MDPF

Cross-searching of datasets by linking repository and vocabulary management in materials data platform

Kosuke Tanabe, Asahiko Matsuda, Masashi Ishii

https://orcid.org/0000-0002-9986-7223

National Institute for Materials Science (NIMS), Japan

October 23rd, 2024

DCMI 2024 Toronto Best practices session

NIMS Materials Data Platform “DICE”

- Research data platform for materials science
- Developed and operated by National Institute for Materials Science (NIMS), Japan
- Consists of several research software and services
  - Materials databases
  - Research data sharing service
  - User authentication / authorization

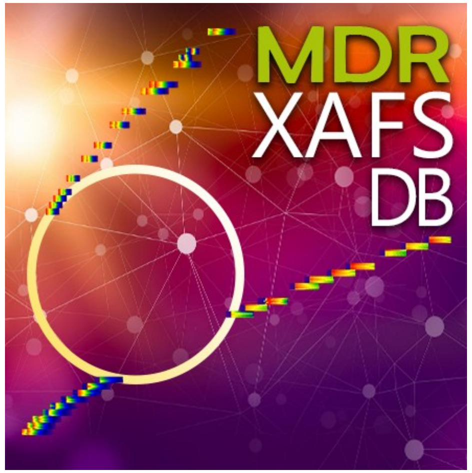
DICE

MatNavi AtomWork Adv. RDE

MDR MATERIALS DATA REPOSITORY MatVoc

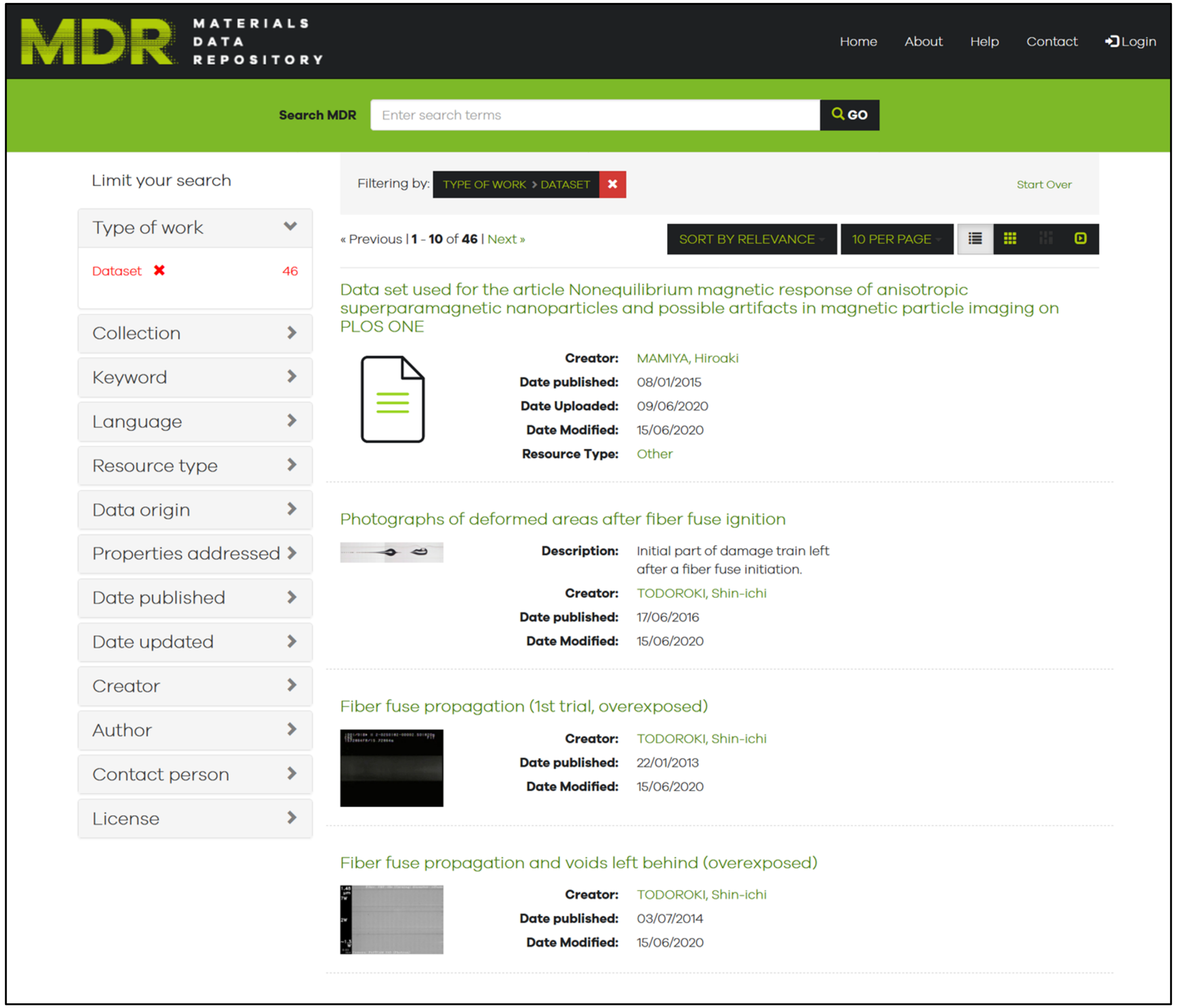
Selected in this presentation

- Materials Data Repository (MDR)
- MDR XAFS DB
- MatVoc / MatVoc Explorer



NIMS Materials Data Repository (MDR)

- Data repository in the DICE
  - https://mdr.nims.go.jp
- released in June 2020
- About 1,000 articles and 12,800 datasets as of October 2024
- provides facet search UI and REST API
- Built on Samvera Hyrax (OSS)
- Project leader: Kosuke Tanabe



MDR MATERIALS DATA REPOSITORY

Home | About | Help | Contact | Login

Search MDR | Enter search terms | GO

Limit your search

Type of work	
Dataset ✖	46
Collection	
Keyword	
Language	
Resource type	
Data origin	
Properties addressed	
Date published	
Date updated	
Creator	
Author	
Contact person	
License	

Filtering by: TYPE OF WORK › DATASET ✖ | Start Over

« Previous | 1 - 10 of 46 | Next » | SORT BY RELEVANCE | 10 PER PAGE

Data set used for the article Nonequilibrium magnetic response of anisotropic superparamagnetic nanoparticles and possible artifacts in magnetic particle imaging on PLOS ONE

Creator: MAMIYA, Hiroaki
Date published: 08/01/2015
Date Uploaded: 09/06/2020
Date Modified: 15/06/2020
Resource Type: Other

Photographs of deformed areas after fiber fuse ignition

Description: Initial part of damage train left after a fiber fuse initiation.
Creator: TODOROKI, Shin-ichi
Date published: 17/06/2016
Date Modified: 15/06/2020

Fiber fuse propagation (1st trial, overexposed)

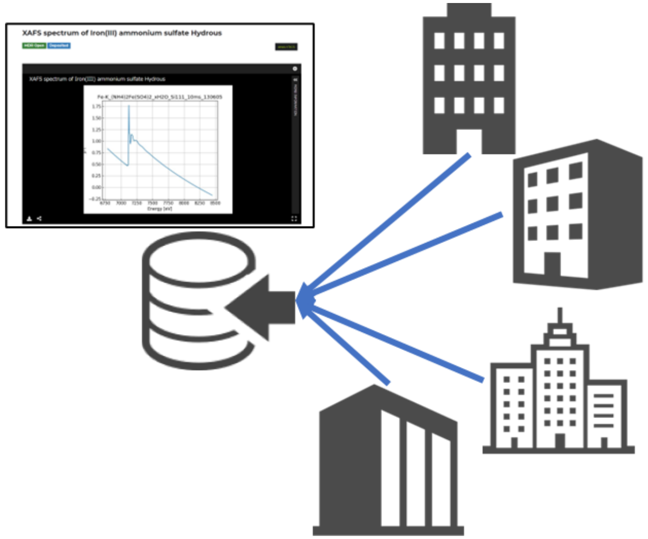
Creator: TODOROKI, Shin-ichi
Date published: 22/01/2013
Date Modified: 15/06/2020

Fiber fuse propagation and voids left behind (overexposed)

Creator: TODOROKI, Shin-ichi
Date published: 03/07/2014
Date Modified: 15/06/2020

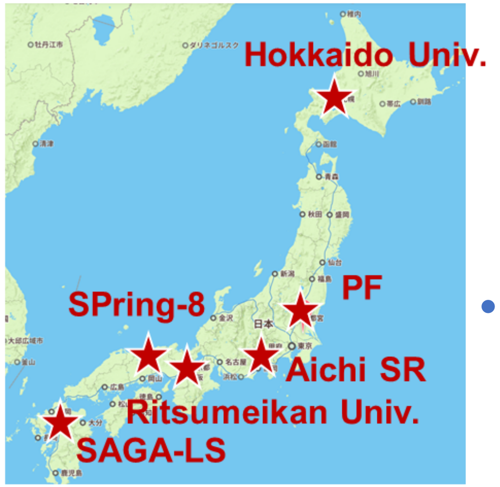
MDR XAFS DB

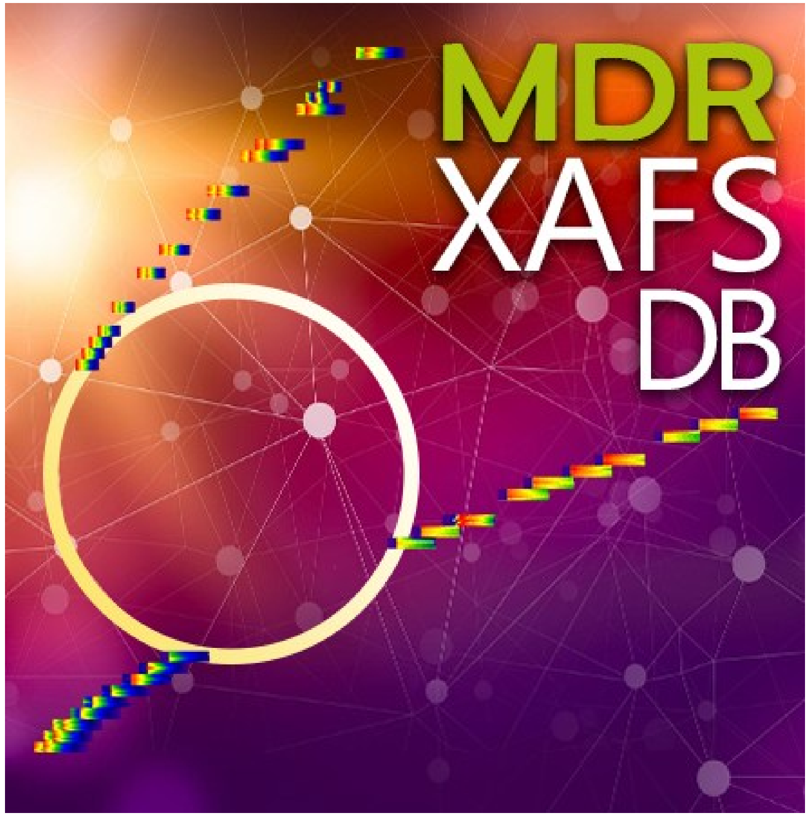
- Collection of X-ray absorption fine structure (XAFS) datasets
  - One of the most popular collections in MDR
- Includes approximately 2600 XAFS spectra datasets collected by six research institutions in Japan
- Project leader: Masashi Ishii



