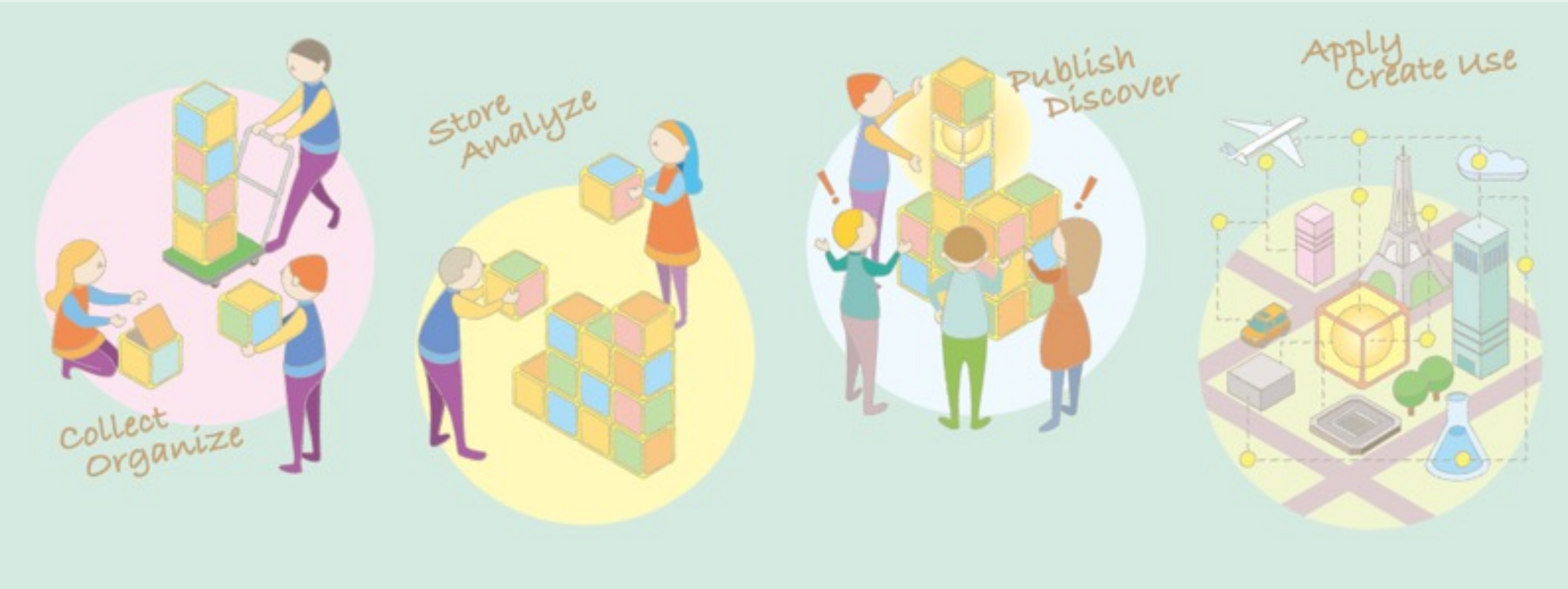
# Materials Data Repository metadata schema and cross-database federation

Asahiko Matsuda*, Kosuke Tanabe, Masashi Ishii, Takuya Kadohira

* https://orcid.org/0000-0001-5989-027X
MATSUDA.Asahiko@nims.go.jp

Materials Data Platform, Research Network and Facility Services Division,
National Institute for Materials Science (Tsukuba, Japan)

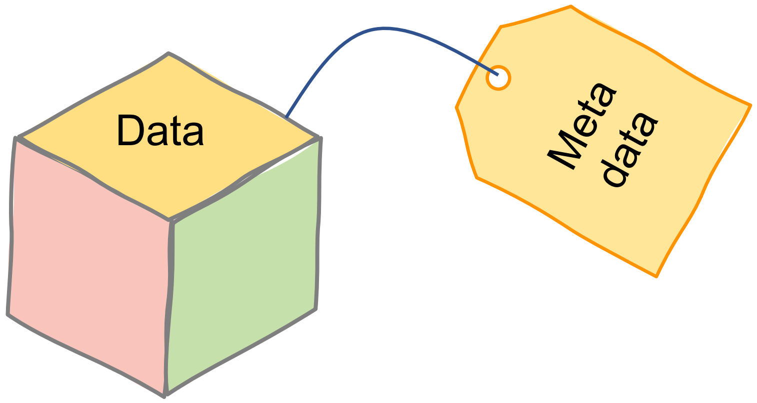
# An artist’s rendition of a materials data lifecycle



Circa 2017: “*We should build a platform to support all four stages of a materials data lifecycle!*”

https://dice.nims.go.jp/

# Every data needs a name tag



Without it, data can easily end up as a random blob.

## Findability

- Basic description
- Instrumentation, methodology
- Sample/Material description

## Experiment reliability

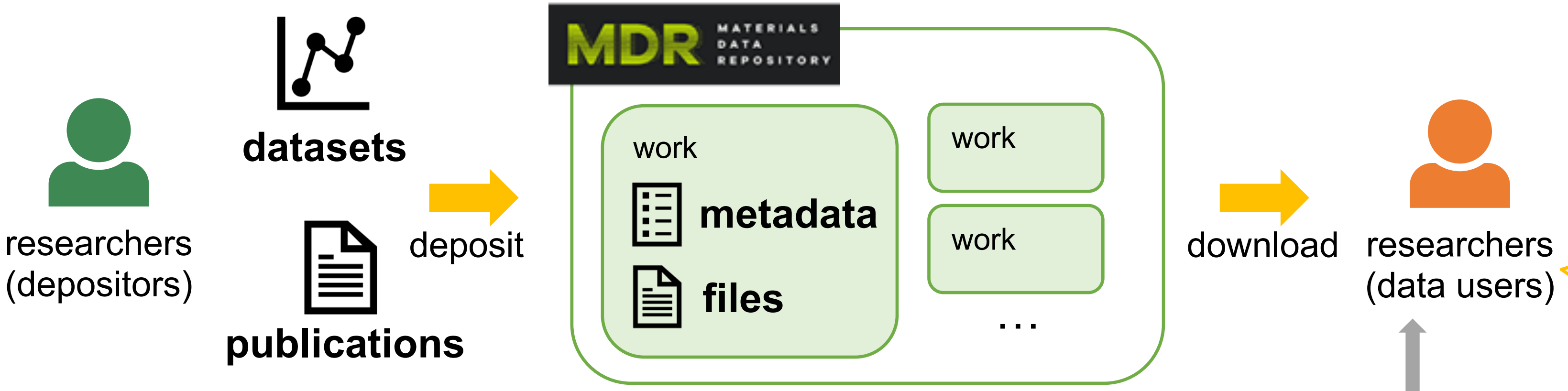
- Instrument conditions
- Experiment parameters
- Experiment environment

