MDR MATERIALS DATA REPOSITORY

The Development of an Integrated Next Generation Data Repository For Materials Science

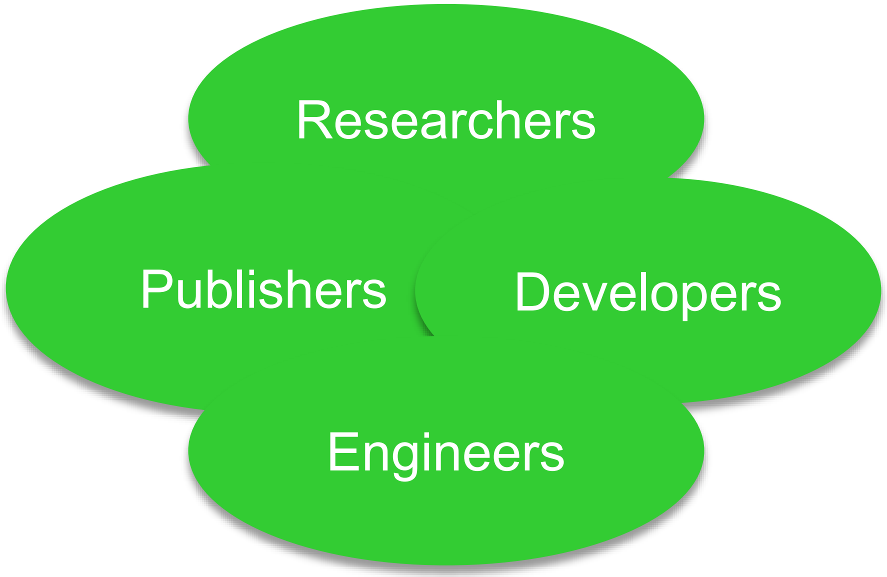
国立研究開発法人物質・材料研究機構
National Institute for Materials Science

·ANTLEAF·

Cottage Labs

MDR Development Project for materials science

- National Institute for Materials Science, Japan
- Cottage Labs, UK
- AntLeaf, UK
- iGroup, Taiwan



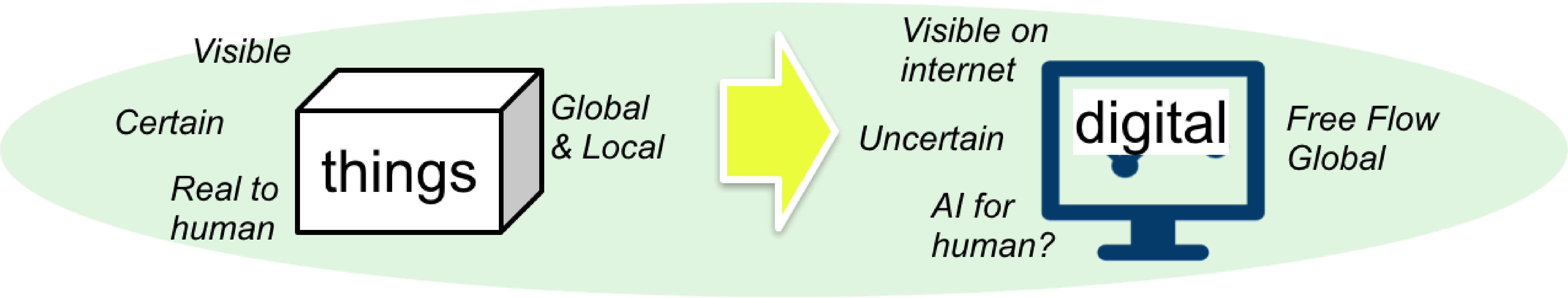
The MDR team: developers, publishers, researchers - at NIMS Library

1. Context: NIMS & the MDR

Mikiko Tanifuji

A landscape of research data – G20 Digital Economy

- G20 - Trade and Digital Economy, June 8, 2019
  - Human Centric Future Society



- “Data Free Flow with Trust” (DFFT concept)
  - Accumulate data for human society
  - Appropriate data management and global consensus for how-to-use

