussion on Diverse RDA Outputs, FAIR Data, & materials data lifecycle
  - Zach Trautt, NIST; Asahiko Matsuda, NIMS; Emanuele Ghedini, UNIBO
    Panel Discussion: 45 minutes
  - General Discussion: 20 minutes
- International Materials Resource Registries WG
  - Ray Plante, NIST : 10 minutes
- Materials Ontologies TG, New Co-Chair & Update
  - Gerhard Goldbeck, Goldbeck Consulting & Cate Brinson, Duke University: 10 minutes

Materials Science and Engineering and Research Data Alliance: An Individual Perspective

Zachary Trautt, Material Measurement Laboratory,
National Institute of Standards and Technology, USA

About Research Data Alliance

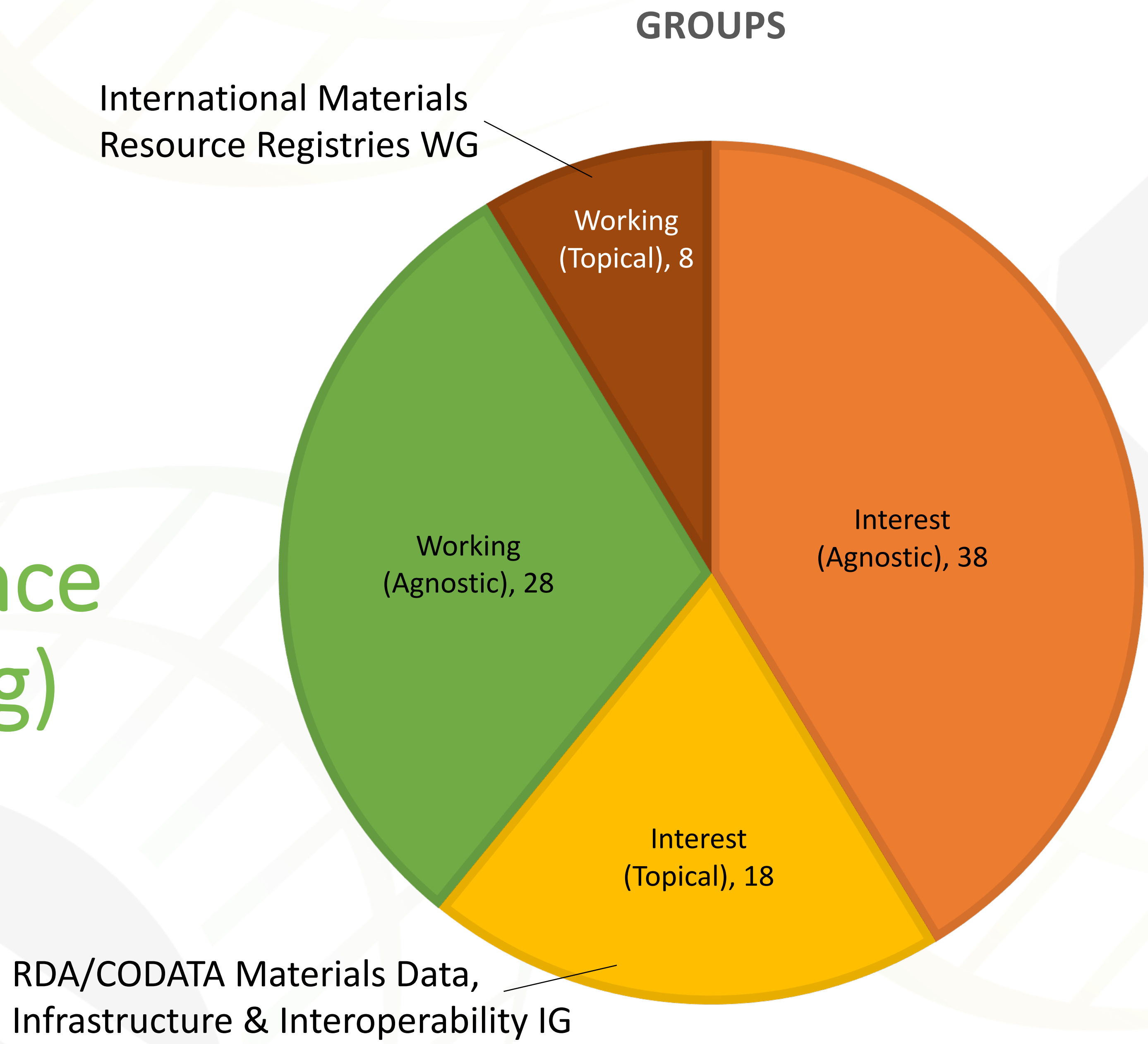
By the Numbers

- Launched as a community-driven initiative in 2013
- More than 8,800 members
- Participation from 137 countries
- 50+ Interest Groups (IGs)
- 30+ Working Groups (WGs)

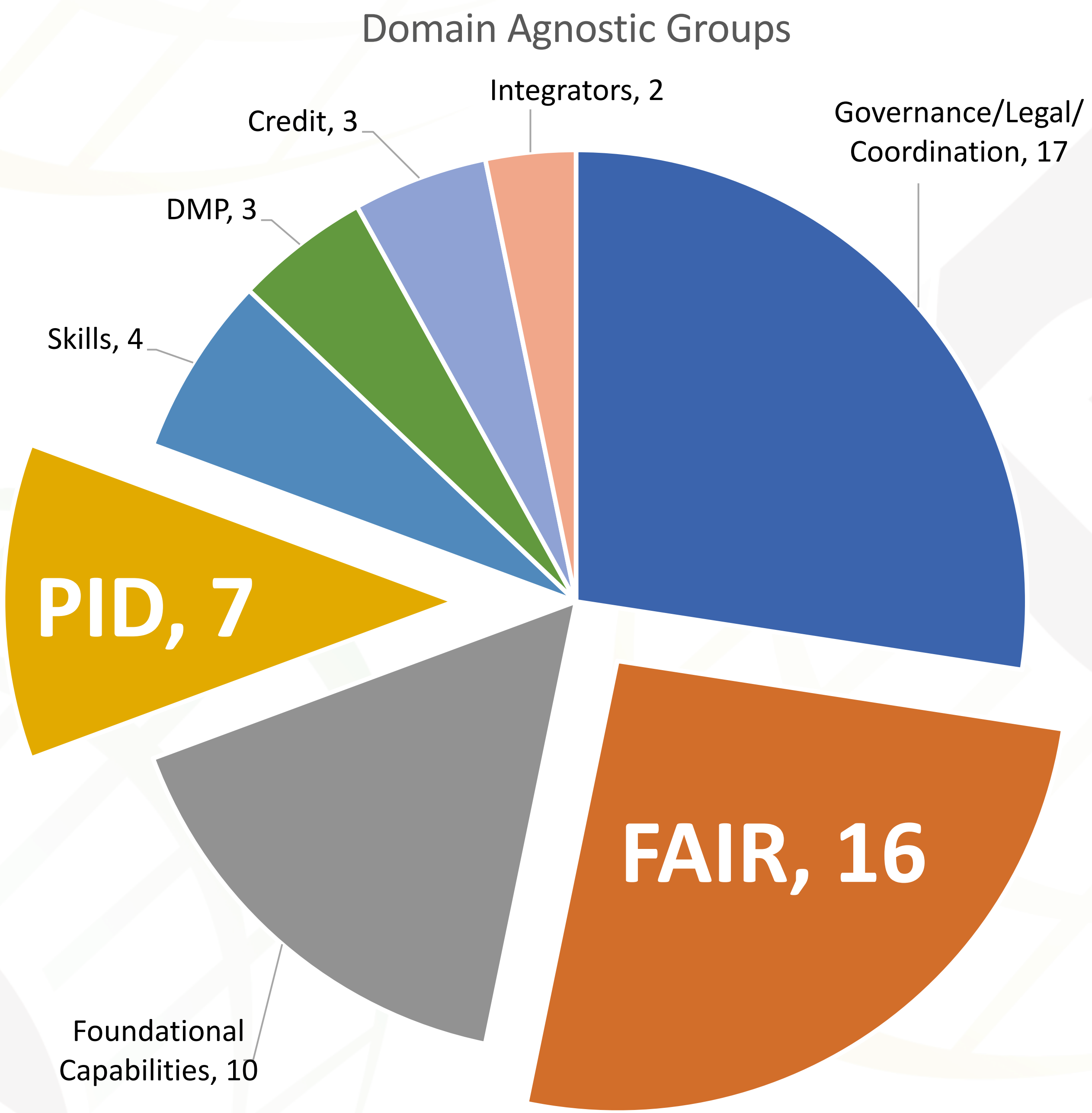
Group Functions

- IGs operate without a time limit, and are committed to enabling data sharing, exchange, or interoperability
- WGs have a lifespan of 12-18 months and are the main vehicle for producing RDA Outputs

Research Data Alliance (my manual tagging)



Research Data Alliance (my manual tagging)



FAIR Focused Groups

Interest Groups

- GO FAIR IG
- Data Discovery Paradigms IG
- Domain Repositories IG
- Open Science Graphs for FAIR Data IG
- Metadata IG
- Data Foundations and Terminology IG
- Vocabulary Services IG
- Software Source Code IG

Working Groups

- FAIR Data Maturity Model WG
- FAIRSharing Registry: connecting data policies, standards & databases WG
- RDA / TDWG Metadata Standards for attribution of physical and digital collections stewardship
- Research Data Repository Interoperability WG
- InteroperAble Descriptions of Observable Property Terminology WG (I-ADOPT WG)
- Data Description Registry Interoperability (DDRI) WG
- Metadata Standards Catalog WG
- Research Metadata Schemas WG

Persistent Identifier Focused Groups

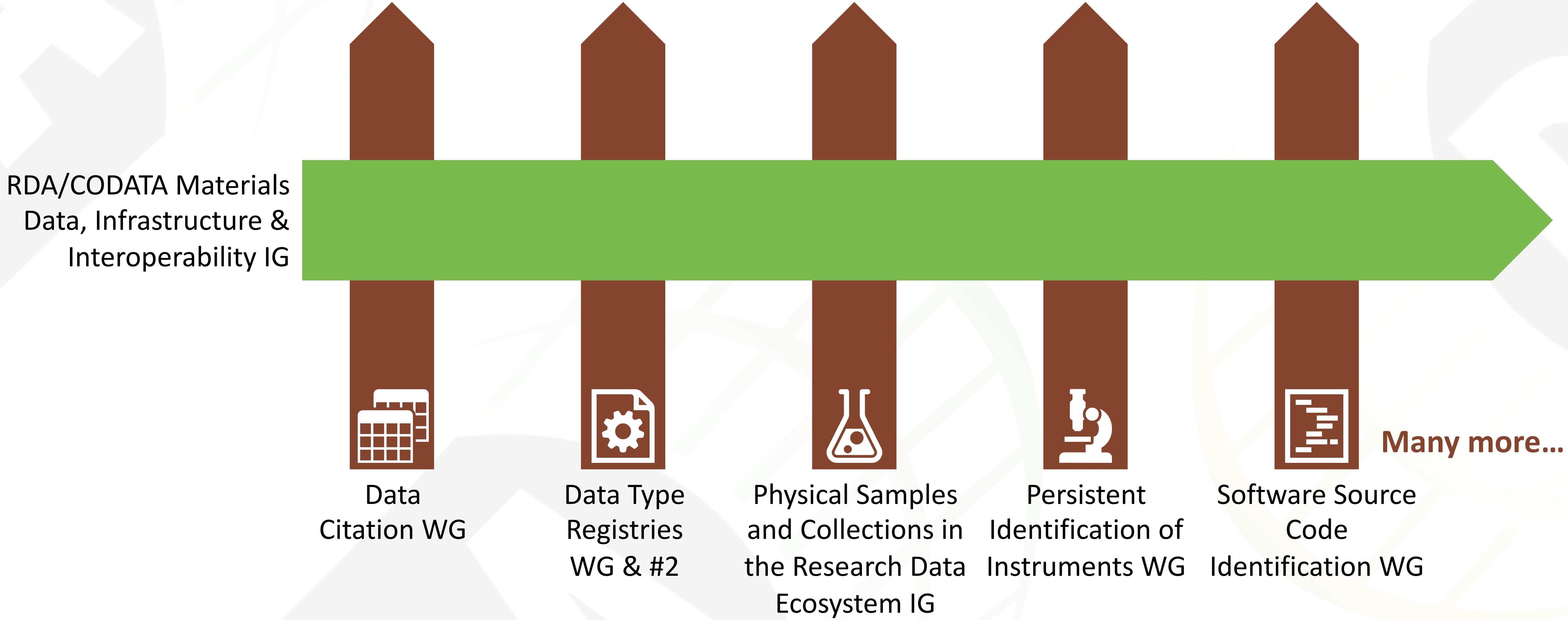
General PID Concerns

- PID IG
- PID Kernel Information Profile Management WG

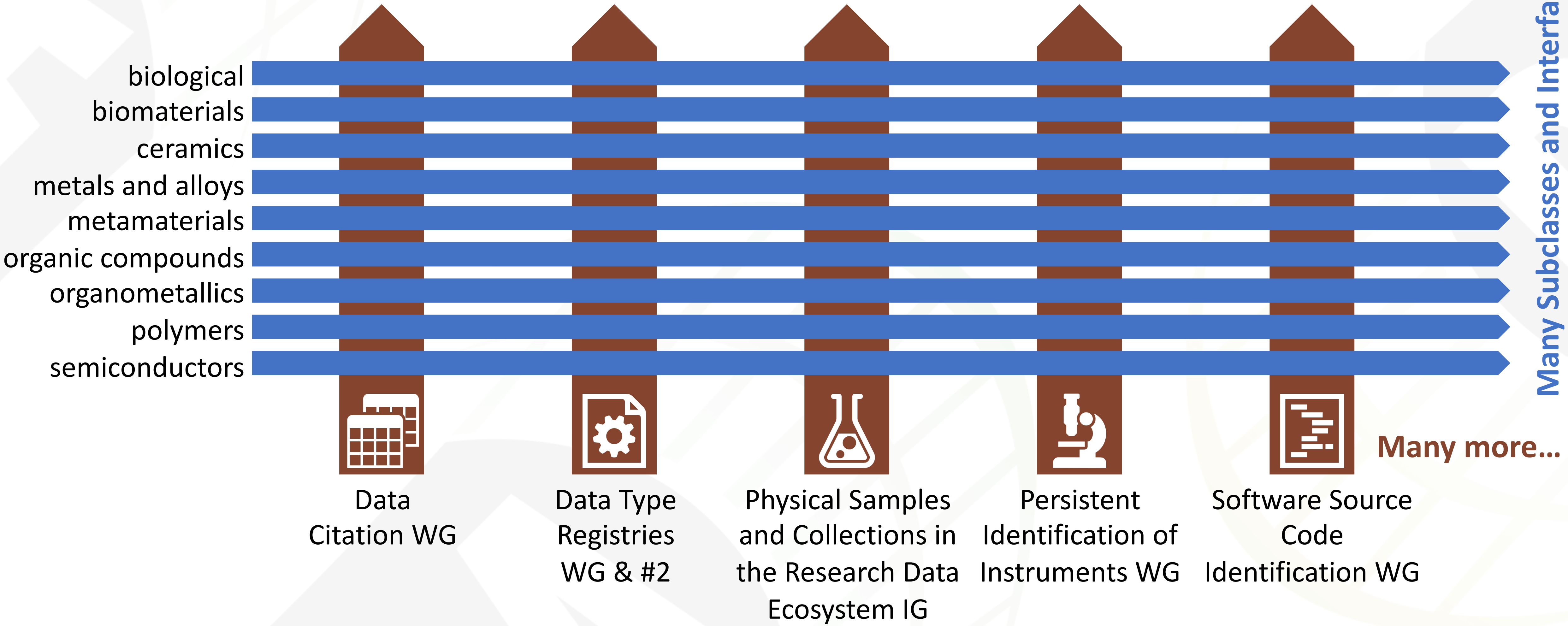
Persistent Identifiers for Things

- Data Citation WG
- Data Type Registries WG & #2
- Physical Samples and Collections in the Research Data Ecosystem IG
- Persistent Identification of Instruments WG
- Software Source Code Identification WG

Materials and Cross-Cutting RDA Groups



Challenge: Diverse Material Types



Our Past Work

International Materials Resource Registries WG

International Materials Resource Registries WG

- First WG established by RDA/CODATA Materials Data, Infrastructure & Interoperability IG
- Develop metadata standards required to establish a network of International Materials Resource Registries
- A resource registry is a system that harvests and makes searchable high-level metadata descriptions of resources held by data repositories, archives, organizations, websites, and services to aid scientists in industry, universities, and government labs in the discovery of data relevant to their research and work interests

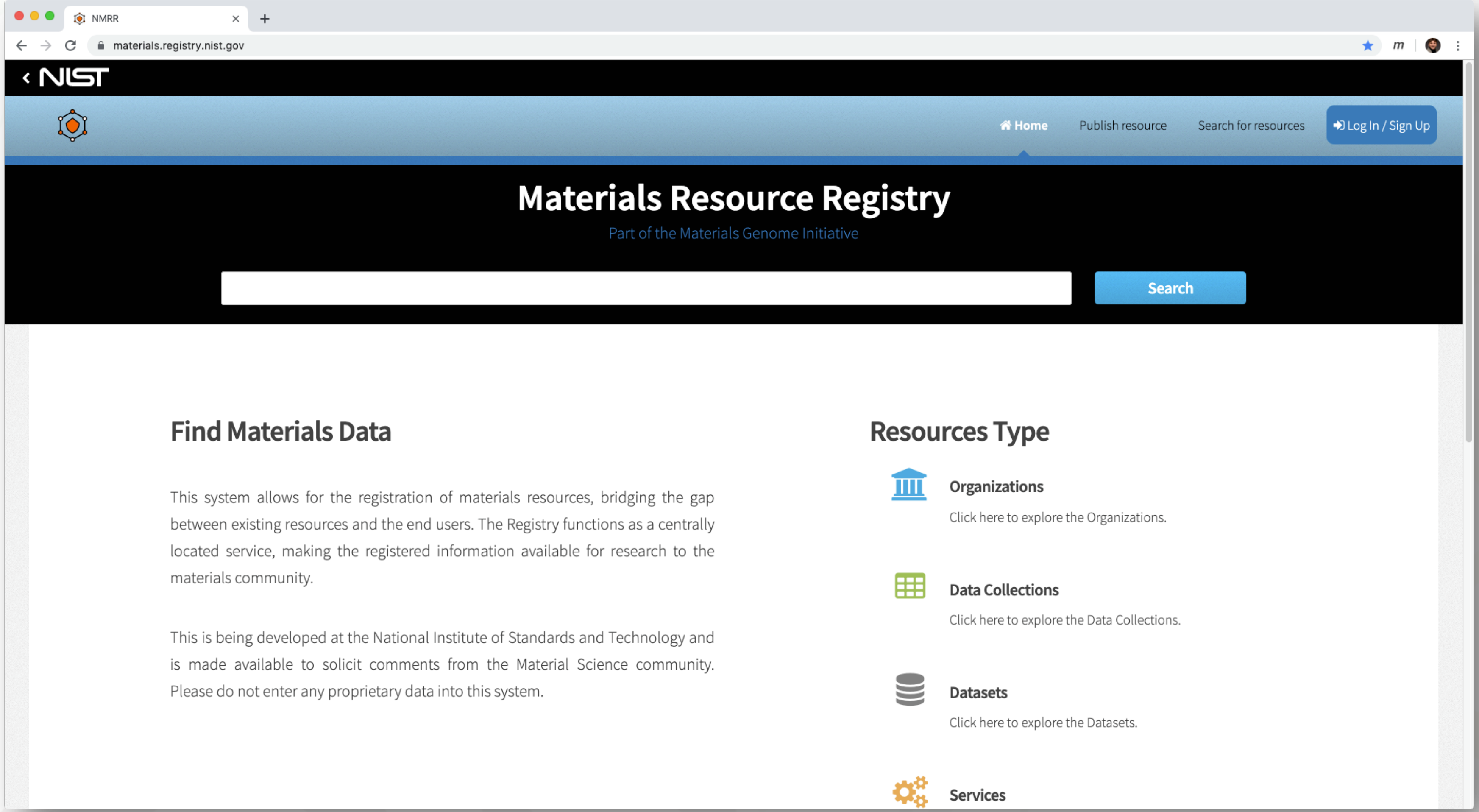
Proposed Timeline in Case Statement

Month: 1-3 2016/01 - 2016/03	Month: 4-8 2016/04 - 2016/08	Month: 9-16 2016/09 - 2017/04	Month: 17-18 2017/05 - 2017/06
• Recruit Subject Matter Experts • Discuss and Survey Existing Service Providers • Meet and Draft First Version of Metadata Schema / Vocab	• Disseminate Draft Schema / Vocab and Solicit Feedback • Meet and Refine Schema / Vocab • Establish Pilots - NIST (U.S.) - MDF (U.S.)	• Implement Pilots - NIST (U.S.) - MDF (U.S.) • Fine Tune Schema / Vocab • Document: - Process - What worked - What didn't - Etc.	• Prepare Final Report