## Data integration

- Dataset format
- Column information

# “Materials Data Repository” for data publishing

https://mdr.nims.go.jp/



researchers (depositors)

datasets

publications

deposit

work

metadata

files

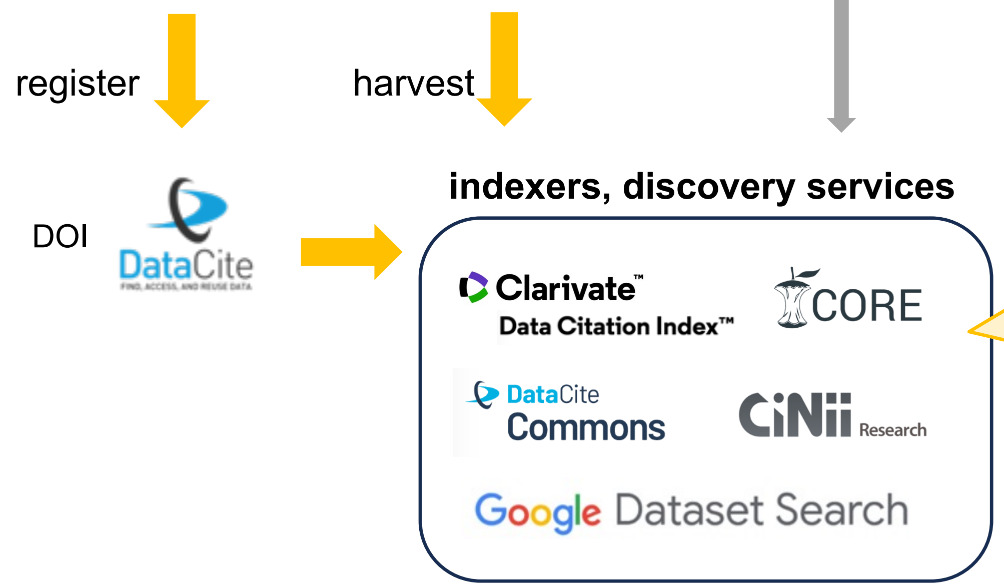
work

work

…

download

researchers (data users)



register

harvest

DOI

indexers, discovery services

Interested in
*“Scientific metadata”*

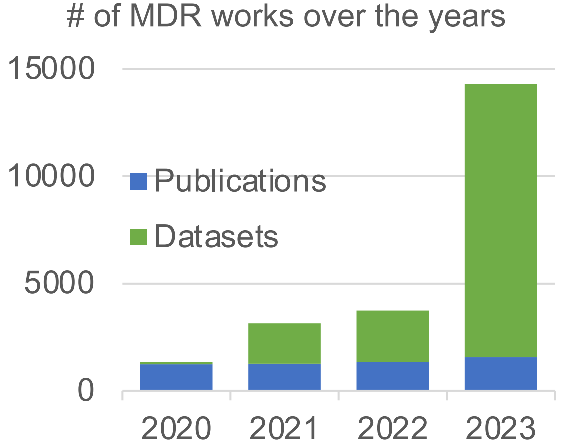
- **Material/Specimen**
- **Sample preparation**
- **Measurement conditions**

etc.

Interested in
*“Bibliographic metadata”*

- **Title**
- **URL**
- **Creators and ORCIDs**

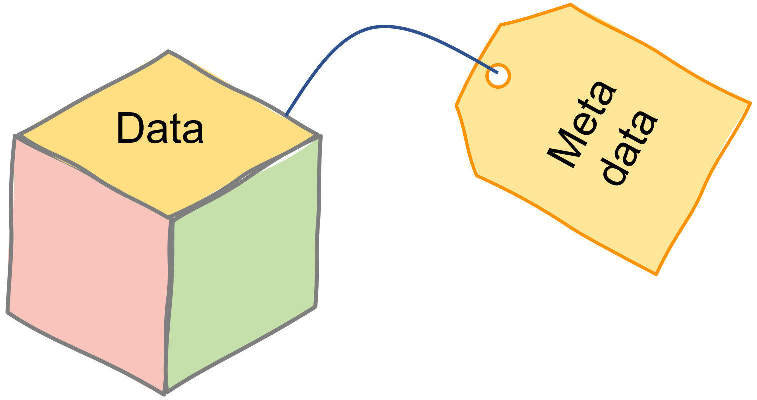
etc.



# of MDR works over the years

K. Tanabe, A. Matsuda, “A development of Materials Data Repository for materials informatics”, *IPSJ SIG Tech. Rep.* **IOT51/SPT39** (2020) (in Japanese)

# Metadata categories



## Bibliographic metadata

- Title
- Creator
- Date created

## Administrative metadata

- Data manager
- License

➡ **Common to all**

## Scientific metadata

- Material name/type
- Characterization method
- Experimental conditions
- Calculation method
- Properties addressed
- Sample preparation process

etc.

➡ **Domain specific**

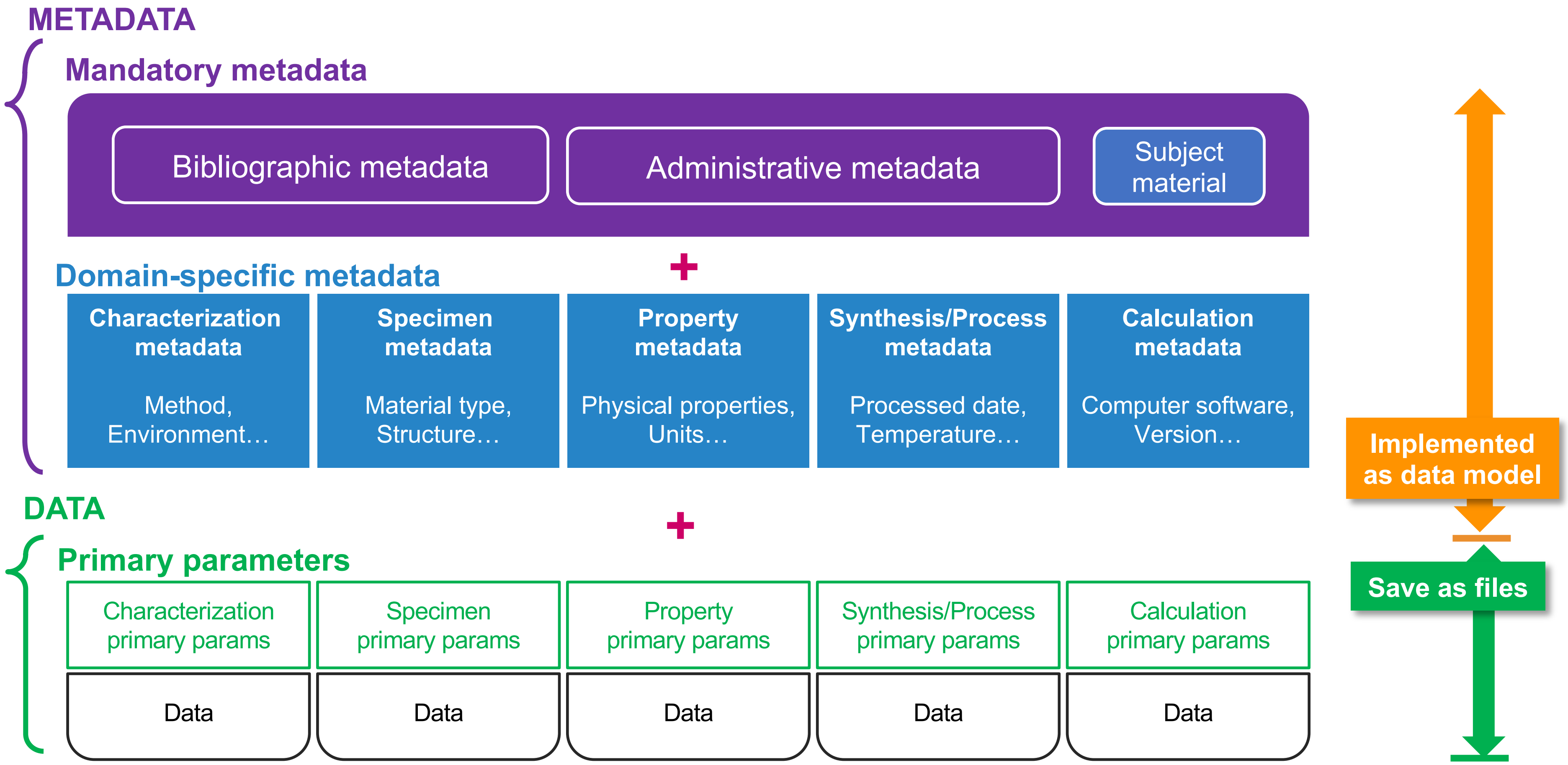
A. Matsuda, “Findability of Materials Research Data”, *Library Fair & Forum 2020* (in Japanese)