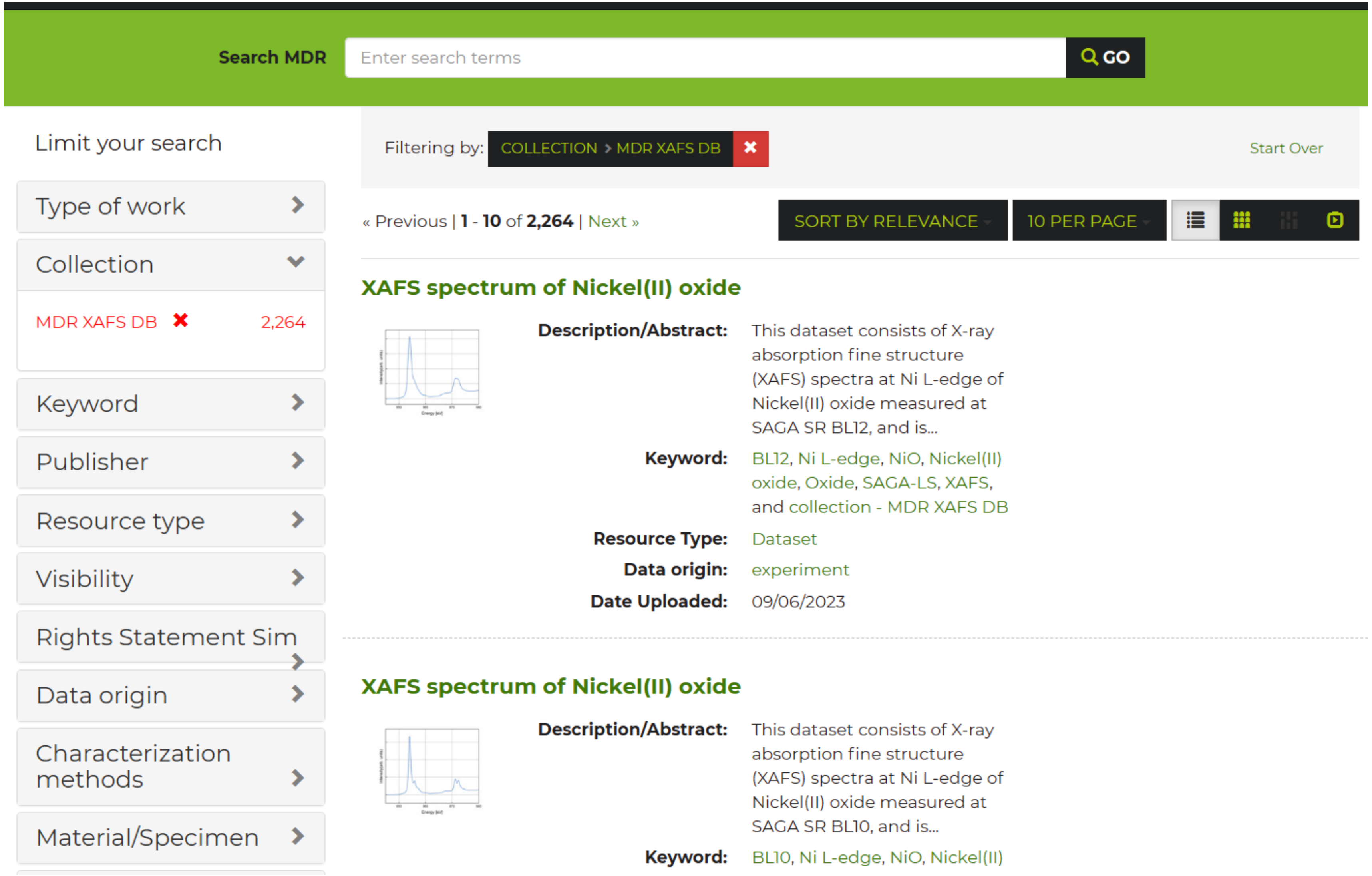
XAFS spectrum of Iron(III) ammonium sulfate Hydrous

XAFS spectrum of Iron(III) ammonium sulfate Hydrous





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



Filtering by: COLLECTION > MDR XAFS DB

MDR XAFS DB 2,264

« Previous | **1** - **10** of **2,264** | Next »

XAFS spectrum of Nickel(II) oxide

Description/Abstract: This dataset consists of X-ray absorption fine structure (XAFS) spectra at Ni L-edge of Nickel(II) oxide measured at SAGA SR BL12, and is...

Keyword: BL12, Ni L-edge, NiO, Nickel(II) oxide, Oxide, SAGA-LS, XAFS, and collection - MDR XAFS DB

Resource Type: Dataset

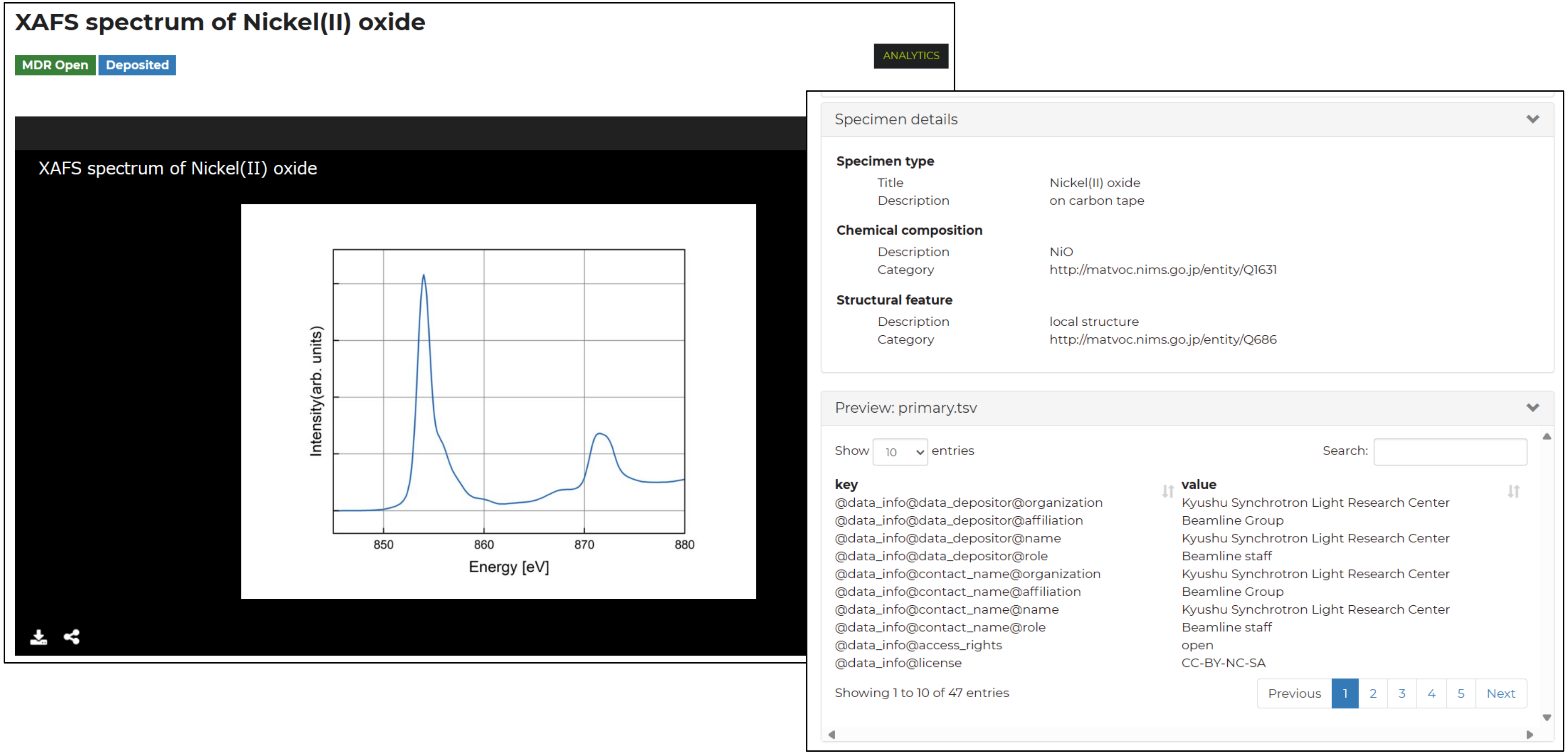
Data origin: experiment

Date Uploaded: 09/06/2023

XAFS spectrum of Nickel(II) oxide

Description/Abstract: This dataset consists of X-ray absorption fine structure (XAFS) spectra at Ni L-edge of Nickel(II) oxide measured at SAGA SR BL10, and is...

Keyword: BL10, Ni L-edge, NiO, Nickel(II)



XAFS spectrum of Nickel(II) oxide

MDR Open Deposited

ANALYTICS

XAFS spectrum of Nickel(II) oxide

Specimen details

Specimen type

Title	Nickel(II) oxide
Description	on carbon tape

Chemical composition

Description	NiO
Category	http://matvoc.nims.go.jp/entity/Q1631

Structural feature

Description	local structure
Category	http://matvoc.nims.go.jp/entity/Q686

Preview: primary.tsv

Show 10 entries Search:

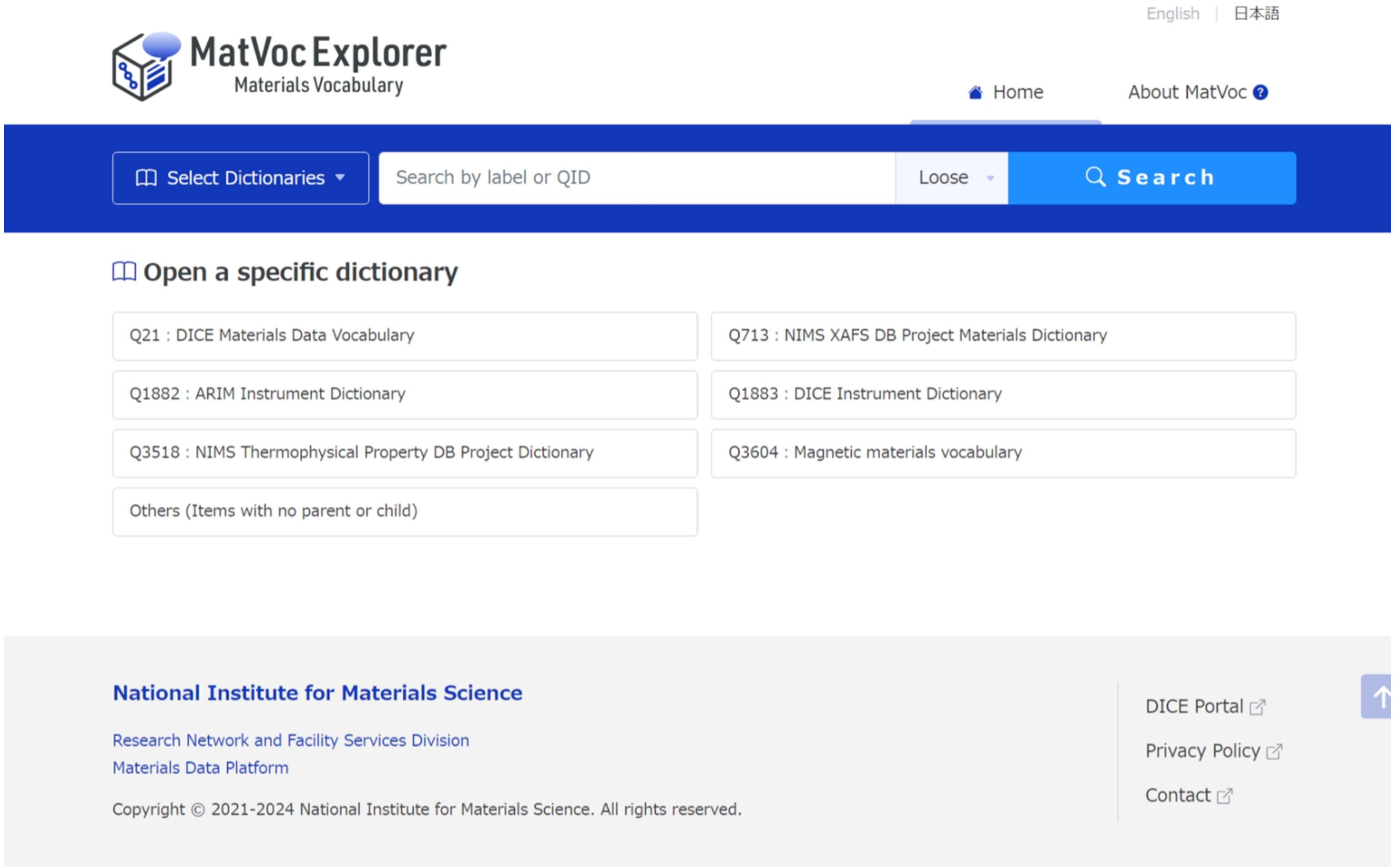
key	value
@data_info@data_depositor@organization	Kyushu Synchrotron Light Research Center
@data_info@data_depositor@affiliation	Beamline Group
@data_info@data_depositor@name	Kyushu Synchrotron Light Research Center
@data_info@data_depositor@role	Beamline staff
@data_info@contact_name@organization	Kyushu Synchrotron Light Research Center
@data_info@contact_name@affiliation	Beamline Group
@data_info@contact_name@name	Kyushu Synchrotron Light Research Center
@data_info@contact_name@role	Beamline staff
@data_info@access_rights	open
@data_info@license	CC-BY-NC-SA

Showing 1 to 10 of 47 entries

MatVoc / MatVoc Explorer

- MatVoc: https://matvoc.nims.go.jp
  - Vocabulary service for materials science
  - Built on Wikibase
  - Includes more than 2400 entities
  - Provides SPARQL endpoint
- MatVoc Explorer: https://matvoc.nims.go.jp/explore/en/home
  - Provides browse and search UI
  - Provides a link to MDR and other external databases associated with each vocabulary
- Project leader: Asahiko Matsuda

MatVoc Explorer
Materials Vocabulary



English | 日本語

MatVoc Explorer
Materials Vocabulary

Home

About MatVoc

Select Dictionaries

Search by label or QID

Loose

Search

Open a specific dictionary

- Q21 : DICE Materials Data Vocabulary
- Q713 : NIMS XAFS DB Project Materials Dictionary
- Q1882 : ARIM Instrument Dictionary
- Q1883 : DICE Instrument Dictionary
- Q3518 : NIMS Thermophysical Property DB Project Dictionary
- Q3604 : Magnetic materials vocabulary
- Others (Items with no parent or child)