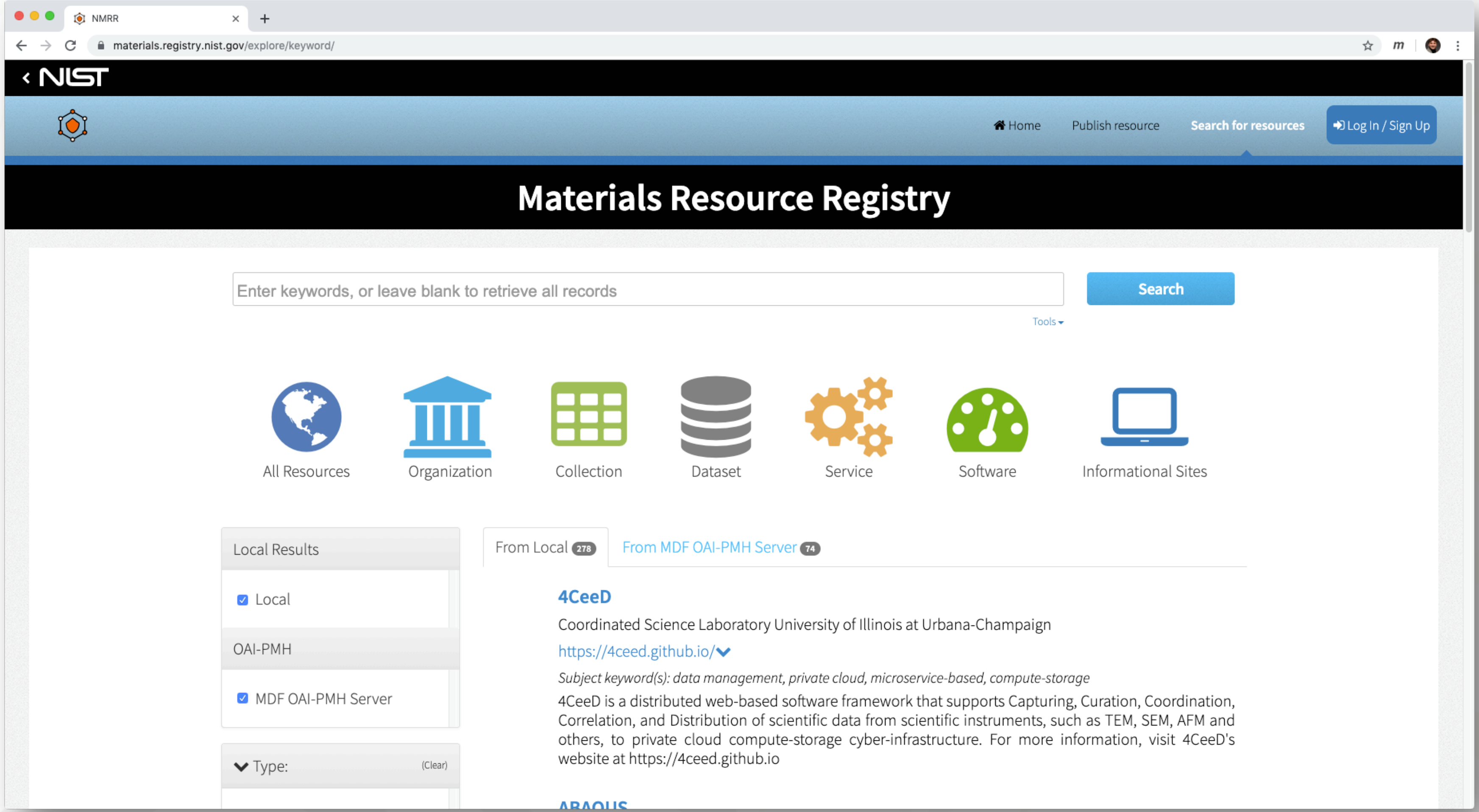
Primary Output

Technology

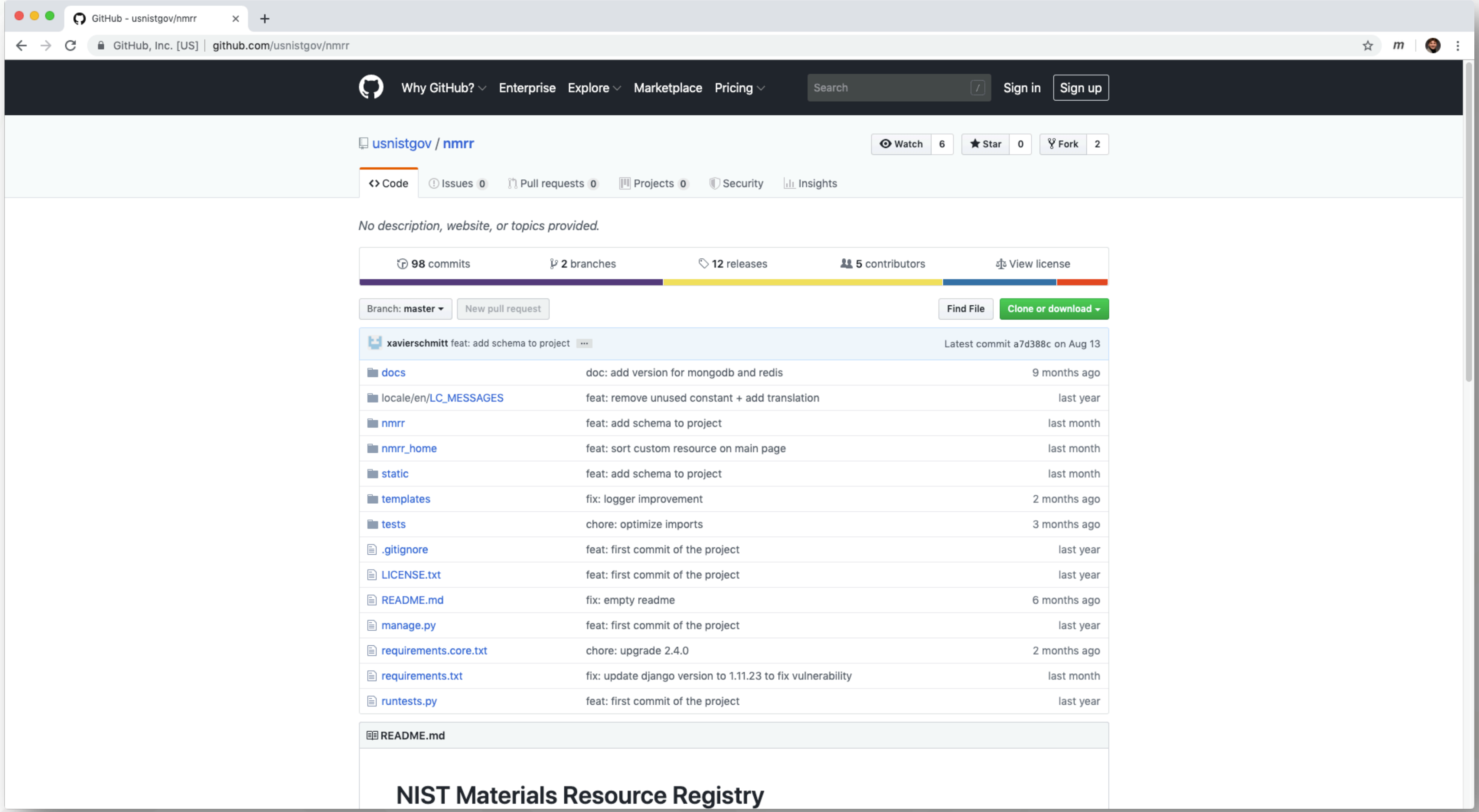
Output: Resource Registry Federation



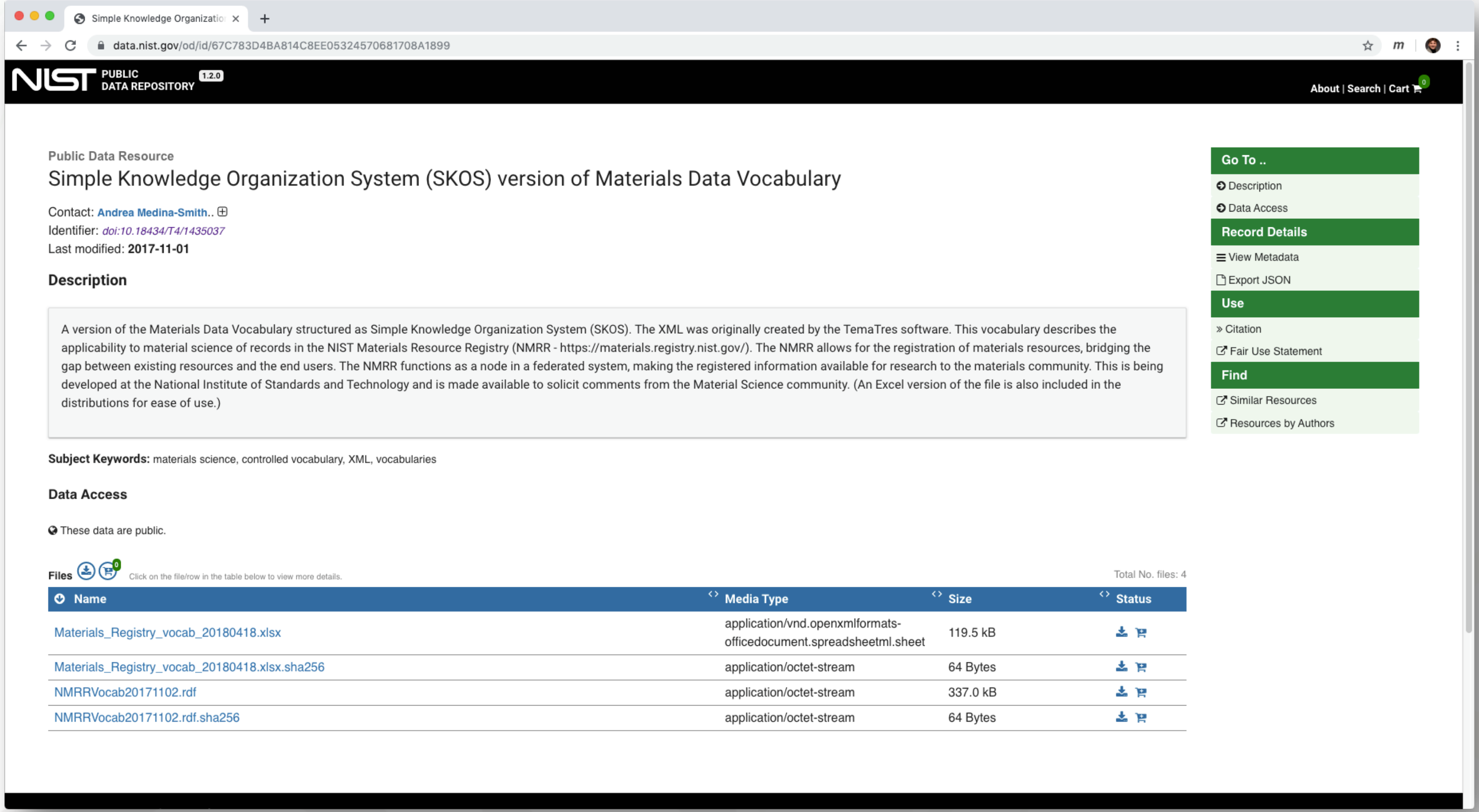
Output: Resource Registry Federation



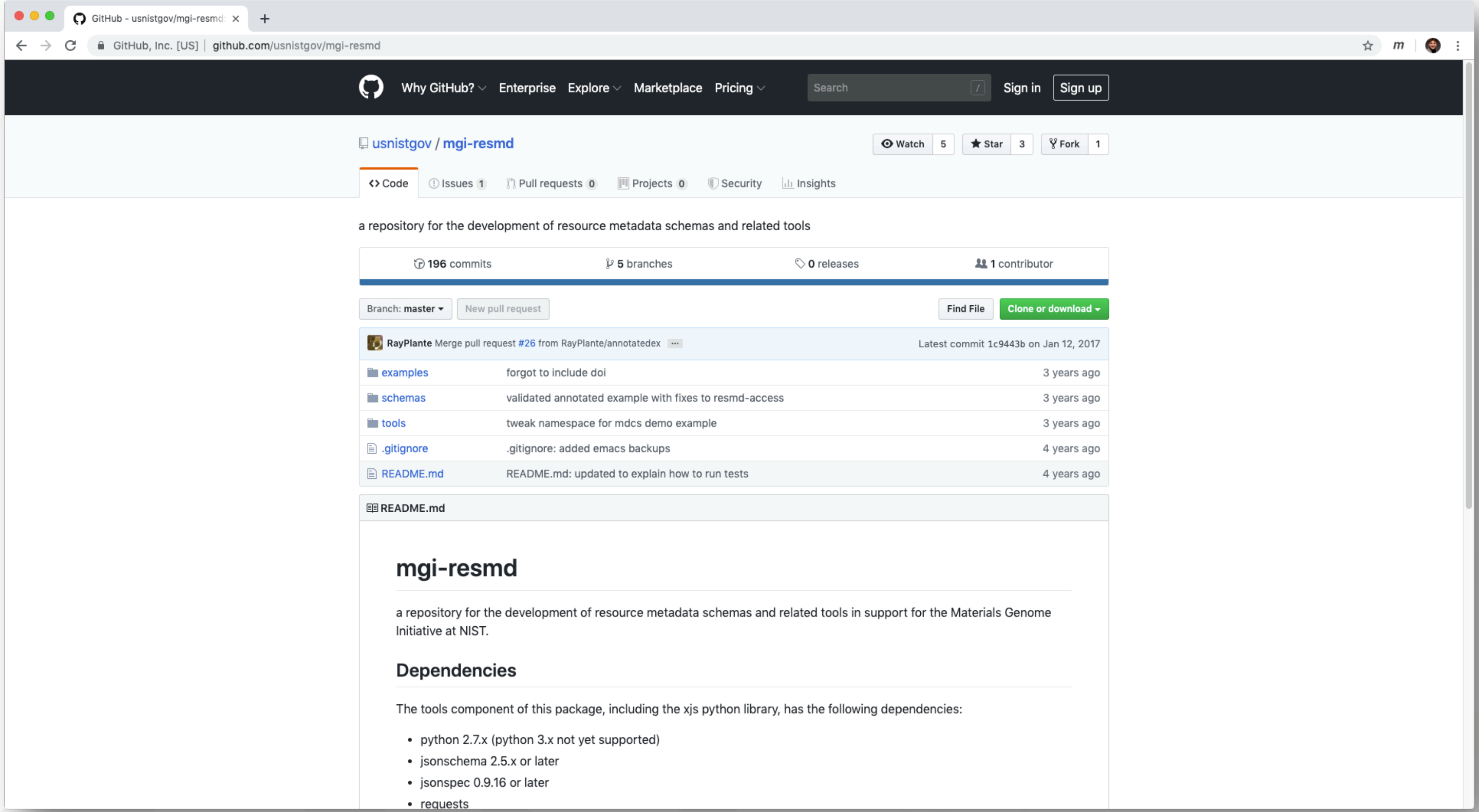
Output: Open Source Software



Output: Vocabulary



Output: Metadata Schema



Primary Output: Technology

Strengths

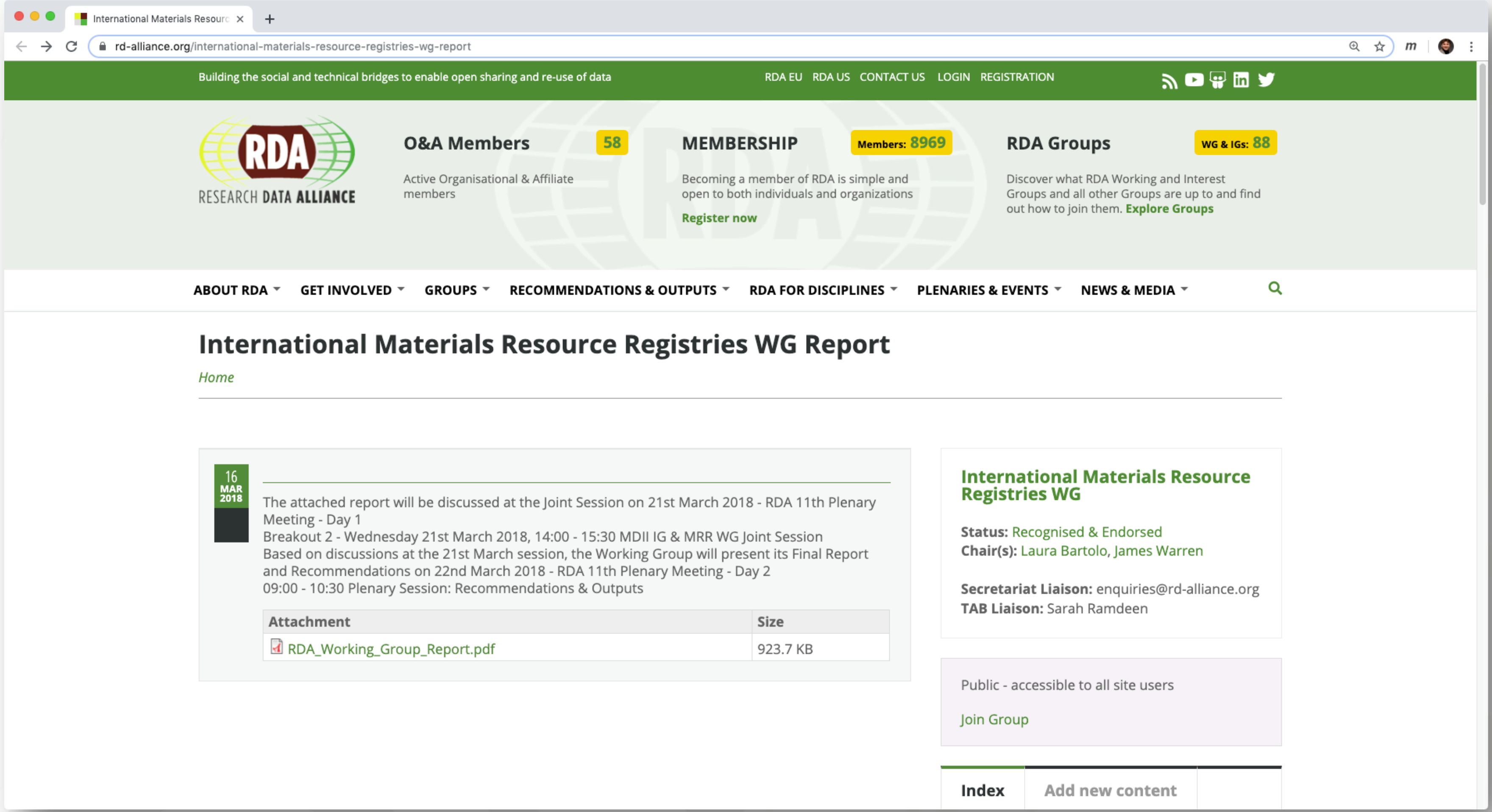
- We Built It, Please Come!
- Working Solution
  - Federation of Services
  - Open Source Software
  - Schema
  - Vocabulary
- Enables a path for growth
  - More registries in more places
  - More users

Weaknesses

- Technology changes rapidly
  - Keeping up with software and security updates/patches
  - **Keeping up with Innovation**
- Requires support model
  - Data producers
  - Data consumers
  - Service providers
- Requires sustainability model
  - Cost Drivers/Funding Sources
  - Demonstrate Value

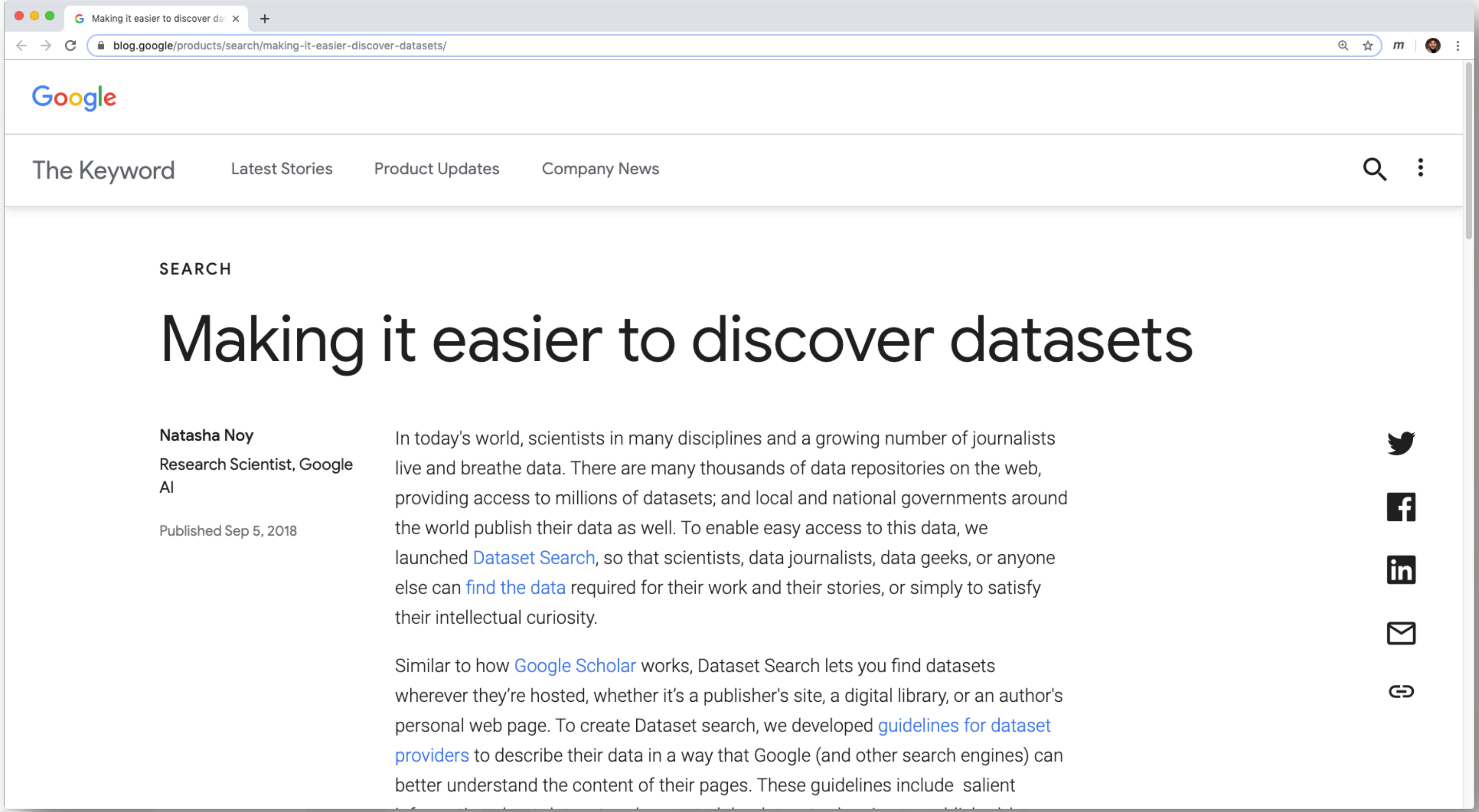
Keeping up with Innovation…

March 2018: Written Report Published



6 Months Later…

September 2018: Google Dataset Search



Making it easier to discover da ×

blog.google/products/search/making-it-easier-discover-datasets/

Google

The Keyword   Latest Stories   Product Updates   Company News

SEARCH

Making it easier to discover datasets

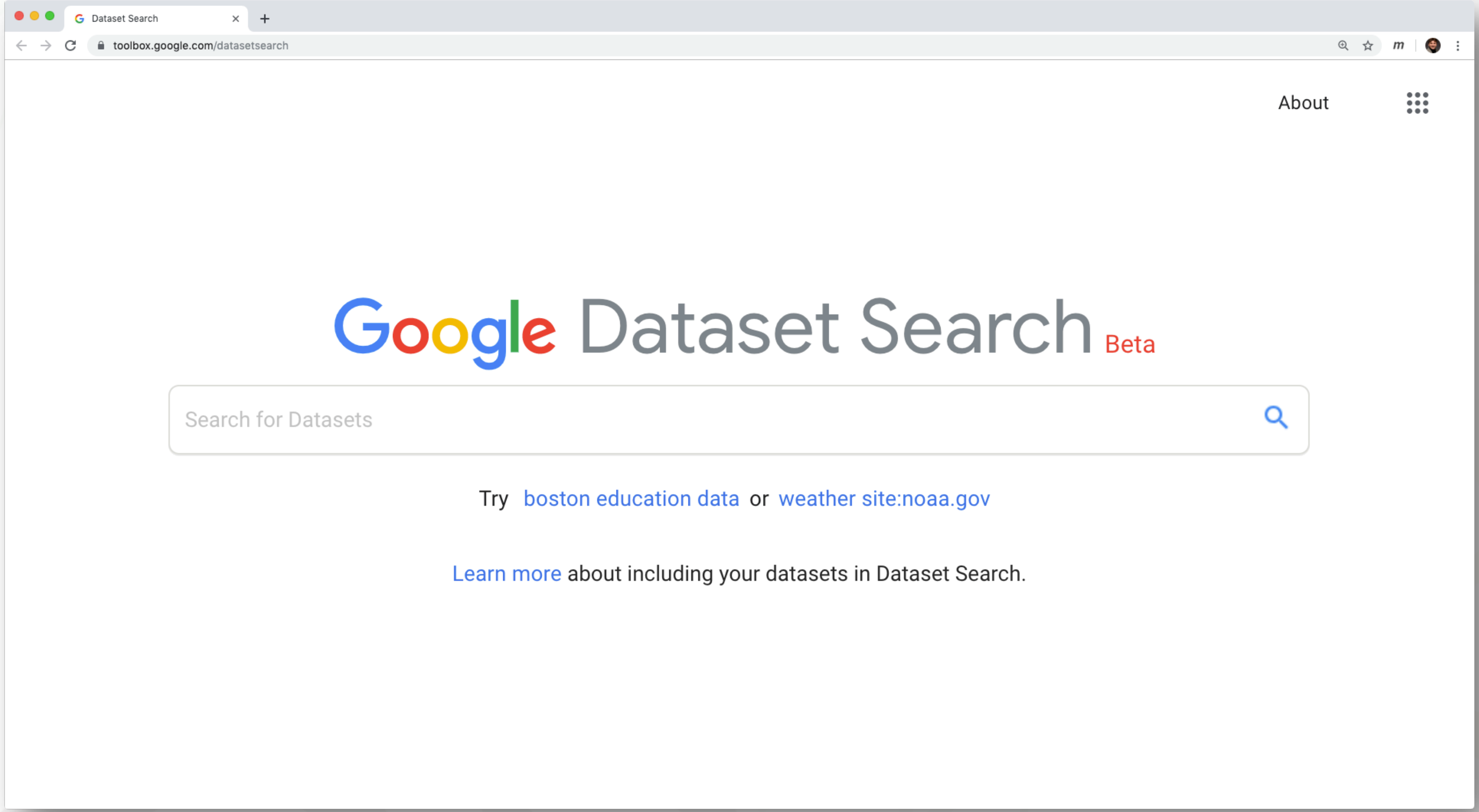
Natasha Noy
Research Scientist, Google AI

Published Sep 5, 2018

In today's world, scientists in many disciplines and a growing number of journalists live and breathe data. There are many thousands of data repositories on the web, providing access to millions of datasets; and local and national governments around the world publish their data as well. To enable easy access to this data, we launched Dataset Search, so that scientists, data journalists, data geeks, or anyone else can find the data required for their work and their stories, or simply to satisfy their intellectual curiosity.

Similar to how Google Scholar works, Dataset Search lets you find datasets wherever they're hosted, whether it's a publisher's site, a digital library, or an author's personal web page. To create Dataset search, we developed guidelines for dataset providers to describe their data in a way that Google (and other search engines) can better understand the content of their pages. These guidelines include salient

September 2018: Google Dataset Search



What should we do next?

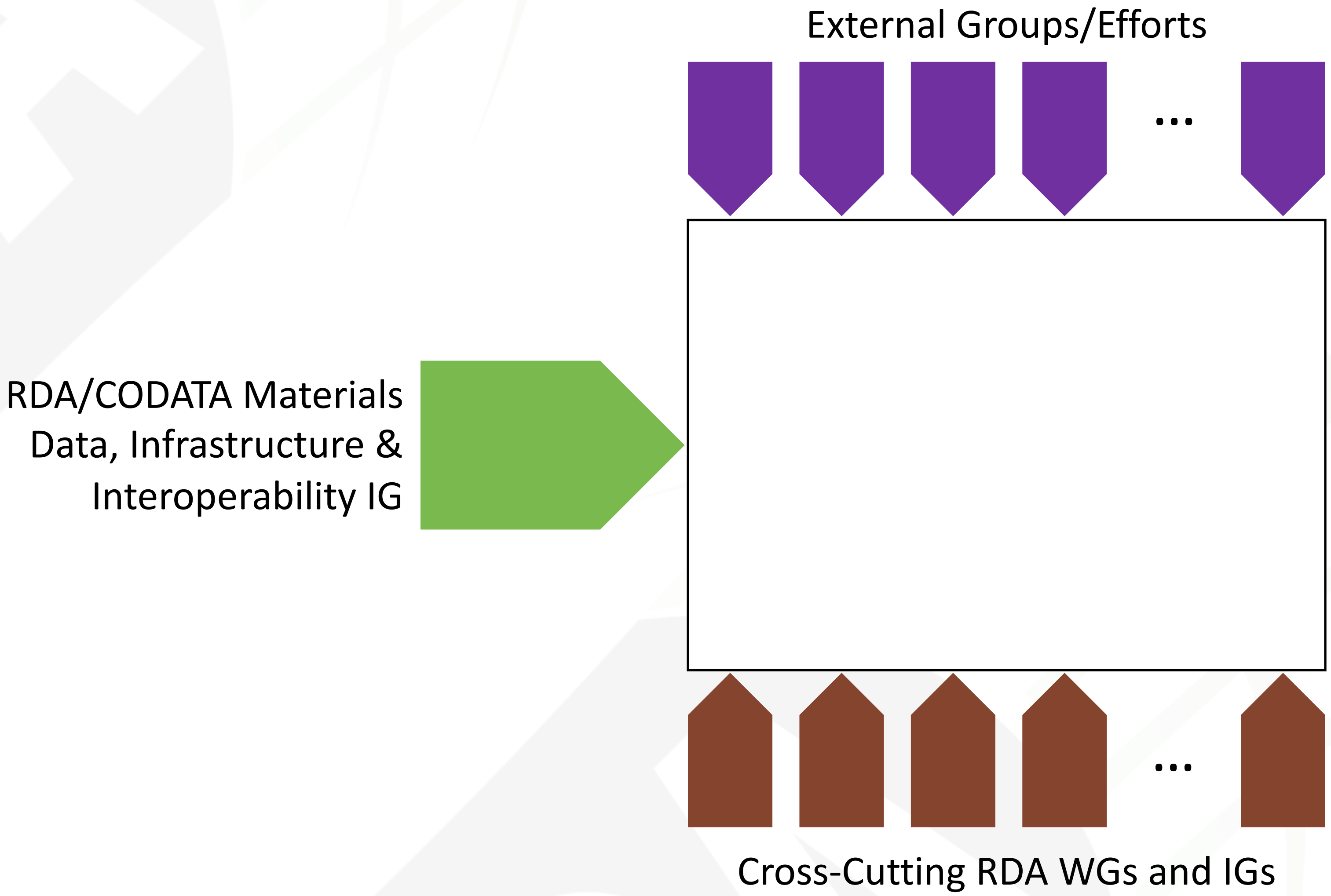
Materials and Cross-Cutting RDA Groups

RDA/CODATA Materials Data, Infrastructure & Interoperability IG

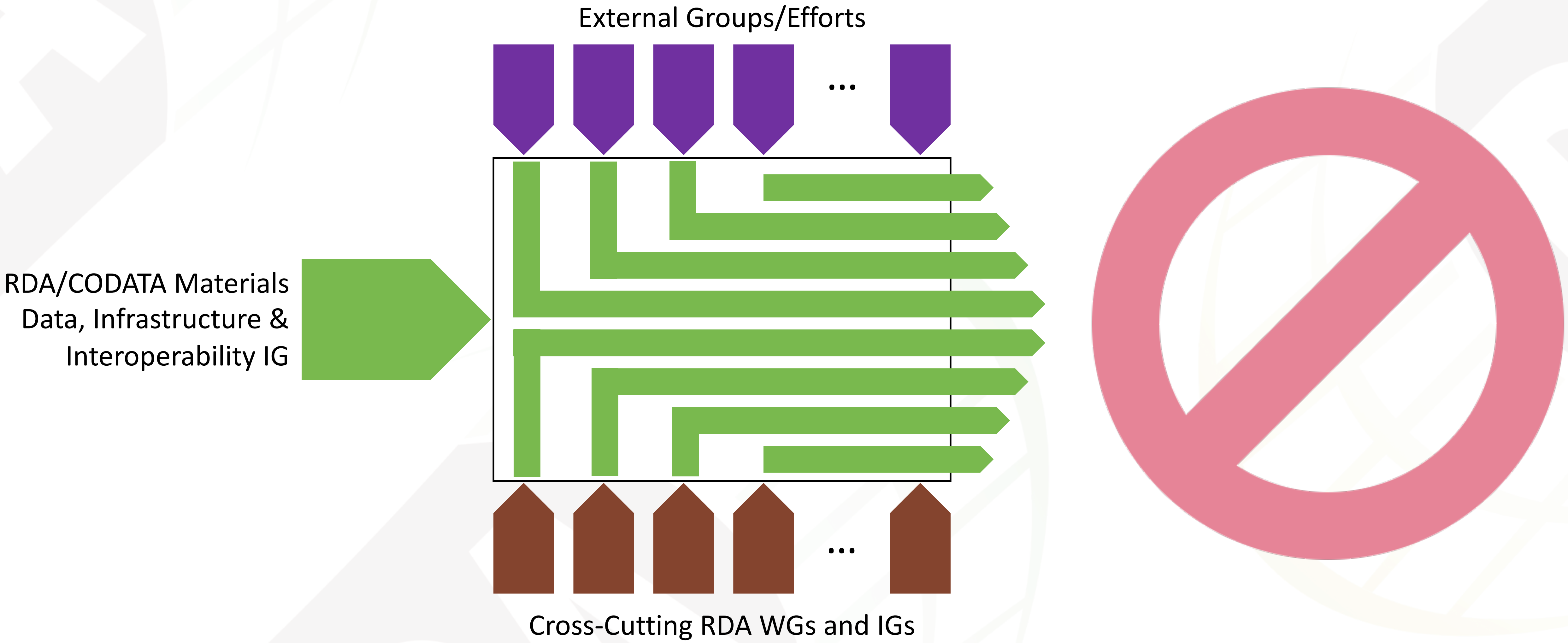
- Data Citation WG
- Data Type Registries WG & #2
- Physical Samples and Collections in the Research Data Ecosystem IG
- Persistent Identification of Instruments WG
- Software Source Code Identification WG