# One schema to rule them all?

## DICE Common Message Format 1.0

S. Kikuchi et al., *IEICE Tech. Rep.* **119** SC2019-2 (in Japanese)
Schema: https://doi.org/10.48505/nims.3240

JSON schema designed for system-to-system communication among DICE systems



METADATA

Mandatory metadata

Bibliographic metadata | Administrative metadata | Subject material

+

Domain-specific metadata

**Characterization metadata** Method, Environment…

**Specimen metadata** Material type, Structure…

**Property metadata** Physical properties, Units…

**Synthesis/Process metadata** Processed date, Temperature…

**Calculation metadata** Computer software, Version…

DATA

+

Primary parameters

Characterization primary params / Data

Specimen primary params / Data

Property primary params / Data

Synthesis/Process primary params / Data

Calculation primary params / Data

**Implemented as data model**

**Save as files**

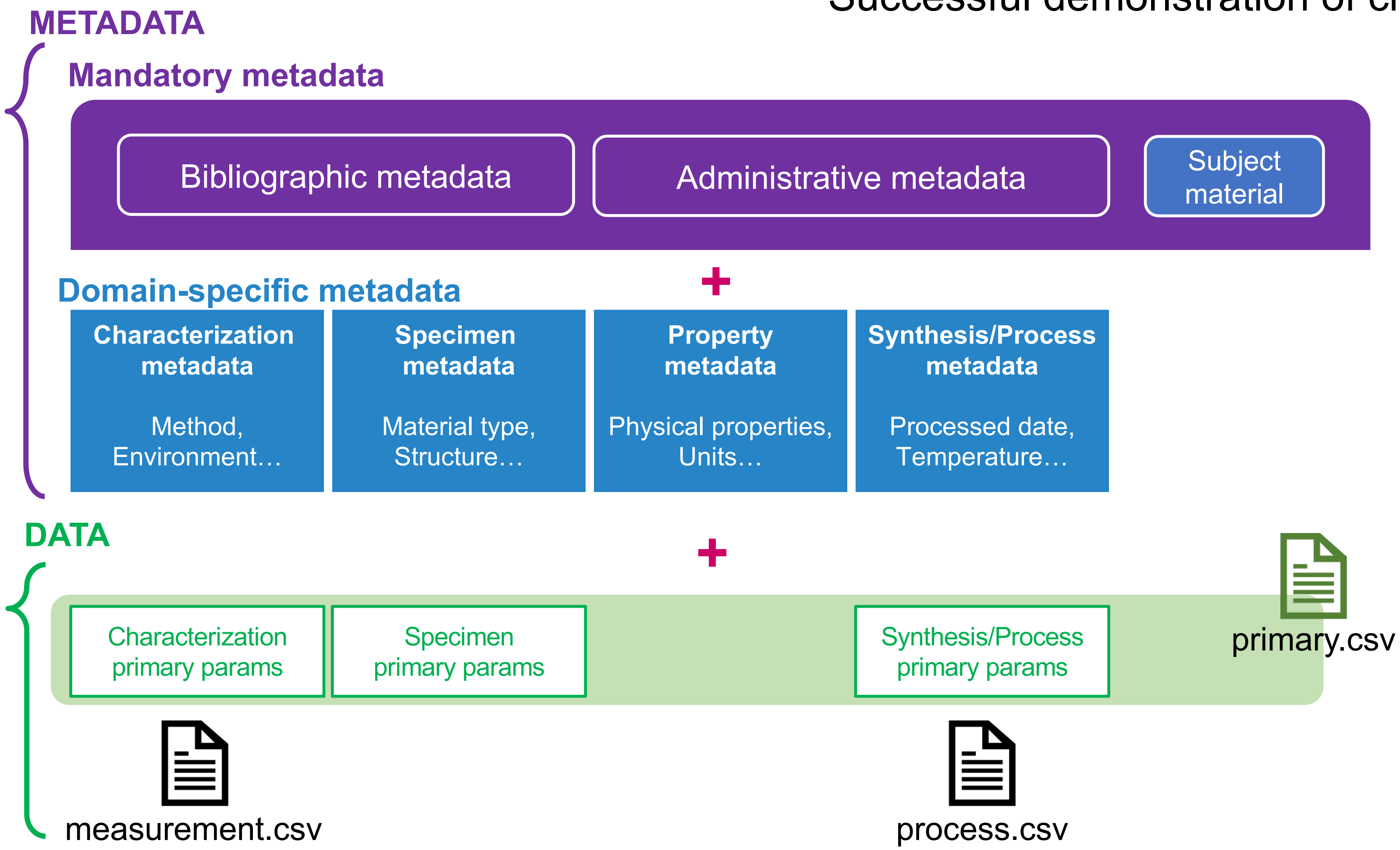
# One schema to rule them all?

DICE Common Message Format 1.0

**Example for a characterization data**

✔ Highly descriptive. High reusability.
Successful demonstration of cross-system messaging.

