NIMS's Materials Data Platform DICE and its future direction

Asahiko Matsuda, Takuya Kadohira, Masahiko Demura

https://orcid.org/0000-0001-5989-027X

National Institute for Materials Science (NIMS)

Research & Services Div. of Materials Data & Integrated System (MaDIS)

Re-introducing ourselves:

国立研究開発法人物質・材料研究機構
National Institute for Materials Science

About 1600 people

Tsukuba Science City
(~1hr from Tokyo)

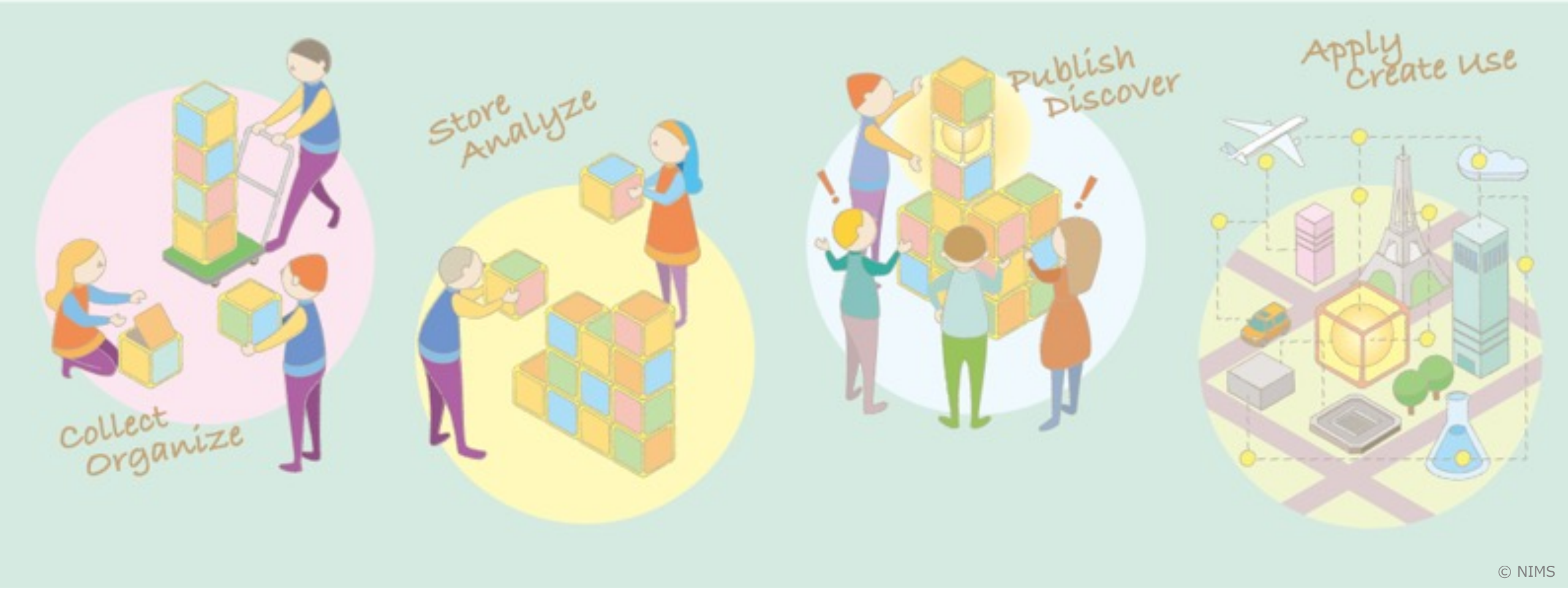
Seven other
research divisions

Research and Services Division of Materials Data and Integrated System

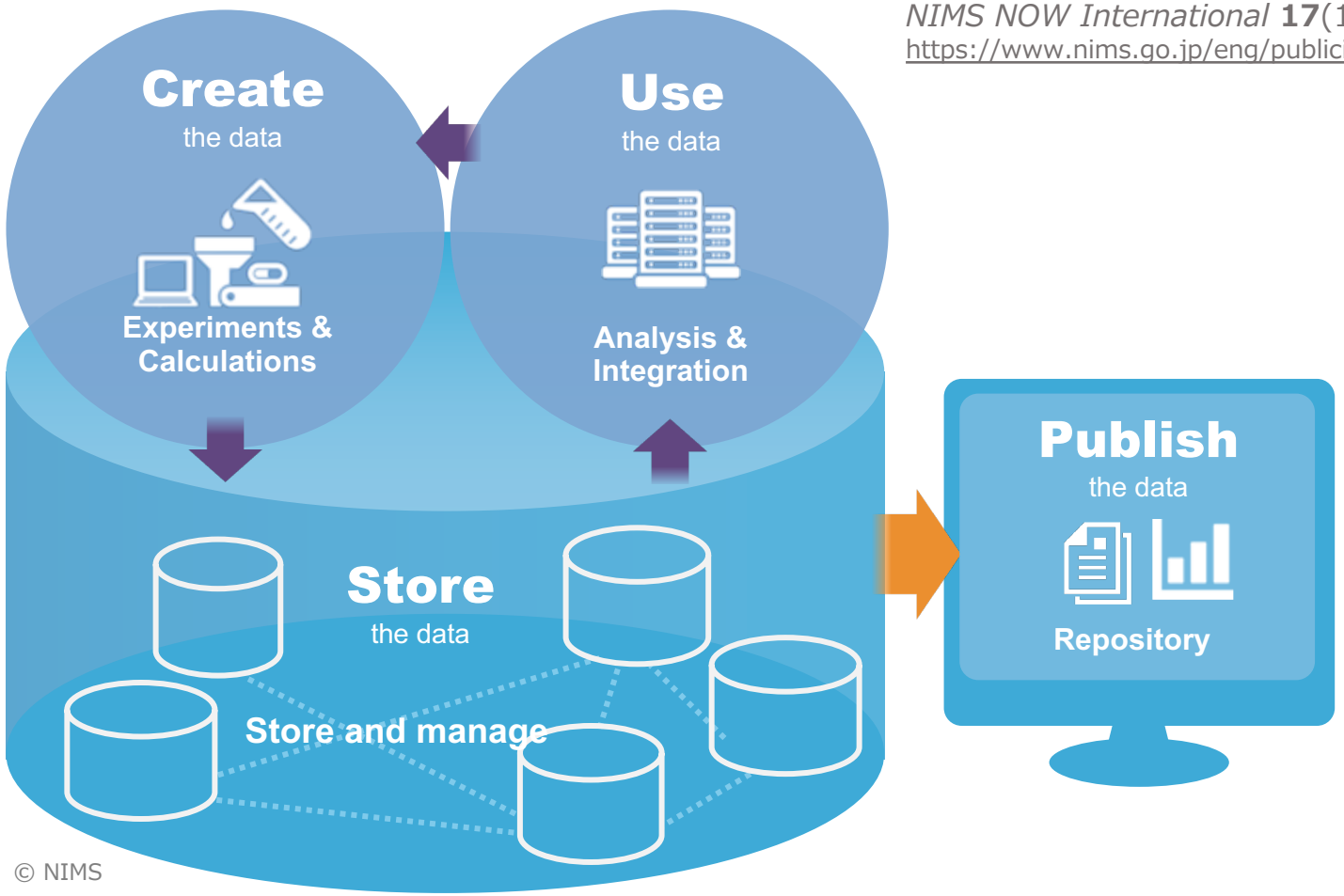
- **Five data-driven research groups**
- **SIP-MI** (Materials Integration) **Laboratory**
- **Materials Data Platform Center**