Many more…

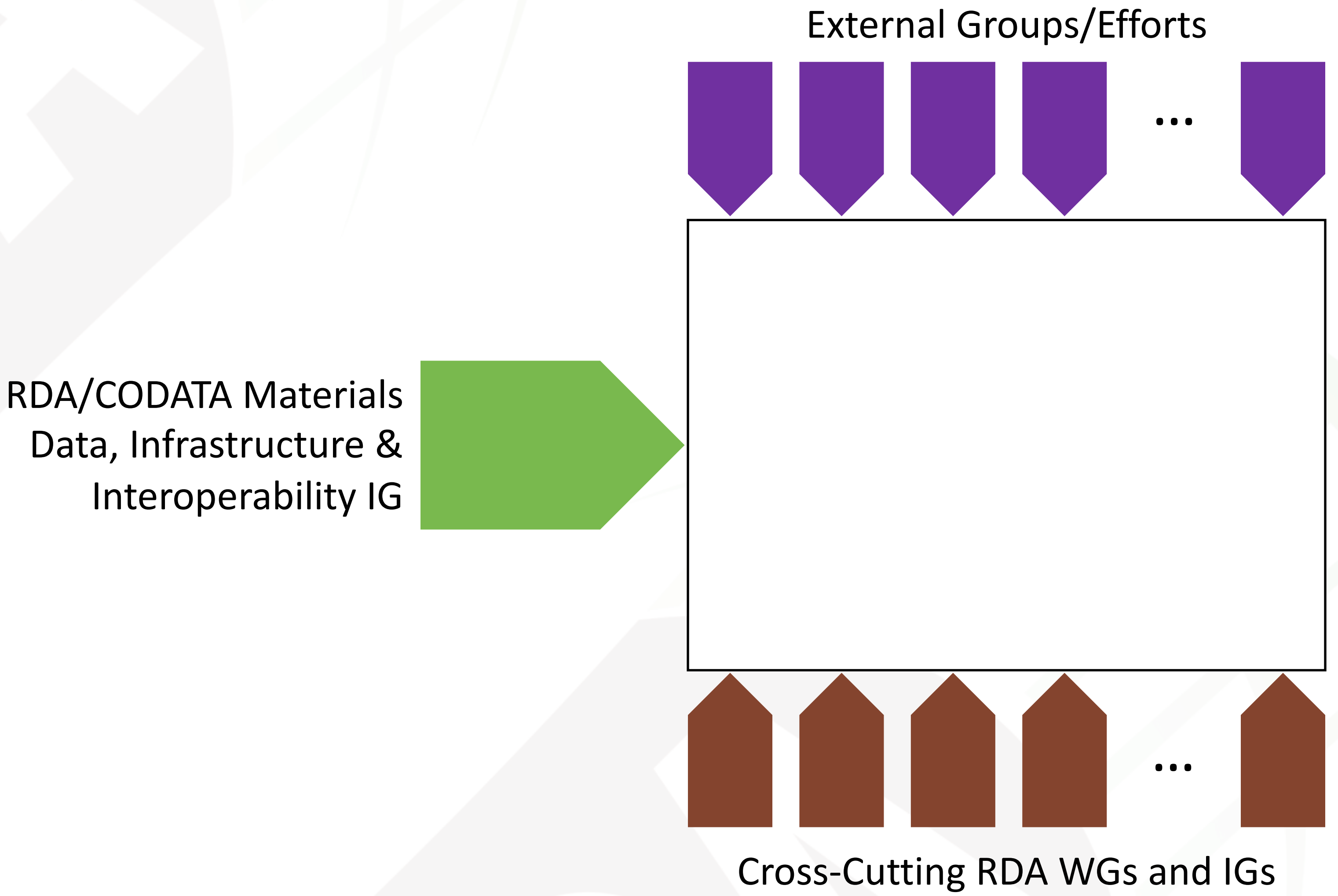
What should we do next?



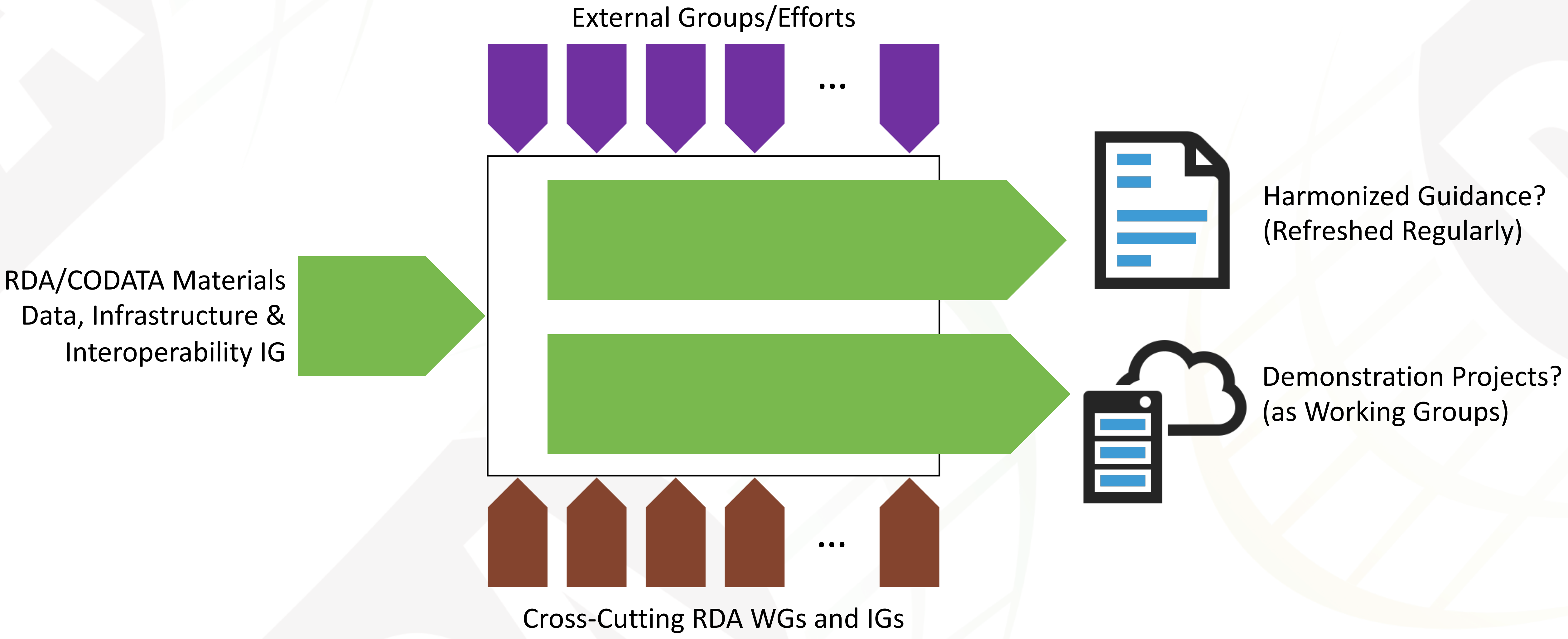
Launch "n" working groups?



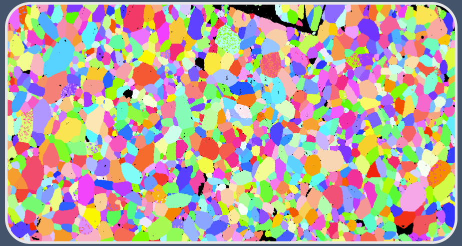
What should we do next?



Harmonization and Demonstration?

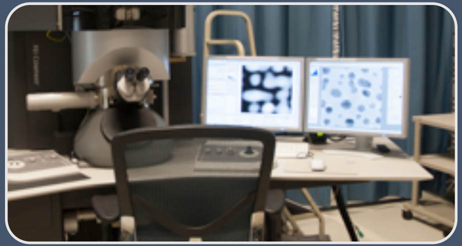


Possible Demonstration Projects



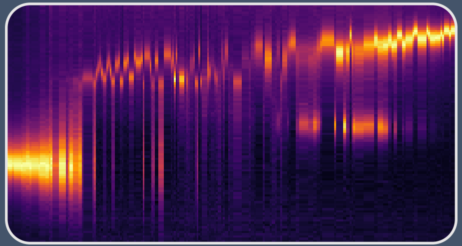
Microstructure Repository

- 1st Workshop: 2018-11-15
- 2nd Workshop: 2019-05-13



Materials Microscopy Data

- 1st Workshop: 2018-10-25
- 2nd Workshop: 2019-05-15



FAIR High Throughput Experimental Data

- Deployed Registry and Repository in 2017
- Tested with Interlaboratory Study (Paper Published 2019-03-19)

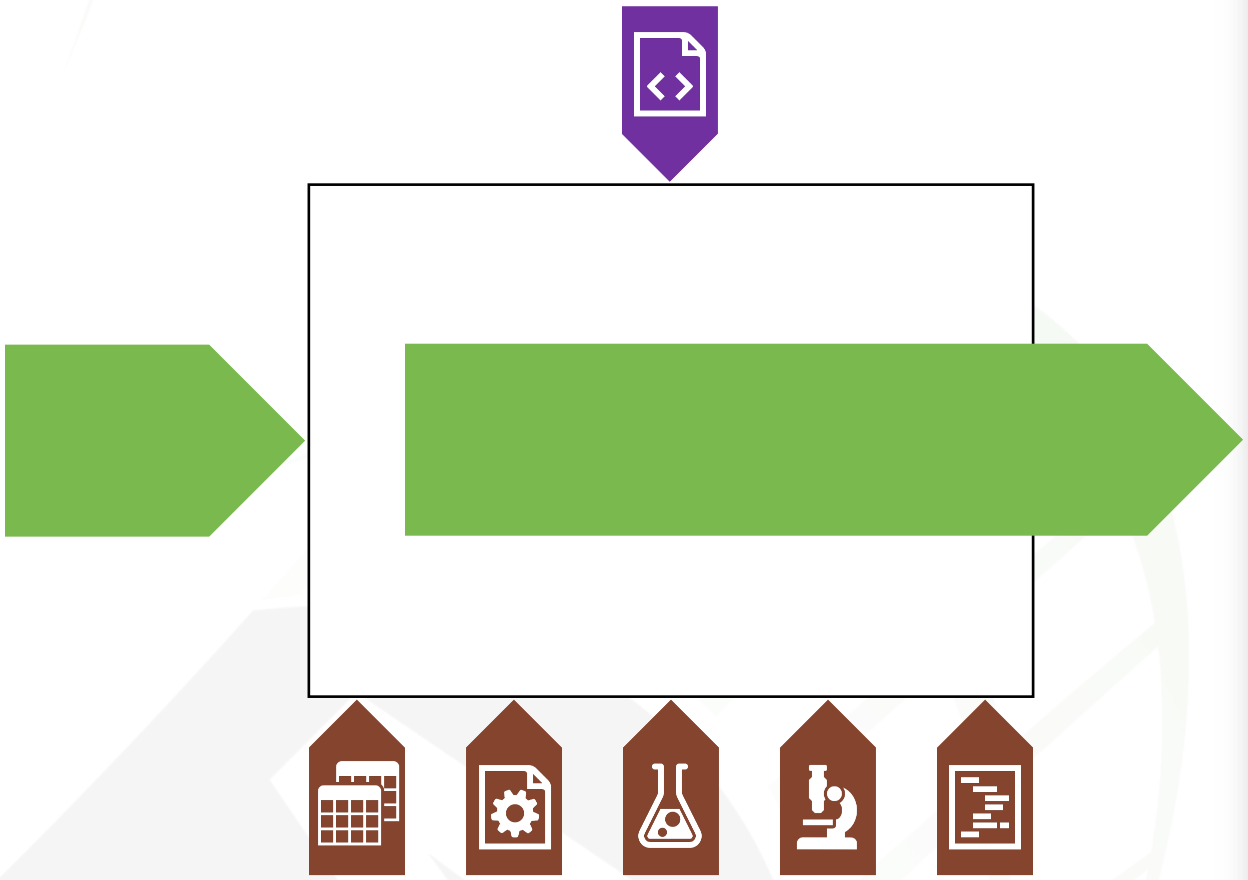
Example Demonstration Project

High Throughput Experimental Materials Registry and Repository

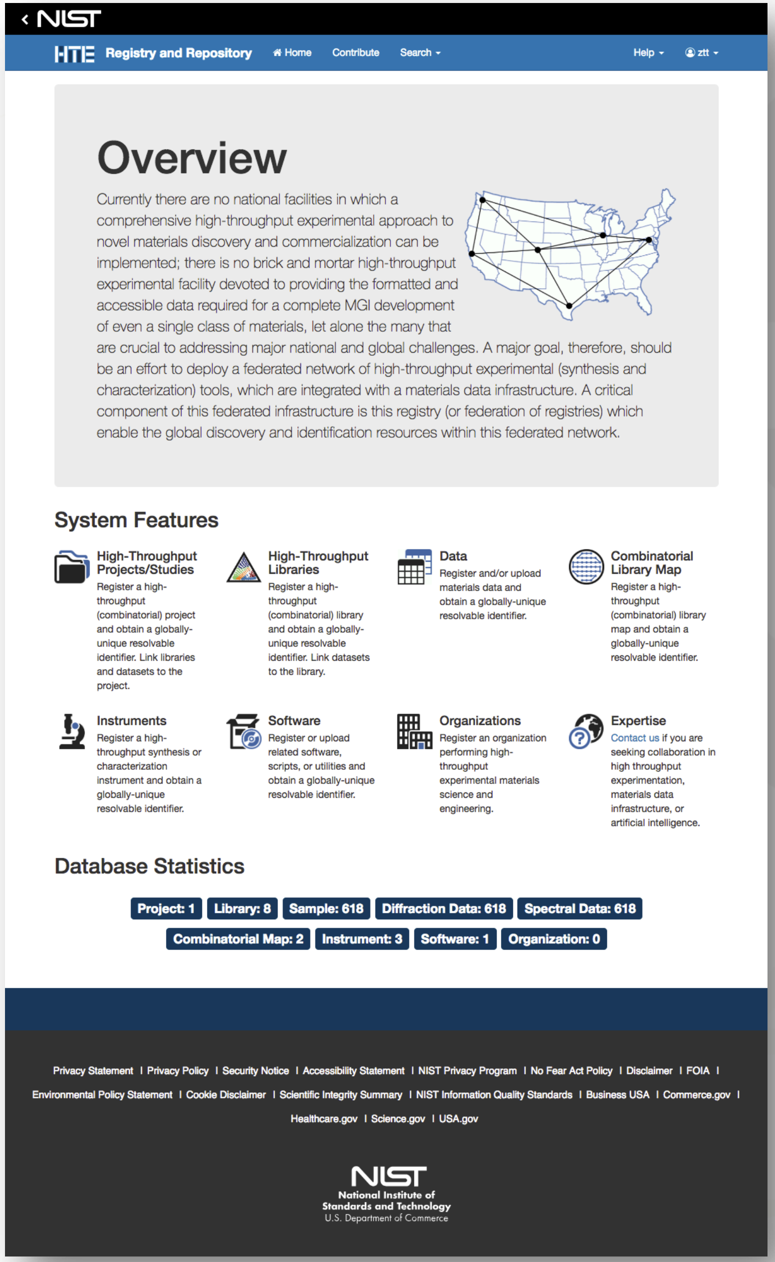
HTE Materials Registry and Repository

Extending Schema.org

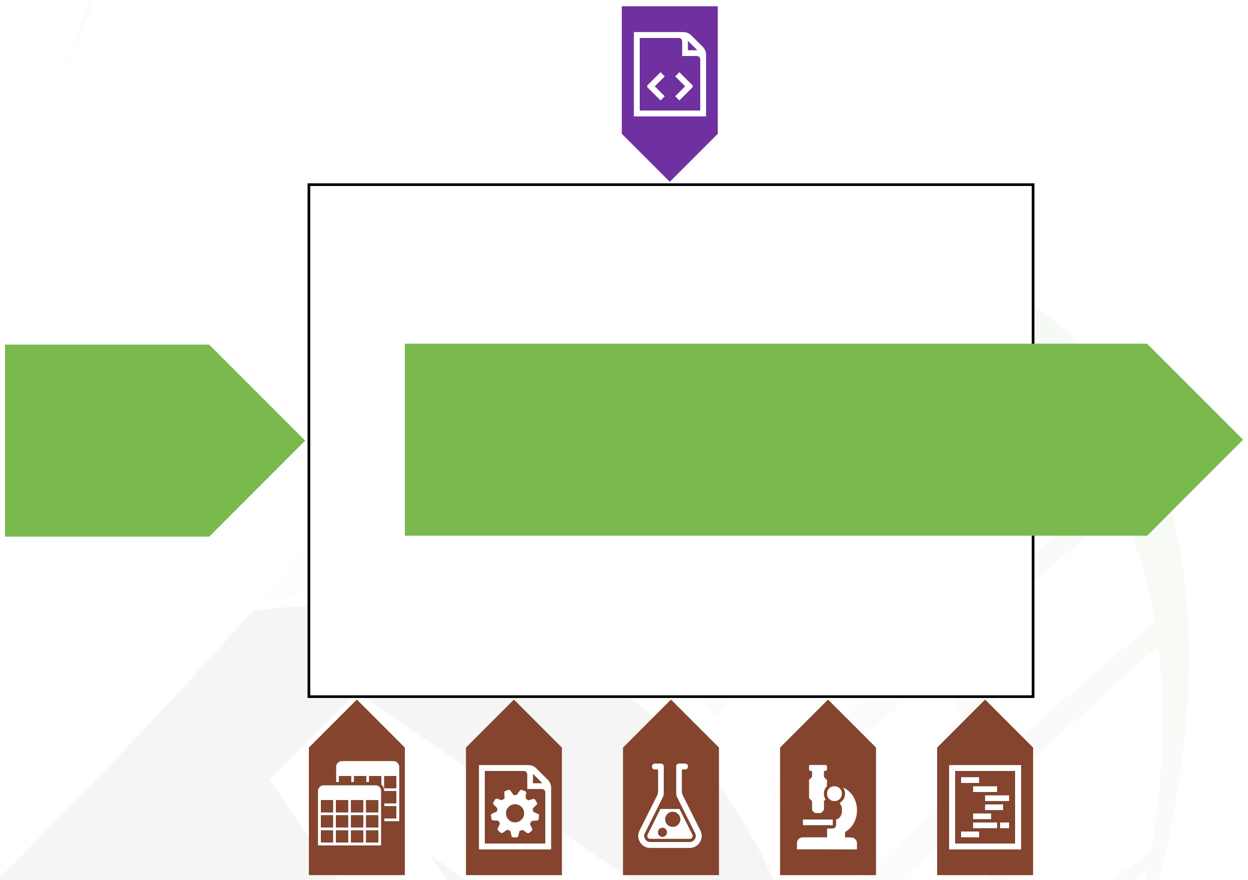
RDA/CODATA Materials Data, Infrastructure & Interoperability IG



PID for Things RDA WGs and IGs



HTE Materials Registry and Repository



Bootstrap

D3.js

Vue.js

Cordra REST API

Cordra

mongoDB

elastic

FAIR Leveraging of the Materials Resource Registry Vocabulary

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

- FAIR Principles
  - F1: (Meta) data are assigned globally unique and persistent identifiers
  - F2: Data are described with rich metadata
  - F3: Metadata clearly and explicitly include the identifier of the data they describe
  - F4: (Meta)data are registered or indexed in a searchable resource
  - A1: (Meta)data are retrievable by their identifier using a standardised communication protocol
  - A1.1: The protocol is open, free and universally implementable
  - A1.2: The protocol allows

What does this mean?

Humans should be able to exchange and interpret each other's data (so preferably do not use dead languages). But this also applies to computers, meaning that data that should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings. Interoperability typically means that each computer system at least has knowledge of the other system's data exchange formats. For this to happen and to ensure automatic findability and interoperability of datasets, it is critical to use (1) commonly used controlled vocabularies, ontologies, thesauri (having resolvable globally unique and persistent identifiers, see F1) and (2) a good data model (a well-defined framework to describe and structure (meta)data).

Examples

- The RDF extensible knowledge representation model is a way to describe and structure datasets. You can refer to the Dublin Core Schema as an example.
- OWL
- DAML+OIL
- JSON LD
- See example data models for **Predicted gene-disease associations from text mining** and **Tissue gene expression**.
- See data models from EBI in the 'documentation' links on this page **http://www.ebi.ac.uk/rdf/**

Links to Resources

- **https://en.wikipedia.org/wiki/Programming_language**

I2: (Meta)data use vocabularie ×

go-fair.org/fair-principles/i2-metadata-use-vocabularies-follow-fair-principles/

I2: (Meta)data use vocabularies that follow the FAIR principles

Home › FAIR Principles › I2: (Meta)data use vocabularies that follow the FAIR principles

- FAIR Principles
  - F1: (Meta) data are assigned globally unique and persistent identifiers
  - F2: Data are described with rich metadata
  - F3: Metadata clearly and explicitly include the identifier of the data they describe
  - F4: (Meta)data are registered or indexed in a searchable resource
  - A1: (Meta)data are retrievable by their identifier using a standardised communication protocol
  - A1.1: The protocol is open, free and universally implementable
  - A1.2: The protocol allows

What does this mean?

The controlled vocabulary used to describe datasets needs to be documented and resolvable using globally unique and persistent identifiers. This documentation needs to be easily findable and accessible by anyone who uses the dataset.

Examples

- Using the FAIR Data Point ensures I2.

Links to resources

- FAIR Data Point specification

Cordra (dev)

cordradev.nist.gov/#objects/20.500.12043/7238c9dcc5fa00de4147

Cordra (dev) Get Started All Objects Show Only ztt [Sign Out]

demo dataset

Type: Dataset

OBJECT | ACL | VERSIONS | RELATIVES

Edit

DO JSON

Dataset

This schema is for describing a Dataset in Cordra.

@id

20.500.12043/7238c9dcc5fa00de4147

Name

demo dataset

Subjects

This is for controlled vocabulary terms.

Subject 1 *

20.500.12043/9c7aecb1-382c-419f-89d5-95392fa32bd6

Name: transmission electron microscopy

Accountable Persons

Cordra (dev)

cordradev.nist.gov/#objects/20.500.12043/9c7aecb1-382c-419f-89d5-95392fa32bd6

Cordra (dev) Get Started All Objects Show Only ztt [Sign Out]

transmission electron microscopy

Type: DefinedTerm

OBJECT ACL VERSIONS RELATIVES

DefinedTerm

This schema is for describing a Defined Term in Cordra.

@id

20.500.12043/9c7aecb1-382c-419f-89d5-95392fa32bd6

Name

transmission electron microscopy

@context

schema

http://schema.org

skos

http://www.w3.org/2004/02/skos/core#

broader

skos:broader

narrower

skos:narrower

Cordra (dev)

cordradev.nist.gov/#objects/20.500.12043/9c7aecb1-382c-419f-89d5-95392fa32bd6

Cordra (dev) Get Started All Objects Show Only ztt [Sign Out]

schema:termCode

inDefinedTermSet *string*

schema:inDefinedTermSet

@type *

DefinedTerm

In Defined Term Sets

In Defined Term Set 1 *

20.500.12043/70638c6ba02ff25247a4

Name: NIST Materials Resource Registry Vocabulary

Broader Terms

Broader Term 1 *

20.500.12043/5073a600-3111-4191-98e4-16ec49bd711f

Name: microscopy

Narrower Terms

Cordra (dev)

cordradev.nist.gov/#objects/20.500.12043/70638c6ba02ff25247a4

Cordra (dev) Get Started All Objects Show Only ztt [Sign Out]

NIST Materials Resource Registry Vocabulary

Type: DefinedTermSet

OBJECT ACL VERSIONS RELATIVES

DefinedTermSet

This schema is for describing a Defined Term Set in Cordra.

@id

20.500.12043/70638c6ba02ff25247a4

Identifers

Identifer 1

@type *

PropertyValue

Identifier Type

DOI

Identifier Value

10.18434/T4/1435037

Name

NIST Materials Resource Registry Vocabulary

Public Data Resource

Simple Knowledge Organization System (SKOS) version of Materials Data Vocabulary

Contact: Andrea Medina-Smith..
Identifier: doi:10.18434/T4/1435037
Last modified: 2017-11-01

Go To ..
Description
Data Access
Record Details
View Metadata
Export JSON
Use
Citation
Fair Use Statement
Find
Similar Resources
Resources by Authors

Description

A version of the Materials Data Vocabulary structured as Simple Knowledge Organization System (SKOS). The XML was originally created by the TemaTres software. This vocabulary describes the applicability to material science of records in the NIST Materials Resource Registry (NMRR - https://materials.registry.nist.gov/). The NMRR allows for the registration of materials resources, bridging the gap between existing resources and the end users. The NMRR functions as a node in a federated system, making the registered information available for research to the materials community. This is being developed at the National Institute of Standards and Technology and is made available to solicit comments from the Material Science community. (An Excel version of the file is also included in the distributions for ease of use.)

Subject Keywords: materials science, controlled vocabulary, XML, vocabularies

Data Access

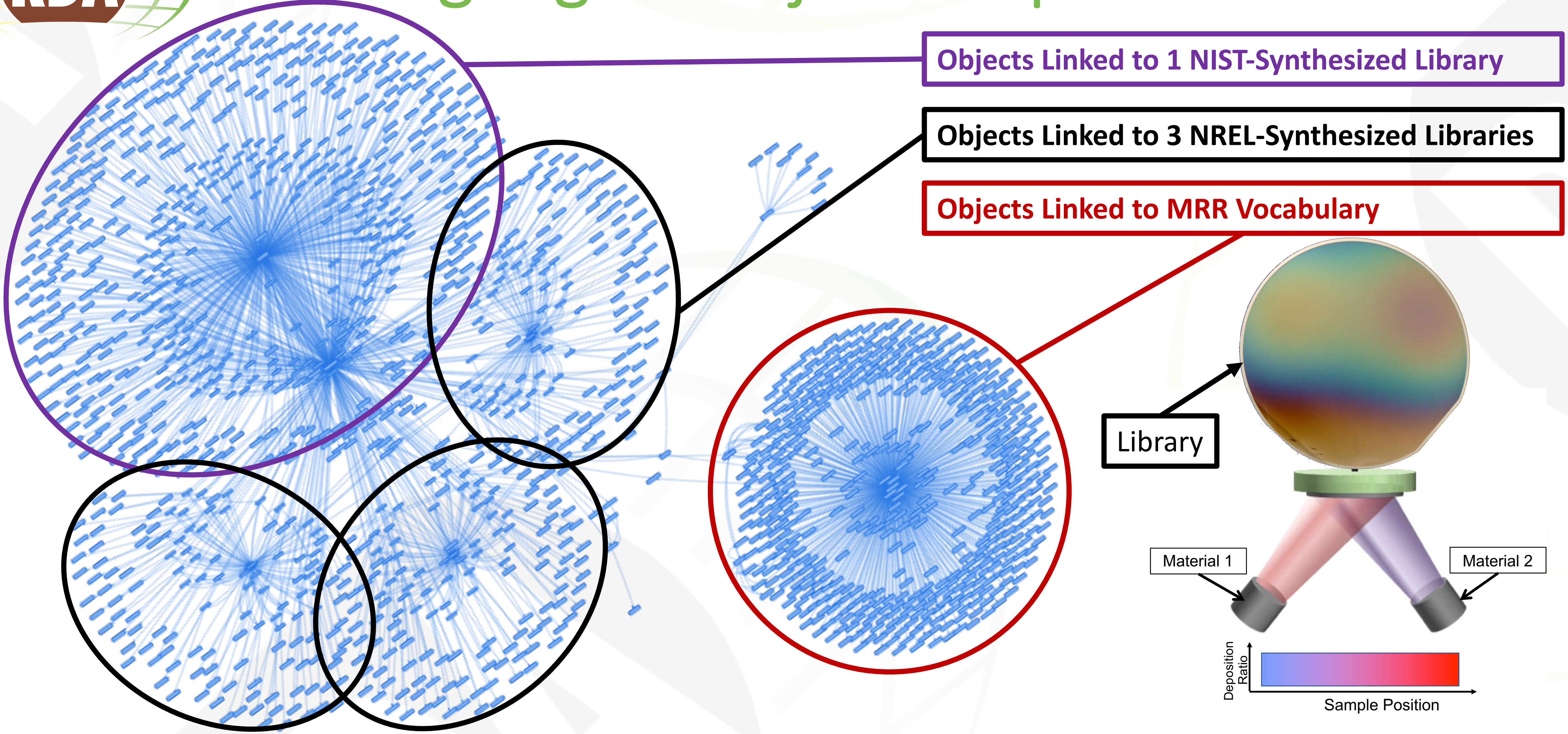
These data are public.