*but…*



METADATA

Mandatory metadata

Bibliographic metadata

Administrative metadata

Subject material

+

Domain-specific metadata

**Characterization metadata**

Method, Environment…

**Specimen metadata**

Material type, Structure…

**Property metadata**

Physical properties, Units…

**Synthesis/Process metadata**

Processed date, Temperature…

DATA

+

Characterization primary params

Specimen primary params

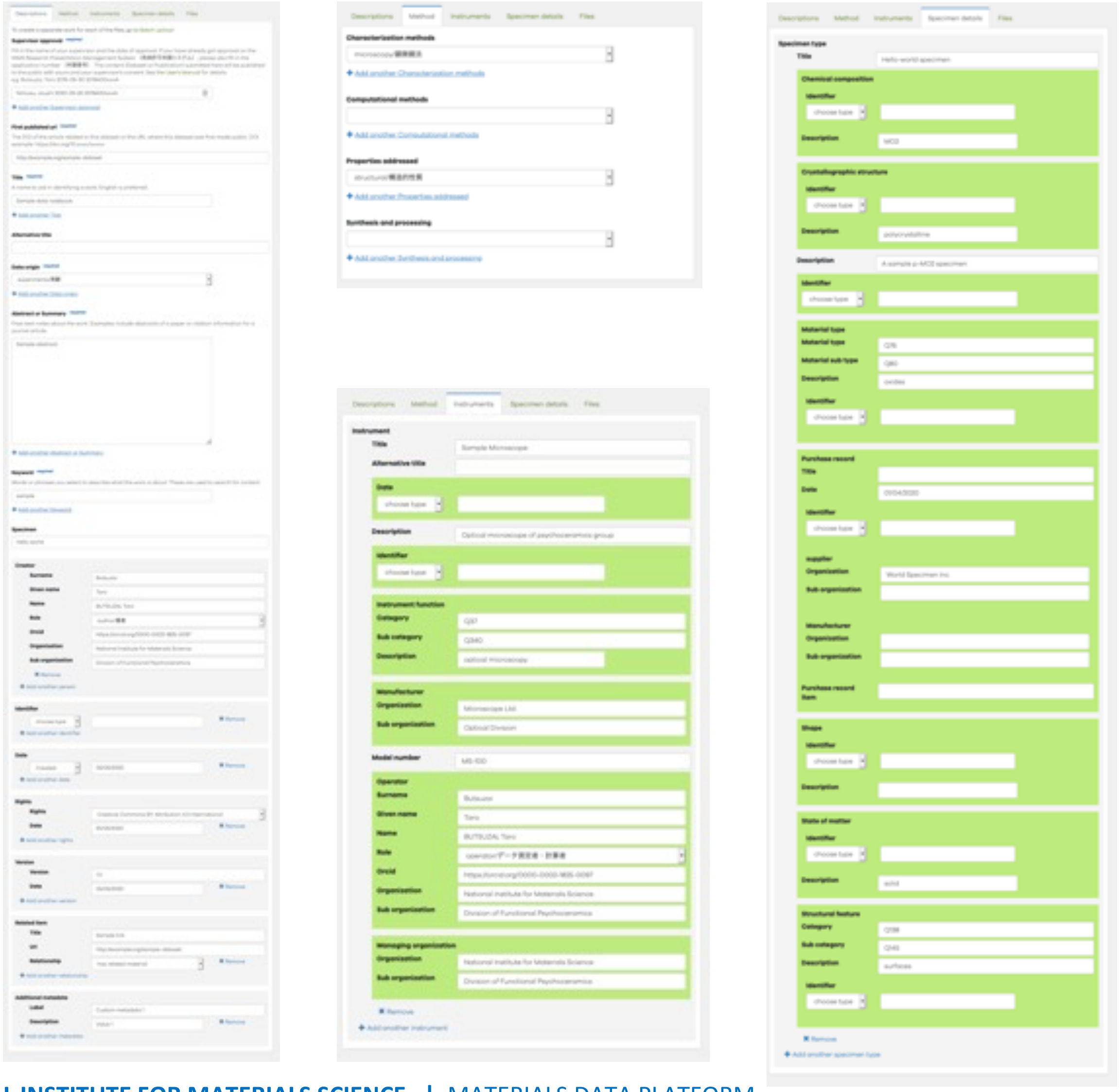
Synthesis/Process primary params

primary.csv

measurement.csv

process.csv

# Nobody's got time for all these!



Descriptions | Method | Instruments | Specimen details | Files

Characterization methods

microscopy/顕微鏡法

✚ Add another Characterization methods

Computational methods

✚ Add another Computational methods

Properties addressed

structural/構造的性質

✚ Add another Properties addressed

Synthesis and processing

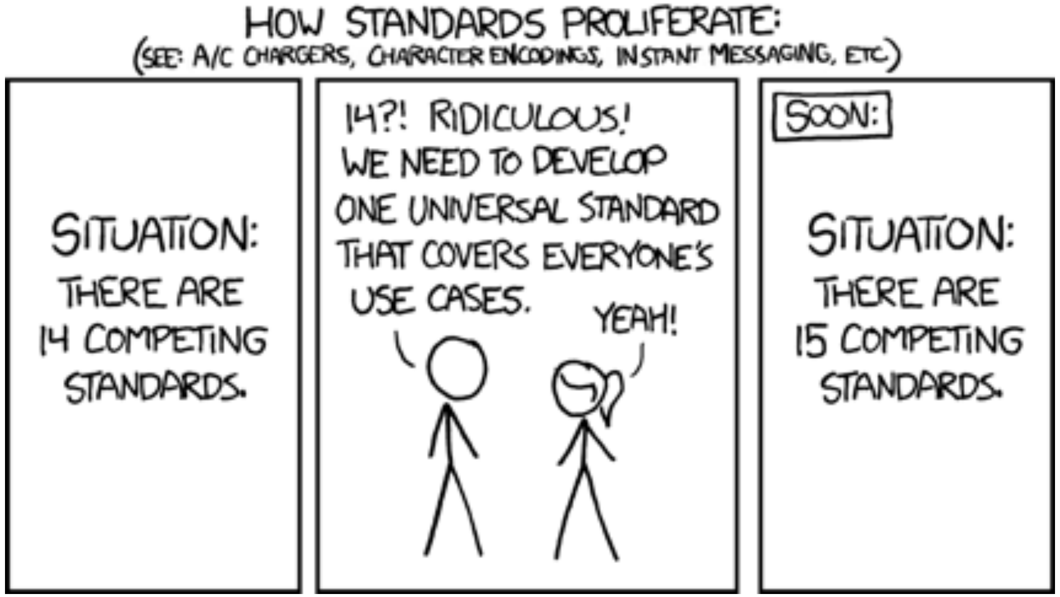
✚ Add another Synthesis and processing

*MDR 1.0 metadata form implementation*

✘ Complex. Not human-readable.
Implementation difficulties.

The full schema defines 300+ fields.
110 of them were implemented in MDR 1.0.

(The schema to cover all uses was
too ambitious to cover practical uses.)

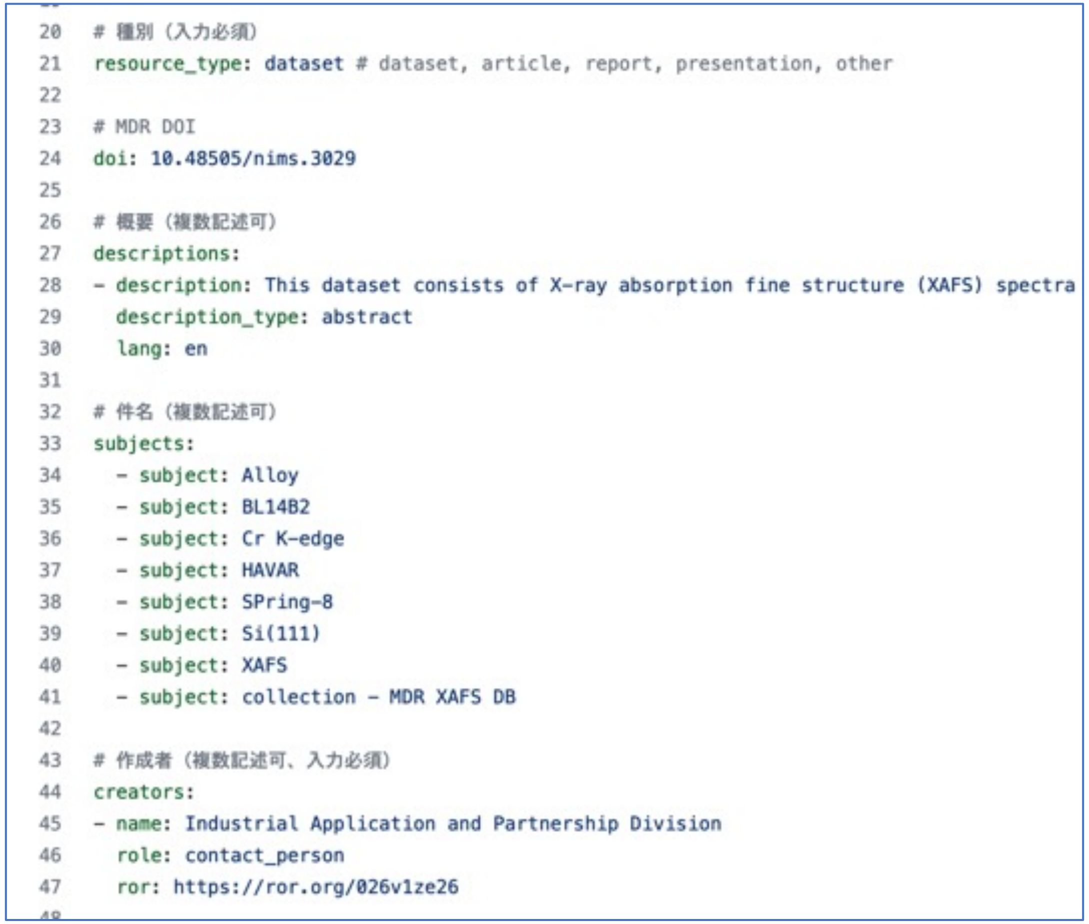

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

https://xkcd.com/927 (cc by-nc)

# MDR Schema 2.0 (Common for all MDR works)

github.com/nims-dpfc/mdr-schema
(doi: 10.48505/nims.3239)

(snippet)



```yaml
# 種別（入力必須）
resource_type: dataset # dataset, article, report, presentation, other

# MDR DOI
doi: 10.48505/nims.3029

# 概要（複数記述可）
descriptions:
- description: This dataset consists of X-ray absorption fine structure (XAFS) spectra
  description_type: abstract
  lang: en

# 件名（複数記述可）
subjects:
  - subject: Alloy
  - subject: BL14B2
  - subject: Cr K-edge
  - subject: HAVAR
  - subject: SPring-8
  - subject: Si(111)
  - subject: XAFS
  - subject: collection - MDR XAFS DB

# 作成者（複数記述可、入力必須）
creators:
- name: Industrial Application and Partnership Division
  role: contact_person
  ror: https://ror.org/026v1ze26
```