National Institute for Materials Science

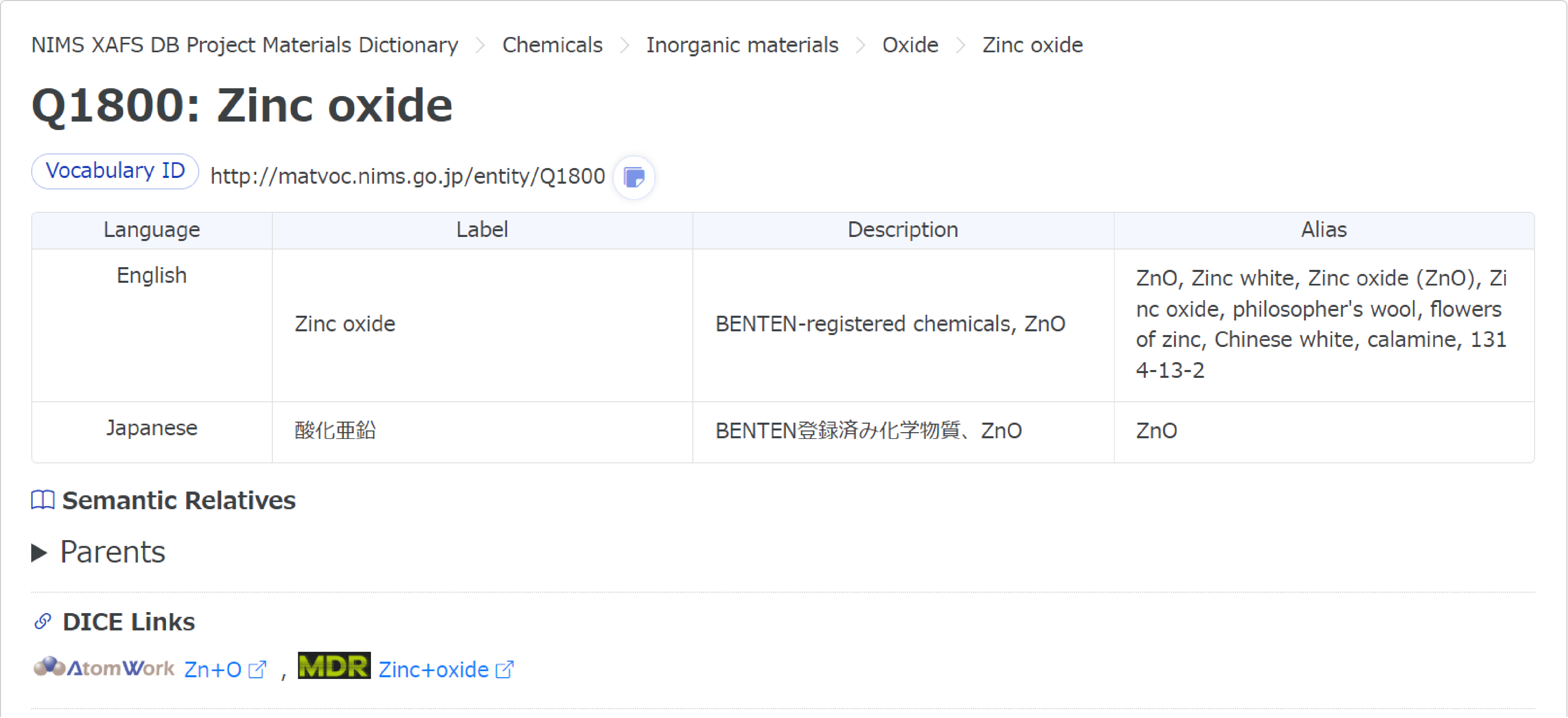
Research Network and Facility Services Division
Materials Data Platform

Copyright © 2021-2024 National Institute for Materials Science. All rights reserved.

DICE Portal

Privacy Policy

Contact



NIMS XAFS DB Project Materials Dictionary › Chemicals › Inorganic materials › Oxide › Zinc oxide

Q1800: Zinc oxide

Vocabulary ID http://matvoc.nims.go.jp/entity/Q1800

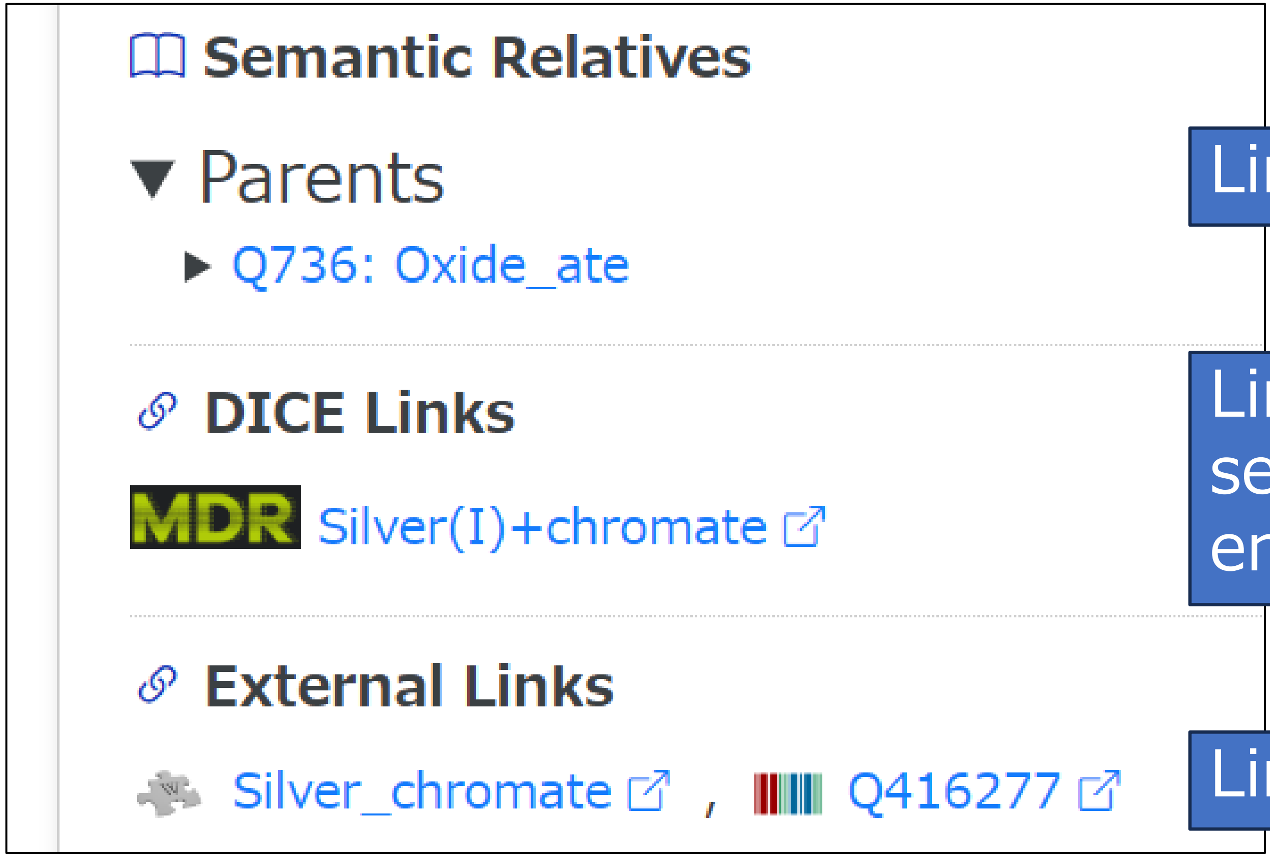
Language	Label	Description	Alias
English	Zinc oxide	BENTEN-registered chemicals, ZnO	ZnO, Zinc white, Zinc oxide (ZnO), Zinc oxide, philosopher's wool, flowers of zinc, Chinese white, calamine, 1314-13-2
Japanese	酸化亜鉛	BENTEN登録済み化学物質、ZnO	ZnO

Semantic Relatives

▶ Parents

DICE Links

AtomWork Zn+O , MDR Zinc+oxide



🕮 Semantic Relatives

▼ Parents

▶ Q736: Oxide_ate

🔗 DICE Links

MDR Silver(I)+chromate ↗

🔗 External Links

Silver_chromate ↗ , Q416277 ↗

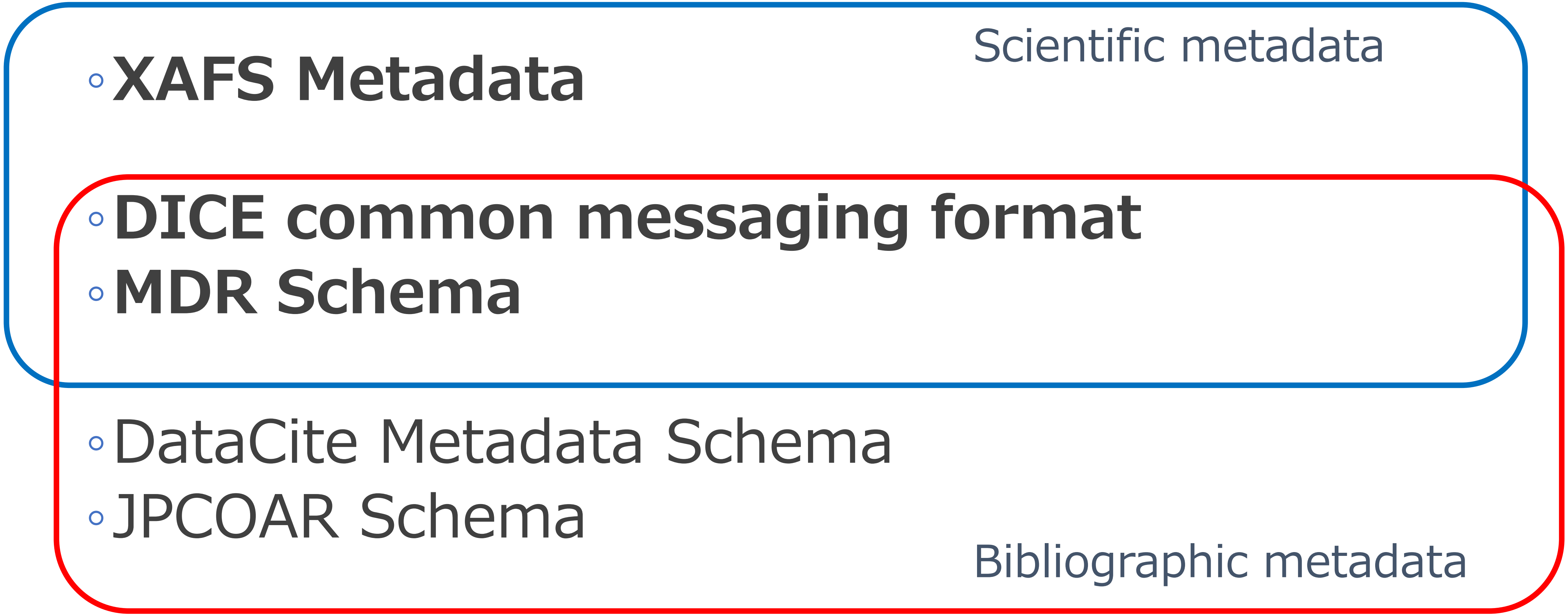
Link to relative MatVoc entities

Link to MDR and other DICE services associated with the entity

Link to Wikipedia and Wikidata

Metadata schemas around MDR and MDR XAFS DB

Metadata (schemas) in MDR and MDR XAFS DB



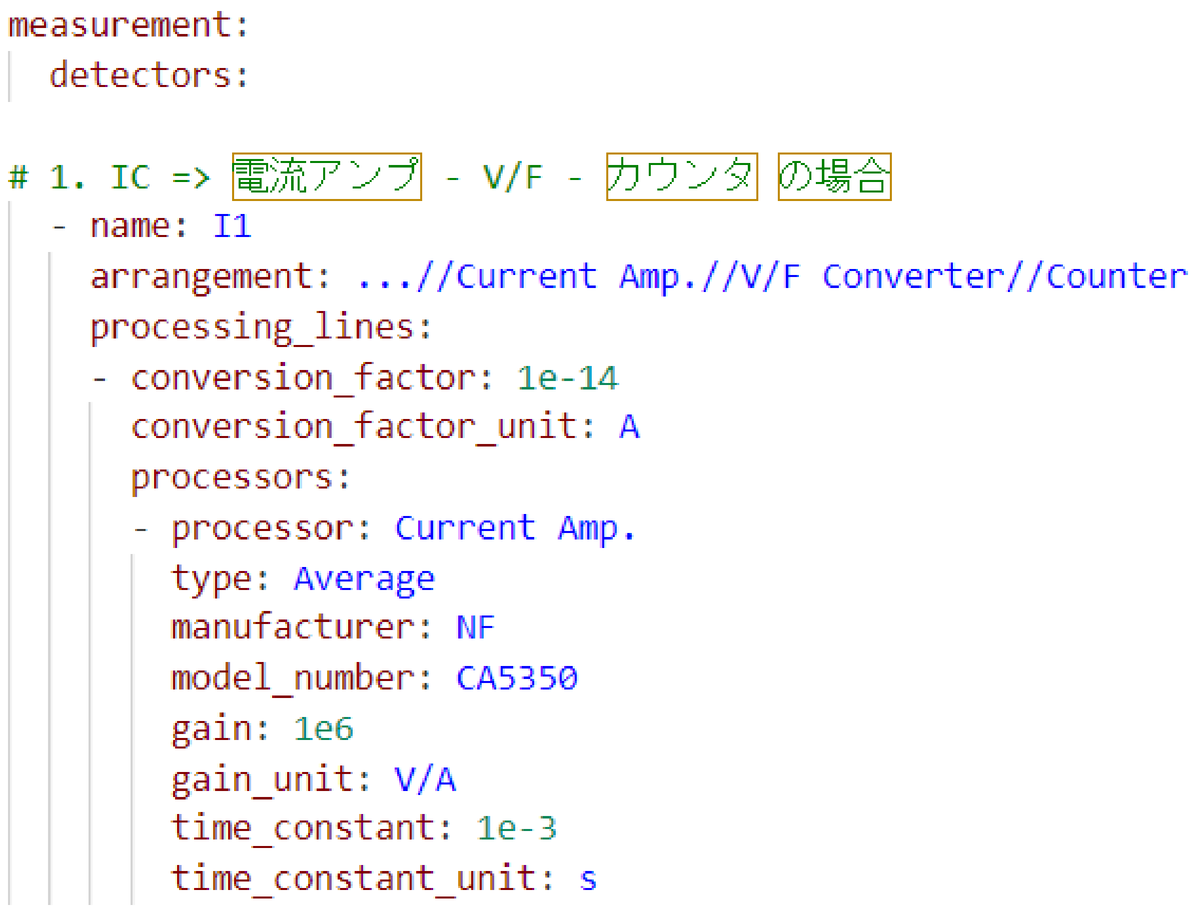
Scientific metadata

- **XAFS Metadata**
- **DICE common messaging format**
- **MDR Schema**
- DataCite Metadata Schema
- JPCOAR Schema