Files 0 Click on the file/row in the table below to view more details. Total No. files: 4

Name	Media Type	Size	Status
Materials_Registry_vocab_20180418.xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	119.5 kB	
Materials_Registry_vocab_20180418.xlsx.sha256	application/octet-stream	64 Bytes	
NMRRVocab20171102.rdf	application/octet-stream	337.0 kB	
NMRRVocab20171102.rdf.sha256	application/octet-stream	64 Bytes	

Building Digital Object Graphs

Building Digital Object Graphs



Thanks!

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

Semantics development and vocabulary platform case studies: PoLyInfo RDF and MatVoc

Asahiko Matsuda, Masashi Ishii, Takuya Kadohira

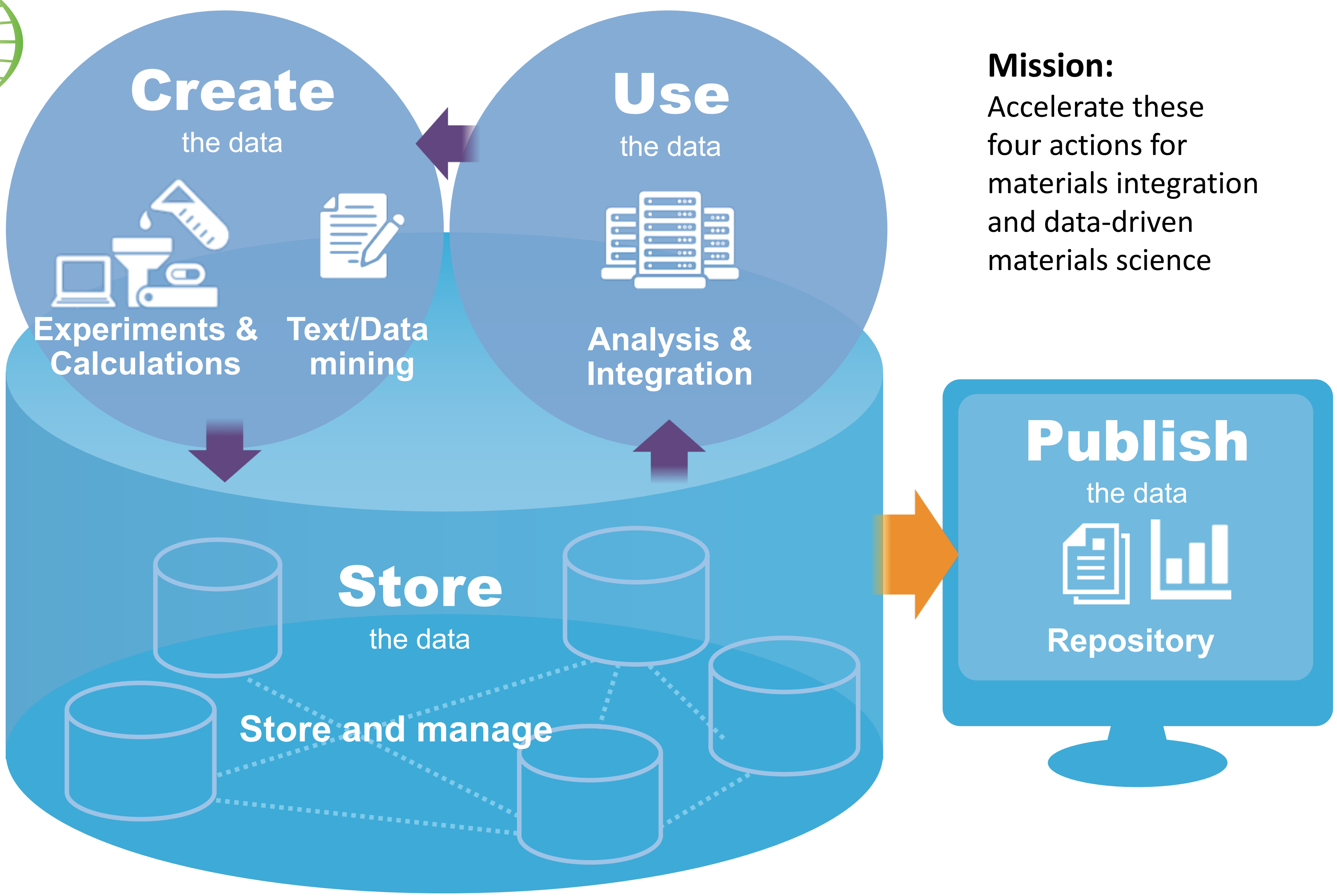
MATSUDA.Asahiko@nims.go.jp

Materials Data Platform Center,

National Institute for Materials Science

Work on PoLyInfo RDF was supported by Cabinet Office, Government of Japan, Cross-ministerial Strategic Innovation Promotion Program (SIP), “Technologies for Smart Bio-industry and Agriculture” (funding agency: Bio-oriented Technology Research Advancement Institution, NARO).
A.M. thanks Mineharu Suzuki (NIMS) for the discussion on metadata.

Materials Data Platform Center at NIMS



Mission:
Accelerate these four actions for materials integration and data-driven materials science

The question

“How can we assemble, integrate, harmonize, and leverage our materials data?”

1. Give PIDs to everything
2. Link them together
3. Query by following those links

But exactly how?

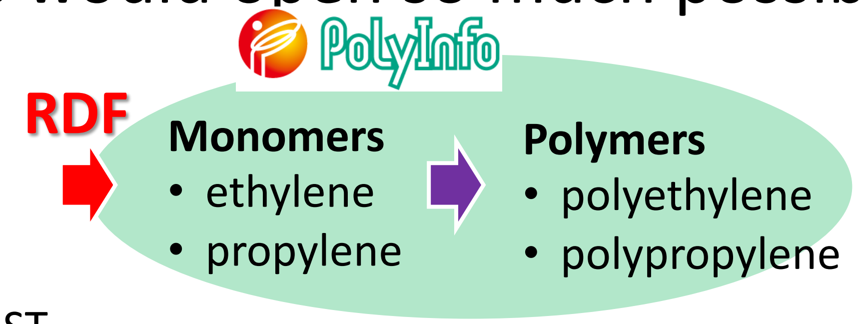
→ Case studies:

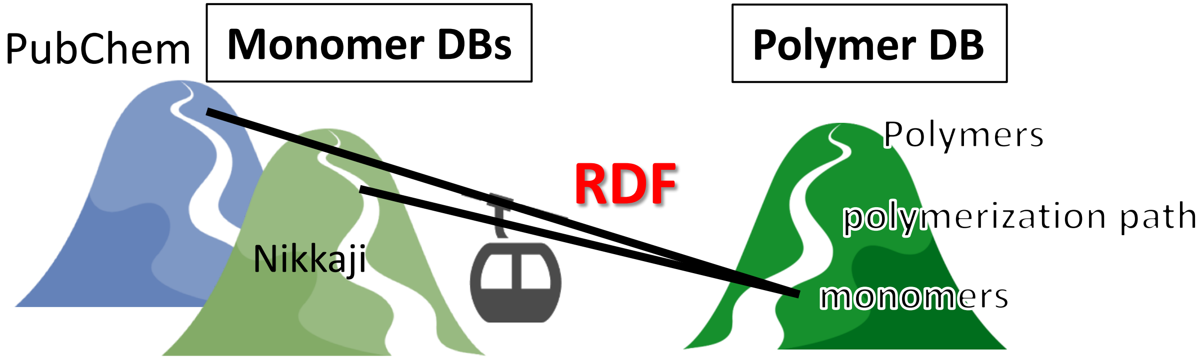
1. **PoLyInfo RDF** for polymer information
2. **MatVoc** for collaborative vocabulary management and distribution

Connecting our polymer database to other DBs

- NIMS has an extensive polymeric materials database PoLyInfo (334k properties), but it is not linked to other data sources.
- Linking to external DBs would open so much possibilities.

Monomer DBs
- PubChem by NIH
- Nikkaji
(Jpn Chem Substance Dict)
by JST





RDF triples and SPARQL queries

RDF triples

"Finland's currency is the Euro"

Subject Predicate Object

Triple

"Finland" "currency" "Euro"

"Finland" "capital" "Helsinki"

"Helsinki" "population" "650058"

SPARQL queries (for DBpedia)

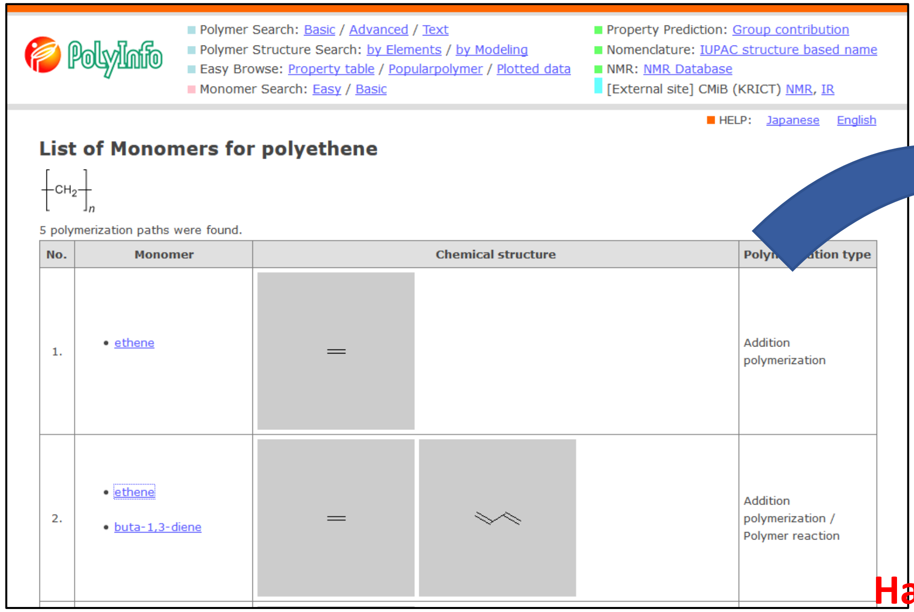
What are the currencies for all the countries?

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?country ?currency
WHERE {
  ?country dbo:currency ?currency .
}
```

Which countries have Euro for their currency, where are their capital cities, and what are the populations of those cities (if data exist)?

```
PREFIX res: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?country ?capital ?population
WHERE {
  ?country dbo:currency res:Euro ;
           dbo:capital ?capital .
  OPTIONAL {?capital
           dbo:populationTotal ?population .}
}
ORDER BY DESC(?population)
```

Polymerization information in PoLyInfo



PolyInfo

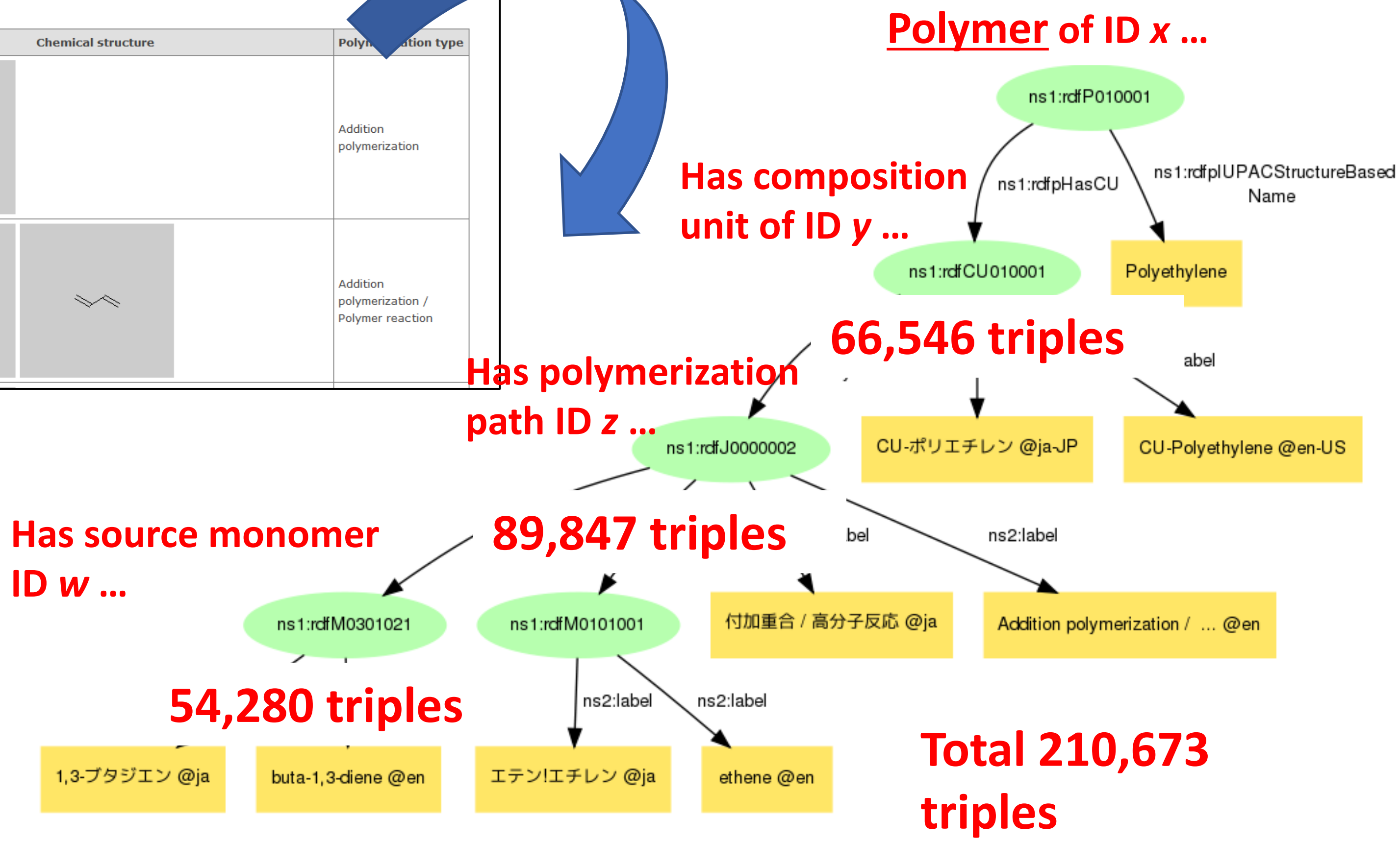
- Polymer Search: Basic / Advanced / Text
- Polymer Structure Search: by Elements / by Modeling
- Easy Browse: Property table / Popularpolymer / Plotted data
- Monomer Search: Easy / Basic
- Property Prediction: Group contribution
- Nomenclature: IUPAC structure based name
- NMR: NMR Database
- [External site] CMiB (KRICT) NMR, IR

HELP: Japanese English

List of Monomers for polyethene

5 polymerization paths were found.

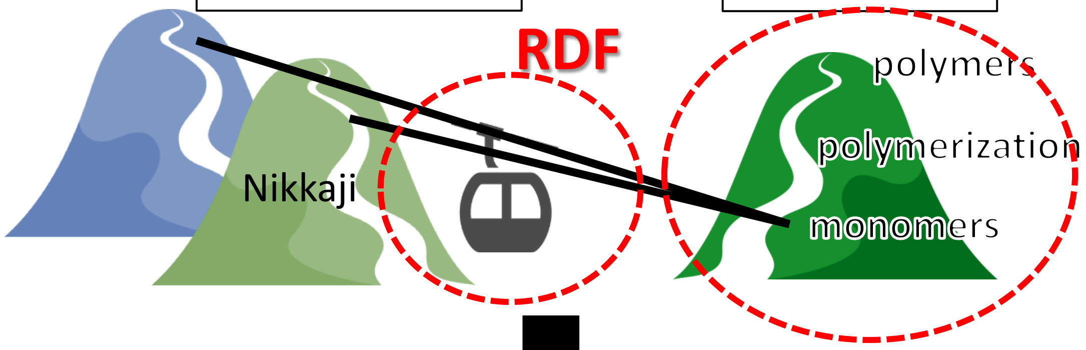
No.	Monomer	Chemical structure	Polymerization type
1.	• ethene		Addition polymerization
2.	• ethene • buta-1,3-diene		Addition polymerization / Polymer reaction

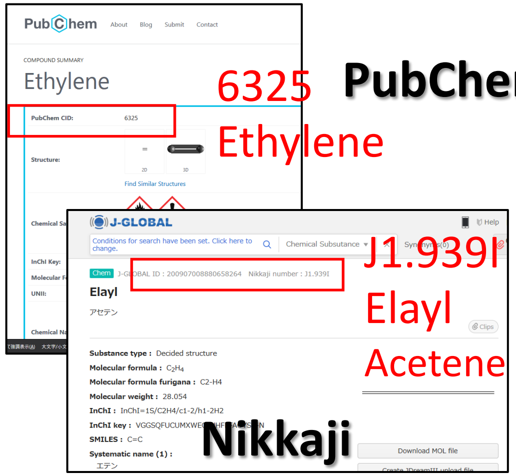


Link to monomer DBs

Conceptual link between PoLyInfo and monomer DBs

PubChem **Monomer DBs** **Polymer DB**



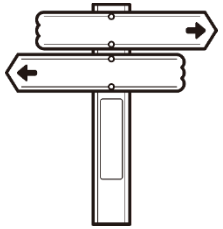


6325 **PubChem**
Ethylene

J1.939I
Elayl
Acetene
Nikkaji

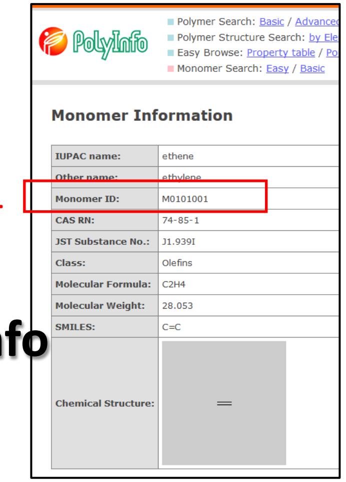
skos:closeMatch

PubChem
Nikkaji



PoLyInfo

M0101001
ethene

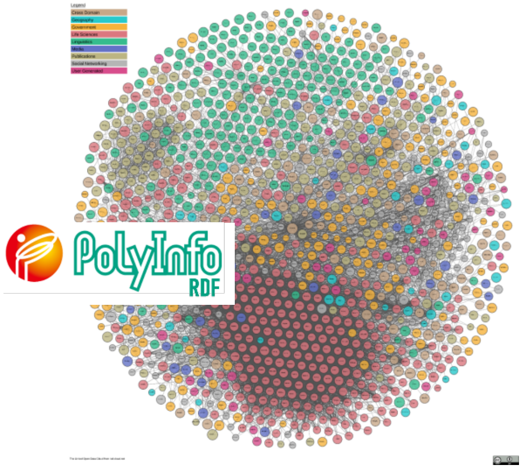


PoLyInfo

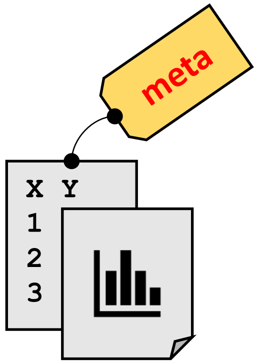
SKOS **16,440 triples**

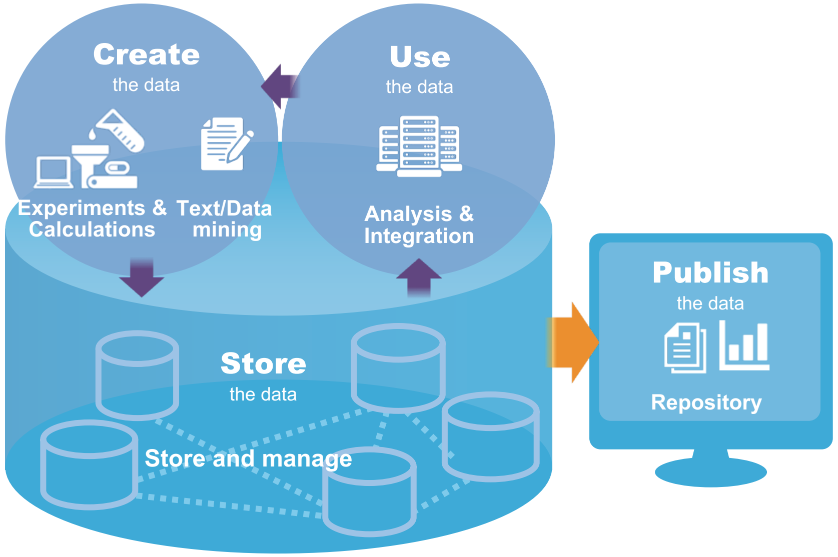
```
@prefix nikkaji: <http://stirdf.jst.go.jp/cde/nikkaji/> .
@prefix rpcsid: <http://rdf.ncbi.nlm.nih.gov/pubchem/substance/SID> .
@prefix ns1: <https://polymer.nims.go.jp/> .