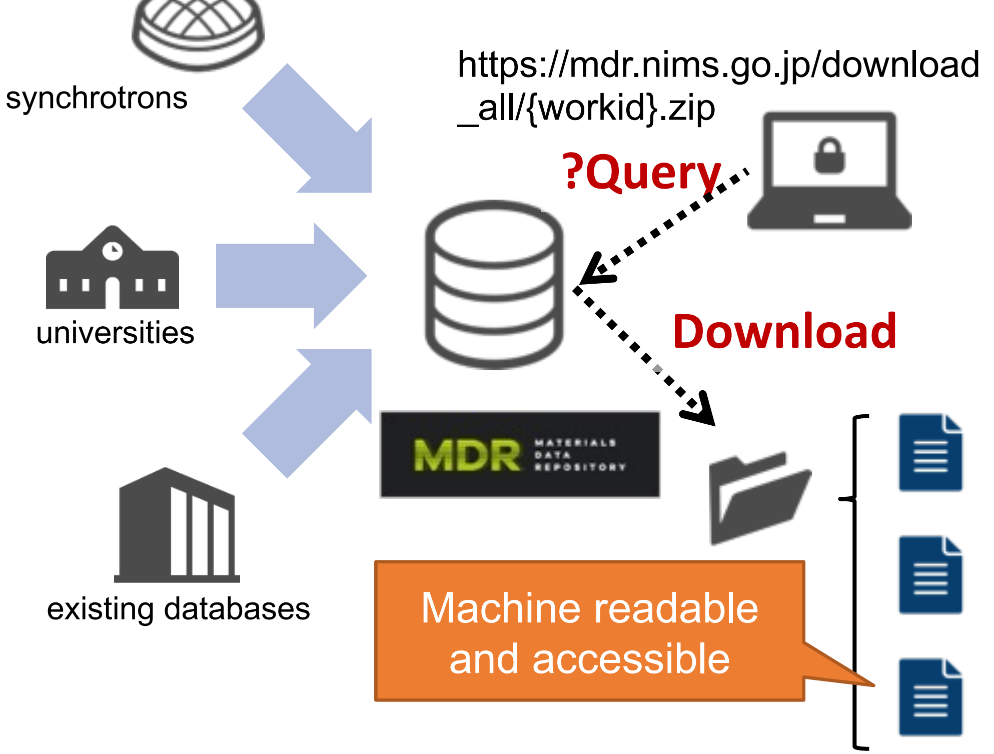
- **Single-layer**
  - no multi-level nesting
- **YAML**
  - for ease of input and readability
- **Deliberately simple**
  - centered around bibliographic metadata for focus on repository use
  - lightweight support for scientific metadata
- Some of the defined fields:
  - Description, Subjects (Keywords), Creator, Rights statement…
  - Instrument, Specimen, Experimental method, Processing, Features…

# **MDR XAFS DB** (X-ray absorption fine structure database)

M. Ishii et al., “Integration of X-ray absorption fine structure databases for data-driven materials science”, *STAM Methods* **3** (2023)

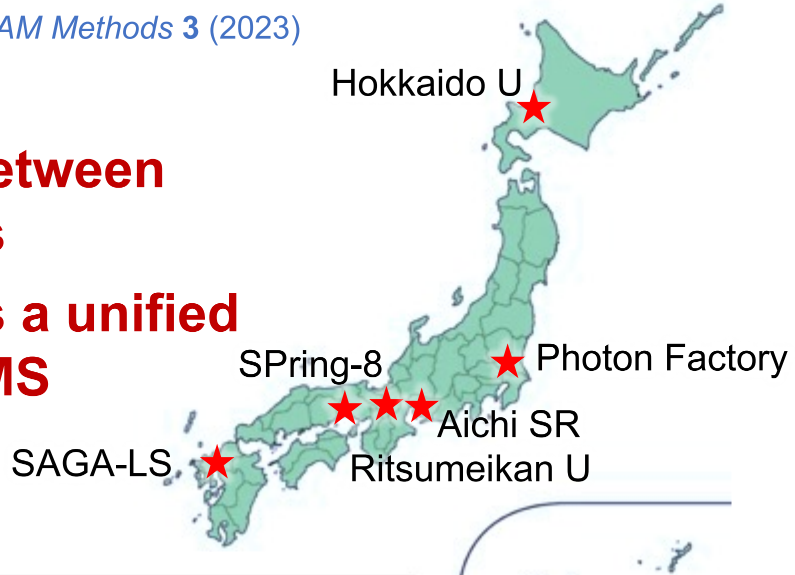
https://doi.org/10.48505/nims.1447

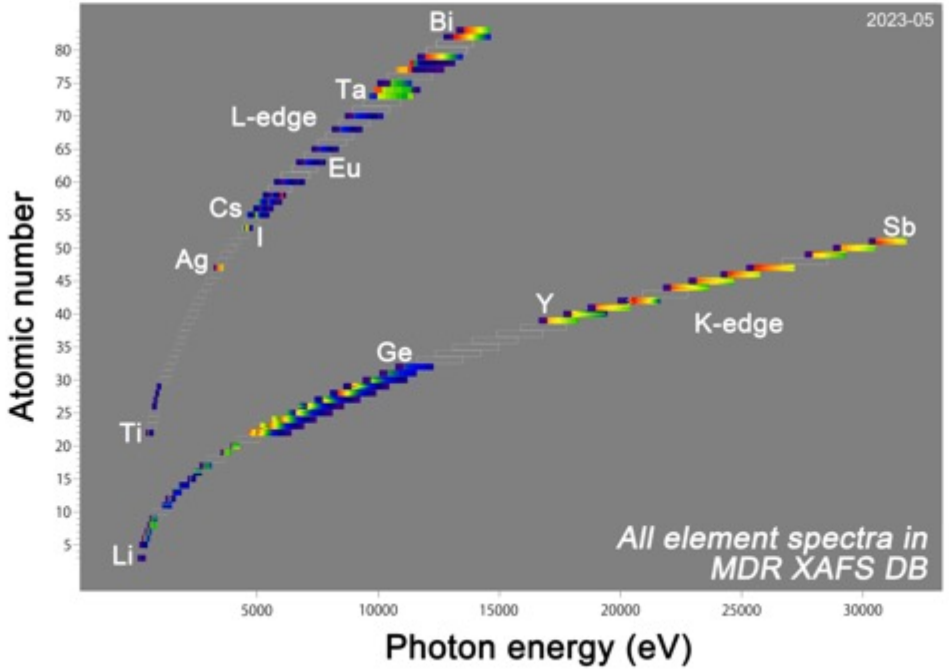
**XAFS spectra**


synchrotrons
universities
existing databases
https://mdr.nims.go.jp/download_all/{workid}.zip
?Query
Download
MDR MATERIALS DATA REPOSITORY

Machine readable and accessible

- **Collaboration between 6 data providers**
- **Consolidated as a unified database by NIMS**