© NIMS

Materials data platform
DICE



Four main functionalities of DICE



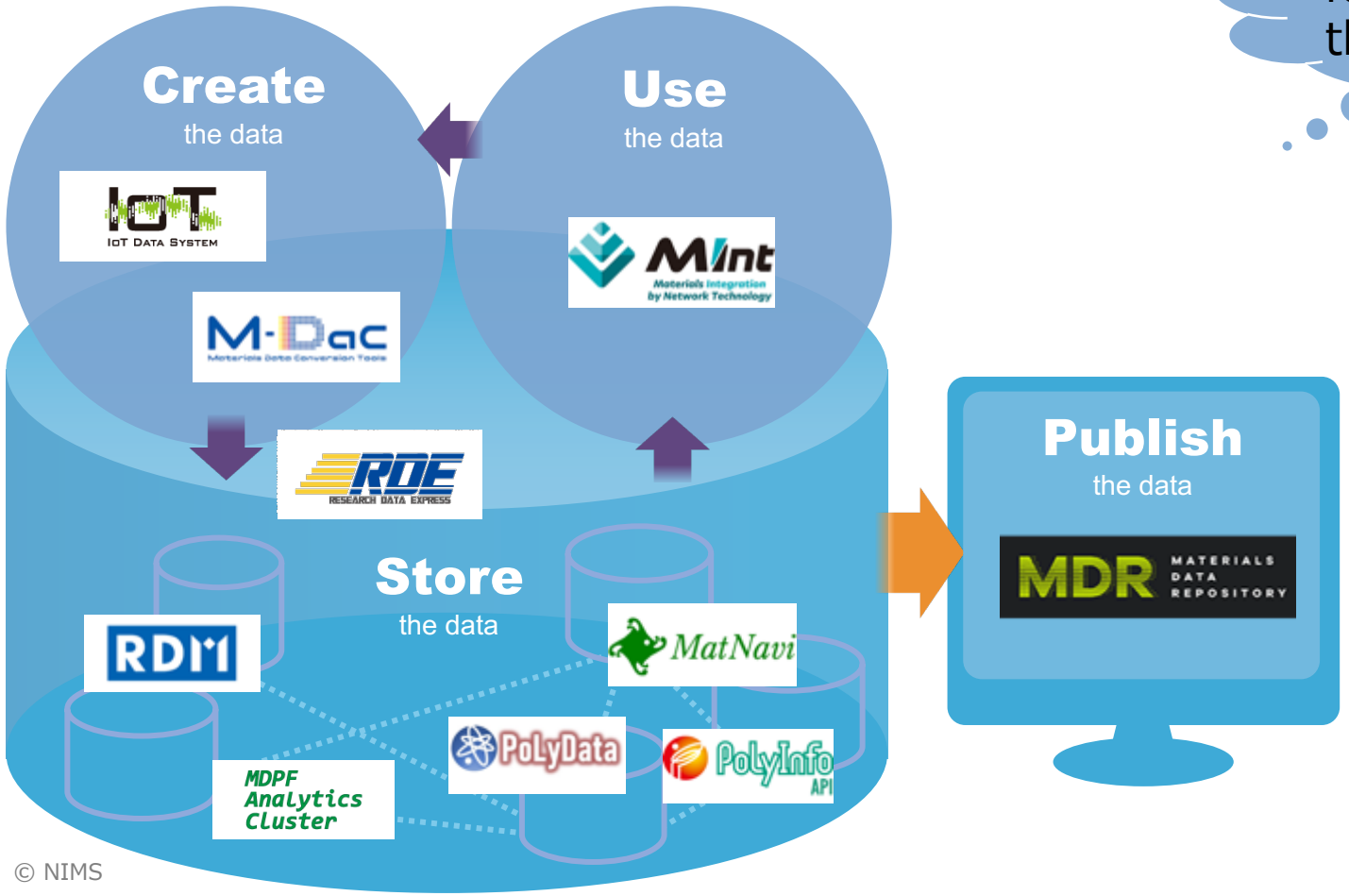
NIMS NOW International **17**(1) 2019
https://www.nims.go.jp/eng/publicity/nimsnow/2019/201901.html