ns1:rdfM0101001 skos:closeMatch rpccid:6325 , idpcc:6325 , pcc:6325 , pcscid:6325 , b2rpcc:6325 .
ns1:rdfM0101001 skos:closeMatch nikkaji:J1.939I .
```

For polymers:



What about other fields?
Or for day-to-day
data management?






NIST
Home
Publish resource
Search for resources
Log In / Sign Up
Materials Resource Registry
Enter keywords, or leave blank to retrieve all records
Search
Tools
All Resources
Organization
Collection
Dataset
Service
Software
Informational Sites
Local Results
Local
OAI-PMH
MDF OAI-PMH Server
Type: (Clear)
Origin of Data: (Clear)
Material Type: (Clear)
biological (8)
From Local 278
4CeeD
Coordinated Science Laboratory University of Illinois at Urbana-Champaign
https://4ceed.github.io/
Subject keyword(s): data management, private cloud, microservice-based, compute-storage
4CeeD is a distributed web-based software framework that supports Capturing, Curation, Coordination, Correlation, and Distribution of scientific data from scientific instruments, such as TEM, SEM, AFM and others, to private cloud compute-storage cyber-infrastructure. For more information, visit 4CeeD's website at https://4ceed.github.io
ABAQUS
Dassault Systemes - Dassault Systemes
http://www.3ds.com/products-services/simulia/portfolio/abaqus/overview/
Subject keyword(s): software
"Sold as part of the SIMULIA package. SIMULIA delivers a scalable suite of unified analysis products that allow all users, regardless of their simulation expertise or domain focus, to collaborate and seamlessly share simulation data and approved methods without loss of information fidelity.The
Show more

NIST Materials Data Vocabulary

- doi:10.18434/T4/1435037
- A vocabulary for browsing records in the NIST Materials Resource Registry
- Top concepts:
  - Data origin
  - Characterization methods
  - Computational methods
  - Material types
  - Properties addressed
  - Structural features
  - Synthesis and processing
- Three-layer hierarchy

Helped us gain a bird's-eye view of materials science

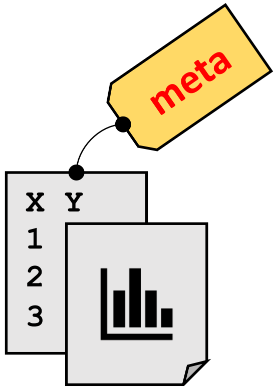
	A	B	C
1			
2	This vocabulary describes the applicability to material science of records in the NIST Materials Resource Registry (NMRR).The NMRR allows for the registration of materials resources, bridging the gap between existing resources and the end users. The NMRR functions as a node in a federated system, making the registered information available for research to the materials		
3	This is being developed at the National Institute of Standards and Technology and is made available to solicit comments from the Material Science community. Please do not enter any proprietary data into this system.		
4	version 16, 7/13/17		
5	Feedback to Chandler Becker, cbecker@nist.gov		
6	See below for additional contributors and people/resources consulted		
7			
8	Data origin	experiments	
9	Data origin	informatics and data science	
10	Data origin	simulations	
11	Data origin	theory	
12			
13	Material types	biological	.
14	Material types	biomaterials	.
15	Material types	ceramics	.
16	Material types	ceramics	carbides
17	Material types	ceramics	cements
18	Material types	ceramics	nitrides
19	Material types	ceramics	oxides
20	Material types	ceramics	perovskites
21	Material types	ceramics	silicates
22	Material types	metals and alloys	.
23	Material types	metals and alloys	Al-containing
24	Material types	metals and alloys	commercially pure metals
25	Material types	metals and alloys	Cu-containing
26	Material types	metals and alloys	Fe-containing
27	Material types	metals and alloys	intermetallics
28	Material types	metals and alloys	Mg-containing
29	Material types	metals and alloys	Ni-containing
30	Material types	metals and alloys	rare earths
31	Material types	metals and alloys	refractories
32	Material types	metals and alloys	steels
33	Material types	metals and alloys	superalloys
34	Material types	metals and alloys	Ti-containing
35	Material types	metamaterials	.
36	Material types	molecular fluids	.
37	Material types	organic compounds	.
38	Material types	organic compounds	alcohols
39	Material types	organic compounds	aldehydes
40	Material types	organic compounds	alkanes
41	Material types	organic compounds	alkenes

Vocabulary used for designing MDPF metadata

NIST MRR Materials Data Vocabulary

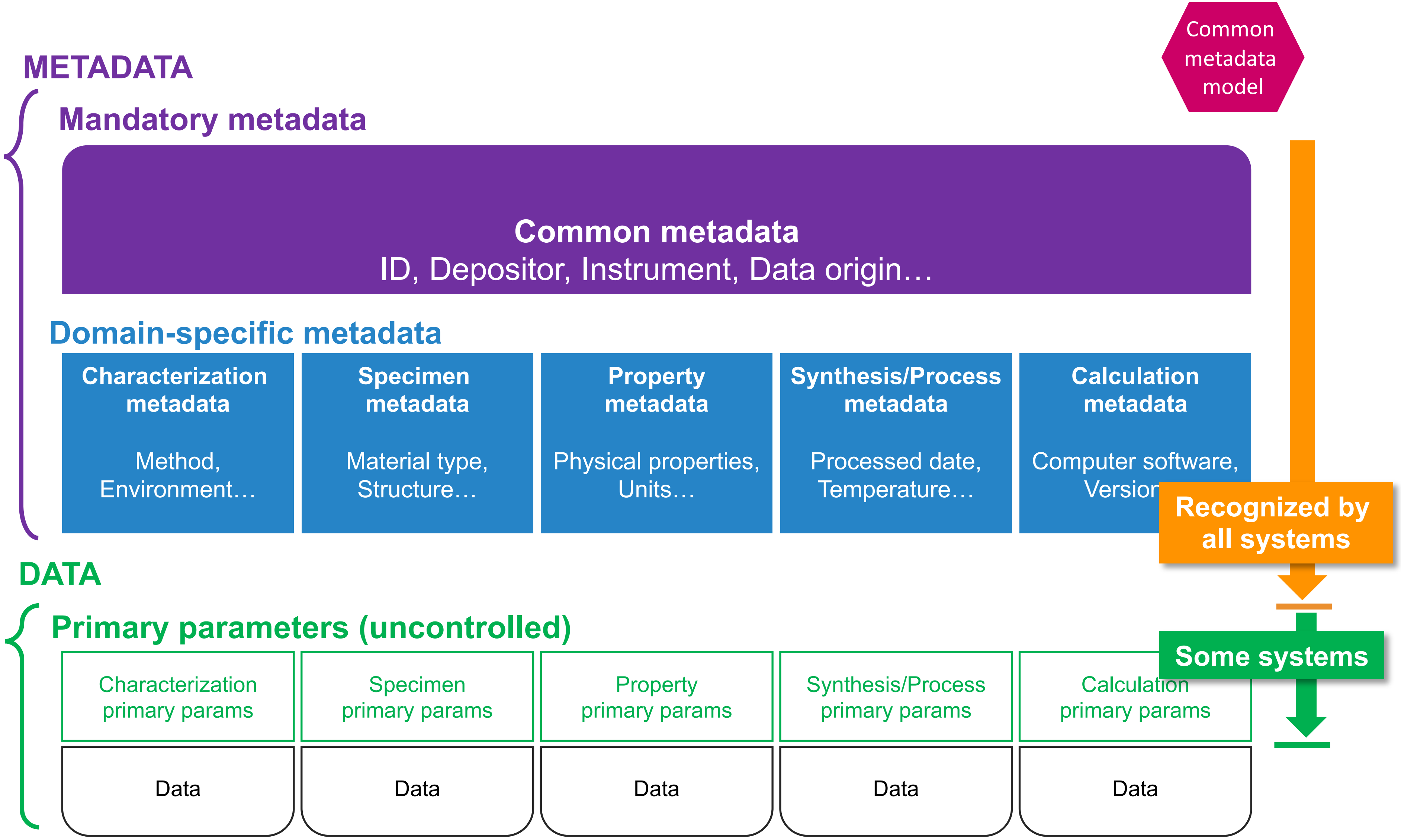
- Data origin →
- Characterization methods →
- Computational methods →
- Material types →
- Structural features
- Properties addressed →
- Synthesis and processing →

Metadata for NIMS MDPF data entries



- Data origin
- Characterization metadata
- Computational metadata
- Specimen metadata
- Properties metadata
- Synthesis and processing metadata

‘Tiered’ metadata model for MDPF data



User feedback on the vocabulary

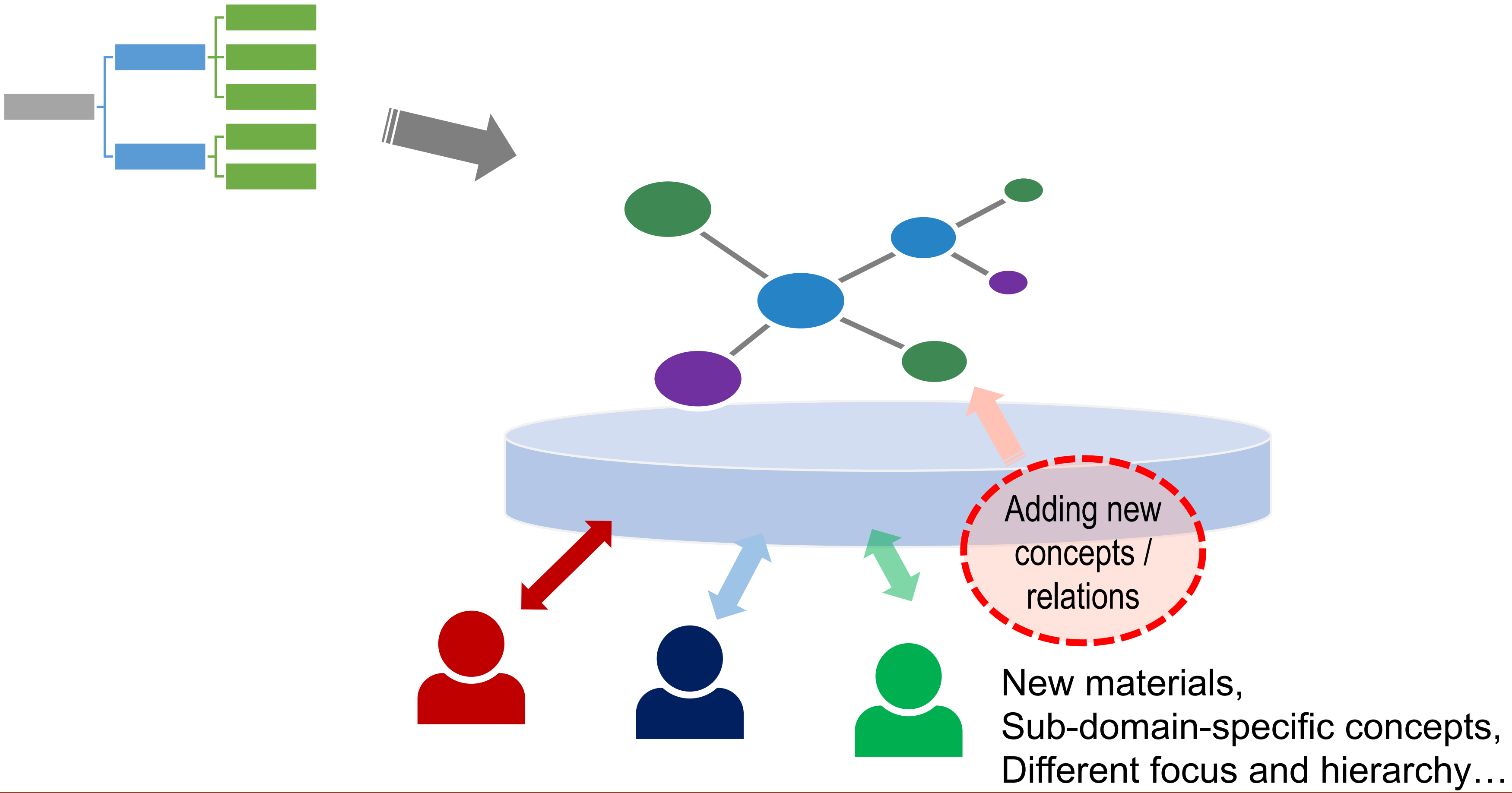
- “The vocabulary is missing this term”
  - E.g., semiconductors should have silicon

Material types	semiconductors	.
Material types	semiconductors	II-VI
Material types	semiconductors	III-V
Material types	semiconductors	extrinsic
Material types	semiconductors	intrinsic
Material types	semiconductors	n-type
Material types	semiconductors	p-type

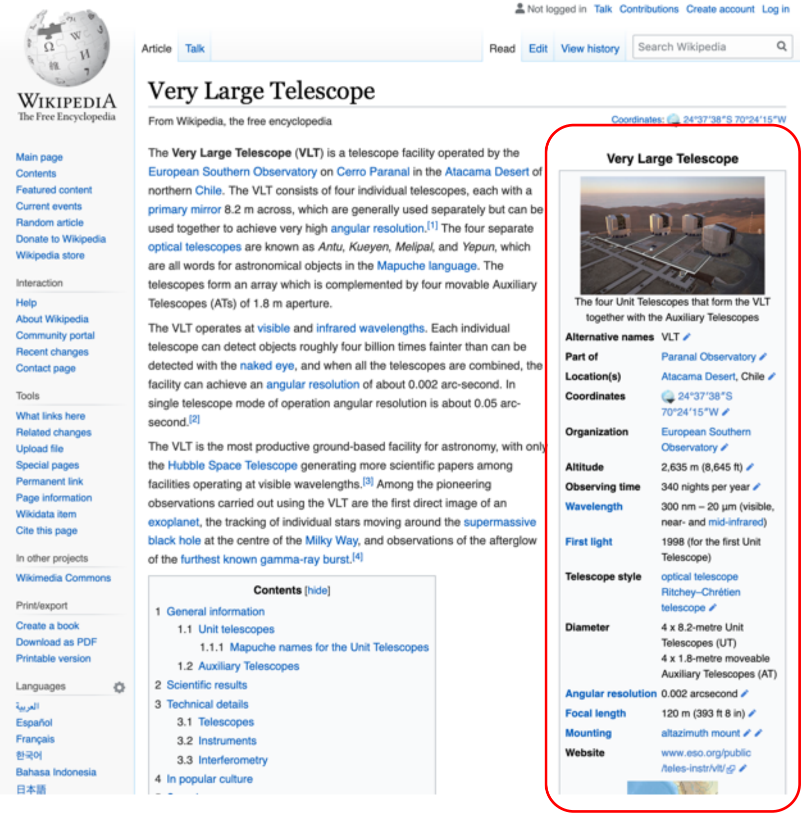
- “I think this characterization method should be classified as spectroscopy, not microscopy.”

etc.

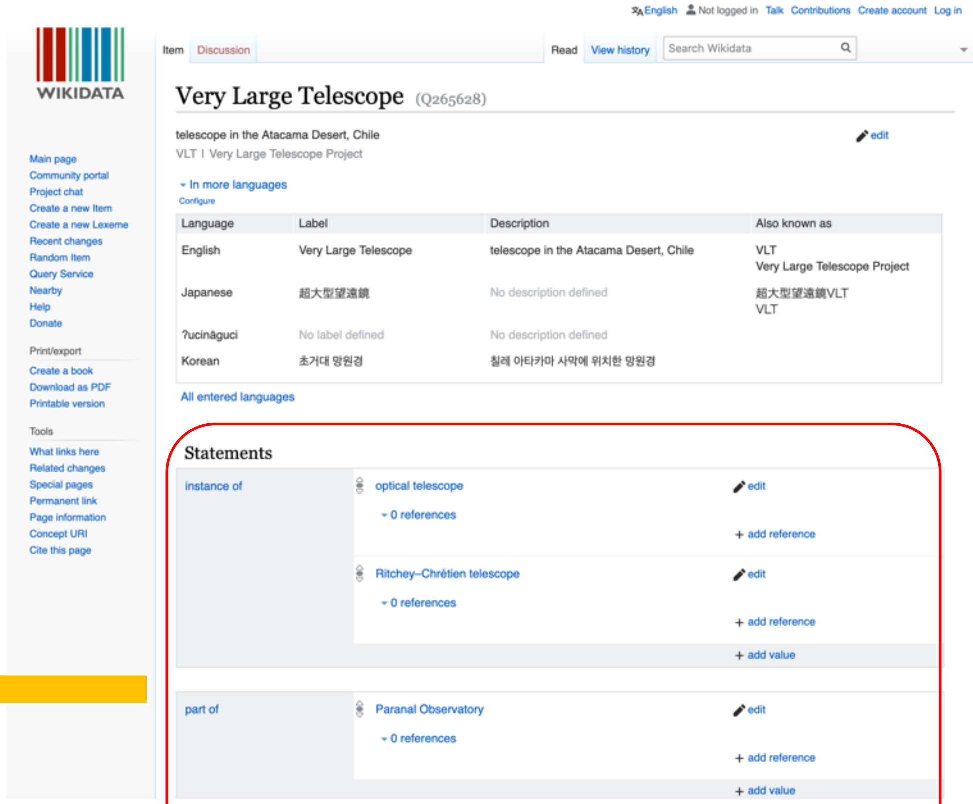
Hierarchy → Graph
Versioned file → Dynamic system



Wikipedia (on MediaWiki) and Wikidata (on Wikibase)



Wikipedia infobox

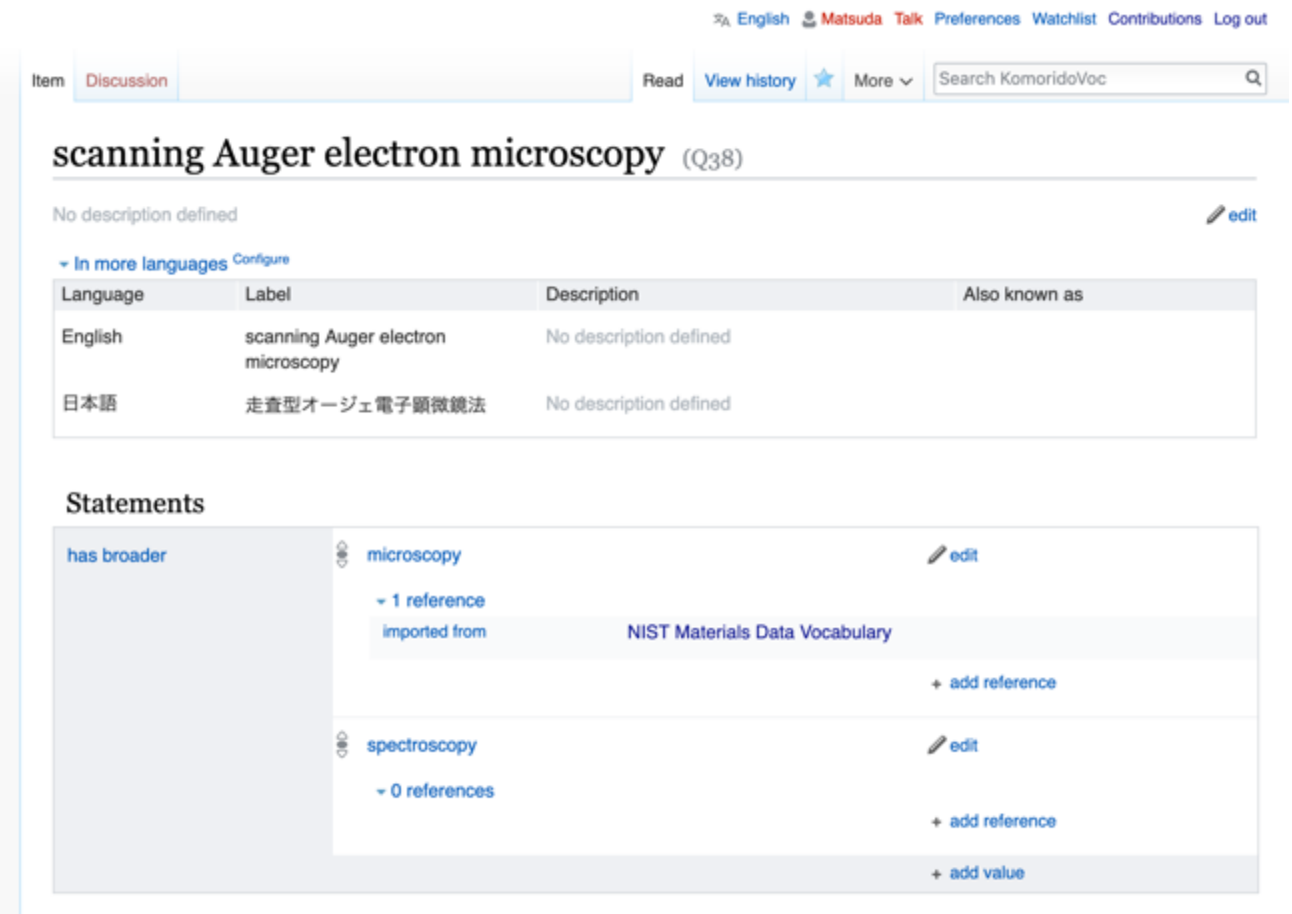


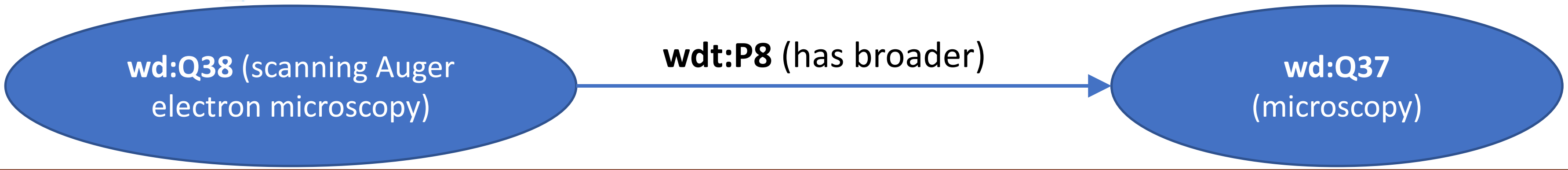
Wikidata statements

MediaWiki

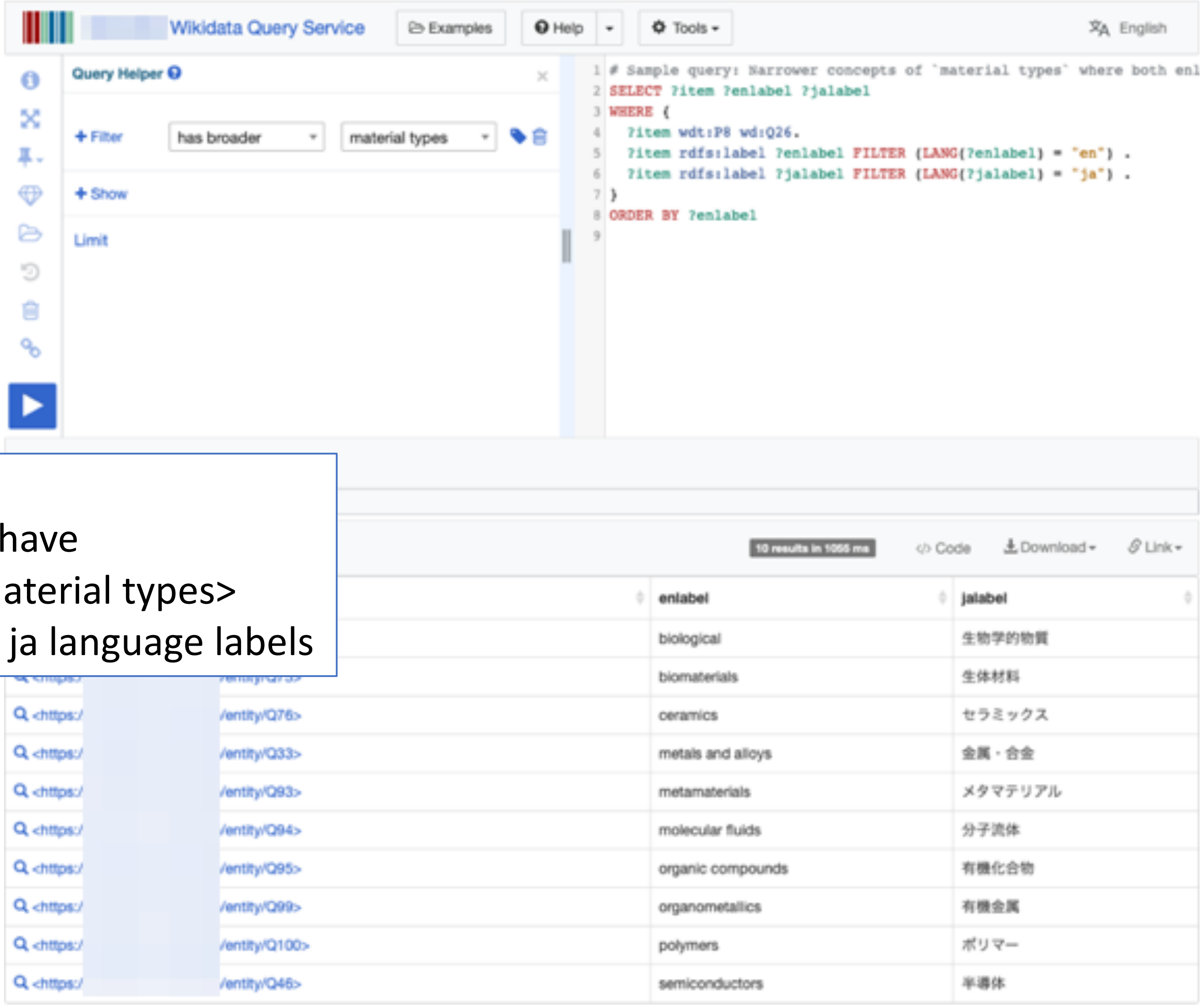
MediaWiki + WIKIBASE

Materials Data Vocabulary in NIMS Wikibase



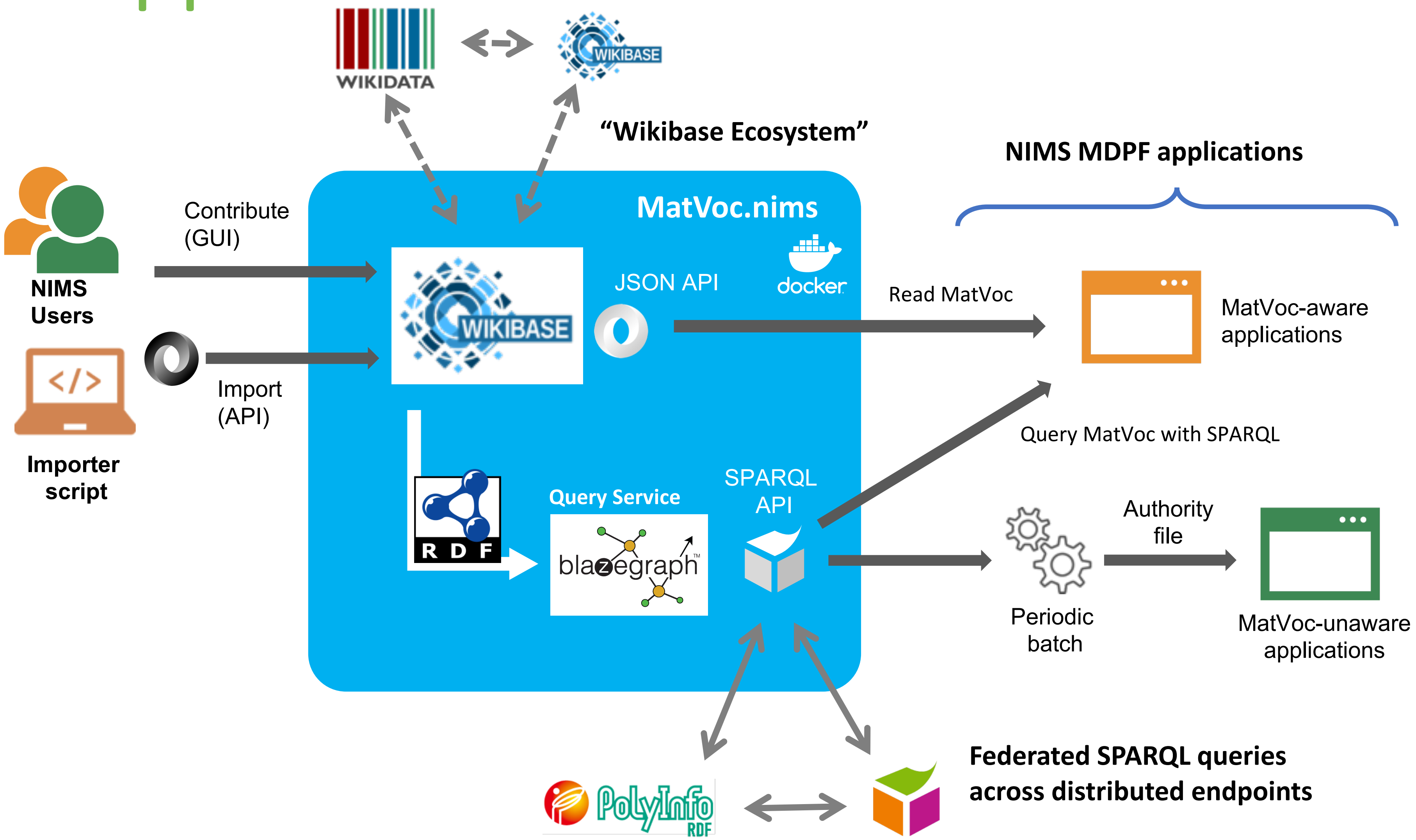


SPARQL query service



e.g.
All items that have
<broader> <material types>
and both en + ja language labels

System overview and distribution to other applications



Conclusion

1. We demonstrated integration between our polymer database and other databases (Nikkaji, PubChem) using Linked Data technologies. Further work is ongoing to align the semantics with even more institutions' databases.

2. For more generic applications and day-to-day heterogeneous data management, we took hints from the Materials Data Vocabulary to design the common metadata schema. We set up a Wikibase system to allow the researchers to add concepts, and designed other applications to read from it.

Both of these represent the concepts in RDF triples, can be queried by SPARQL, and can be easily federated with other databases.

EMMO

An Ontology for Physical Sciences

Emanuele Ghedini

University of Bologna

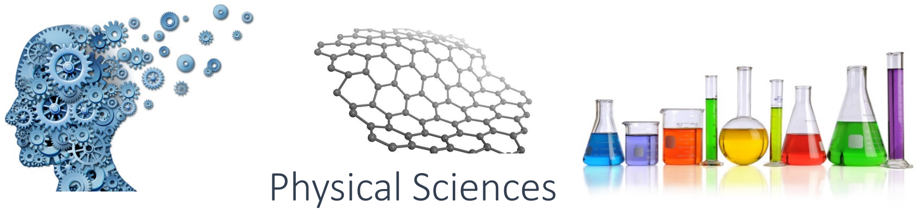
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

SimDOME

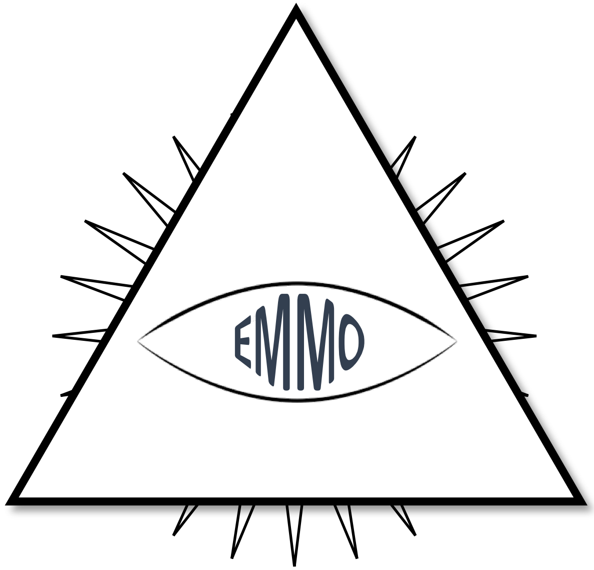
EMMC

What is the EMMO?

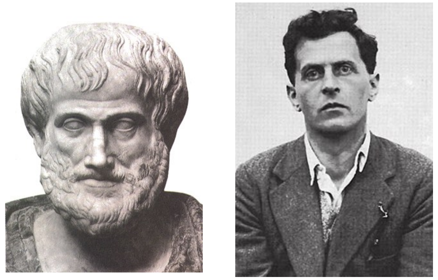
The **EMMO** is a multidisciplinary efforts within the **EMMC** and a **network of H2020 projects** aimed to the development of a **standard representational framework**, in the form of an **ontology**, based on current **materials modelling and characterization knowledge**.



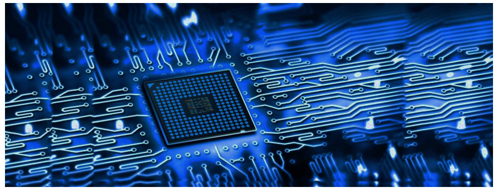
Physical Sciences
(e.g. physics, chemistry, material science, engineering)



European
Materials
Modelling
Ontology



Analytical Philosophy
(e.g. mereotopology, semiotics, logic)



Information and Communication Technologies
(e.g. reasoners, platforms, formats)

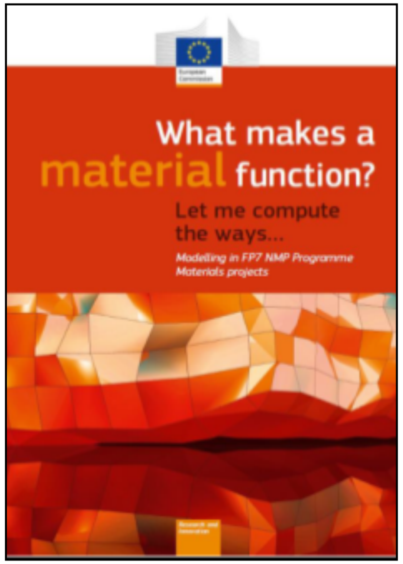
What is the EMMO?

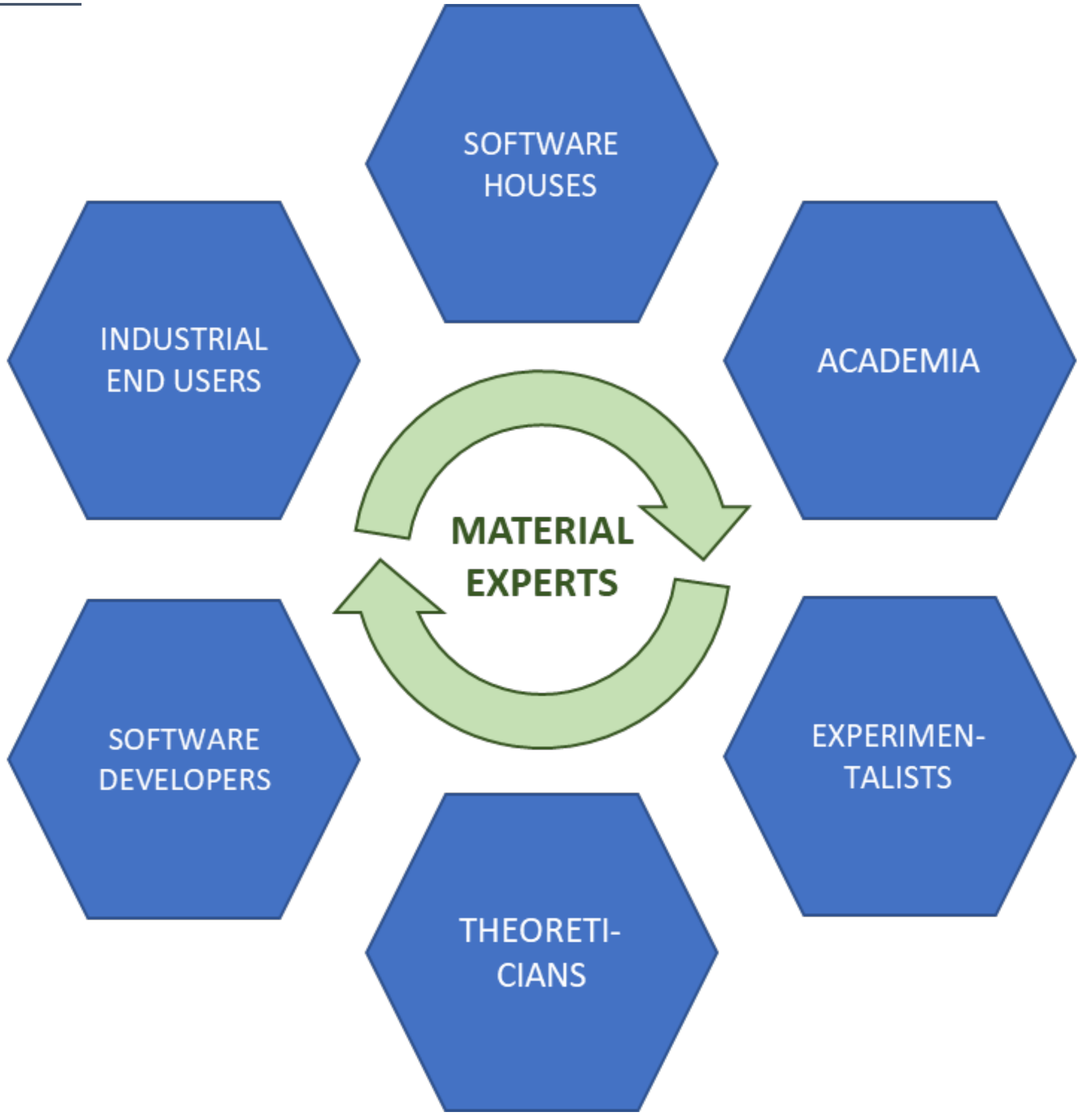
The **EMMO** aims to facilitate the work of **materials experts** in **connecting stakeholders** in the field of **materials development** and **characterization**, making use of the **standardization efforts** already performed within the EU, and facilitating the development of **materials modelling software tools** (i.e. OSP).