Bibliographic metadata

XAFS Metadata

- https://github.com/xafs-db/xafs-schema
- Metadata schema developed by the Japan XAFS Society
  - https://www.jxafs.org/xafs-database
- Includes domain-specific properties (e.g., instruments, measurement)



```
measurement:
  detectors:

# 1. IC => 電流アンプ - V/F - カウンタの場合
  - name: I1
    arrangement: ...//Current Amp.//V/F Converter//Counter
    processing_lines:
    - conversion_factor: 1e-14
      conversion_factor_unit: A
      processors:
      - processor: Current Amp.
        type: Average
        manufacturer: NF
        model_number: CA5350
        gain: 1e6
        gain_unit: V/A
        time_constant: 1e-3
        time_constant_unit: s
```

DICE common messaging format

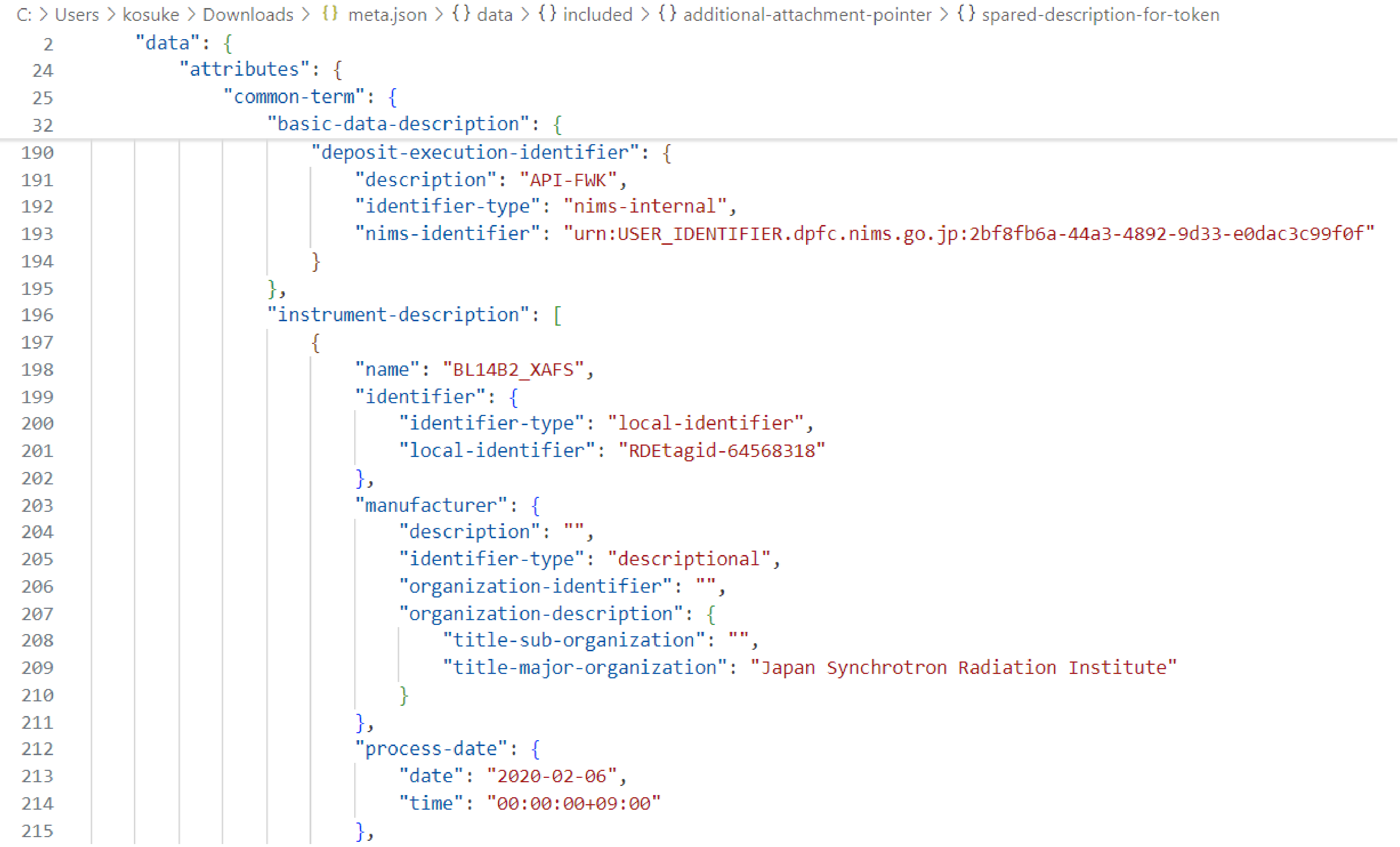
- https://doi.org/10.48505/nims.3240
- Designed for entire datasets and research areas on the data platform DICE
- Supports many scientific entities and properties including instruments, specimens and experimental methods
- Consists of many deeply-nested properties
- *"We have defined a common metadata schema for research dataset distribution/exchange/storage, with features for not only system-to-system communication but also description of the datasets. The metadata were initially defined by a XML Schema, and has been converted into JSON Schema. The schema has a complex structure, which includes the dataset's bibliographic metadata as well as description of the dataset from the materials science viewpoint."*

The structure of DICE common messaging format

Mandatory	**Common metadata** ID, Depositor, Specimen, Instrument, Data origin…				
Domain-specific	**Characterization metadata** Method, Environment…	**Specimen metadata** Material type, Structural features	**Property metadata** Characteristic properties	**Synthesis/Process metadata** Processed date Temperature	**Calculation metadata** Computer Software
Parameters (uncontrolled)	**Characterization primary parameters**	**Specimen primary parameters**	**Property primary parameters**	**Synthesis/Process primary parameters**	**Calculation primary parameters**
Arbitrary data	Arbitrary data	Arbitrary data	Arbitrary data	Arbitrary data	Arbitrary data

S. Kikuchi et al., IEICE Tech. Rep. vol. 119, no. 66, SC2019-2, pp. 7-17, 2019, in Japanese
https://ken.ieice.org/ken/paper/20190531k1nc/
(also in Ranganathan et al., 14th Int. Conf. Open Repositories, 2019. https://doi.org/10.5281/zenodo.3553963)

An example of the DICE common messaging format

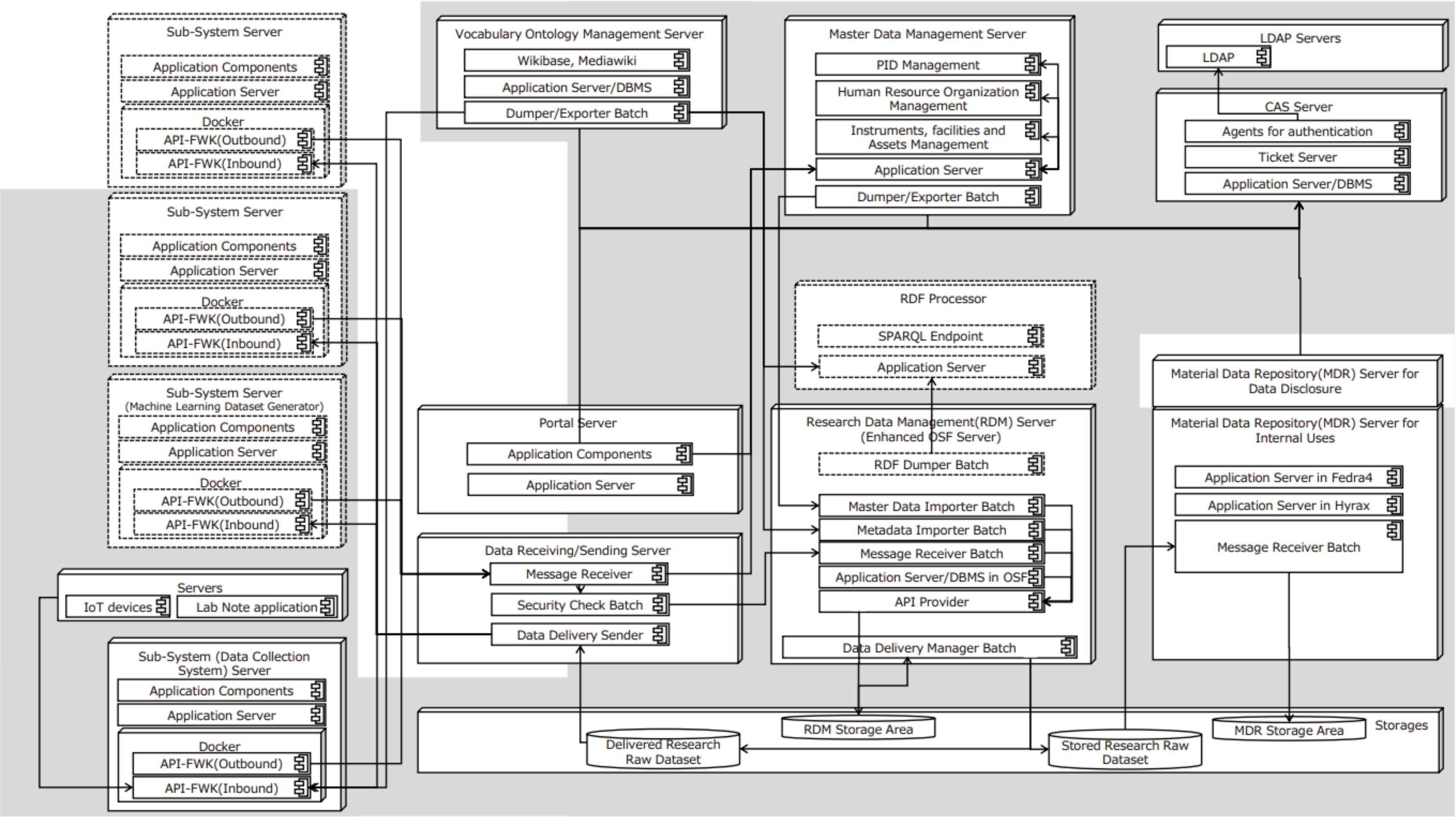


C: > Users > kosuke > Downloads > {} meta.json > {} data > {} included > {} additional-attachment-pointer > {} spared-description-for-token

```
    "data": {
        "attributes": {
            "common-term": {
                "basic-data-description": {
                    "deposit-execution-identifier": {
                        "description": "API-FWK",
                        "identifier-type": "nims-internal",
                        "nims-identifier": "urn:USER_IDENTIFIER.dpfc.nims.go.jp:2bf8fb6a-44a3-4892-9d33-e0dac3c99f0f"
                    }
                },
                "instrument-description": [
                    {
                        "name": "BL14B2_XAFS",
                        "identifier": {
                            "identifier-type": "local-identifier",
                            "local-identifier": "RDEtagid-64568318"
                        },
                        "manufacturer": {
                            "description": "",
                            "identifier-type": "descriptional",
                            "organization-identifier": "",
                            "organization-description": {
                                "title-sub-organization": "",
                                "title-major-organization": "Japan Synchrotron Radiation Institute"
                            }
                        },
                        "process-date": {
                            "date": "2020-02-06",
                            "time": "00:00:00+09:00"
                        },
```

Initial plan in depositing XAFS datasets to MDR

- All research datasets and applications are expected to use the DICE common messaging format to describe and share its metadata



Sub-System Server
Application Components
Application Server
Docker
API-FWK(Outbound)
API-FWK(Inbound)

Sub-System Server
Application Components
Application Server
Docker
API-FWK(Outbound)
API-FWK(Inbound)

Sub-System Server
(Machine Learning Dataset Generator)
Application Components
Application Server
Docker
API-FWK(Outbound)
API-FWK(Inbound)

Servers
IoT devices
Lab Note application

Sub-System (Data Collection System) Server
Application Components
Application Server
Docker
API-FWK(Outbound)
API-FWK(Inbound)

Vocabulary Ontology Management Server
Wikibase, Mediawiki
Application Server/DBMS
Dumper/Exporter Batch

Portal Server
Application Components
Application Server

Data Receiving/Sending Server
Message Receiver
Security Check Batch
Data Delivery Sender

Master Data Management Server
PID Management
Human Resource Organization Management
Instruments, facilities and Assets Management
Application Server
Dumper/Exporter Batch