M-Dac
FigResourceMiner
TDM
PolyInfo API
MatNavi
PoLyData
Labnotes
DICE
RDE
IoT
MDR
MDPF Analytics Cluster
RDM
Mint

DICE 1.0 birds-eye view

Materials data platform DICE

Further development for DICE 2.0, to spread this effort nationwide



Creating/Converting data

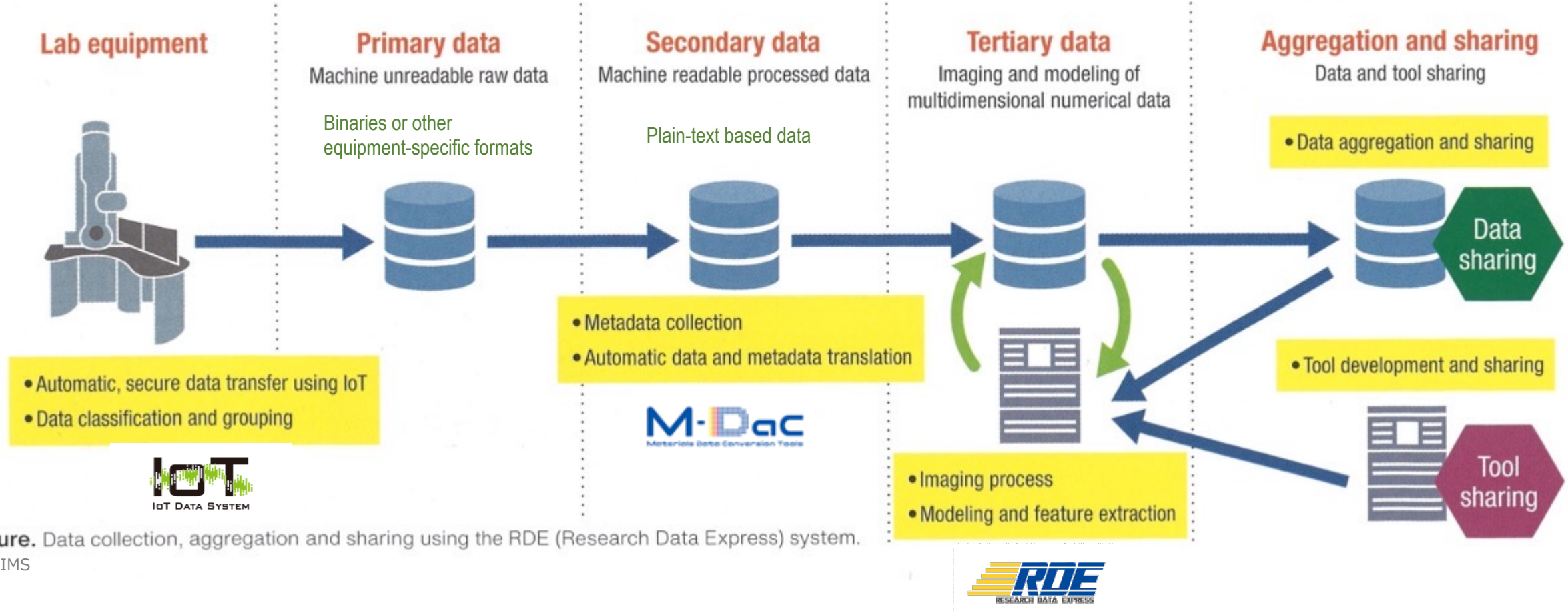
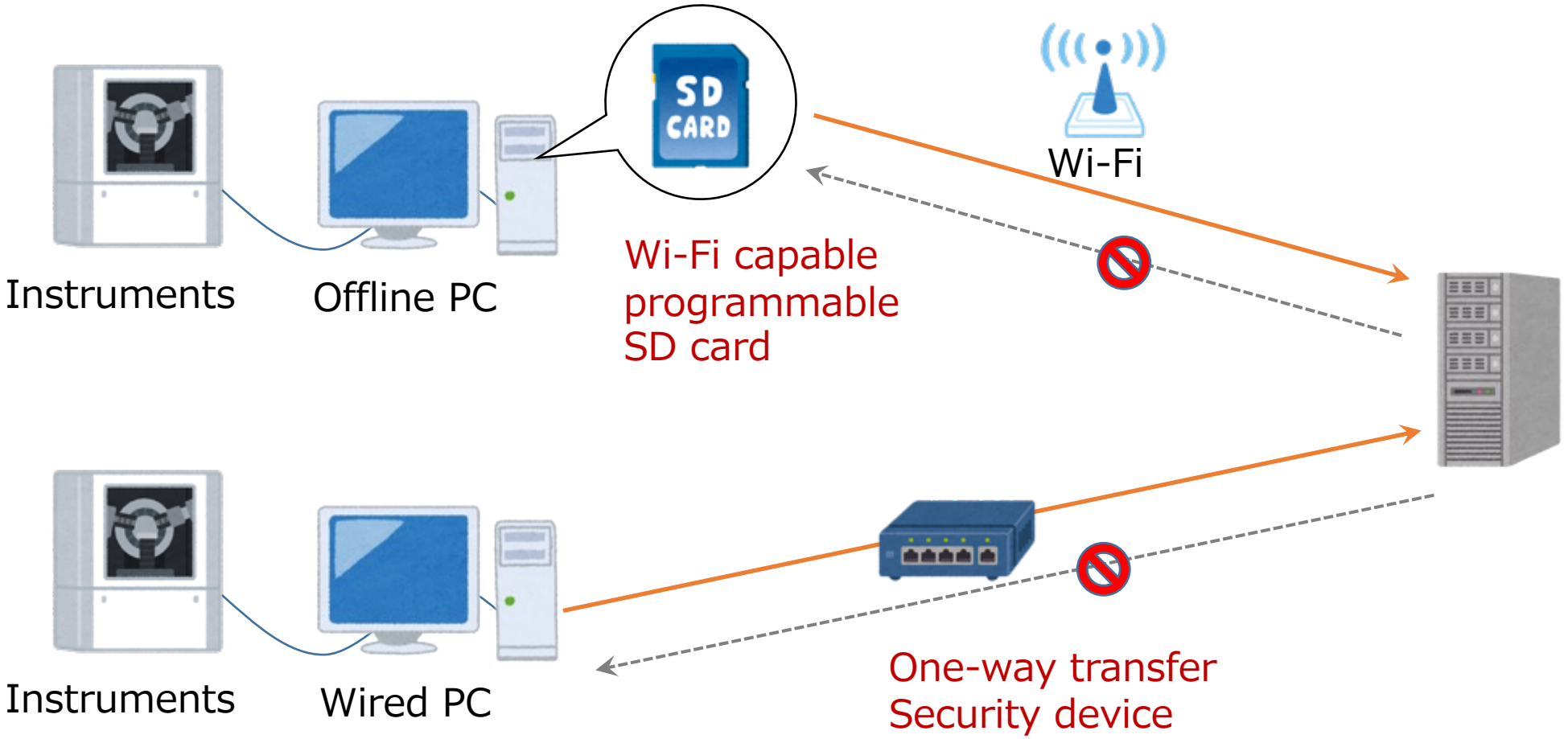