CEN WORKSHOP AGREEMENT

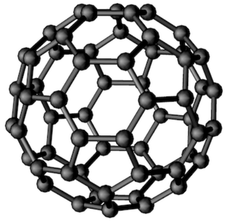
CWA 17284
April 2018

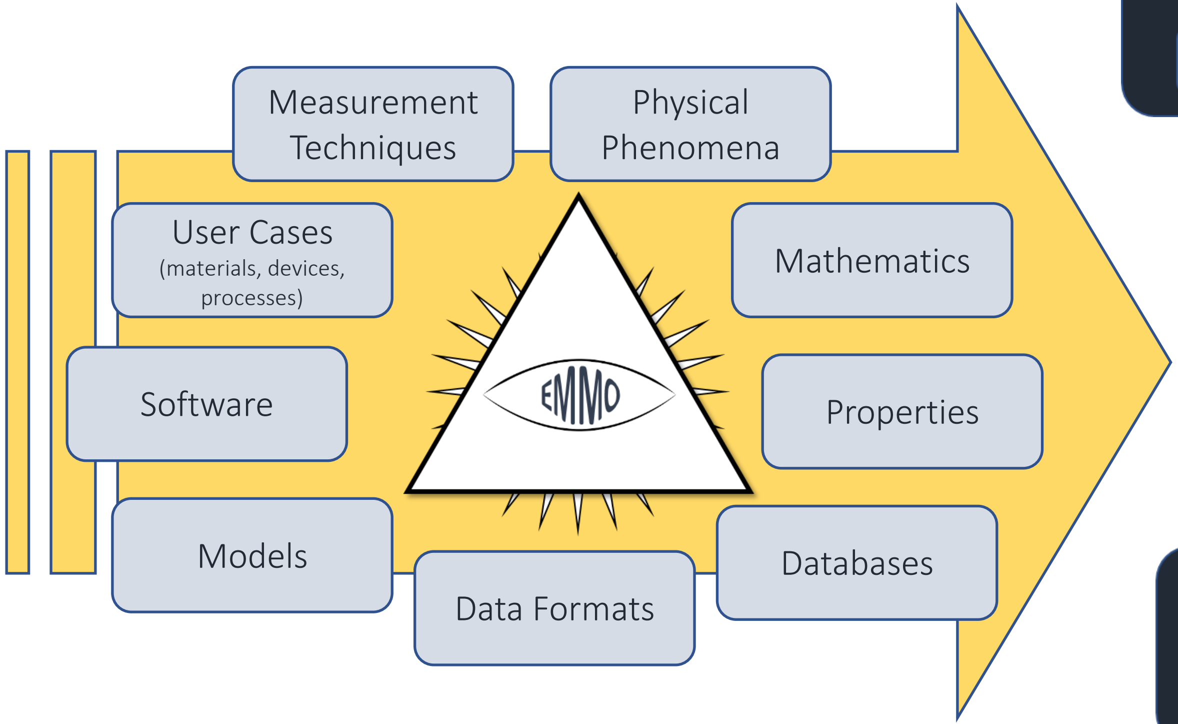
Materials modelling – Terminology, classification and metadata

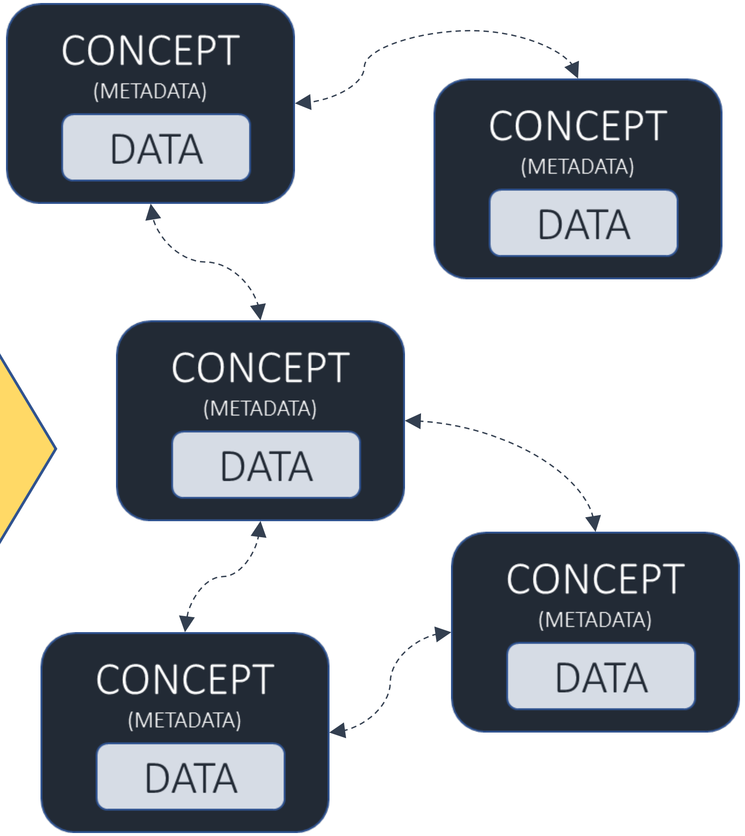




What is the EMMO?







USER CASE

From real world entities…

ONTOLOGY

…trough a formal knowledge-based representational system…

INFORMATION

…to a digital representation.

EMMO application fields

Science
standard reference concepts to facilitate understanding between scientific communities (multi-disciplinarity)

AI
formalized knowledge system ready to be used in AI applications

Modelling
connections between real world entities and available physical models (OSP, translation) at different scales

BigData
data schematics for specific applications and facilitate semantic extraction for data harvesting

Characterization
formalization of the entity-measurement-property connection to facilitate data exchange between experimentalists and modellers

Industry
formalization of the manufacturing process and product, connection with material databases and modelling software to facilitate business decisions

What EMMO is and what is not

Charles S. Peirce
semiotic theory

ONTOLOGY WORLD

SIGN

a physical entity that stands for another physical entity
(an icon, a digital twin, a symbol, a word, a C++ class)

USER WORLD

INTERPRETER

a human being or a machine that can recognize object attributes (e.g. color, shape)

INTERPRETANT

the mental/internal representation determined by a sign or an object (interpreter dependent)

PHYSICAL WORLD

OBJECT

any physical entity

EMMO is a formalized system of signs (representation)

The **EMMO** helps users (interpreters) providing them a way to communicate their interpretations but is not itself an automatic connection between real and ontological world.
(Dear user, you can lie about what you see, so you responsibilities!!!) have great

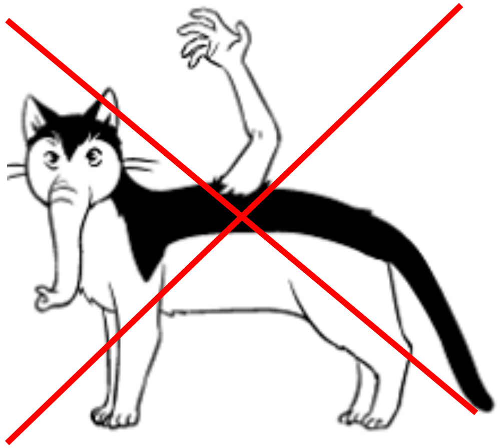
While any system of signs is made of physicals that stand for other physical, it is clear that the object of interpretation is not and is not part of the **EMMO**

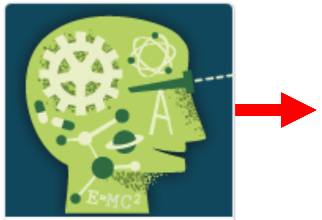
EMMO is not the truth about the world, but only a tool!

What EMMO is and what is not

The **EMMO** helps you to provide **sign**s that represents correctly what a real world entity is, using formal logic.

The **EMMO** prevents you from giving unrealistic representations of real world entities.





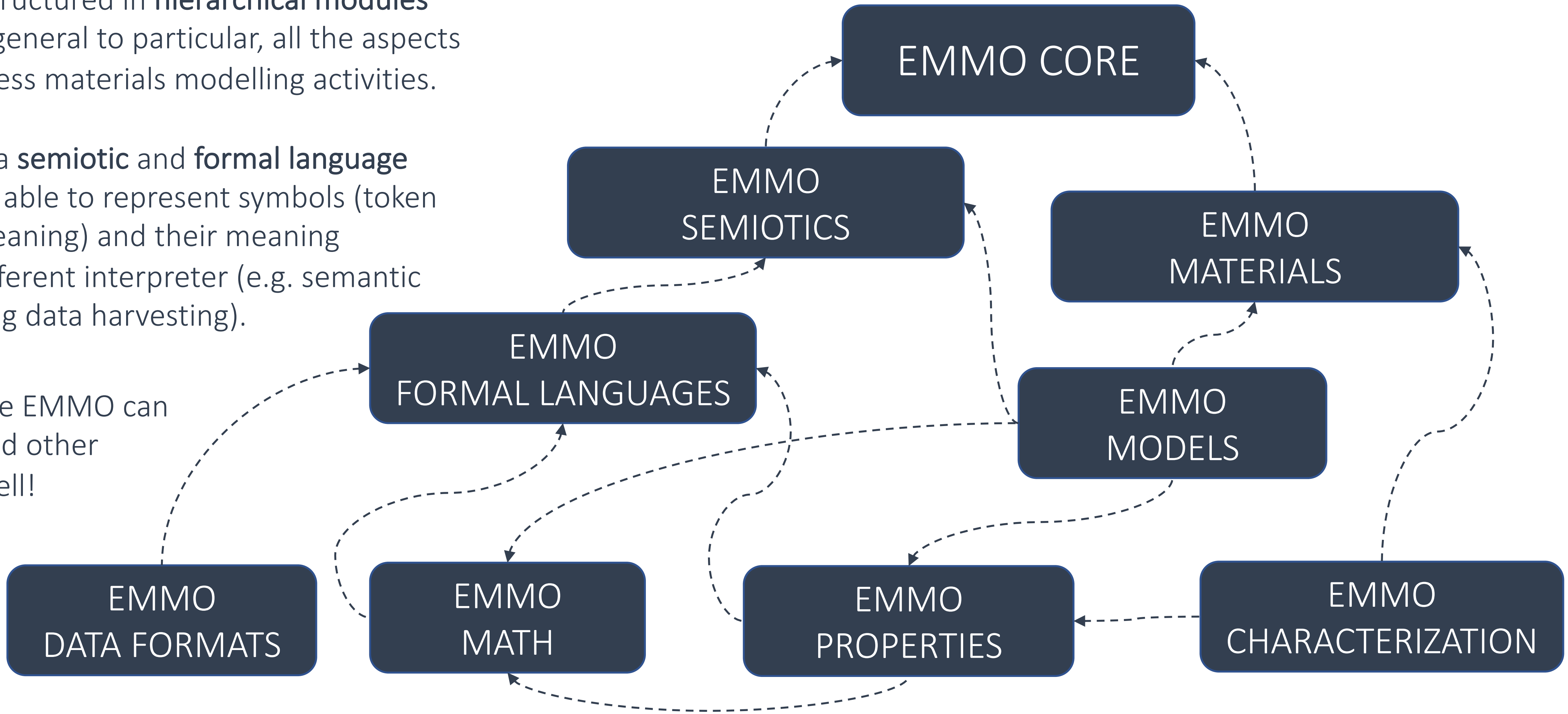
The **EMMO** tells you nothing about the existence of the **object** that stands for the **sign**.
Is up to the **interpreter** to connect ontology world to real entities world.

EMMO structure

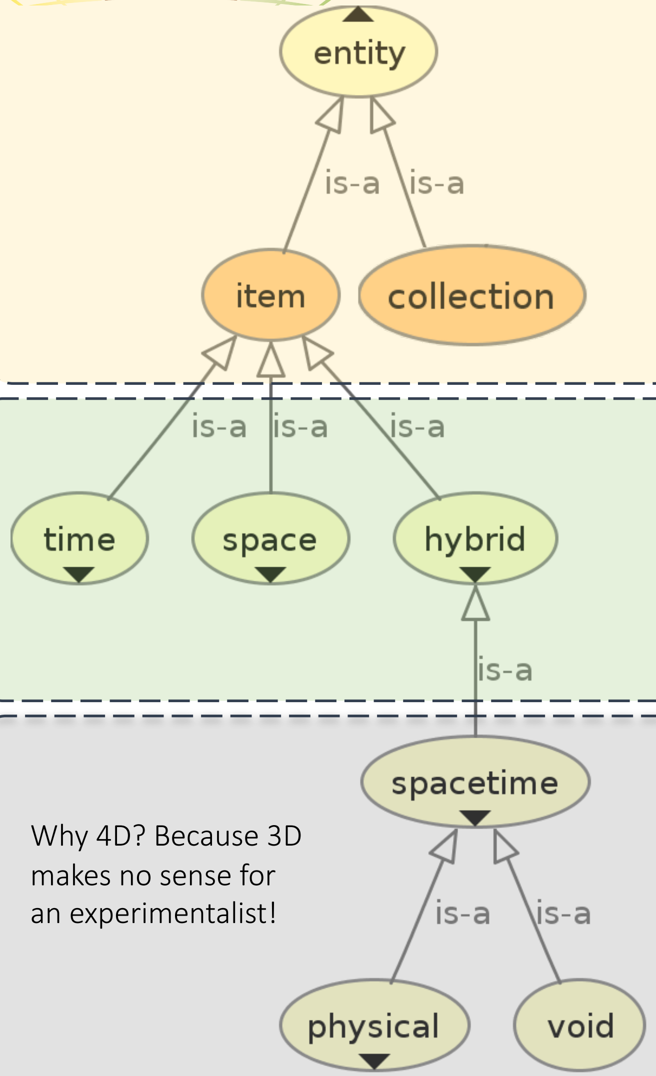
The **EMMO** is structured in **hierarchical modules** covering, from general to particular, all the aspects needed to address materials modelling activities.

It includes also a **semiotic** and **formal language** branches, to be able to represent symbols (token that have no meaning) and their meaning according to different interpreter (e.g. semantic extraction during data harvesting).

In this sense, the EMMO can contain itself and other ontologies as well!



EMMO Core



ABSTRACT CONCEPTUAL LEVEL

Clear separation between **collection** and **item** (based on mereotopology). **collection** individuals are collection of **item**s according to defined concepts (e.g. red entities). **item**s individuals stand for something that is 'real', i.e. a 4D portion of the universe.

In the EMMO abstract concepts are represented as the **collection**s that concretize them (e.g. friendship is the collection of all friendship acts) embracing a rigorous nominalistic view.

GEOMETRIC/TOPOLOGICAL LEVEL

items unfolds in space (3D) and time (1D) and can be sliced in pure **time**, pure **space** or **hybrid** space and time entities.

PHYSICAL LEVEL

Real world entities exists only in full 4D **spacetime** (3D space and 1D time), i.e. you can't partition a cake in infinitely thin slices!

A **spacetime** that can be perceived by (interact with) the interpreter is a **physical**. If the **spacetime** entity is empty in terms of perception, is a **void**.

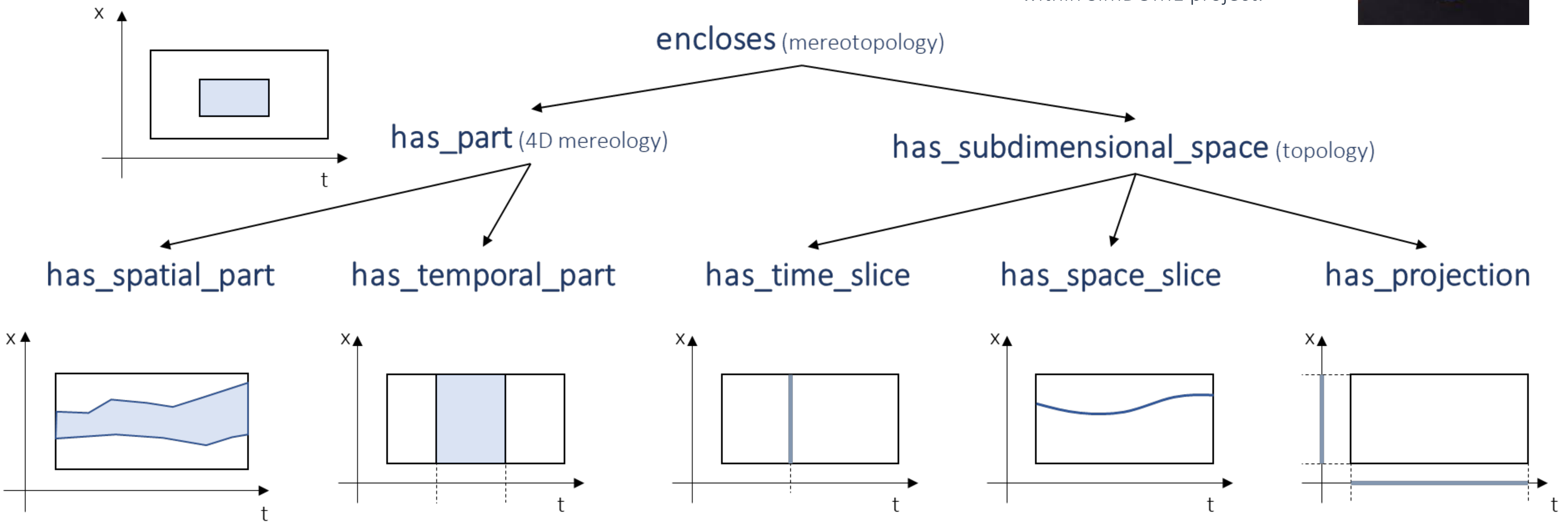
Why 4D? Because 3D makes no sense for an experimentalist!

EMMO Core Mereotopology

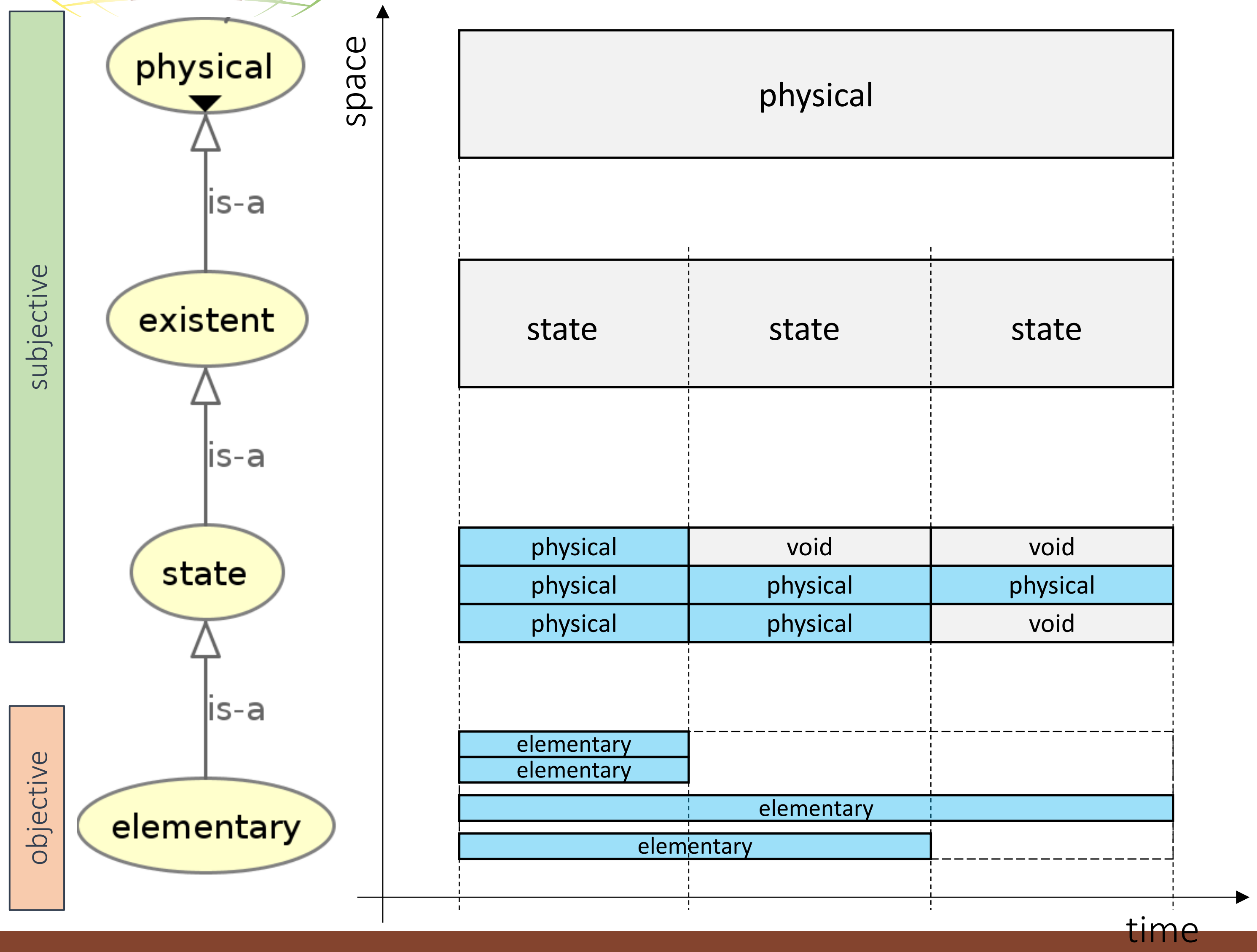
R. Casati, A. Varzi, *“Parts and Places”*, MIT Press

The **EMMO** makes use of **mereotopology** set of logical axioms, extended to 4D entities, to represent formally the evolution in time and space of entities.

Prof. Achille Varzi (Columbia University, NY) is one of the top mereologists and will act as advisor for EMMO development within SimDOME project.



EMMO Core Mereotopology



The **EMMO** identifies a parthood hierarchy in **physical**s, by introducing the concept of:

- **elementary** as the fundamental, non-divisible, constituent of entities (i.e. atomistic mereology)
- **state** as a **physical** whose parts have a constant cardinality during its life time (similar to endurants)
- **existent** as a succession of **state**s (similar to perdurants)

so that a **physical** entity can be defined using a multiscale perspective.

An elementary particle, that expresses some fundamental physical properties (e.g. mass, charge, spin) can be represented by an **elementary** in a physics ontology.

However, in another material ontology an **elementary** can be something else, depending on the perspective (e.g. a brick for a LEGO ontology, a furniture component in a IKEA ontology)

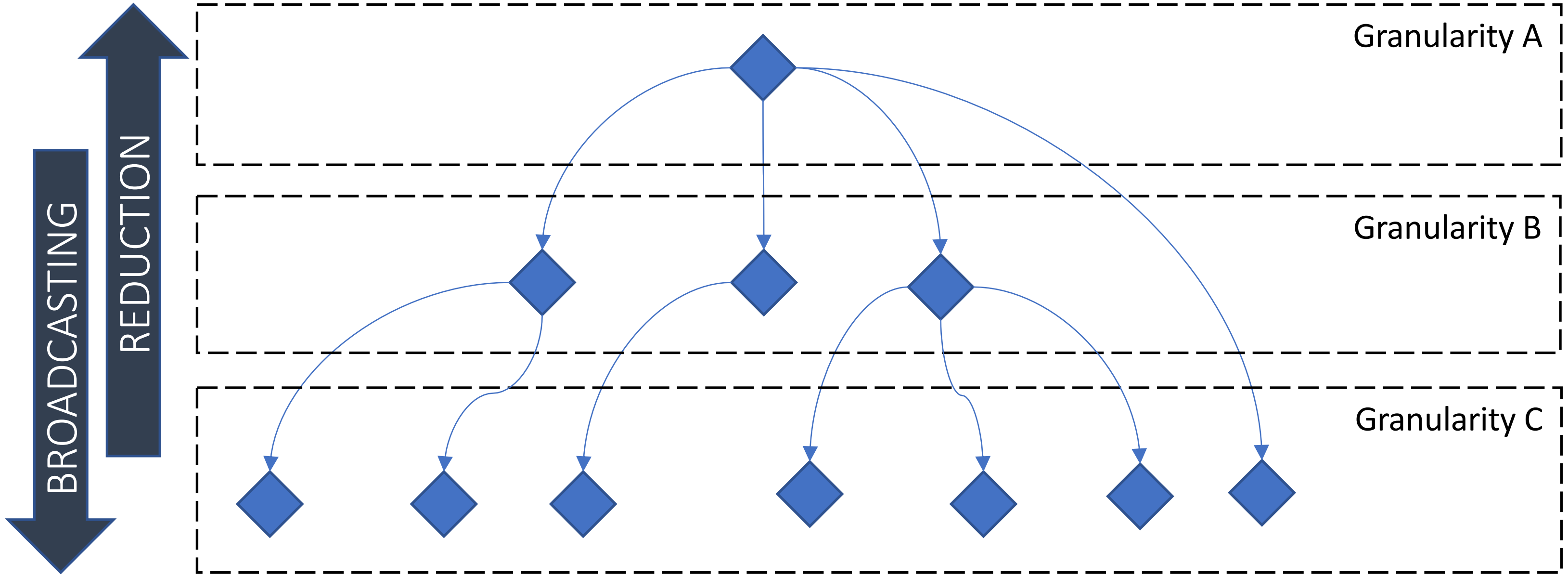
EMMO Core Mereotopology

By defining the mereological relation of **direct parthood**, the **EMMO** is able to describe entities as made of parts at different level of **granularity**.

The individuals are forming a **directed rooted tree**:

this is paramount for cross scale interoperability (vertical interoperability) that is the basis for multi-scale modelling.

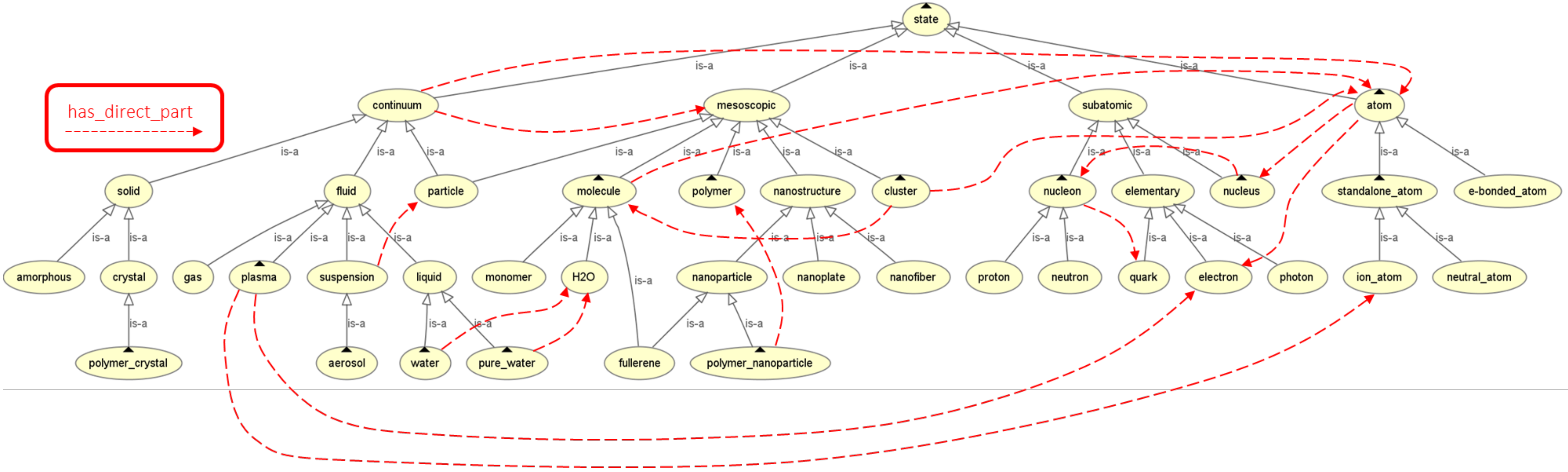
Reduction and broadcasting can be easily implemented by navigating in this type of tree.



EMMO Material

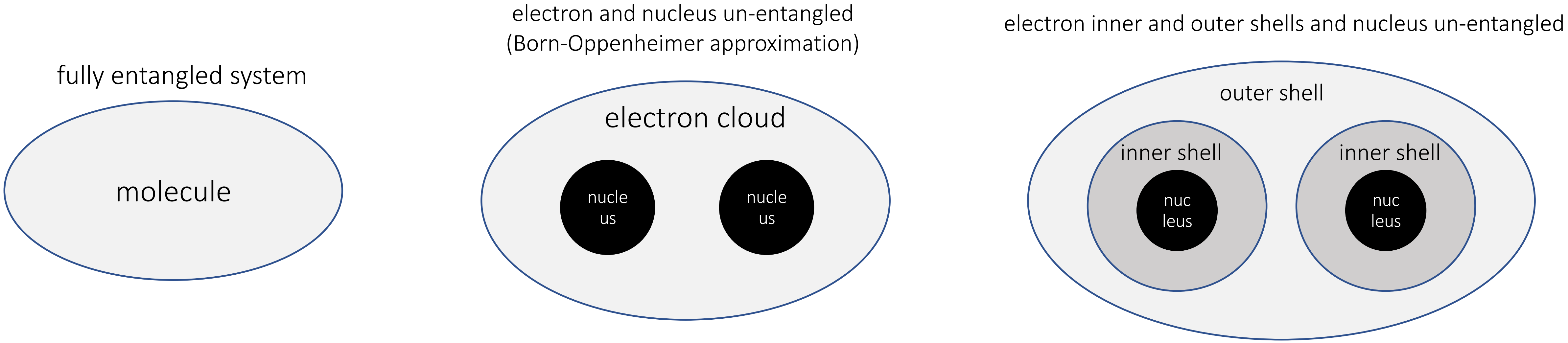
A first draft of a material ontology branch has been developed within the **EMMO** to demonstrate the powerful expressiveness of direct parthood in identifying granularity levels.

The material branch is defined with large use of axioms with the **has_direct_part** relation that put constraints about the attributes of each individual that will be declared in material classes (e.g. a molecule can't have part crystal).



EMMO Material

The **EMMO** material branch is also generic and flexible enough to represent **quantum systems** in a way that is compatible with different interpretations (i.e. Copenhagen, De Broglie-Bohm) and approximations (e.g. Born–Oppenheimer).



Hamiltonian parameters can be derived by axioms that define the specific quantum system class (i.e. the sub-parts).

Wave function collapse can also be represented within the **EMMO** mereological framework.

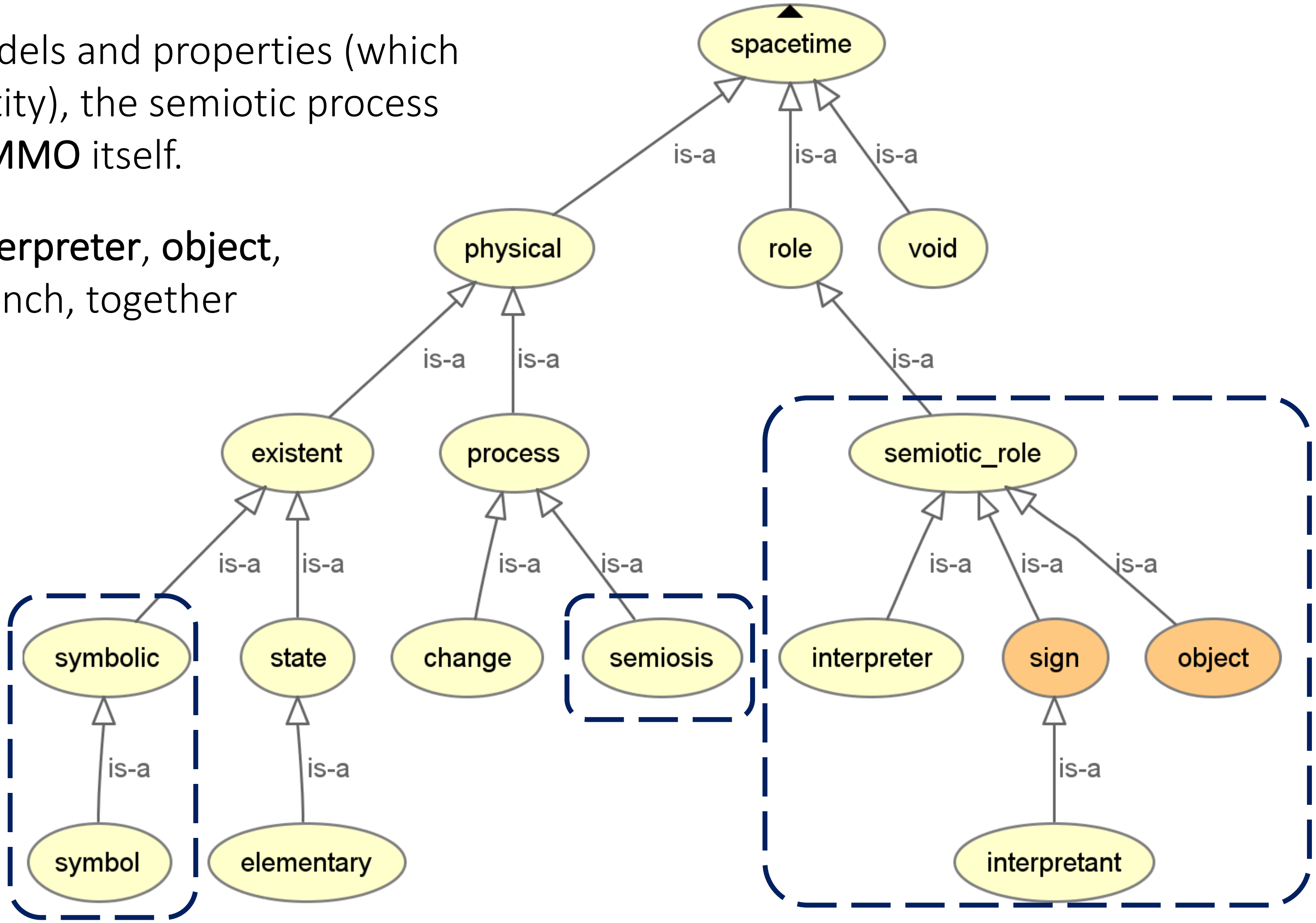
EMMO Semiotic

Since the **EMMO** must represent models and properties (which are signs that stand for a physical entity), the semiotic process must be described also within the **EMMO** itself.

The concepts of Peirce semiotics (**interpreter**, **object**, **sign**) are included in the semiotic branch, together with the **semiosis** process.

Besides that, a branch for representing **symbol**s and **symbolic** entities (e.g. characters, numbers, words) has been introduced, based on formal languages approach.