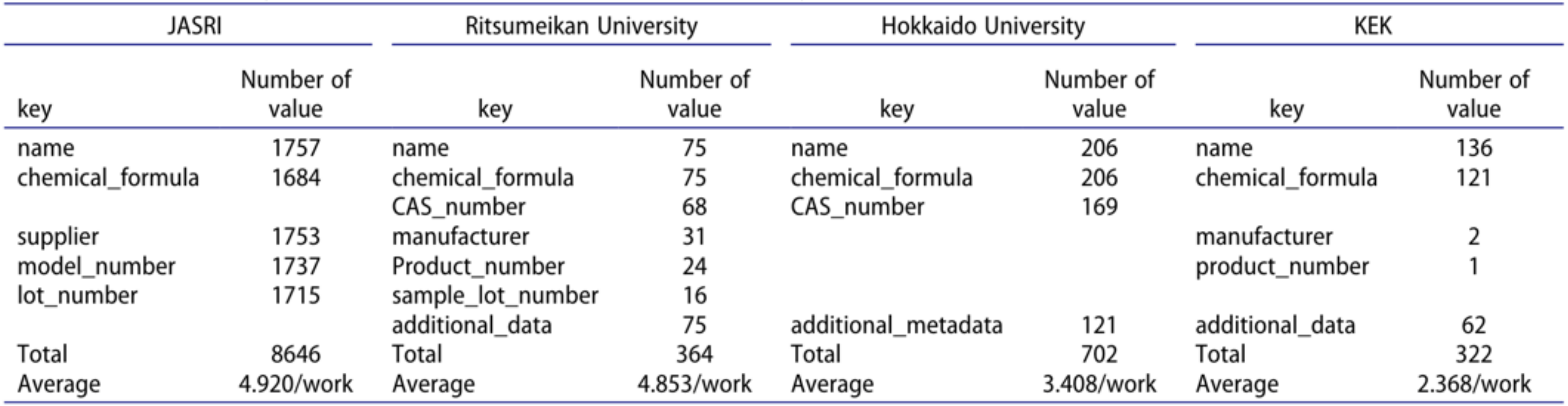




*All element spectra in MDR XAFS DB*

# Metadata alignment within the community

**Table 2.** Metadata keys related to samples and the number of keys. (from the *STAM Methods* paper)



| JASRI | | Ritsumeikan University | | Hokkaido University | | KEK | |
|---|---|---|---|---|---|---|---|
| key | Number of value | key | Number of value | key | Number of value | key | Number of value |
| name | 1757 | name | 75 | name | 206 | name | 136 |
| chemical_formula | 1684 | chemical_formula | 75 | chemical_formula | 206 | chemical_formula | 121 |
| | | CAS_number | 68 | CAS_number | 169 | | |
| supplier | 1753 | manufacturer | 31 | | | manufacturer | 2 |
| model_number | 1737 | Product_number | 24 | | | product_number | 1 |
| lot_number | 1715 | sample_lot_number | 16 | | | | |
| | | additional_data | 75 | additional_metadata | 121 | additional_data | 62 |
| Total | 8646 | Total | 364 | Total | 702 | Total | 322 |
| Average | 4.920/work | Average | 4.853/work | Average | 3.408/work | Average | 2.368/work |

Common info among participating data providers:

- **Metadata according to MDR Schema**
  - keywords for querying within MDR
- **Primary parameters CSV**
- **Structured experimental metadata YAML**

schema

csv

yaml

main data

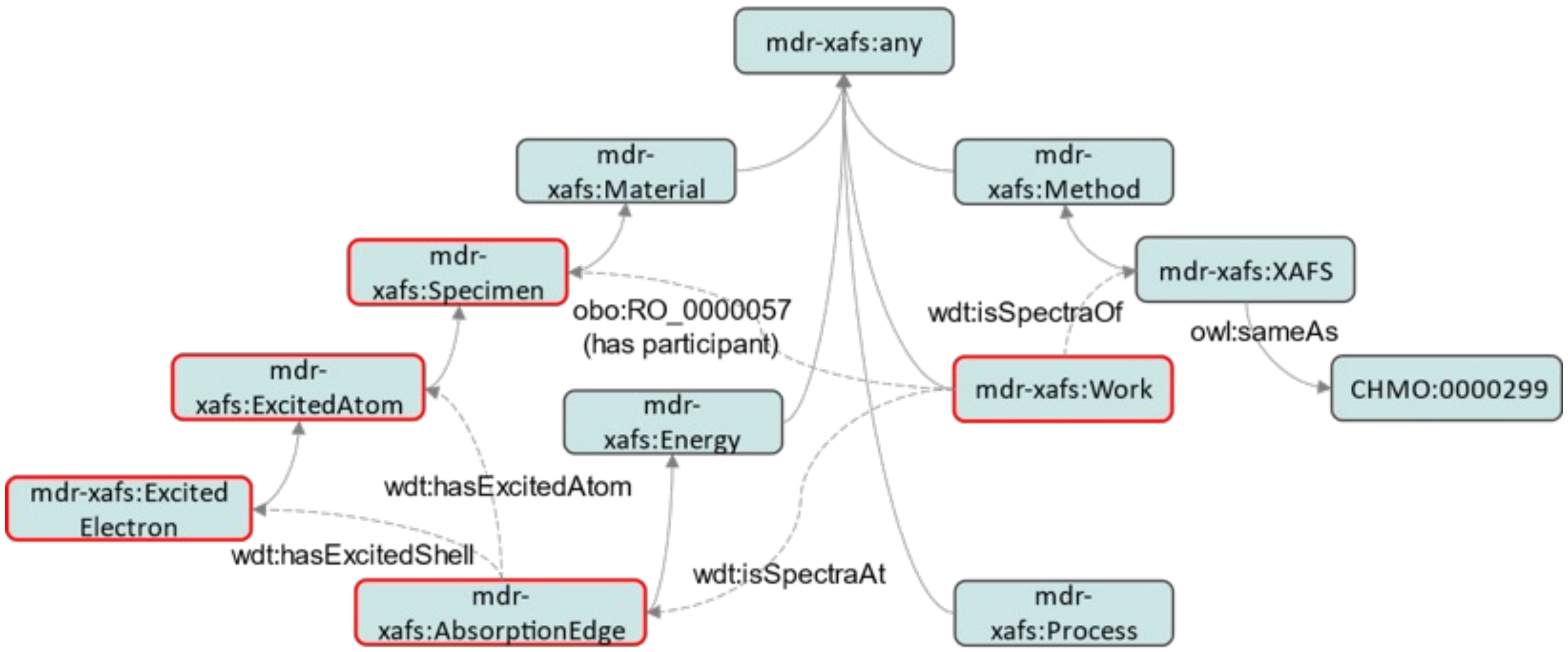
M. Ishii et al., "Integration of X-ray absorption fine structure databases for data-driven materials science", *STAM Methods* **3** (2023)

# RDF for MDR XAFS DB

In addition to the metadata and files in the repository system itself, MDR XAFS DB defines its own ontology.

See https://dice.nims.go.jp/ontology/about.html by M. Ishii (Docs and Turtle available)

MDR XAFS DB Project Ontology



*Side note:*

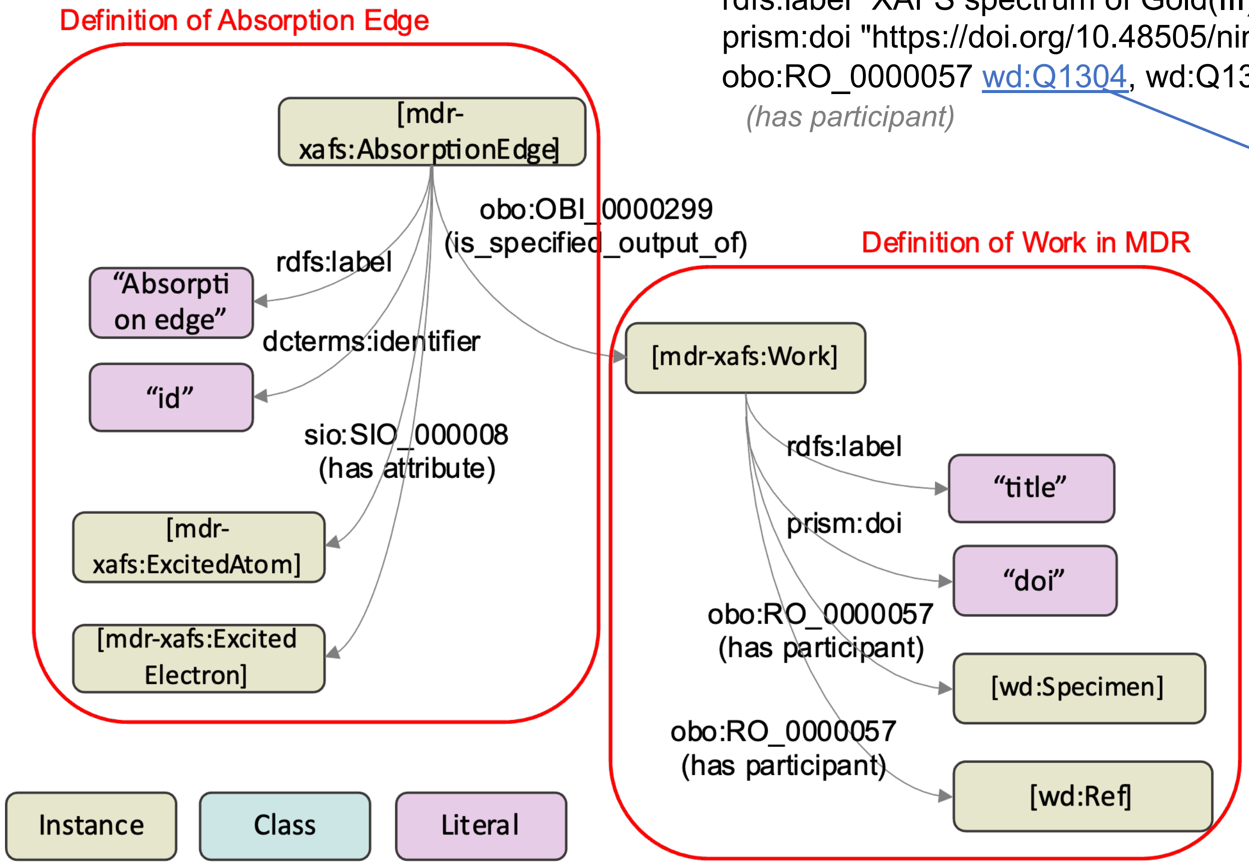
MDR runs on a customized version of Samvera Hyrax software, which internally stores all metadata as RDF.