Figure. Data collection, aggregation and sharing using the RDE (Research Data Express) system.

NIMS NOW International **19**(5) 2021
https://www.nims.go.jp/eng/publicity/nimsnow/2021/202105.html

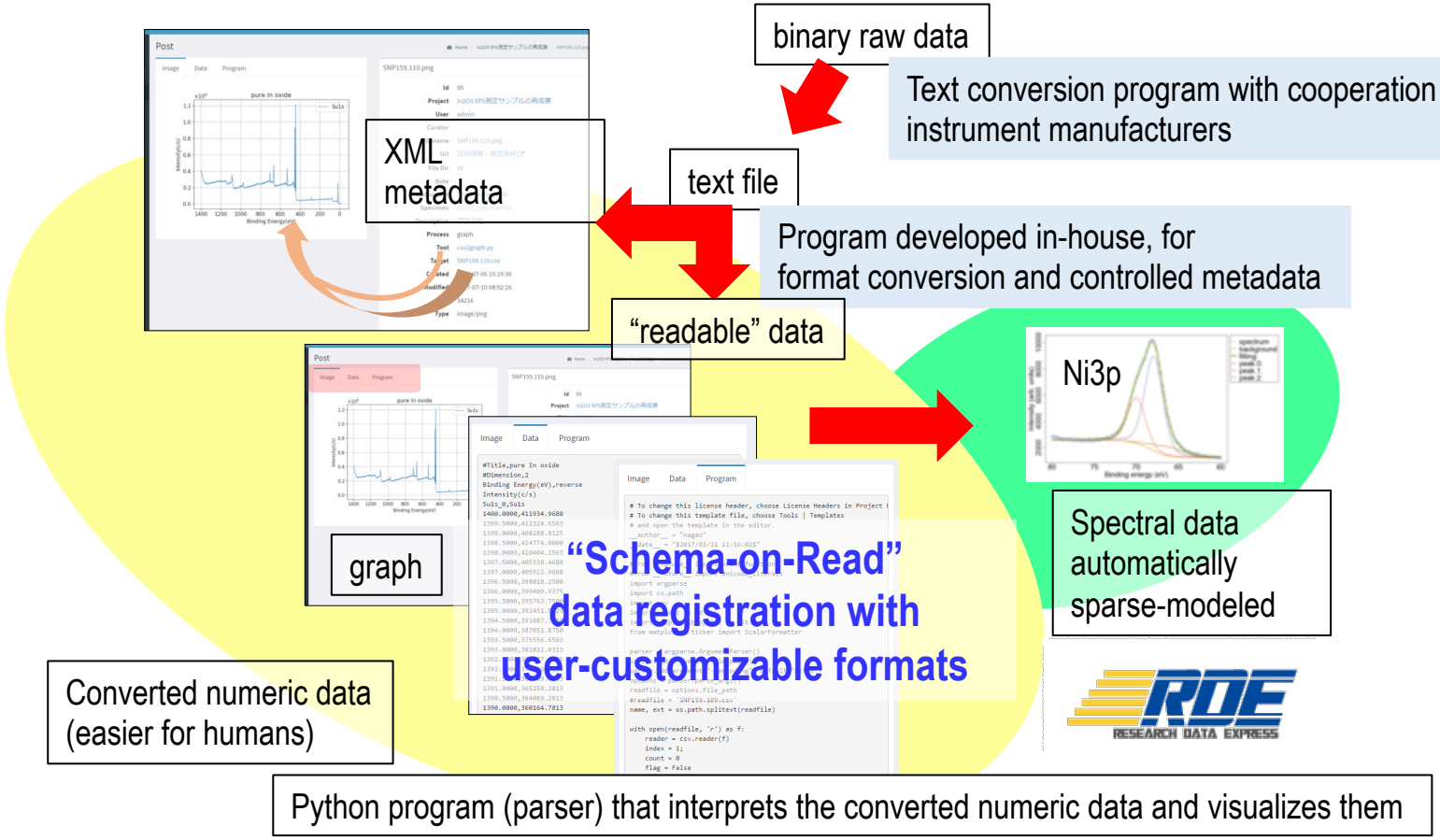
Automatic data collection from labs



M.Matsunami, A.Matsuda, T.Chikyow, Y.Harada, H.Yoshikawa, *IPSJ Digital Practice* **2**(2) 80-89 (2021)
https://doi.org/10.48505/nims.3050 (full text Japanese)