RDF Processor
SPARQL Endpoint
Application Server

Research Data Management(RDM) Server
(Enhanced OSF Server)
RDF Dumper Batch
Master Data Importer Batch
Metadata Importer Batch
Message Receiver Batch
Application Server/DBMS in OSF
API Provider
Data Delivery Manager Batch

LDAP Servers
LDAP

CAS Server
Agents for authentication
Ticket Server
Application Server/DBMS

Material Data Repository(MDR) Server for Data Disclosure

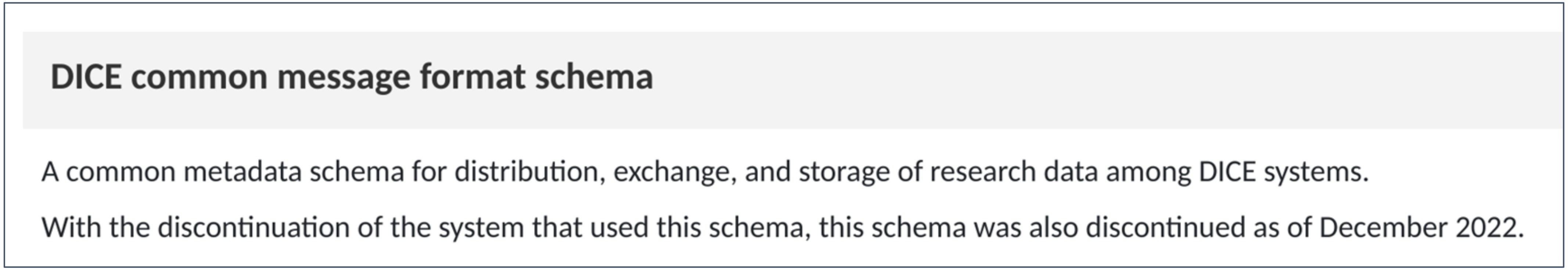
Material Data Repository(MDR) Server for Internal Uses
Application Server in Fedra4
Application Server in Hyrax
Message Receiver Batch

Storages
Delivered Research Raw Dataset
RDM Storage Area
Stored Research Raw Dataset
MDR Storage Area

S. Kikuchi et al., IEICE Tech. Rep. vol. 119, no. 66, SC2019-2, pp. 7-17, 2019, in Japanese https://ken.ieice.org/ken/paper/20190531k1nc/

Difficulties in applying the huge metadata schema

- Some of the XAFS datasets were deposited to MDR using the schema, but it took much more efforts than expected
- It was extremely difficult for researchers and application engineers to understand its complex specification
  - Encountered implementation incompatibilities in handling the metadata and schema between the application software
  - Encountered many limitations and bugs in the software (customized OSS)
- After some operations, it was officially abandoned in 2022



DICE common message format schema

A common metadata schema for distribution, exchange, and storage of research data among DICE systems.

With the discontinuation of the system that used this schema, this schema was also discontinued as of December 2022.

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

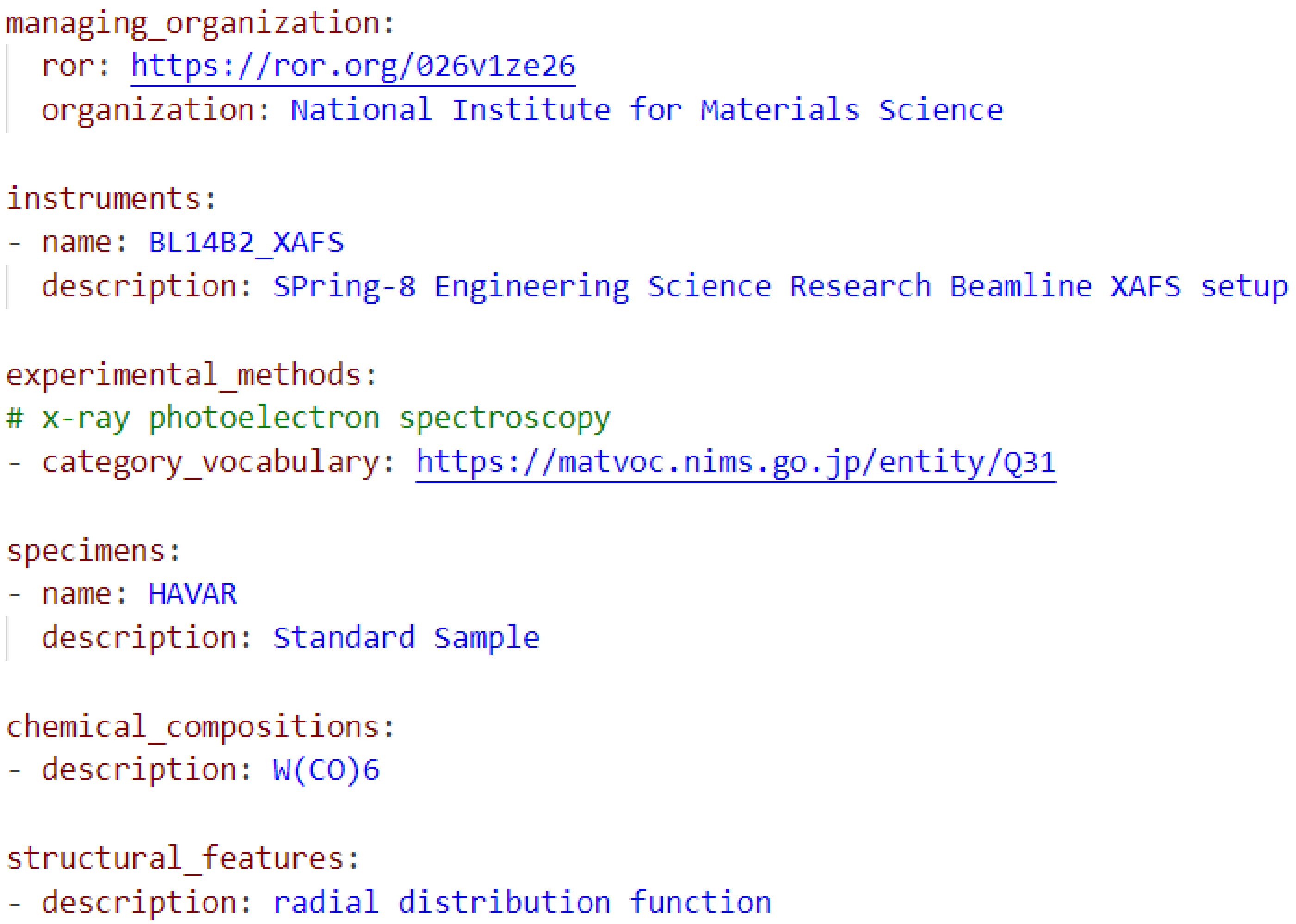
Technical limitations in MDR

- MDR basically supports only keyword search
- MDR doesn’t fully support nested metadata structure in the DICE common messaging format
    - e.g., cannot update a value in a nested property
- These came from technical limitations in MDR’s software stack
- “…What should we do?”

MDR Schema

- https://github.com/nims-dpfc/mdr-schema
- Simplified version of the DICE common messaging format
- Solely designed for depositing datasets to MDR
  - Mostly compatible with the existing metadata on MDR
- Eliminates deeply-nested hierarchy structure
- Expects to be written in YAML

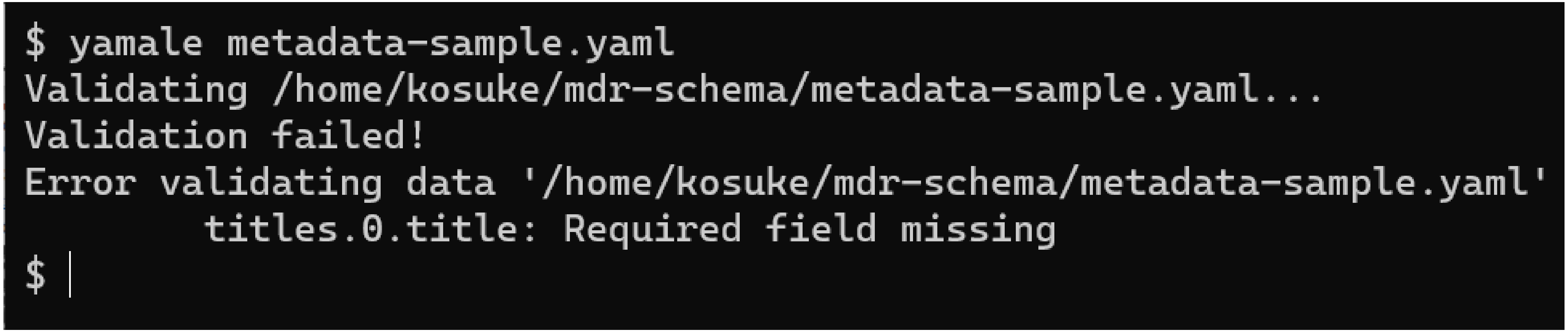
An example of metadata in MDR Schema



```yaml
managing_organization:
  ror: https://ror.org/026v1ze26
  organization: National Institute for Materials Science

instruments:
- name: BL14B2_XAFS
  description: SPring-8 Engineering Science Research Beamline XAFS setup

experimental_methods:
# x-ray photoelectron spectroscopy
- category_vocabulary: https://matvoc.nims.go.jp/entity/Q31

specimens:
- name: HAVAR
  description: Standard Sample

chemical_compositions:
- description: W(CO)6

structural_features:
- description: radial distribution function
```

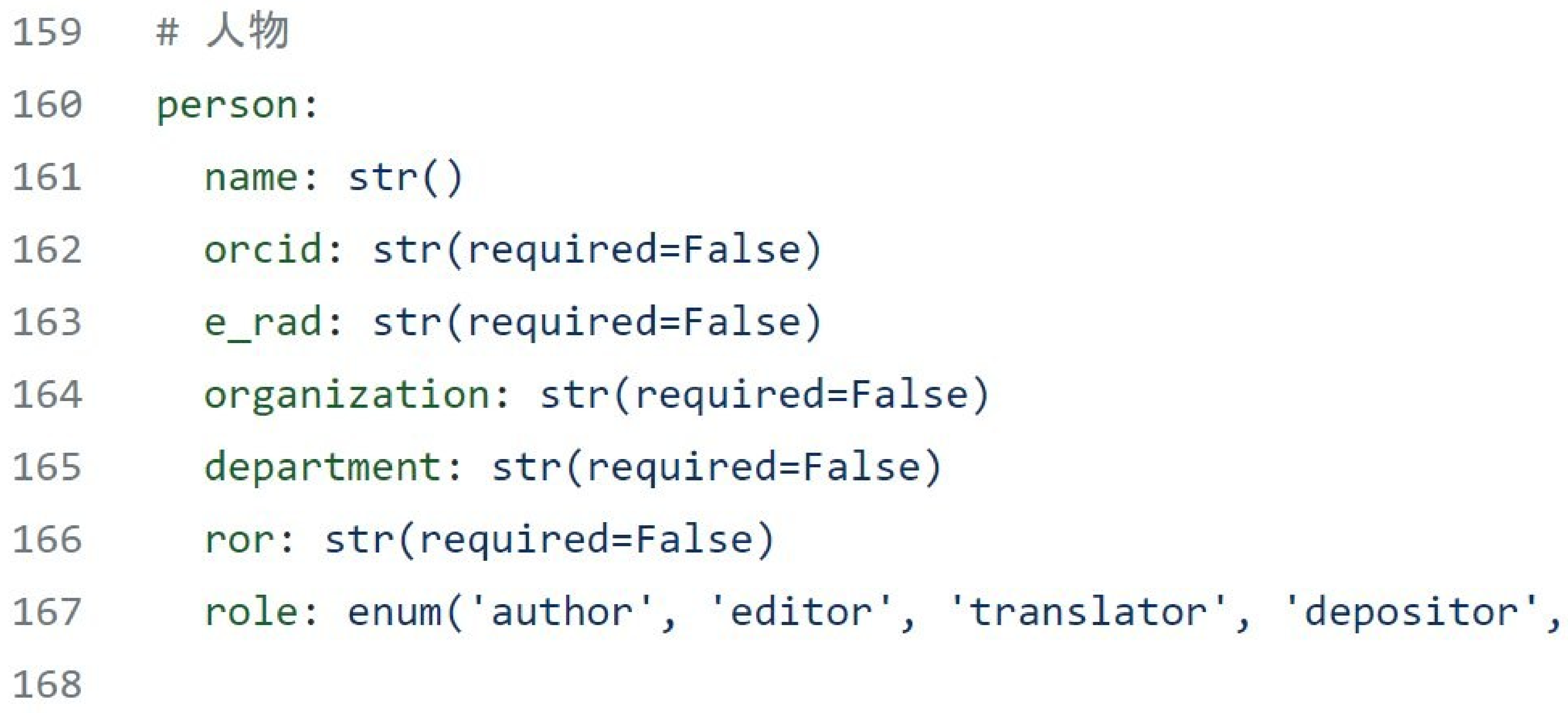
Validating MDR Schema YAML

- Using Yamale to create and validate the metadata schema
  - https://github.com/23andMe/Yamale