Symbols of a formal language need not be symbols of anything.

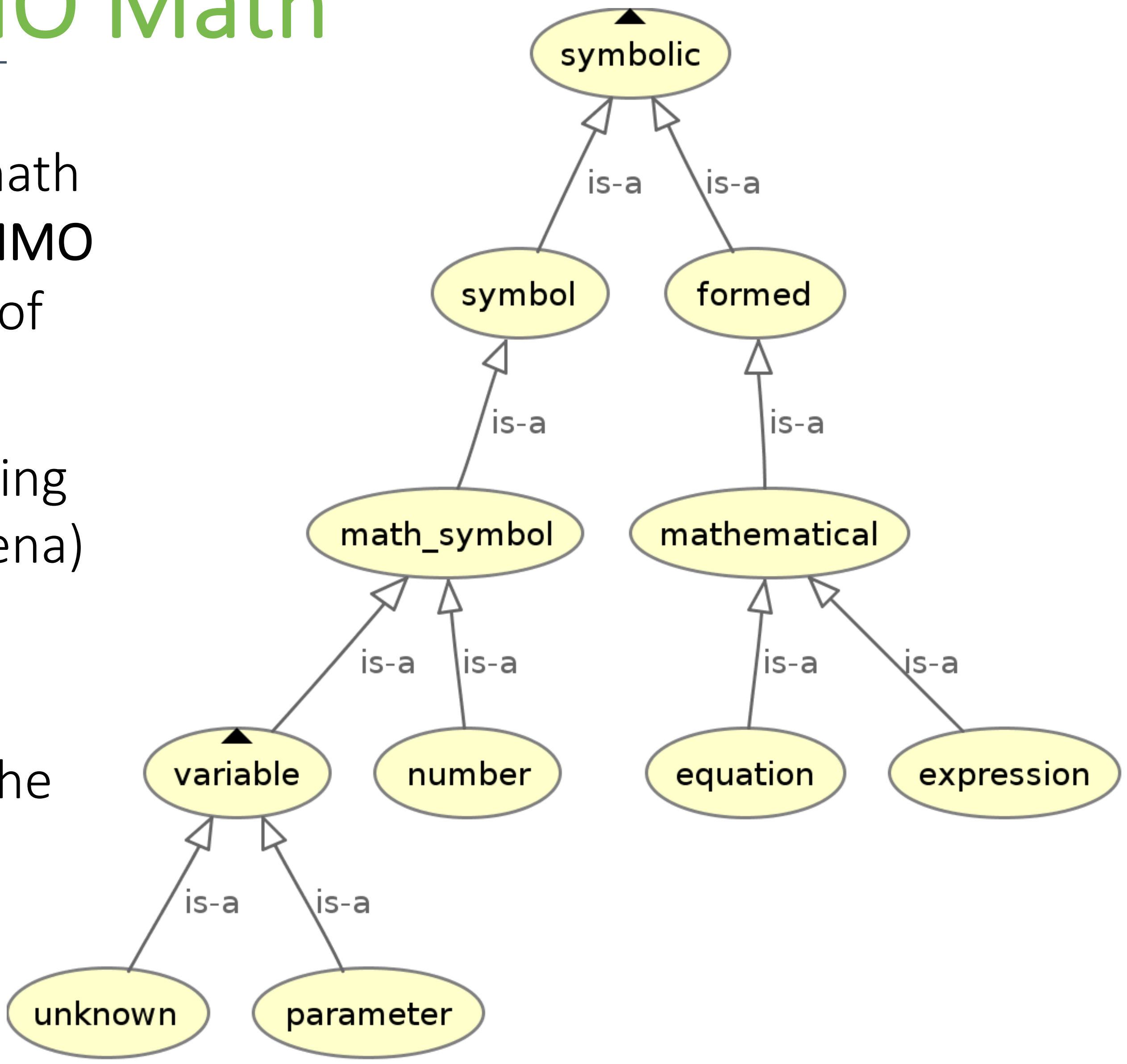


EMMO Math

The **symbolic** class is the superclass of the math branch, since mathematics is seen in the **EMMO** as a formal language, based on an alphabet of **mathematical symbols**.

Mathematical expressions that have a meaning (i.e. are used to represent physical phenomena) are also **sign**s (e.g. physics equations).

The **formed** class includes formal languages constructs (i.e. list of symbols) that follows the rules of a specific language.



EMMO Properties

In the **EMMO**, a **property** is a **sign** that stands for an **object** that the **interpreter** perceived through a well defined **observation** process.

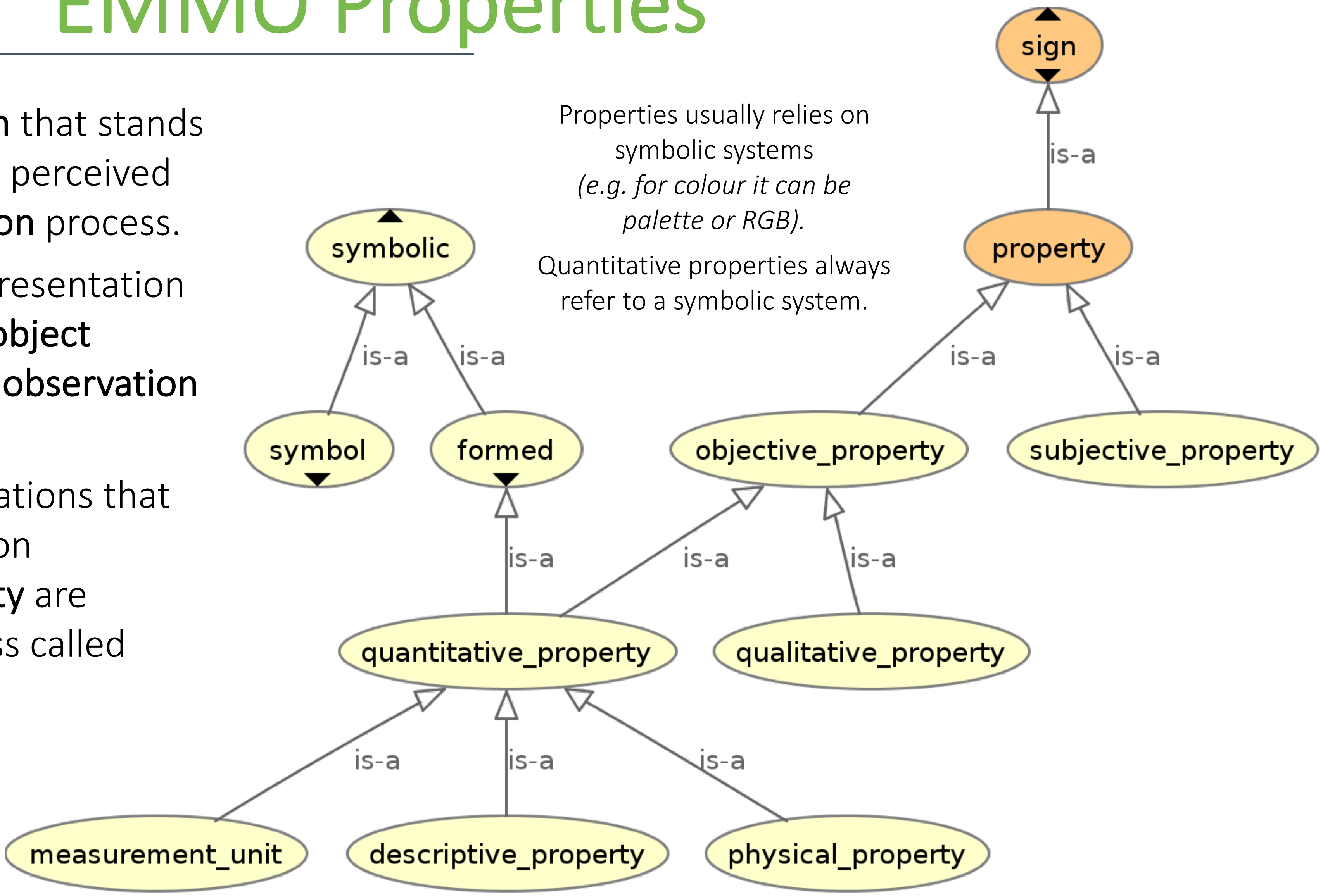
A **property** is always a partial representation of an **object** since it reflects the **object** capability to be part of a specific **observation** process.

Property subclasses are specializations that depend on the type of observation processes. A **quantitative property** are related to an **observation** subclass called **measurement**.

e.g. the property 'colour' is related to a process that involves emission or interaction of photon and an observer who can perceive electromagnetic radiation in the visible frequency range.

Properties usually relies on symbolic systems *(e.g. for colour it can be palette or RGB).*

Quantitative properties always refer to a symbolic system.



EMMO Properties

3.4 kg

The semiotic branch paves the way for the inclusion in the **EMMO** of formal languages and **data recognition**.

Change **raw data** into **information** through **interpretation** of the format.

Semantic extraction is represented within the **EMMO** at the same time for several interpreters!

How to represent the 'thing' on the left within the **EMMO**?
It depends on the interpreter:

physical: it is a physical object, i.e. the black and white pixels on the screen

existent: its a physical that unfolds in time retaining its meaning (i.e. does not change class)

symbolic: is made of symbols coming from a code (i.e. math and western alphabet) for an interpreter used to this alphabet

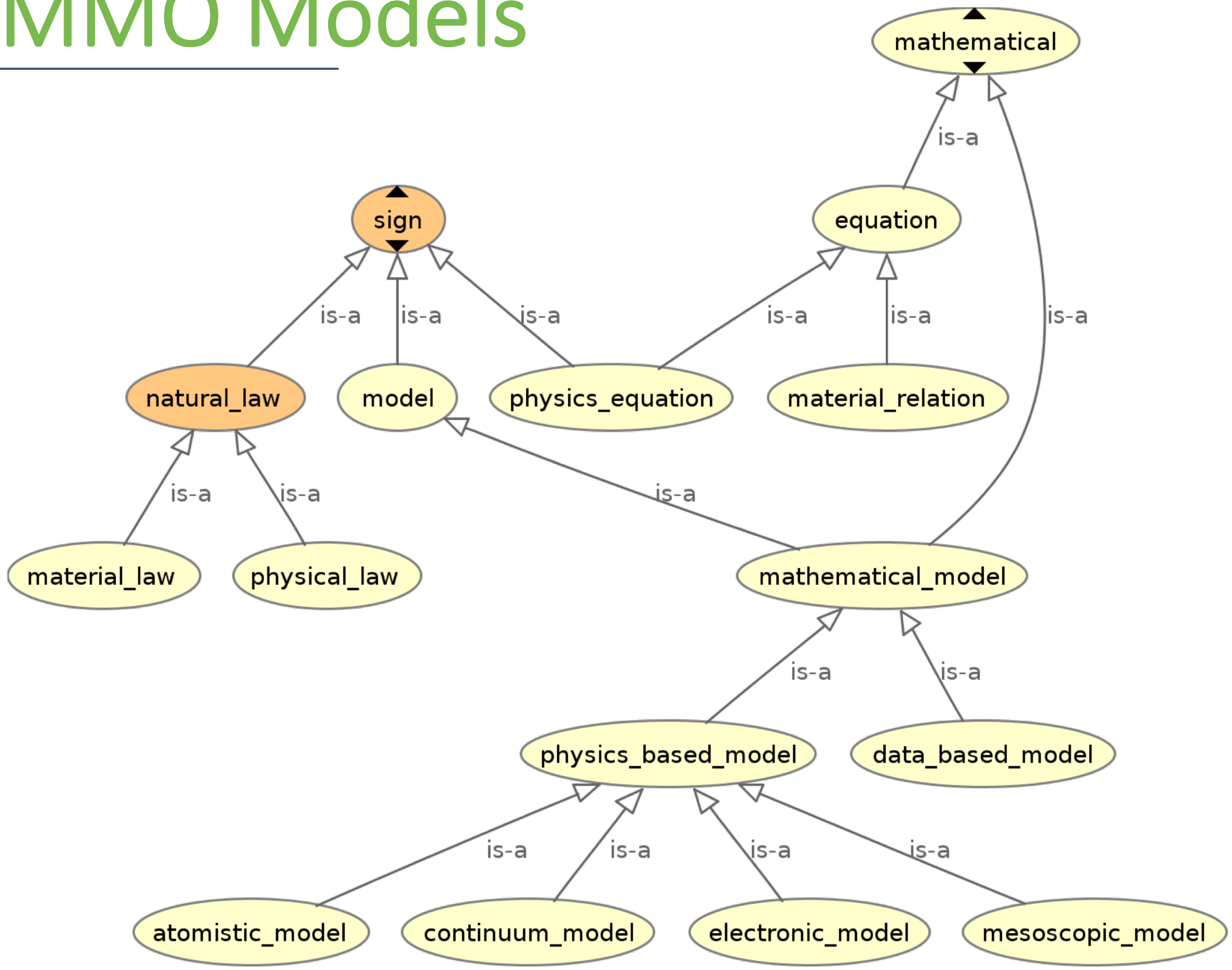
sign/property: has a meaning for an interpreter who is skilled in numbers measurement units

physical property: stands for a physical property of another physical entity according to an interpreter who knows a bit of physics

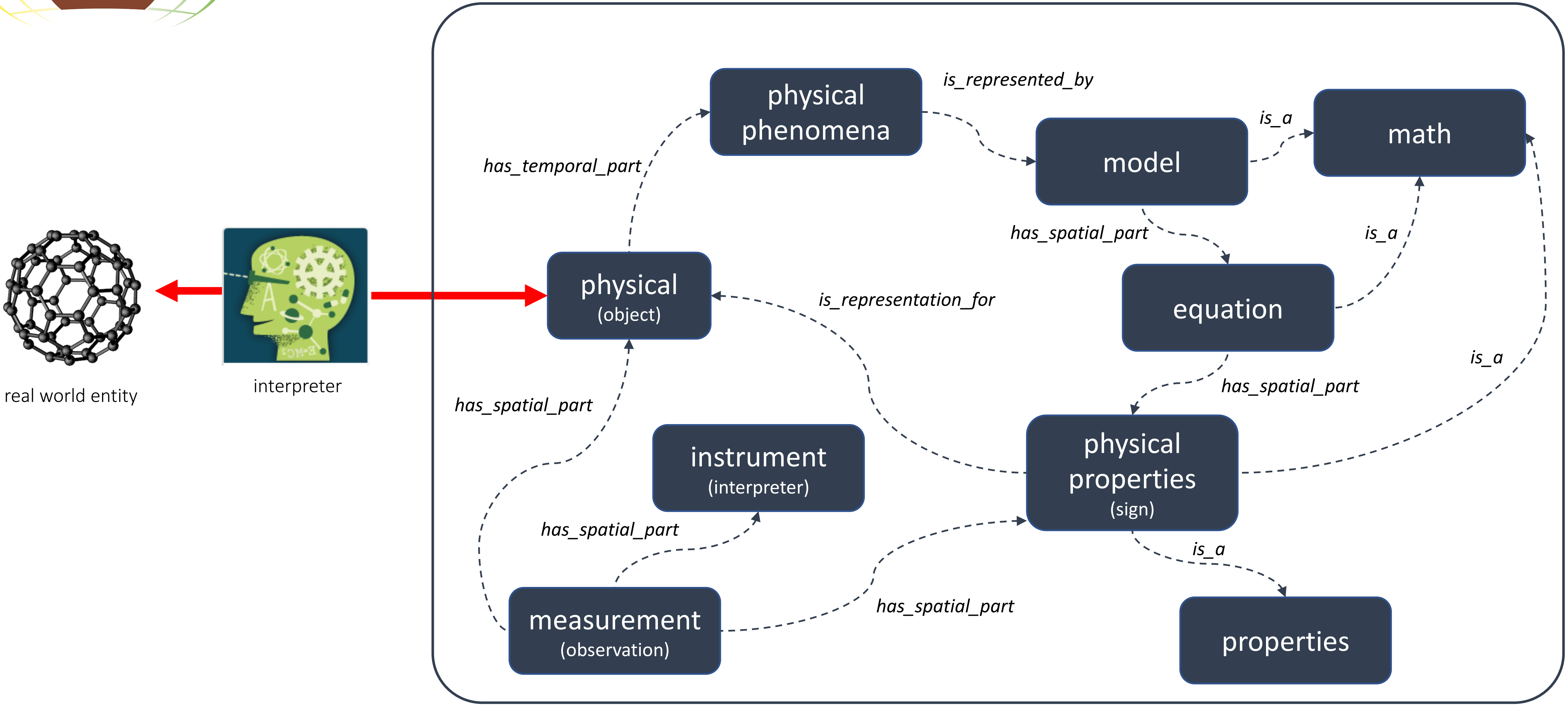
EMMO Models

A **model** is a **sign** that not only stands for a **physical** or a **process**, but it is also a simplified representation, aimed to assist calculations for its description or for predictions of its behaviour.

A **model** represents a **physical** or a **process** by direct similitude (e.g. small scale replica) or by capturing in a logical framework the relations between its properties (e.g. mathematical model).

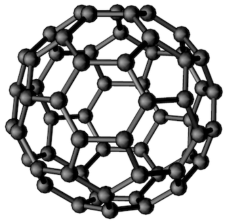


EMMO Models


real world entity
interpreter
physical
(object)
has_temporal_part
physical
phenomena
is_represented_by
model
is_a
math
has_spatial_part
is_a
equation
is_representation_for
has_spatial_part
is_a
has_spatial_part
instrument
(interpreter)
physical
properties
(sign)
has_spatial_part
is_a
measurement
(observation)
has_spatial_part
properties

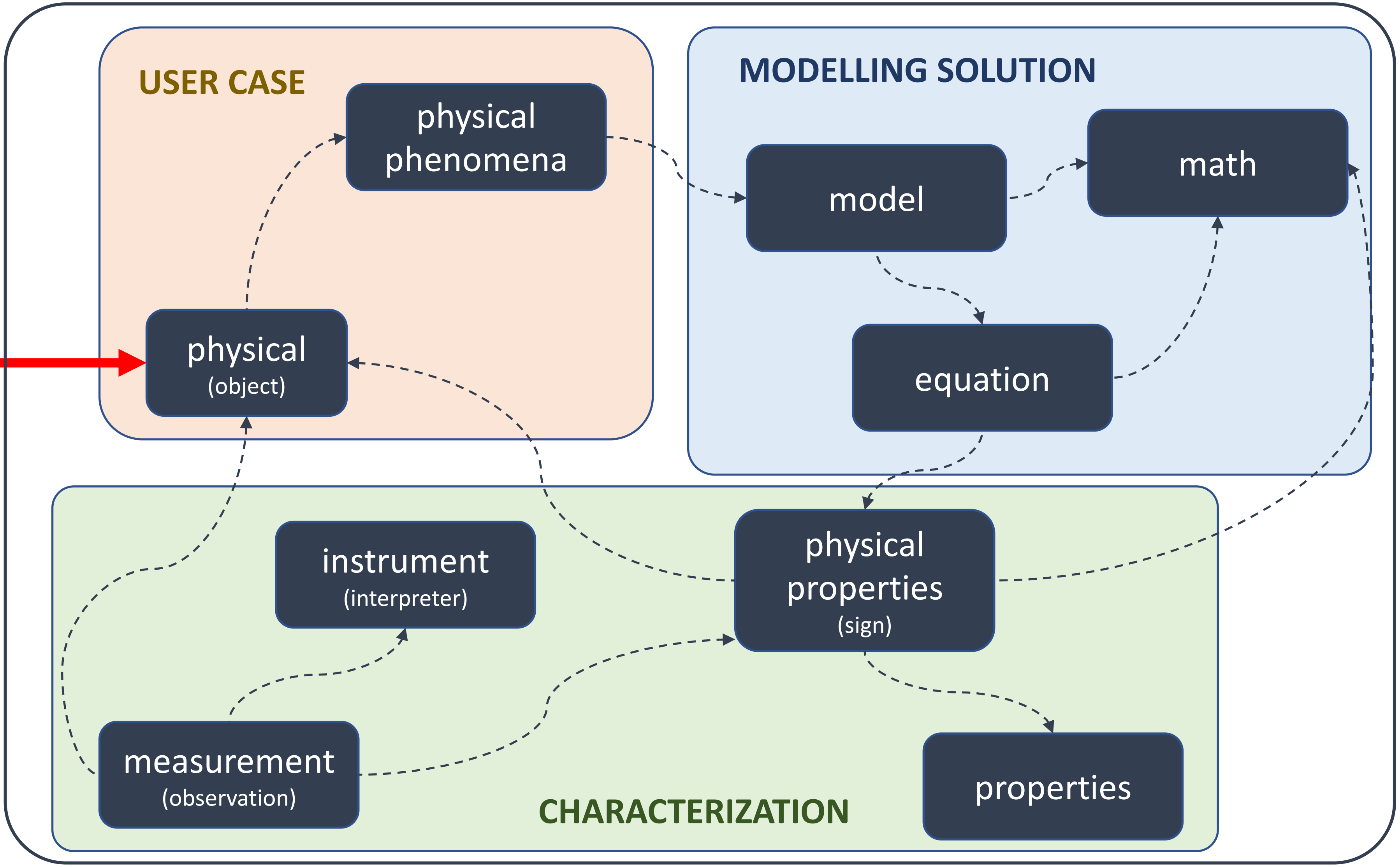
EMMO Models

Horizontal interoperability: one user case, multiple modelling solutions.



real world entity

interpreter



Linking between properties database, models and user cases to facilitate **validation** and **data collection**.

EMMO relations

EMMO has very limited and strictly categorized relations, easy to use, understand and maintain. All goes down to **two primitive relations families**:

MEREOTOPOLOGY

Parthood and *Slicing*

SEMIOTIC

Representation

Relations such as participation to a process falls under mereology.

e.g. you have to be part of a 4D process in order to participate to it

Mereology is also used to declare symbols that constitute symbolic entities.

e.g. unit of measurement as part of a physical property

EMMO taxonomy is strongly based on reasoning, up to level of expressivity allowed by OWL-DL.

(EMMO concepts would be better expressed in FOL or even Second Order Logic)

EMMO maintainers

Who will ensure a constant development and testing of the **EMMO** in the next years?

NMBP-24-2016

European Materials Modelling Council - CSA

2019

EMMO foundations laid within this CSA project.

DT-NMBP-09-2018

Digital Ontology-based Modelling Environment for Simulation of materials

2022

EMMO applications cases and integration within a OSP expected within 2020-2021.

Team of philosophers, ICT experts and applied scientists.

NMBP-25-2017

Materials Modelling Marketplace for Increased Industrial Innovation

2022

VIMMP

Virtual Materials Market Place

2021

EMMO applied to larger materials modelling communities and marketplaces infrastructures.

… more **existing projects** to involve and more to come in the **next DT-NMBP calls** (hopefully)!!!

EMMO: where to find it?

- https://emmc.info/emmo-info/ or https://emmo.tech
- EMMO v0.9.9 available on: https://github.com/emmo-repo/emmo
- EMMO Authors (IP Owners): Emanuele Ghedini (UNIBO), Gerhard Goldbeck (GCL), Adham Hashibon (Fraunhofer), Georg Schmitz (ACCESS), Jesper Friis (SINTEF)

- Manuscripts to be submitted in peer review journals
  - Foundations of EMMO
  - EMMO: an ontology for applied sciences
  - Authors: Emanuele Ghedini, Gerhard Goldbeck, Adham Hashibon, Georg J. Schmitz, Jesper Friis

EMMO NEXT STEPS

- **Publish** the EMMO v1.0 version with more consistent mereotopological foundations
- **Provide documentation** for the EMMO and its specific approaches implemented in the modules (e.g. papers, reports)
- Looking for **taxonomical compatibility** (at least) with other important ontologies (e.g. BFO, CHEBI, IAO)
- Use it on the field within **other H2020 projects**

THANKS FOR YOUR ATTENTION

EMMO authors:

Emanuele Ghedini	*(University of Bologna)*
Gerhard Goldbeck	*(Goldbeck Consulting)*
Adham Hashibon	*(Fraunhofer IWM)*
Georg Schmitz	*(ACCESS)*
Jesper Friis	*(SINTEF)*

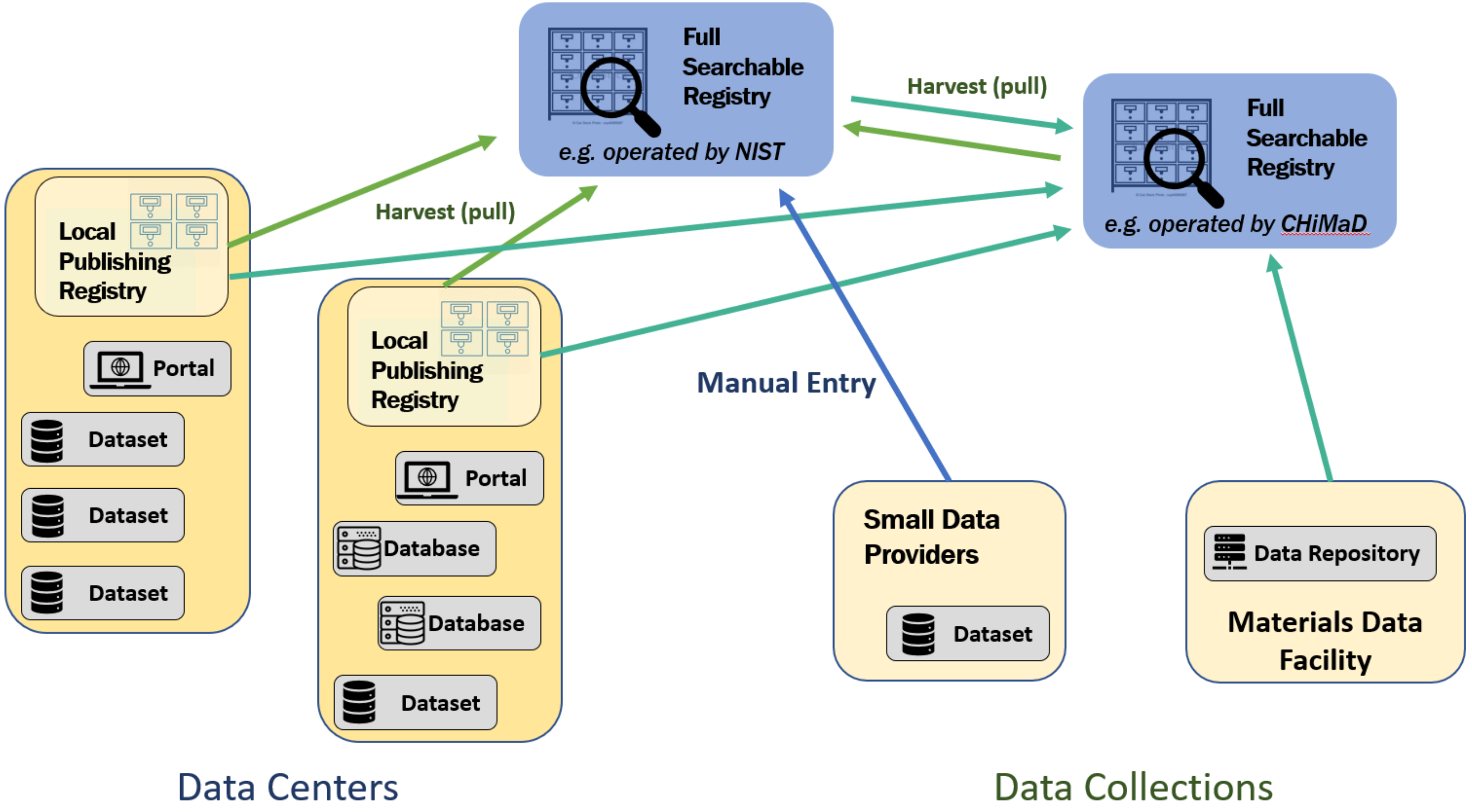
International Materials Resource Registry:
Outputs and Current Activities

Raymond Plante -- NIST

Final Report of Outputs

- Presented at P11 (Berlin)
- Internal WG review of Final Report Document:
  - Available in Working Group Area: see Folder, “**International Materials Resource Registries WG Report: Drafts**”
  - Current version is RC4
  - Please send comments to working group mailing list (imrr-wg@rd-groups.org) by 20 November 2019
- To be submitted to secretariat for RDA-wide RFC afterward.

Output Summary:
Registry Federation Framework



Output Summary:
Registry Federation Framework

- Requirements
  - Resource metadata exchange protocol
    - Identifiers
    - Distinguish between own records and those harvested from other sources
    - Communicate when resources are no longer available
    - Require minimal record validation
    - *Our implementation: OAI-PMH. (Others: Linked Data Platform (LDP), ResourceSync, …)*
  - Common Metadata Schema/Format
    - Openly defined
    - Associate a globally unique identifier
    - Validate-able
    - Low-impact evolution mechanism (e.g. extensions)
    - *Our implementation: XML, XML Schema*
- Architecture, Recipe for registry interaction

Output Summary:
Metadata

- XML Schema
  - Extension mechanisms
  - Different Types of Resources:
    Data Collections Databases Software Informational Sites Organizations