System and tools for measurement data collection and automatic conversion

M. Suzuki et al., *J. Vac. Sci. Technol. A* **38** 023204 (2020) https://doi.org/10.1116/1.5128408



Materials Data Conversion tools

- **Converts instrument-specific data format into readable formats**
  - Plain-text formats and graphs
  - Converters were developed through cooperation of instrument manufacturers
    - XPS (survey, narrow, depth) by ULVAC-PHI
    - XRD, 2D-XRD by Rigaku
    - AES by JEOL
    - AES (survey, depth) by ULVAC-PHI
- **github.com/nims-dpfc**
  - Runs on Jupyter Notebook
- **Wouldn't it be nice if…**
  - **we start doing this more?** Convert all of your data in a stable, readable format
  - **we can get more techniques from more manufacturers on board?**

Storing/Searching data: Metadata

- Four different styles, even within the Center (not mutually exclusive)
  - Multi-layer nested grand metadata schema
  - Simplified metadata with a focus on data repository
  - Metadata as directory names
  - Metadata as plain-text files

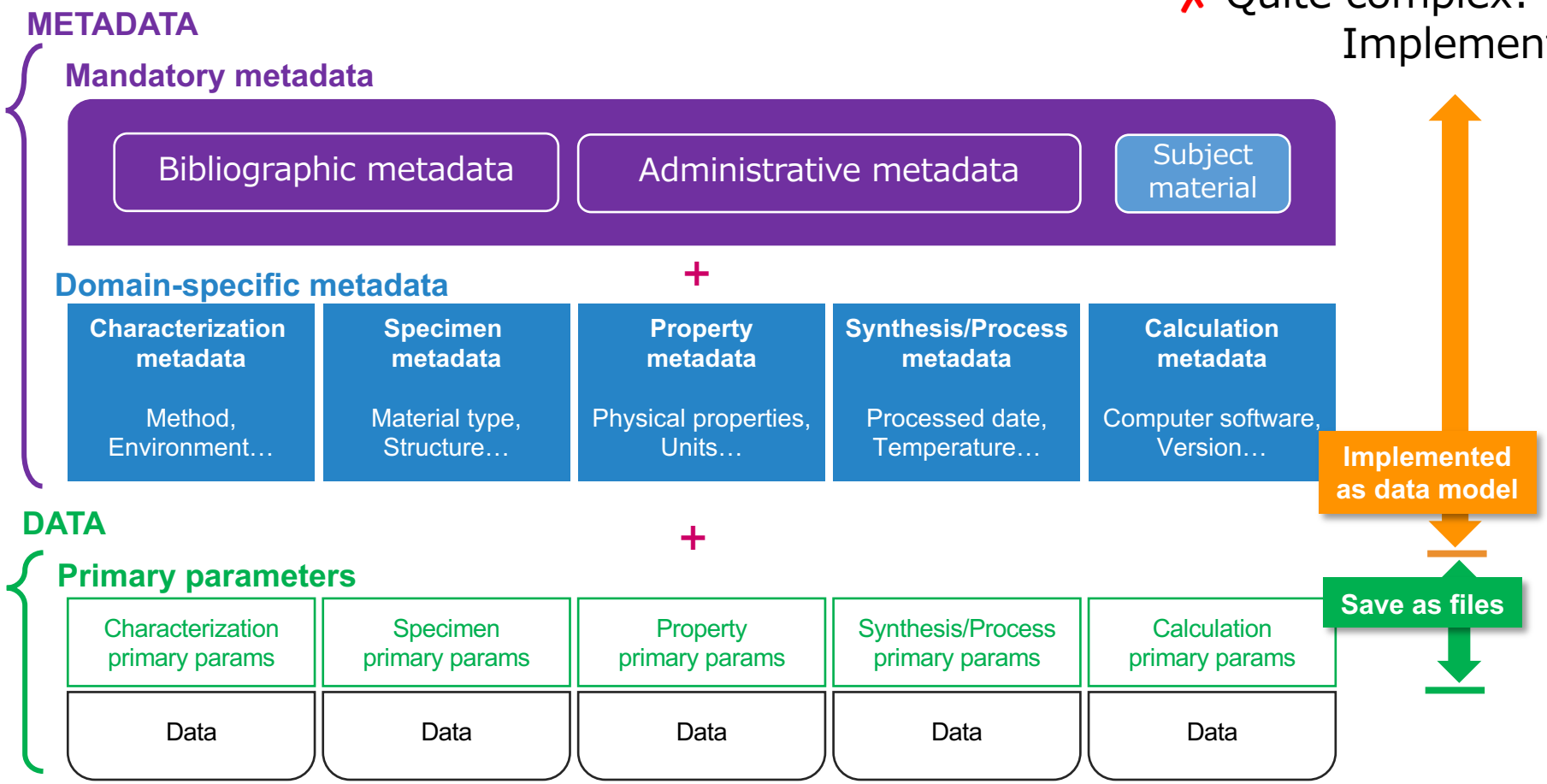
Matsuda et al. “Materials metadata: as a custom schema, as directories, or in a data package” RDA Virtual Plenary 15 (2020) https://doi.org/10.48505/nims.3031

DICE Common Message Format 1.0

S. Kikuchi, T. Kadohira, M. Suzuki, H. Naito, *IEICE Tech. Rep.* **119** SC2019-2
Full text (Japanese): https://mdr.nims.go.jp/concern/publications/m039k565z
Schema: https://doi.org/10.48505/nims.3240