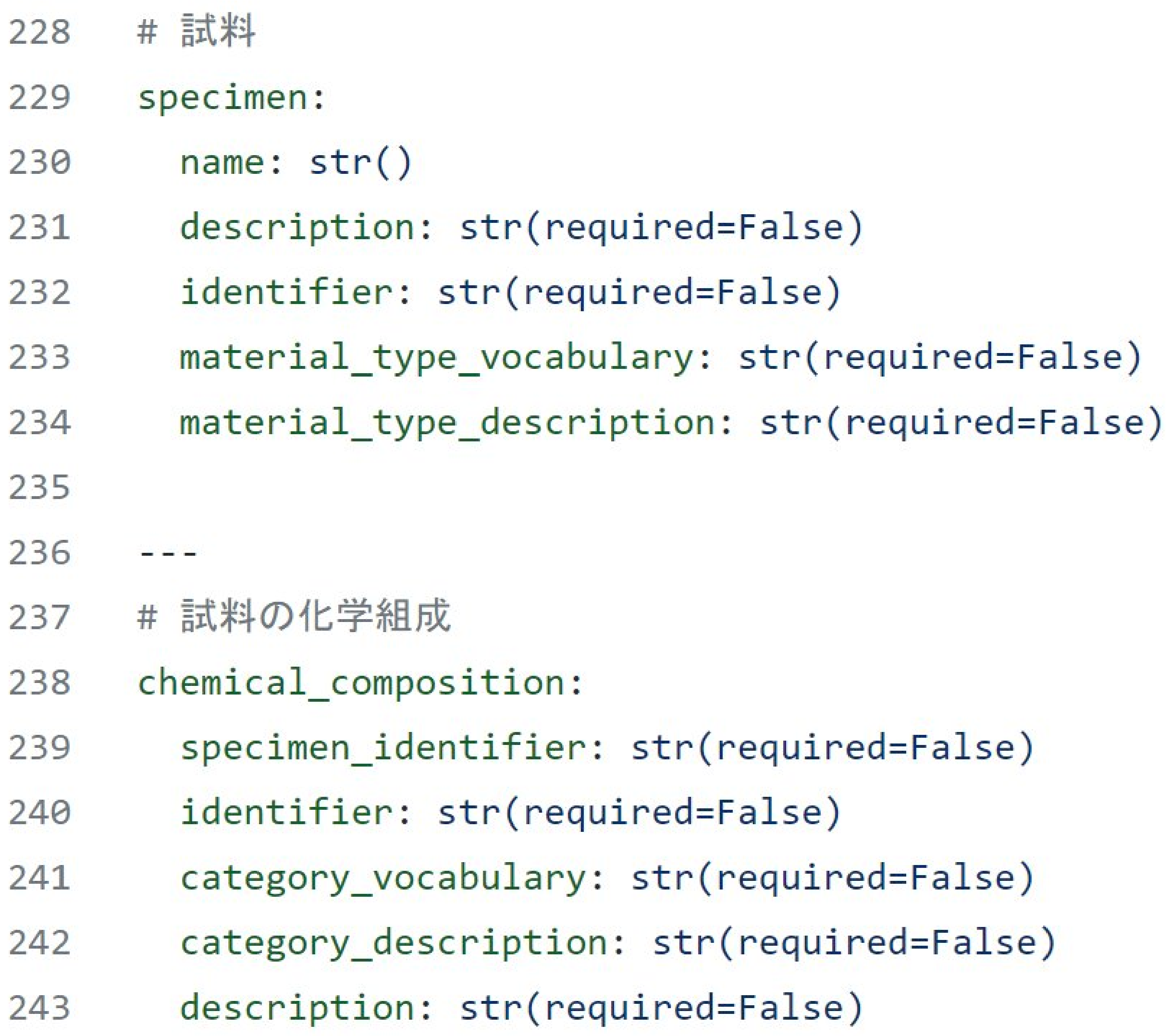
```
$ yamale metadata-sample.yaml
Validating /home/kosuke/mdr-schema/metadata-sample.yaml...
Validation failed!
Error validating data '/home/kosuke/mdr-schema/metadata-sample.yaml'
        titles.0.title: Required field missing
$ |
```

Schema definition of MDR Schema



```
# 人物
person:
  name: str()
  orcid: str(required=False)
  e_rad: str(required=False)
  organization: str(required=False)
  department: str(required=False)
  ror: str(required=False)
  role: enum('author', 'editor', 'translator', 'depositor',
```

Schema definition of MDR Schema



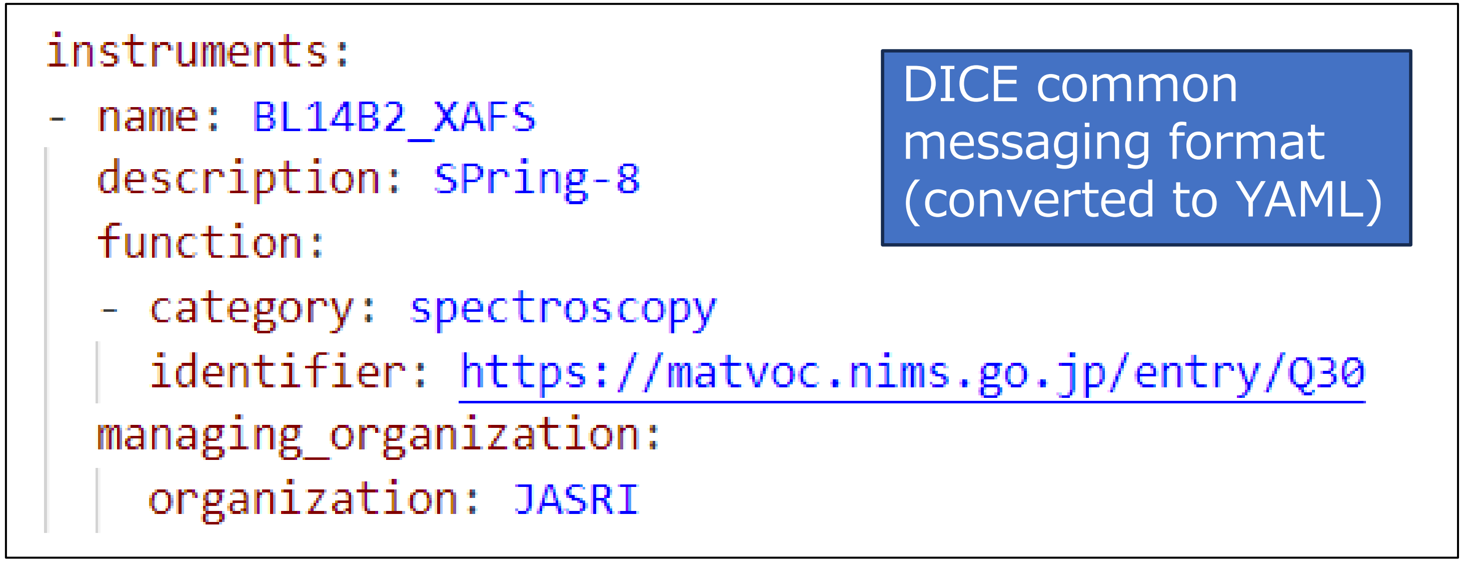
```
# 試料
specimen:
  name: str()
  description: str(required=False)
  identifier: str(required=False)
  material_type_vocabulary: str(required=False)
  material_type_description: str(required=False)

---
# 試料の化学組成
chemical_composition:
  specimen_identifier: str(required=False)
  identifier: str(required=False)
  category_vocabulary: str(required=False)
  category_description: str(required=False)
  description: str(required=False)
```

Why YAML?

- human-readable and (almost) writable
  - Easy to create and distribute a metadata template file since it supports single-line and inline comments
  - It is painful for human to read and write a spreadsheet that has a lot of columns
- It can be converted to JSON or other formats easily

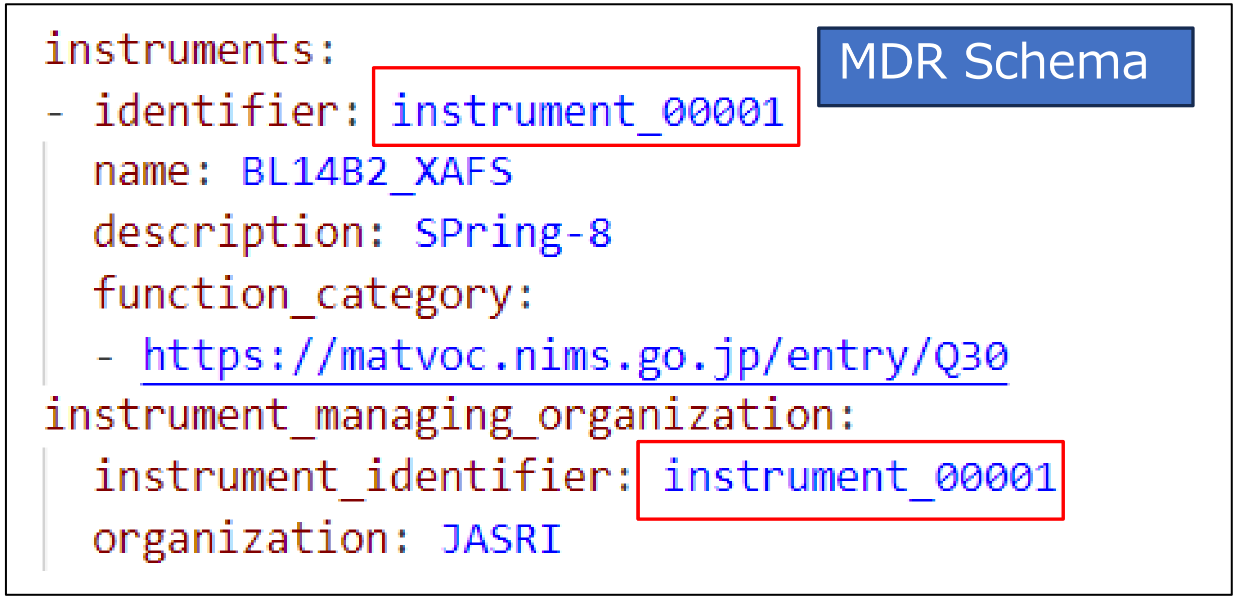
An example of MDR metadata migration

- “instruments”->”function” ->”category” is moved to “function_category”
- “instruments” ->“managing_organization” is moved to “instrument_management_organization”
- Assigns local identifier “instrument_00001” to the instrument and refers it from the “instrument_management_organization”



DICE common messaging format (converted to YAML)

```
instruments:
- name: BL14B2_XAFS
  description: SPring-8
  function:
  - category: spectroscopy
    identifier: https://matvoc.nims.go.jp/entry/Q30
  managing_organization:
    organization: JASRI
```



MDR Schema

```
instruments:
- identifier: instrument_00001
  name: BL14B2_XAFS
  description: SPring-8
  function_category:
  - https://matvoc.nims.go.jp/entry/Q30
instrument_managing_organization:
  instrument_identifier: instrument_00001
  organization: JASRI
```

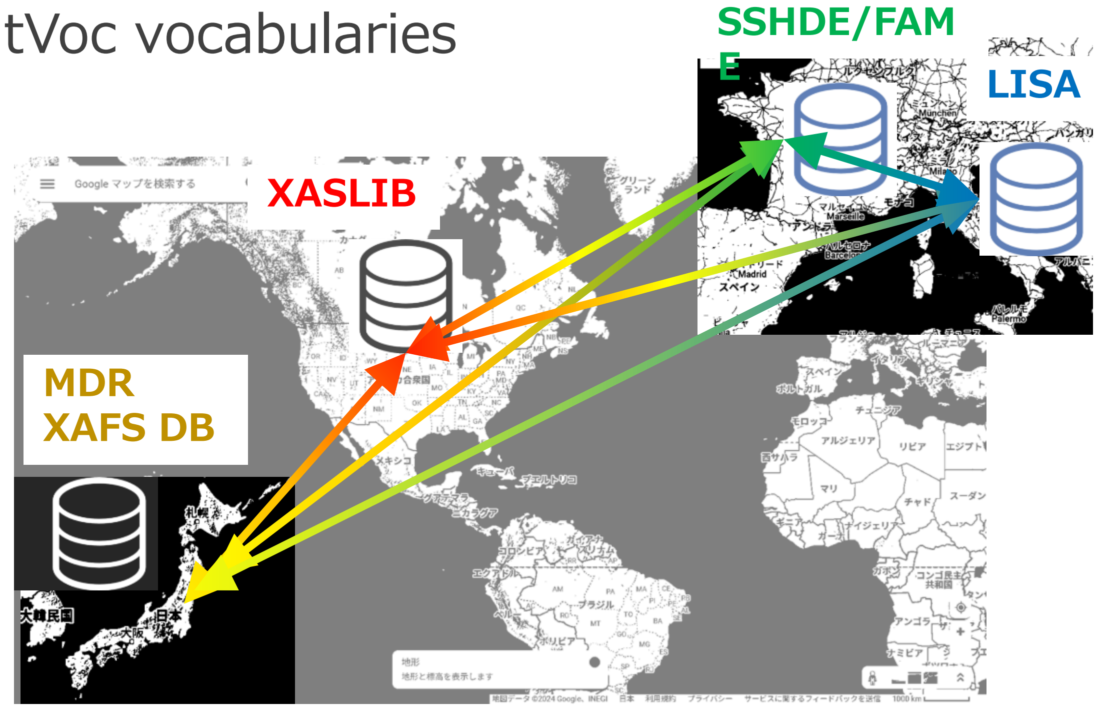
Comparison of metadata schemas around MDR XAFS DB