However, customizing its RDF requires rewrite, rebuild, and restart of the whole repository software. Not suited for user-side RDF like this.

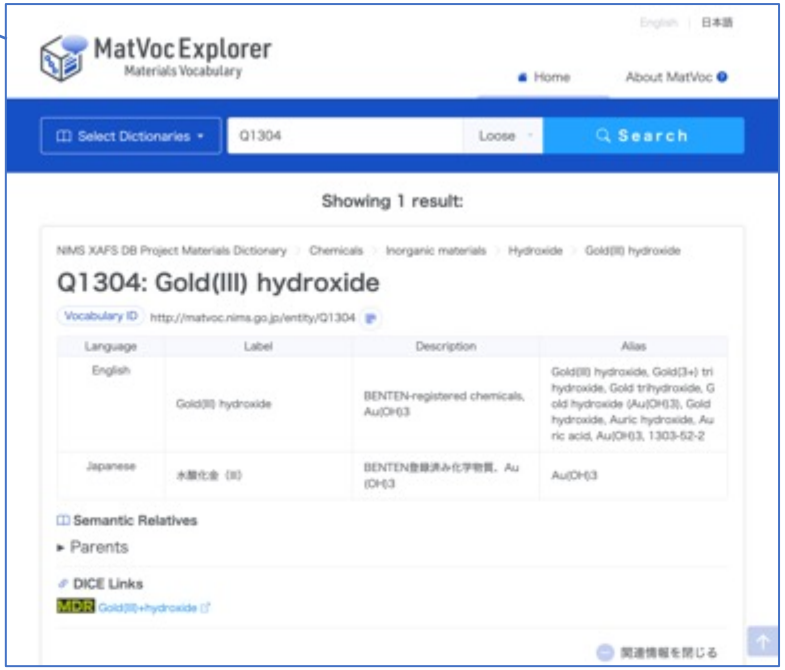
MDR XAFS DB's RDF lives outside the MDR system.

# RDF for MDR XAFS DB

```
@prefix mdr-xafs: <http://dice.nims.go.jp/ontology/mdr-xafs-ont/Schema#>
@prefix obo: <http://purl.obolibrary.org/obo/>
@prefix prism: <http://prismstandard.org/namespaces/1.2/basic/>
@prefix wd: <http://matvoc.nims.go.jp/entity/>
<http://dice.nims.go.jp/ontology/mdr-ont#8a02714e-46a7-4fdc-95a5-1acb18338d7d> a mdr-xafs:Work ;
   rdfs:seeAlso <https://mdr.nims.go.jp/concern/datasets/h128nh653> ;
   rdfs:label "XAFS spectrum of Gold(III) hydroxide"@en ;
   prism:doi "https://doi.org/10.48505/nims.1602"^^xsd:string ;
   obo:RO_0000057 wd:Q1304, wd:Q1308 .
     (has participant)
```



DICE's vocabulary service **MatVoc** (beta ver.)



Showing 1 result:

NIMS XAFS DB Project Materials Dictionary › Chemicals › Inorganic materials › Hydroxide › Gold(III) hydroxide

Q1304: Gold(III) hydroxide

Vocabulary ID http://matvoc.nims.go.jp/entity/Q1304

| Language | Label | Description | Alias |
|---|---|---|---|
| English | Gold(III) hydroxide | BENTEN-registered chemicals, Au(OH)3 | Gold(III) hydroxide, Gold(3+) trihydroxide, Gold trihydroxide, Gold hydroxide (Au(OH)3), Gold hydroxide, Auric hydroxide, Auric acid, Au(OH)3, 1303-52-2 |
| Japanese | 水酸化金（III） | BENTEN登録済み化学物質、Au(OH)3 | Au(OH)3 |