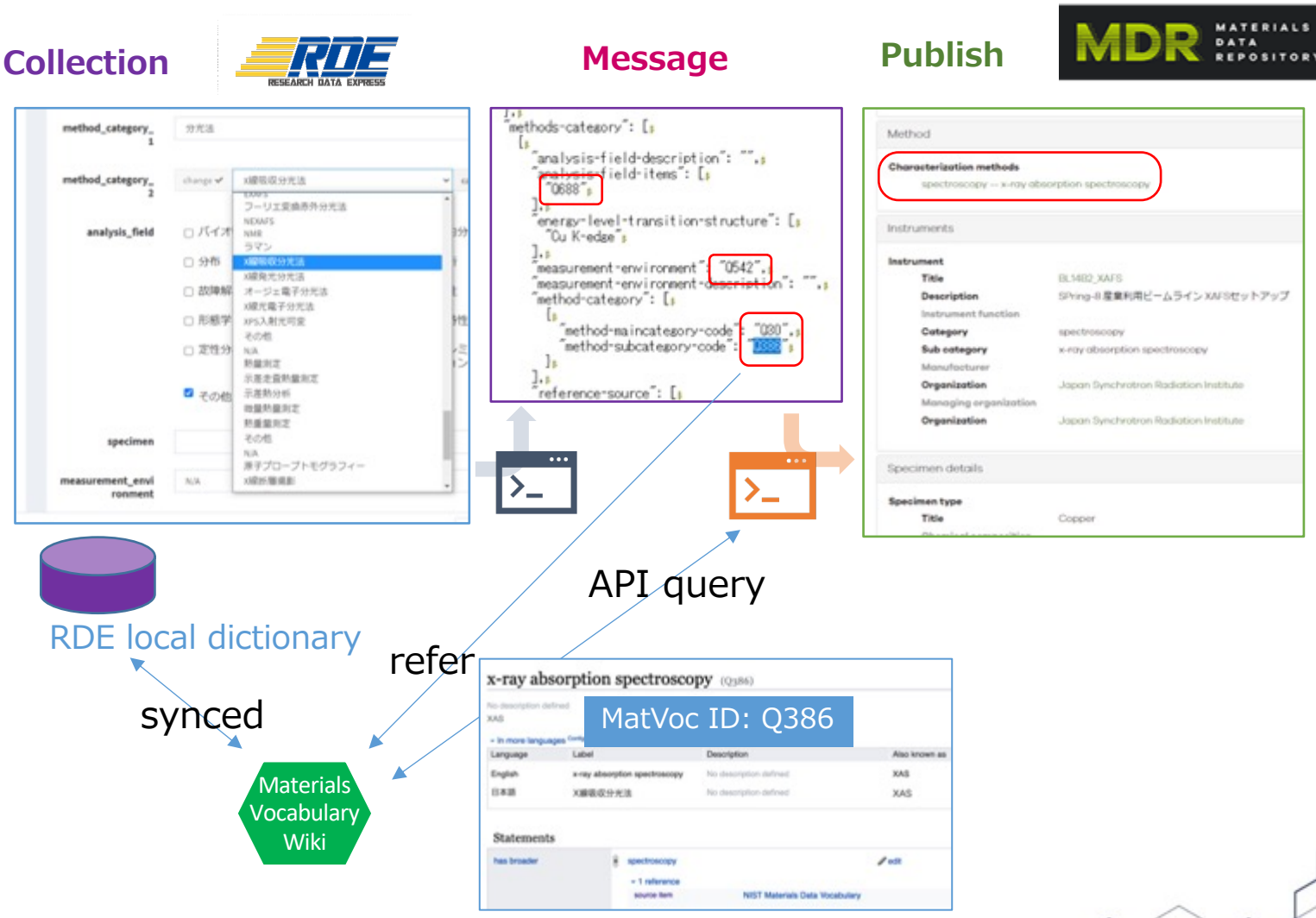
- JSON Schema
- System-to-system communication

✓ Highly descriptive, high reusability.
✗ Quite complex: Not human-readable. Implementation difficulties.



Vocabulary-based metadata transfer between systems

M. Ishii, H. Nagao, A. Matsuda, K. Tanabe, H. Yoshikawa, *23rd XAFS Forum* (2020) (Japanese)



MDR Schema 2.0

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

Materials data platform DICE

(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
  - ✓ no brackets!
  - ✓ in-file comments!
- Deliberately simple (relatively)
  - focus on repository use
- Support for common IDs
- Can be validated with “Yamale”

Metadata as directories

instruments &
lab PCs

Pre-defined structure:

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

Implemented as part of our
IoT-assisted data collection system.

platform systems

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

- ✓ Easy to use in labs.
- ✓ Network not required.
- ✗ Prone to human errors.
- ✗ Only simple common metadata.
- ✗ Different mapping for every research project.

Metadata in a file / data package

root directory: /

RO-Crate

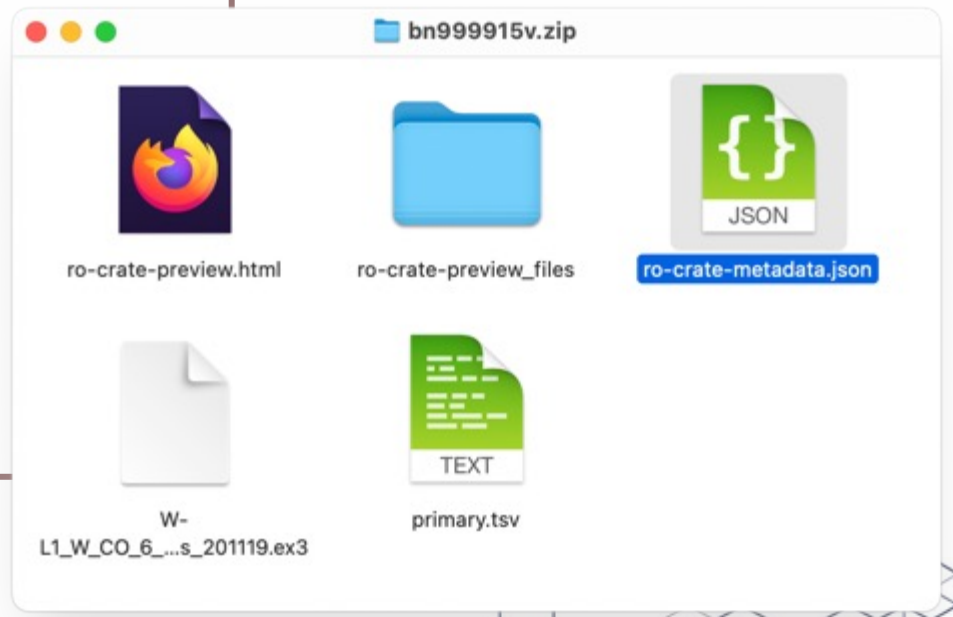
https://www.researchobject.org/ro-crate/

ro-crate-metadata.json
JSON-LD mandatory metadata

ro-crate-preview.html
HTML preview for humans (optional)

ro-crate-preview_files/
Files for HTML

Payload data



K.Tanabe & A.Matsuda, *Joho Chishiki Gakkaishi* **31**(4) 497-502 (2021)
https://doi.org/10.2964/jsik_2021_067 (full text Japanese)