    - Different types can have different data associated with them
    - New types can be defined. (Semantic Asset)
  - Applicability to different domains
    - Place to include domain-specific metadata
    - Can support multiple domains simultaneously
- Materials Science Vocabulary
  - 3-tiered subject terms
  - Drives faceted browsing
  - SKOS definition available

Output Summary:

Working, Populated Registry Federation

- Implementation: NIST Materials Resource Registry
  - Adaptation of the NIST Configurable Metadata Curation System (CDCS)
  - https://github.com/usnistgov/MaterialsResourceRegistry
- Two instances:
  - NIST: https://materials.registry.nist.gov/
  - CHiMaD/MDF: https://mrr.materialsdatafacility.org/
- Over 350 records

Post-WG Activity:

Supporting Semantic Assets for MSE

- Adding “Semantic Asset” as a resource type to MMR
  - Cover vocabularies, ontologies, types, registries, …
  - To encourage sharing use across continents
- MSE Vocabulary Use & Maintenance
  - Elsevier pilot: considering tagging MSE journal articles with vocabulary terms
- MDII Task Group
  - Expanding impact of vocabularies and registries

Post-WG Activity:
Enhancing Discovery

- Additional registry instances in the world
- NIST: Software improvements
  - Improving usability
  - Stronger support for PIDs and PID resolution
- Deep Discovery: leveraging data provider search tools
  - Register search services
  - Tools can pass search queries to a repository's or database's search service for retrieving dataset/measurement-level results

Task Group on Materials Ontologies

https://sites.google.com/view/rda-materials-ontologies-tg/home

RDA/CODATA Materials Data IG

Gerhard Goldbeck (GCL, UK)

Cate Brinson (Duke, US)

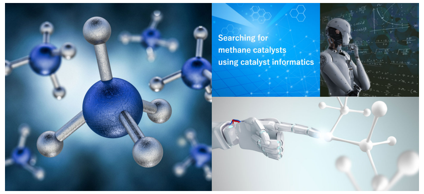
Clare Paul (AFRL, US)

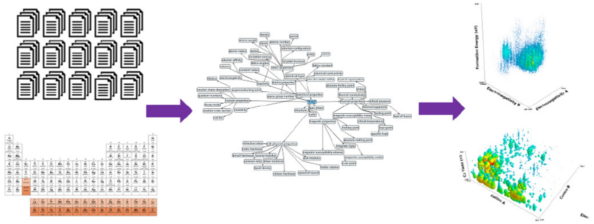
Scope

- Materials includes any substance in any state at any scale.
- Any ontology about materials or directly linked to materials.
  - E.g. characterisation, modelling, processing, safety etc
- Building ontologies as well as application of ontology.

Examples/Use of Ontologies

- Database integration
  - Connected data!
  - Discover new trends
  - New materials candidates





Takahashi, et al (2018). Redesigning the Materials and Catalysts Database Construction Process Using Ontologies. J Chem Inf Mod 58, 1742.

- Easier Database queries
  - Ontology organises data by domain knowledge: contrast to database which is organised by IT need.
  - Querying can be done by scientist using script!

SCHOTT glass made of ideas *presentation, EMMC Workshop, Vienna 2019*

- Materials Modelling: interoperability, translation; EMMO
- Characterisation, analytical: standardisation
- Materials Modelling/Data Marketplaces

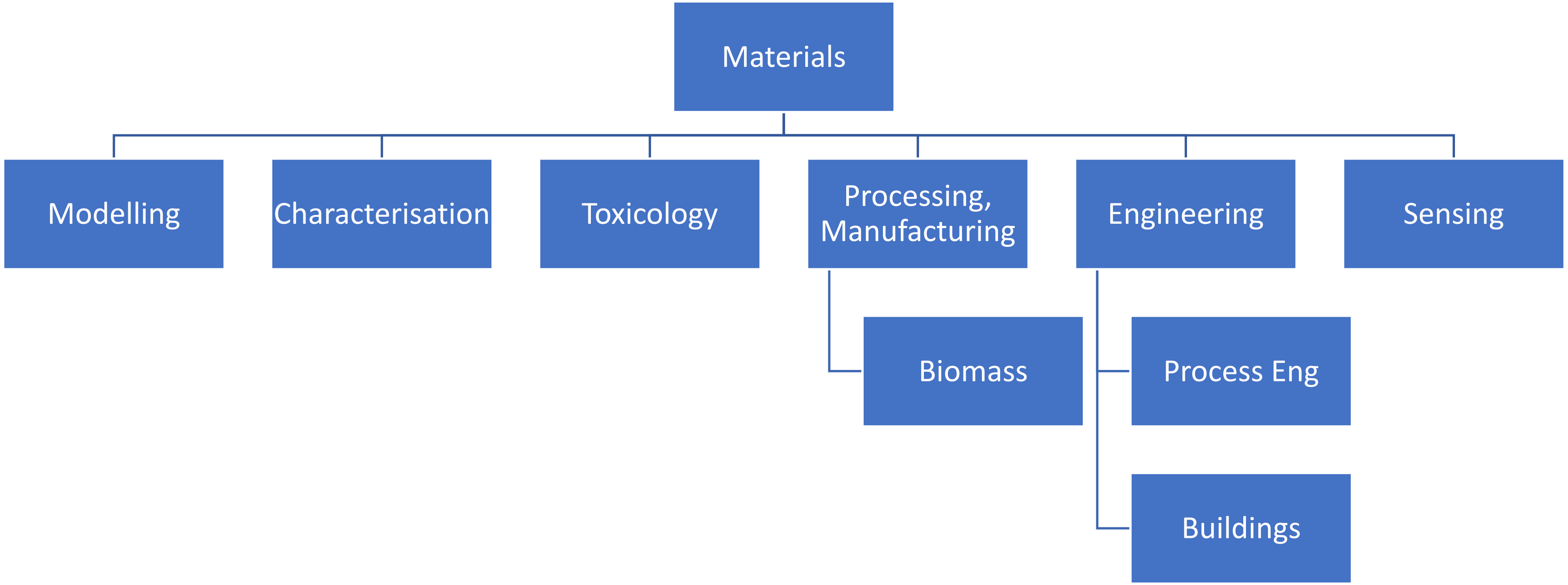
EMMC

OYSTER Allotrope Foundation

VIMMP VIRTUAL MATERIALS MARKETPLACE MarketPlace

Where are we today?

- Range of disciplines related to materials with their own efforts in taxonomies and ontologies

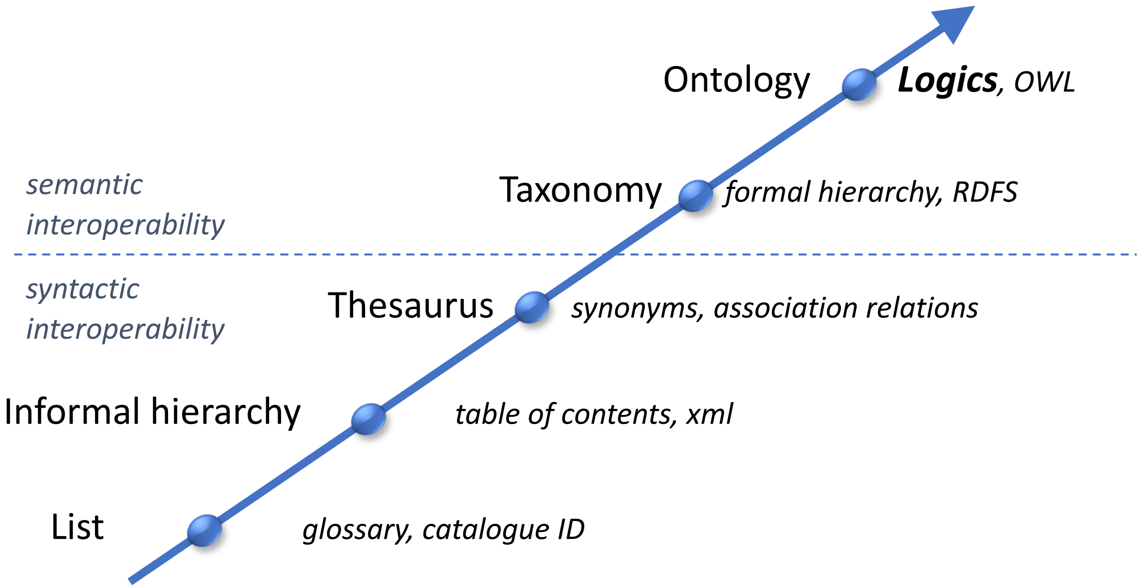


Where do we want to get to?

- A materials focus on ontologies
- A connected eco-system
- Interoperability
- Widest possible agreement about top-level concepts and relations

Knowledge Organization Systems: Semantic Spectrum

Semantics allows a resource to be understood by both humans and machines → promote interoperability.



Machine can **interpret** information and **reason**.

Machine can process information due to **compatible syntax**

Adapted from:
Leo Obrst "The Ontology Spectrum". Book section in of Roberto Poli, Michael Healy, Achilles Kameas "Theory and Applications of Ontology: Computer Applications". Springer Netherlands, 17 Sep 2010.

Ontology levels

- Express fundamental concepts of physics and materials science
- Models, Properties, Processes, Materials (and their structure/granularity), Measurements etc
- Specific taxonomies and ontologies

Upper Level Ontology

Mid Level Ontologies

Domain Level Ontology

Status of Materials Ontologies

- No common Upper Ontology
- Most have mid level or domain focus, e.g. Structural Materials, Composites, Steels, Catalysts, etc
- Lack of connection between chemistry and materials efforts

Objective: Mapping

- *What* groups *are working on ontologies*?
- Which topics *and what are their projects*?
- Establish improved communication.
- Document and Categorise
  - International Federation of Materials Resource Registries https://materials.registry.nist.gov/ & https://www.rd-alliance.org/groups/working-group-international-materials-resource-registries.html
  - TaxOnDa https://emmc.info/document-your-semantic-asset/

Further Objectives

- Making materials ontologies FAIR.
  - repositories for Materials Ontologies etc
- Collect use cases and requirements
  - Set up online means for collection (e.g. github).
  - Create problem statements and sample instance graph related to use case
- Recommendations on a governance system
- Ontology interoperability framework
- Ontology alignment
- Interfacing with other domains (e.g. engineering)

An active field

- EMMO European Materials & Modelling Ontology
- US grants/activities
- "A Method for Extending Ontologies with Application to the Materials Science Domain" https://datascience.codata.org/articles/10.5334/dsj-2019-050/
  - *In materials science,…, a large number of research groups and communities are building and developing data-driven workflows. However, much of the data and knowledge is stored in different heterogeneous data sources maintained by different groups. This leads to a reduced availability of the data and poor interoperability between systems in this domain.* ***Ontology-based techniques are an important way to reduce these problems and a number of efforts have started.***

Towards Standardised Documentation of Data through taxonomies and ontologies (CSA)

ID: DT-NMBP-39-2020

Focus area: Digitising and transforming European industry and services (DT)

Bundesanzeiger
Herausgegeben vom Bundesministerium der Justiz und für Verbraucherschutz
www.bundesanzeiger.de

Bekanntmachung
Veröffentlicht am Freitag, 20. September 2019
BAnz AT 20.09.2019 B4
Seite 1 von 7

Bundesministerium für Bildung und Forschung

Bekanntmachung zur Förderung von Zuwendungen von Vorhaben im Rahmen der Initiative zur Digitalisierung der Materialforschung in Deutschland (Material*Digital*)

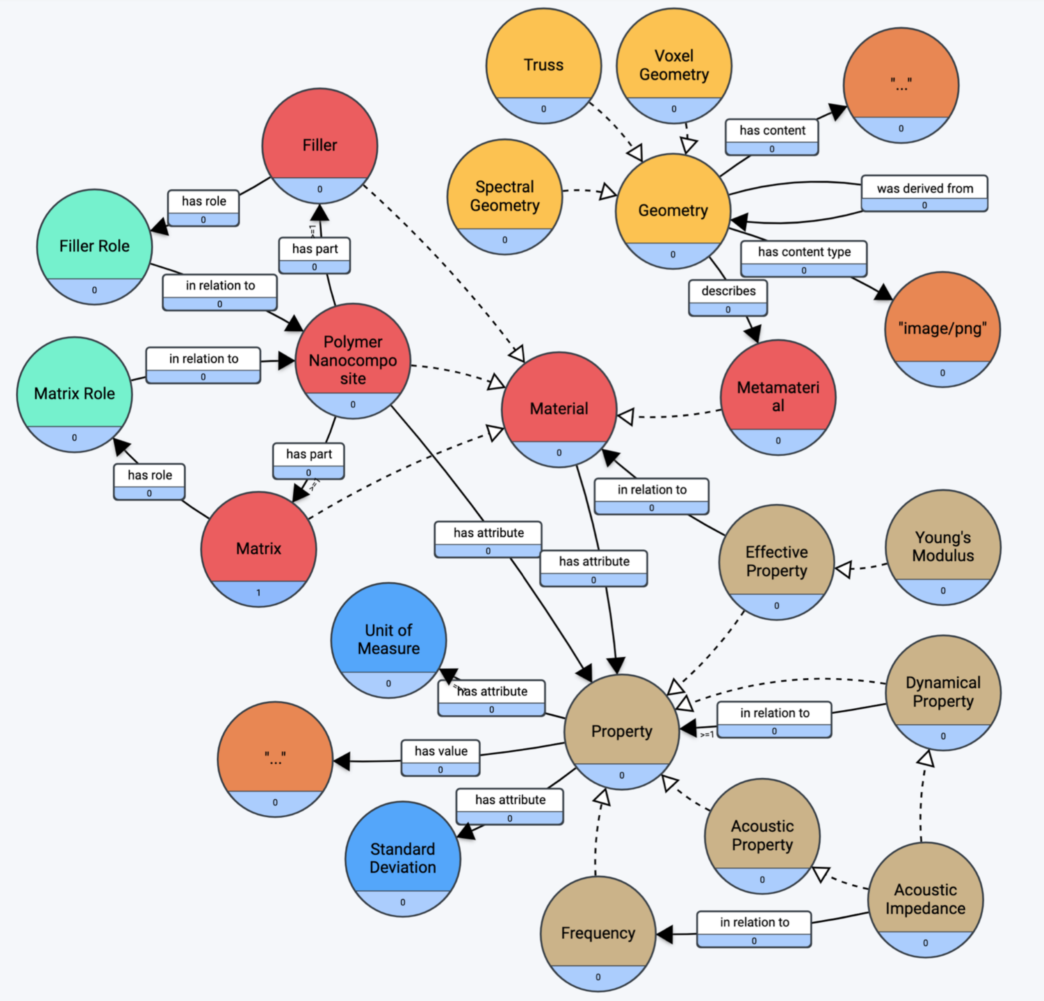
Vom 20. August 2019

- EC and Germany calls for proposals

An active field: example

Materials Ontology & Knowledge graph

- Created an **ontology** and **knowledge graph** that provide accessibility and compatibility between parallel material standards and provide a platform for data storage and search, customized visualization, and machine learning tools for material discovery and design.
- The initial Nanomine knowledge graph for polymer nanocomposites has been deployed: https://materialsmine.org/wi/home
- Extension to metamine ontology and generalization to materials ontology ongoing
- Presented at Federated KBs & the Open Knowledge Network Workshop at AKBC 2019: *"A Provenance-Aware Knowledge Graph Framework for Open Knowledge Network Settings."*
- *NanoMine Schema: A Data Representation for Polymer Nanocomposites*, APL Materials, 2018, doi.org/10.1063/1.5046839
- PIs: Brinson, McGuinness, Chen, Schadler, Rudin, Daraio, McCusker



Acknowledgements

https://vimmp.eu/

https://www.the-marketplace-project.eu/

http://www.oyster-project.eu/

- VIMMP: received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760907
- MarketPlace: received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760173
- Oyster: received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760827