Semantic Relatives

▸ Parents

DICE Links

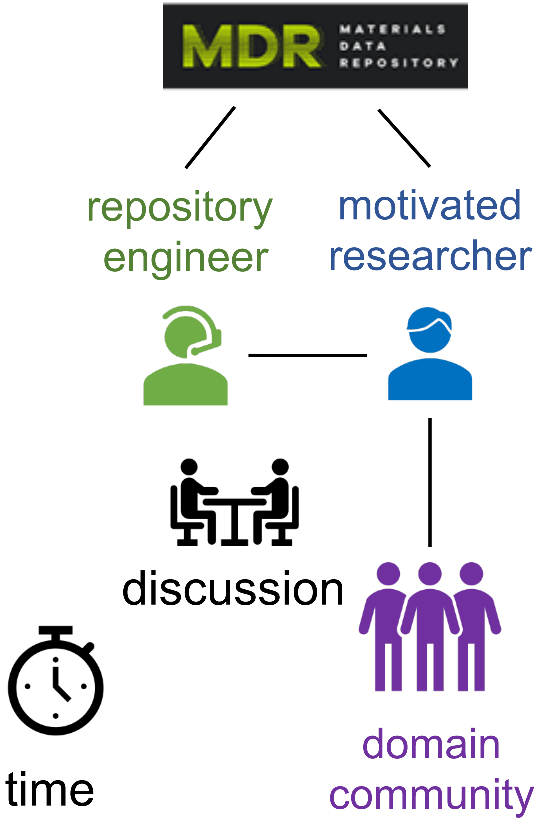
MDR Gold(III)+hydroxide

関連情報を閉じる

https://matvoc.nims.go.jp/explore/

# That was (hopefully) a nice example, but…

## Setting up MDR XAFS DB required all of these:



MDR MATERIALS DATA REPOSITORY

repository engineer

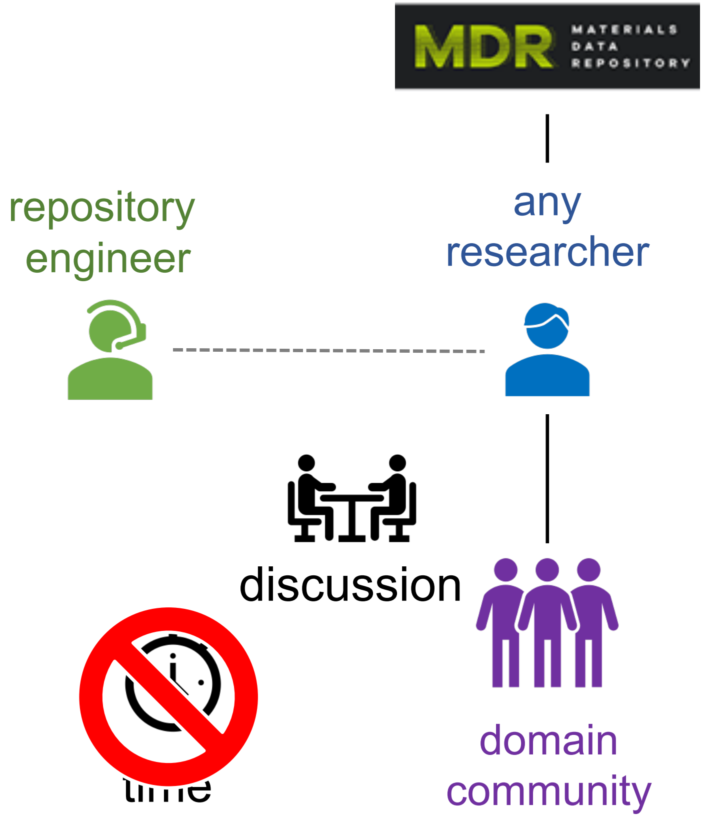
motivated researcher

discussion

time

domain community

## We want to make our next ones work with lighter effort:



MDR MATERIALS DATA REPOSITORY

repository engineer

any researcher

discussion

time

domain community

# Typical situation in labs: Metadata as directories

instruments &
lab PCs

Pre-defined structure:

```
📁 OSC-100      (Instrument ID)
└ 📁 u01234      (User Name/ID)
  └ 📁 ProjABC (Project Name)
    └ 📄 data.csv
```

platform systems

**Mapped to appropriate metadata fields**
Instrument, User, Project...

Implemented as part of our
IoT-assisted data collection system

- ✔ Exact alignment with each lab’s modus operandi
- ✘ Only simple common metadata possible
- ✘ Different mapping for every research project, customization effort required
- ✘ **May not cover all necessary metadata**

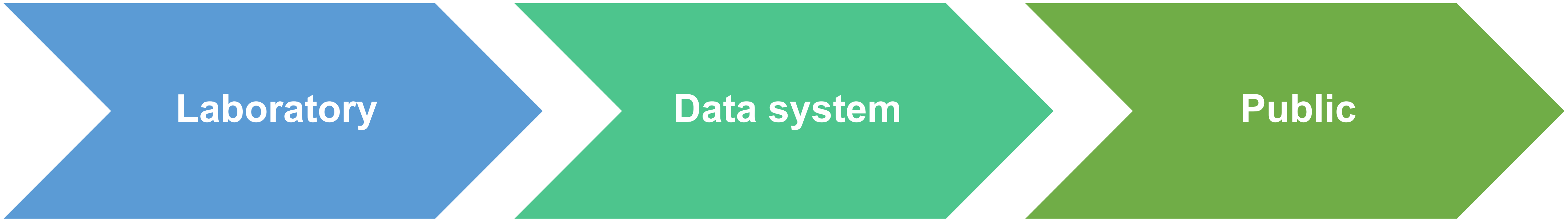
A. Matsuda et al., “Materials metadata: as a custom schema, as directories, or in a data package”, *RDA Virtual Plenary 15* (2020) https://doi.org/10.48505/nims.3031
S. Matsunami et al., “Data Architecture for IoT Data Collection System”, *Digital Practice* **2**(2) 80-89, IPSJ (2021) (in Japanese)

# Towards integrating different types of data

RDE

MDR MATERIALS DATA REPOSITORY

Data journals
Public data repositories



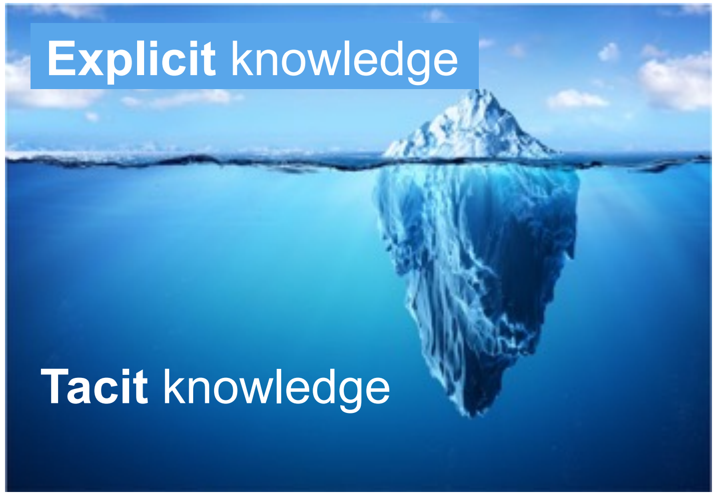
- Raw instrument output

- **What needs to be stored at this stage?**

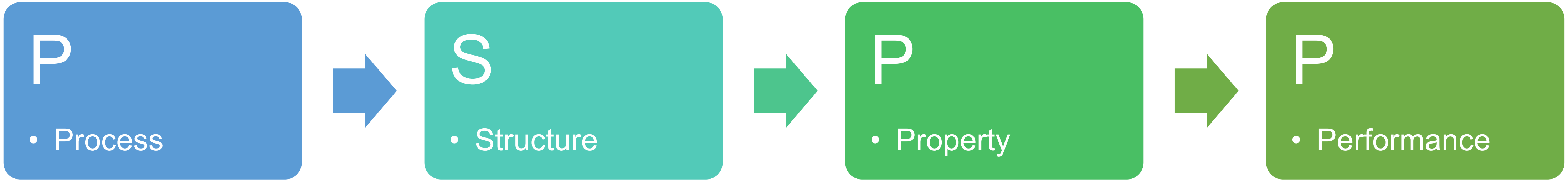
- Bibliographic metadata

Only the researchers themselves can provide this information.

(But frequently, this part can be tacit knowledge!)



# Metadata and the 'PSPP' model



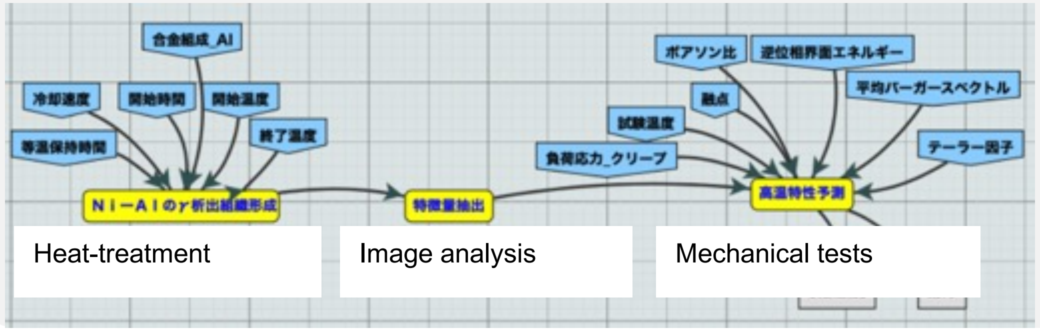
Common sense for some,
New framework for others

**Questions I ask the researchers:**

**What are the key parameters that define your process?**
**What structure are you measuring?**
**What property are you measuring?**
**What performance are you shooting for?**

**Key metadata for that group**

cf. DICE's data integration system



MInt Materials Integration by Network Technology

- In-silico link among process, structure, property, performance
- Each module is aware of what it takes as input and output

# Summary

- The iterations of our metadata models coincides with how our data platform has been evolving.
  - First, we tried a single mega-schema to cover all use cases.
  - A more simplified schema centered around bibliographic metadata was adopted for our data repository.
- A cross-institutional XAFS database built on MDR used a combination of MDR's simplified metadata schema, community-defined YAML, and RDF.
- Spreading awareness of metadata management is an ongoing effort. We've begun to ask researchers about their key parameters using the PSPP model and store them in our system.
  - We hope this leads to better integration of heterogeneous datasets, and lead to accelerated materials research and engineering.

*Thank you for your attention!*

Contact: matsuda.asahiko@nims.go.jp