JSON-LD metadata for RO-Crate package

Materials data platform DICE

MDR MATERIALS DATA REPOSITORY (test environment)

Package info

Dataset description

Scientific metadata

Material/Substance

Data

Additional metadata file reference

Bioschemas

```
{
  "@context": ["https://w3id.org/ro/crate/1.1/context", {"bio": "http://schema.org"}],
  "@graph": [
    {
      "@type": "CreativeWork",
      "@id": "ro-crate-metadata.json",
      "conformsTo": {"@id": "https://w3id.org/ro/crate/1.1"},
      "about": {"@id": "./"}
    },
    {
      "@id": "./",
      "@type": "Dataset",
      "url": "https://mdr.nims.go.jp/concern/datasets/bn99991
      "identifier": "https://doi.org/10.48505/nims.2563",
      "name": "XAFS spectrum of Tungsten carbonyl",
      "datePublished": "2021",
      "description": "This dataset consists of X-ray absorption fine structure (XAFS) spectra at W
      L21-edge of Tungsten carbonyl measured at SPring-8 BL14B2",
      "hasPart": [
        {"@id": "W-L1_W(CO)6_Si311_50ms_201119.ex3"},
        {"@id": "primary.tsv"}
      ],
      "material": [ {"@id": "#material-0"} ]
    },
    {
      "@id": "#material-0",
      "@type": "Material",
      "name": "Tungsten carbonyl",
      "description": "Standard Sample",
      "bio:ChemicalSubstance": {
        "bio:chemicalComposition": "W(CO)6",
        "identifier": "https://matvoc.nims.go.jp/wiki/Item:Q1765"
      }
    },
    {
      "@id": "W-L1_W(CO)6_Si311_50ms_201119.ex3",
      "@type": "File",
      "encodingFormat": "text/plain",
      "url": "https://mdr.nims.go.jp/downloads/73666690b",
      "mainEntityOfPage": {"@id": "https://mdr.nims.go.jp/concern/parent/bn999915v/file_sets/73666690b"}
    },
    {
      "@id": "primary.tsv",
      "@type": "File",
      "encodingFormat": "text/tab-separated-values",
      "url": "https://mdr.nims.go.jp/downloads/k930c032p",
      "mainEntityOfPage": {"@id": "https://mdr.nims.go.jp/concern/parent/bn999915v/file_sets/k930c032p"}
    }
  ]
}
```

K.Tanabe & A.Matsuda, *Joho Chishiki Gakkaishi* **31**(4) 497-502 (2021)
https://doi.org/10.2964/jsik_2021_067 (full text Japanese)

Metadata in a file: Some notes

- RO-Crate's main metadata file is JSON-LD
  - … but the package can include **any other file as metadata**
  - e.g., aforementioned MDR Schema YAML
- A package doesn't have to be RO-Crate
  - … but at the least, it's nice to base it on **plain text**

✓ Files are relatively stable, separated from a system.

✗ Will not be easily searchable on a system.

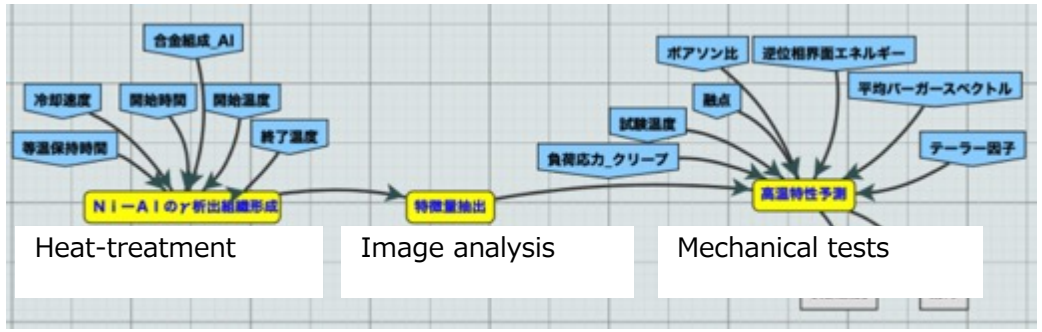
What are the essential metadata we want across various data?

(I'm far from the first person mentioning this question, but this is still a key issue from my view!)

Using the data

Demura, RDA P18

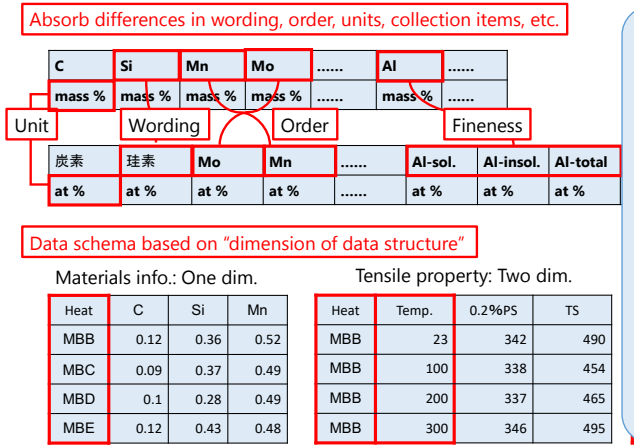
Materials Integration by Network Technology