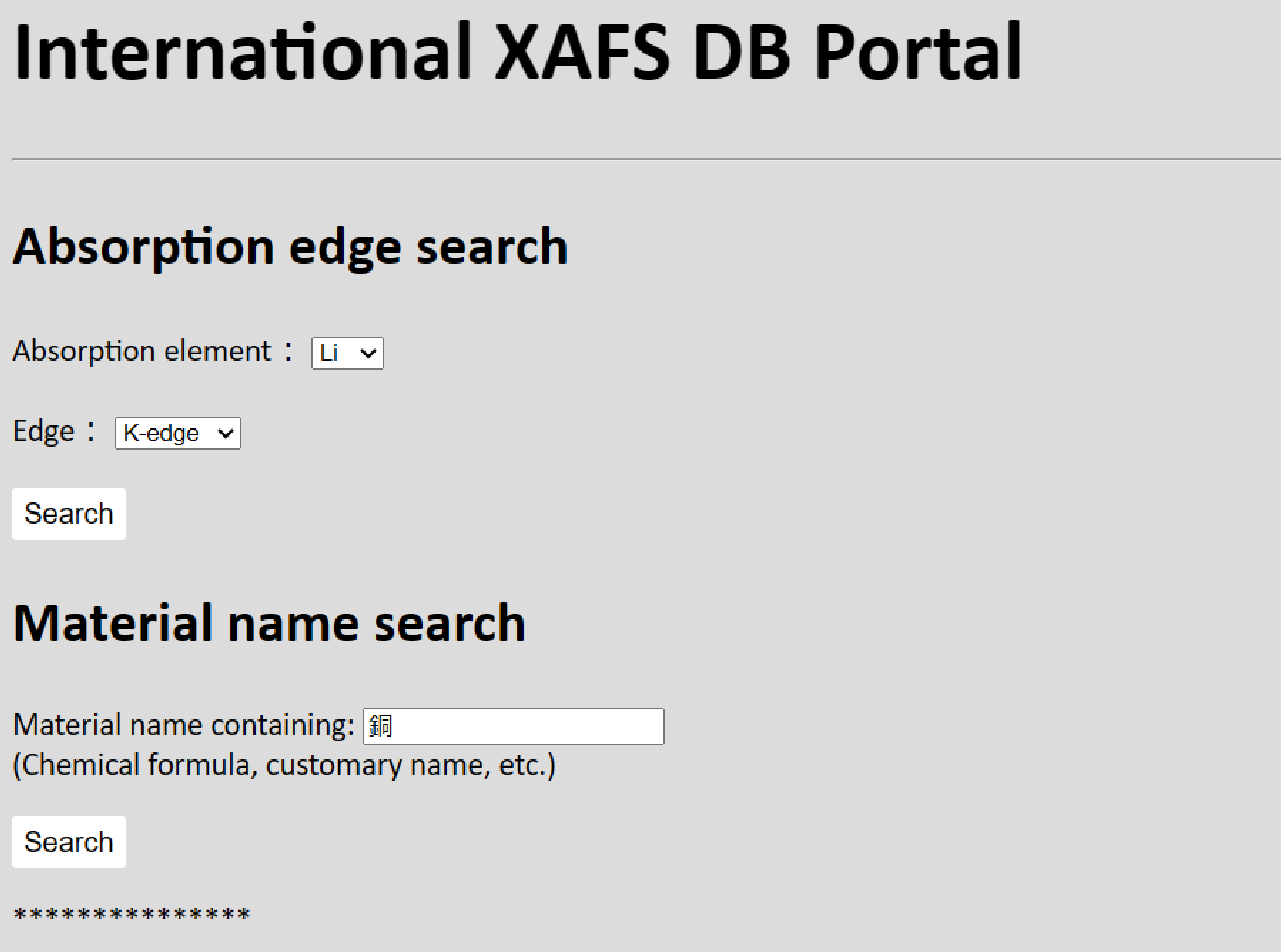
	XAFS Metadata	DICE common messaging format	MDR Schema
Support scientific metadata properties?	Yes	Yes	Partially yes
Research domain specific?	Yes	No	No
Format	YAML, JSON or TSV	XML, JSON	YAML, JSON
Metadata structure	deeply-nested	deeply-nested	single-nested
Human-friendly?	?	No	Yes(?)

advanced case study:
International XAFS DB Portal

International XAFS DB Portal

- Cross-search for MDR XAFS DB and other three XAFS databases in the world
  - SSHADE/FAME (France)
  - XASLIB (US)
  - LISA XAS Database (Italy)
- All datasets are linked with MatVoc vocabularies
- https://ixdb.jxafs.org/





International XAFS DB Portal

Absorption edge search

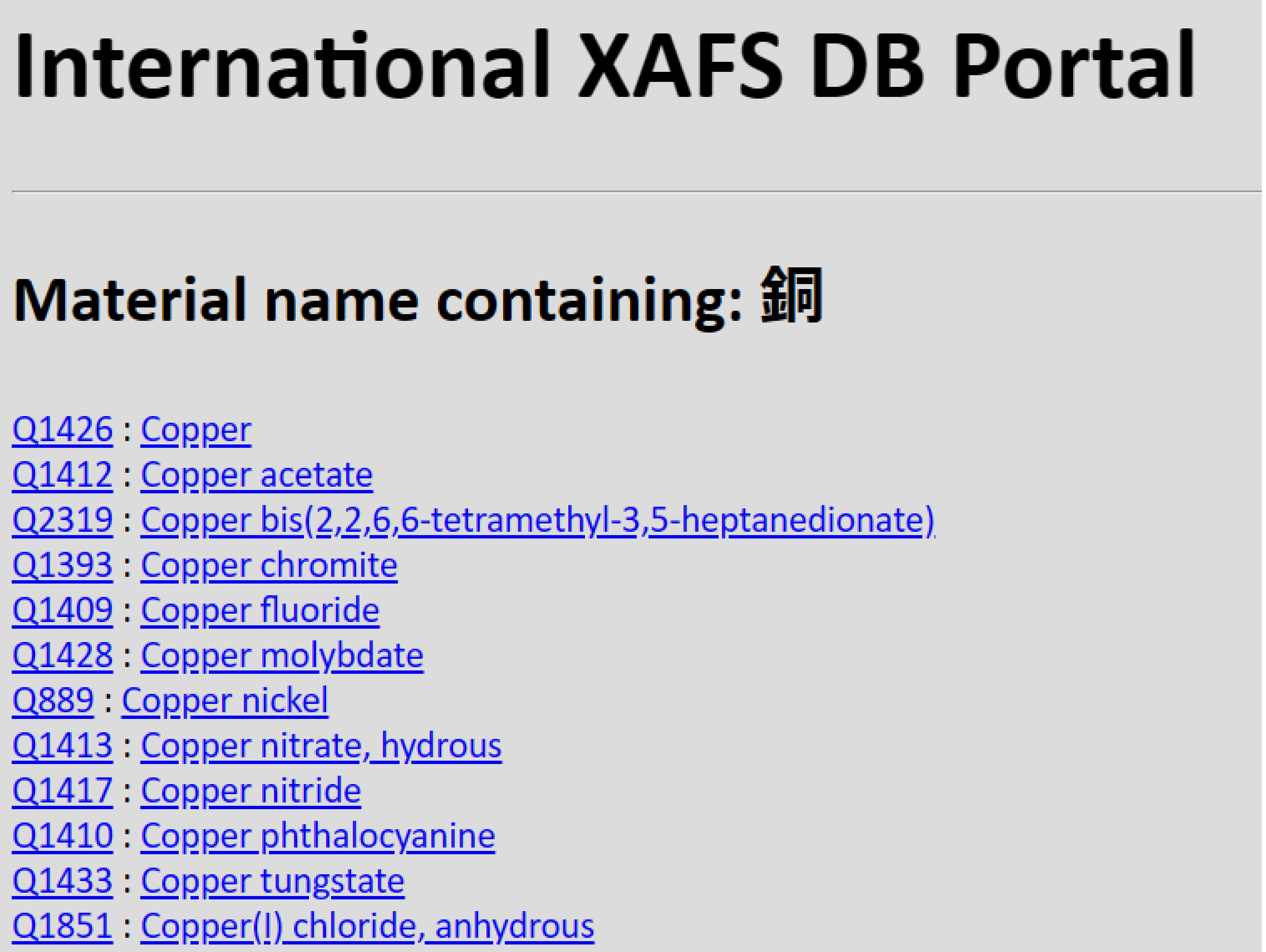
Absorption element： Li

Edge： K-edge

Material name search

Material name containing: 銅
(Chemical formula, customary name, etc.)

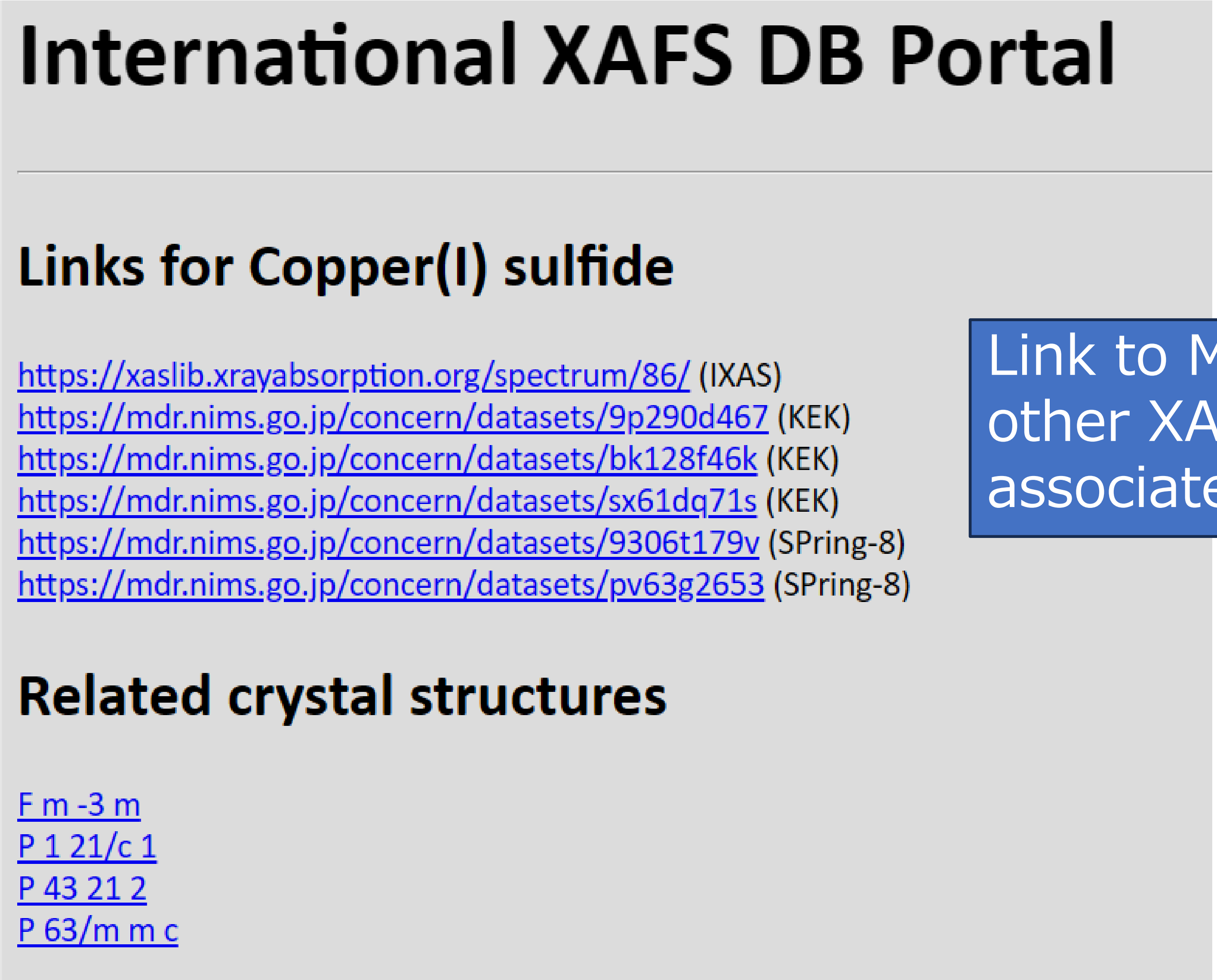
****



International XAFS DB Portal

Material name containing: 銅

Q1426 : Copper
Q1412 : Copper acetate
Q2319 : Copper bis(2,2,6,6-tetramethyl-3,5-heptanedionate)
Q1393 : Copper chromite
Q1409 : Copper fluoride
Q1428 : Copper molybdate
Q889 : Copper nickel
Q1413 : Copper nitrate, hydrous
Q1417 : Copper nitride
Q1410 : Copper phthalocyanine
Q1433 : Copper tungstate
Q1851 : Copper(I) chloride, anhydrous