- **In-silico link among process, structure, property, performance**
- **Integrate experiments and computations by data driven approaches**

Share well-defined data schema

Example of designed data structure



XSD

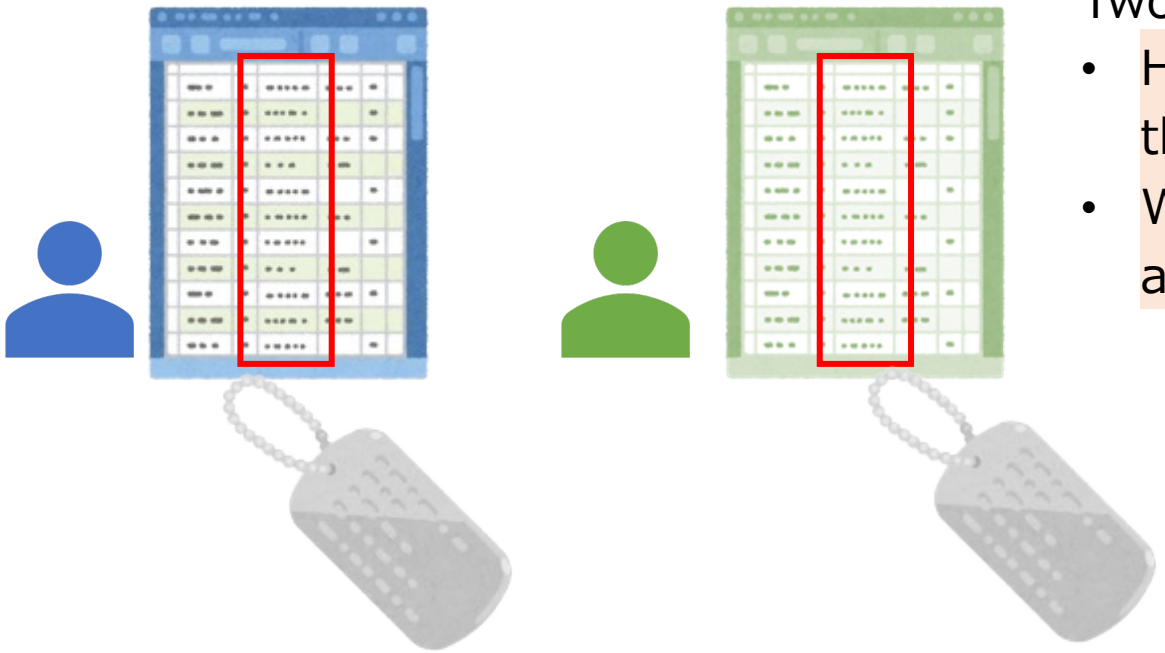
```
<?xml version="1.0"?>
<xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSchema
"
  elementFormDefault="qualified"
  version="1.0">
 <xs:element name="data">
  <xs:complexType>
   <xs:all>
    <xs:element name="material">
     <xs:complexType>
      <xs:all>
```

We've designed data structure for our materials problem, which will be published.

Wouldn't it be in the public interest to **increase the number of data schema**?

Sharing data structures

- Data can be made available and readable
- But there's a barrier in *actually* reading (understanding what they represent) and utilizing data



Two data about the same topic (supposedly)

- How can we be sure that we can mash up these two data?
- What would tell me that these two columns are directly comparable?

Attach data structure descriptions to all data and share them somehow? Ideas?

Conclusion

- We built “DICE” as a platform where our researchers can create, store, use, and publish their data (the four actions of the platform), supported through various subsystems
- There are many areas where we all have benefited from and will benefit from continued discussions and cooperation:
  - Making it easier to store data in a stable manner
  - Metadata, how and what
  - Describing and sharing data structure
  - *and probably more…*

Further reading

- *NIMS NOW International*
  - MaDIS special issue (2019) https://www.nims.go.jp/eng/publicity/nimsnow/2019/201901.html
  - Materials DX special issue (2021) https://www.nims.go.jp/eng/publicity/nimsnow/2021/202105.html
- M. Tanifuji, A. Matsuda, H. Yoshikawa: Materials Data Platform - a FAIR System for Data-Driven Materials Science. Proc. 2019 8th Int. Congress on Advanced Applied Informatics (IIAI-AAI). 2019
- T. Takemura, M. Ishii, M. Tanifuji: PoLyInfo RDF: A Semantically Reinforced Polymer Database for Materials Informatics. CEUR Workshop Proceedings. 2019, 69-72.
- L. Foppiano, L. Romary, M. Ishii, M. Tanifuji: Automatic Identification and Normalisation of Physical Measurements in Scientific Literature. DocEng '19 Proc. ACM Symp. on Document Eng 2019.
- M. Suzuki, H. Nagao, Y. Harada, H. Shinotsuka, K. Watanabe, A. Sasaki, A. Matsuda, K. Kimoto, H. Yoshikawa: Raw-to-Repository Characterization Data Conversion for Repeatable, Replicable, and Reproducible Measurements. J. Vac. Sci. Technol. A **38** [2] 023204 (2020).
- X. Liu, L. Yang, Z. Hou, Bo Da, K. Nagata, H. Yoshikawa, S. Tanuma, Y. Sun, Z. Ding: Machine learning approach for the prediction of electron inelastic mean free paths. Phys. Rev. Mater. **5** [3] (2021).
- H. Oka, A. Yoshizawa, H. Shindo, Y. Matsumoto, M. Ishii: Machine extraction of polymer data from tables using XML versions of scientific articles. Sci. Technol. Adv. Mater. Methods **1**, 12 (2021).
- S. Dieb, K. Amano, K. Tanabe, D. Sato, M. Ishii, M. Tanifuji: Creating Research Topic Map for NIMS SAMURAI Database Using Natural Language Processing Approach. Sci. Technol. Adv. Mater. Methods **1**, 2 (2021).