International XAFS DB Portal

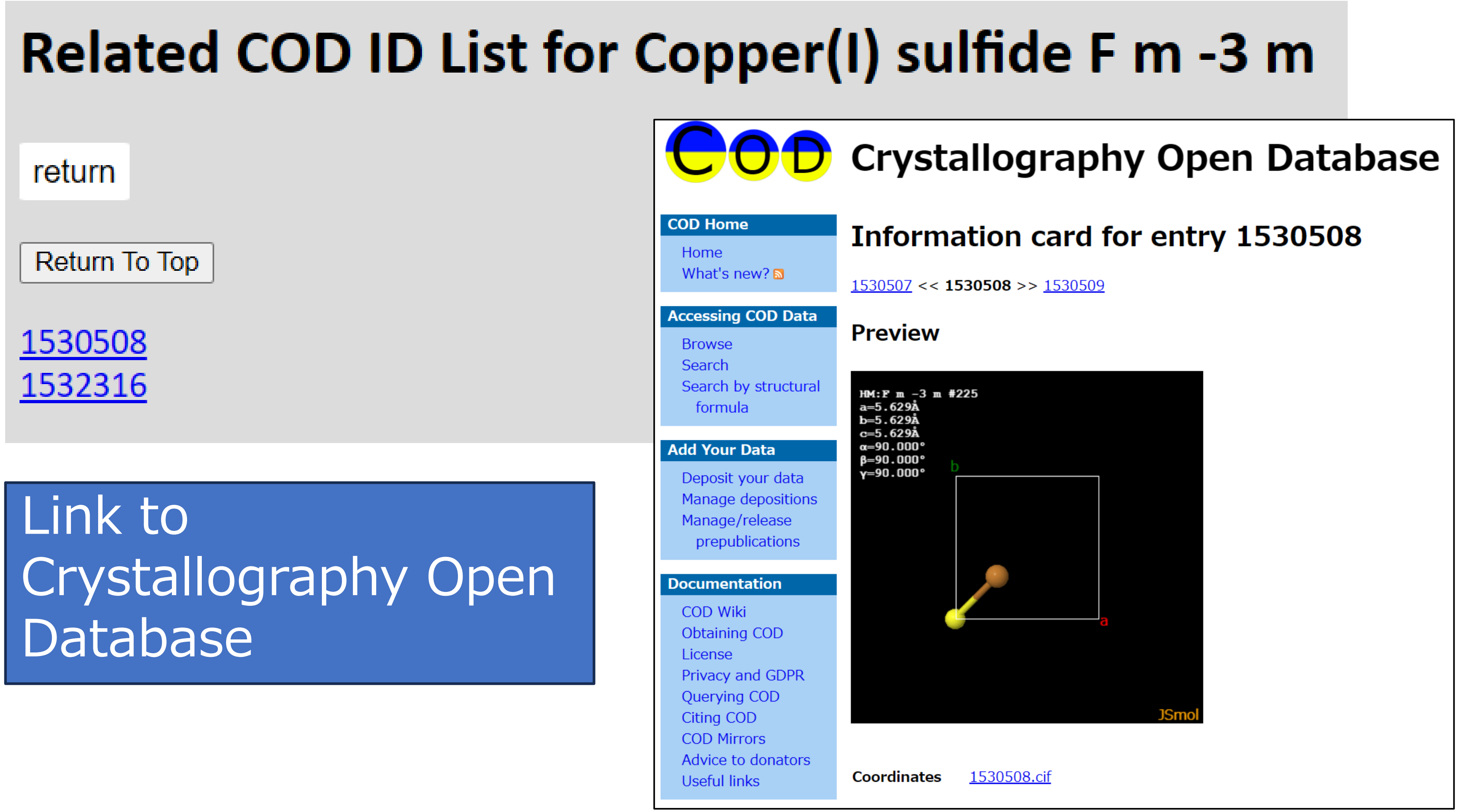
Links for Copper(I) sulfide

https://xaslib.xrayabsorption.org/spectrum/86/ (IXAS)
https://mdr.nims.go.jp/concern/datasets/9p290d467 (KEK)
https://mdr.nims.go.jp/concern/datasets/bk128f46k (KEK)
https://mdr.nims.go.jp/concern/datasets/sx61dq71s (KEK)
https://mdr.nims.go.jp/concern/datasets/9306t179v (SPring-8)
https://mdr.nims.go.jp/concern/datasets/pv63g2653 (SPring-8)

Related crystal structures

F m -3 m
P 1 21/c 1
P 43 21 2
P 63/m m c

Link to MDR XAFS DB and other XAFS databases associated with the entity



Related COD ID List for Copper(I) sulfide F m -3 m

[1530508]
[1532316]

Crystallography Open Database

Information card for entry 1530508

1530507 << **1530508** >> 1530509

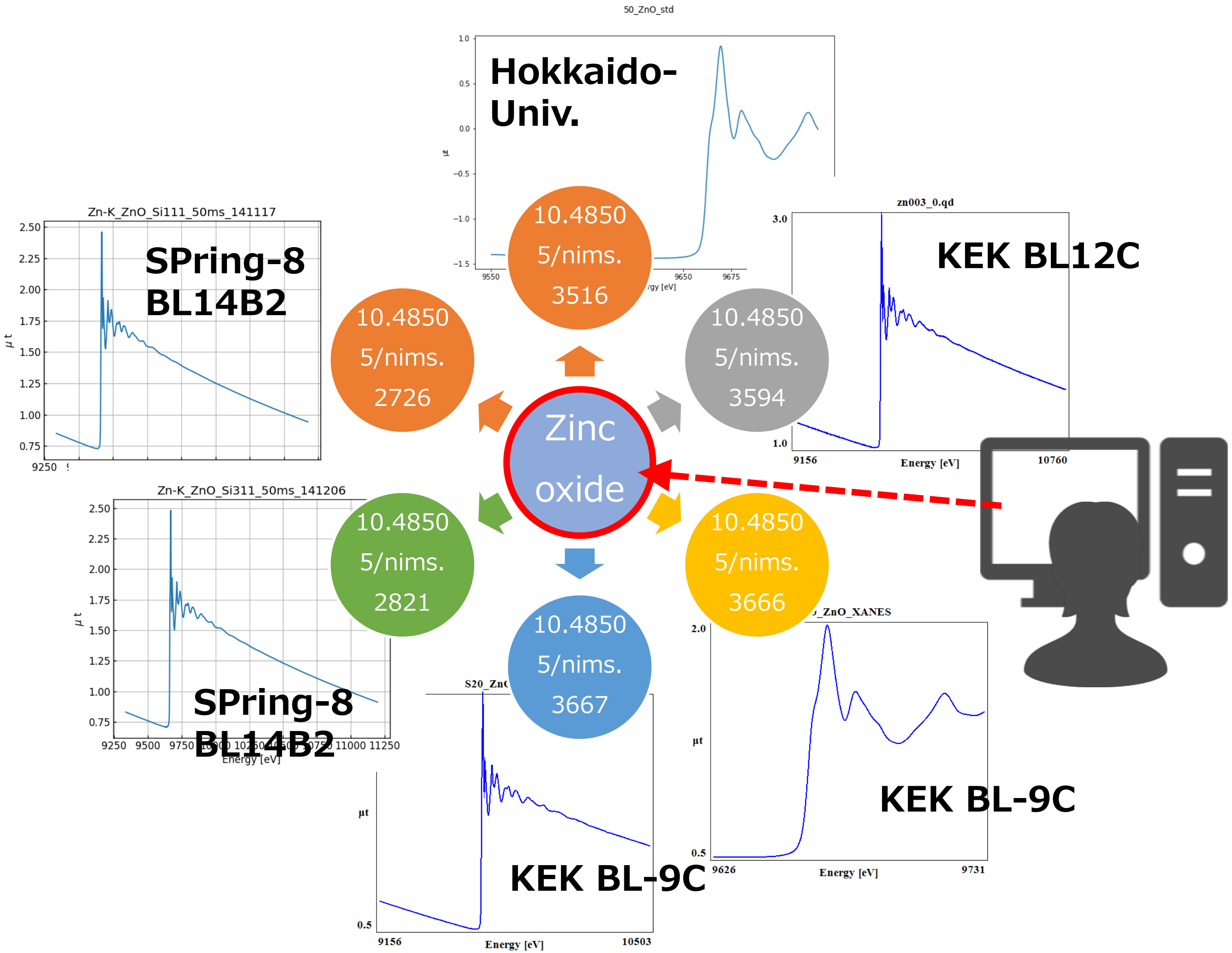
Preview

Coordinates 1530508.cif

Link to Crystallography Open Database

https://www.crystallography.net/cod/1530508.html

Linking metadata using MatVoc



Zinc oxide

10.48505/nims.3516

Hokkaido-Univ.

10.48505/nims.3594

KEK BL12C

10.48505/nims.3666

KEK BL-9C

10.48505/nims.3667

KEK BL-9C

10.48505/nims.2821

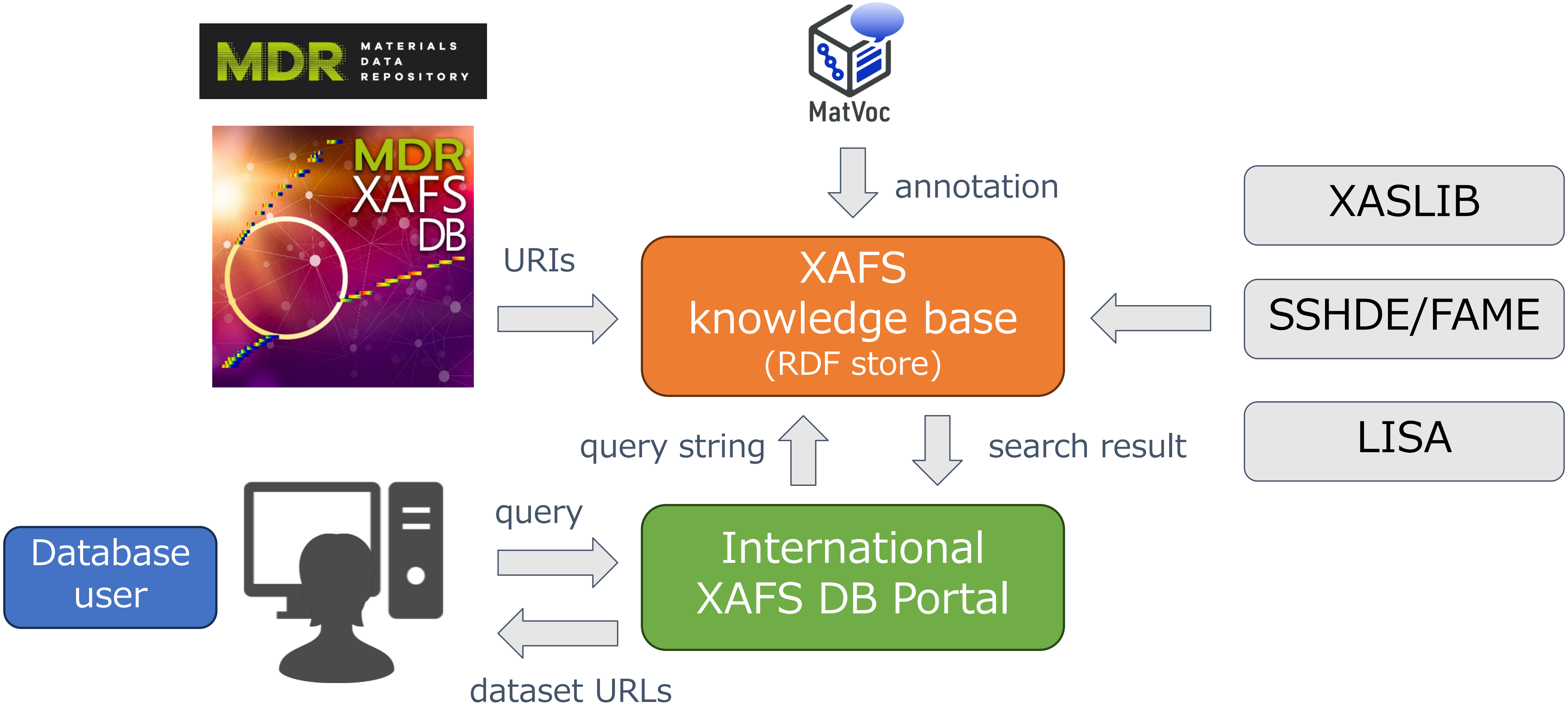
SPring-8 BL14B2

10.48505/nims.2726

SPring-8 BL14B2

MatVoc

XAFS knowledge base



MDR MATERIALS DATA REPOSITORY

MDR XAFS DB

MatVoc

annotation

URIs

XAFS knowledge base (RDF store)

XASLIB

SSHDE/FAME

LISA

query string

search result

Database user

query

International XAFS DB Portal

dataset URLs

Lessons learned

- One big metadata schema does not fit all actually
- Everyone says, “Metadata is important”, but the meaning depends on the context
  - Researchers, engineers, and librarians have different viewpoints on metadata and its use
  - Especially we should consider researcher-friendly metadata creation environment
- Browse and search features are essential in research data management even if they are basic implementations and have basic metadata
  - It is difficult for even researchers and data curators to grasp the whole picture of their datasets
- Identifier (for datasets and vocabularies) is the key to tame diverse research data and their domain-specific metadata

Further plans

- Support querying MatVoc URIs in MDR REST API
- Add MatVoc URIs to datasets previously deposited on MDR
- Add MatVoc URIs to datasets on other DICE services
- Expand MatVoc vocabulary for a focused domain of materials and enhance findability of those data
- Linking materials to their crystallographic information through MatVoc
- **Cross-linking more XAFS databases around the world**

TANABE.Kosuke@nims.go.jp