G20 Ministerial Meeting on Trade and Digital Economy
June 8-9, 2019, Tsukuba, Ibaraki, Japan
G20 2019 JAPAN

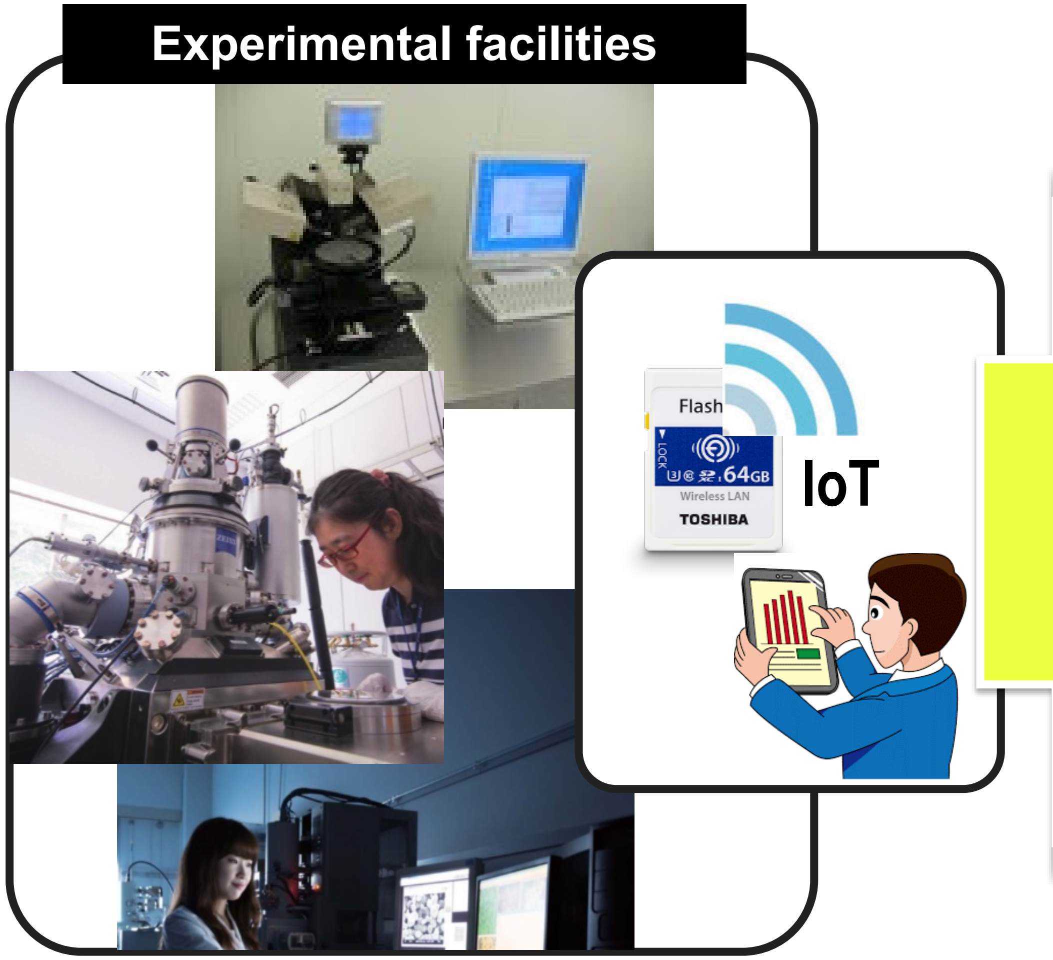
MDR Development Project – Why?

1. A new trend “Data-driven science” >> data science/scientists
2. Not just “machine-readable”, move to machine-actionable >> really FAIR
3. Incentives of “machine-learning” >> must WebAPI, with metadata
4. Not just a database >> semantic-aware database
5. Not just an archive >> metadata, machine-readable formats, analytics tools

1. Next Generation Repository (NGR) must have machine-actionable data
2. NGR must have researchers’ trust-based quality data
3. NGR should/could be repository-tenant concept Example: res project repository

MDR Development Project - What?

Experimental facilities



DMP

RDM

Vocabulary

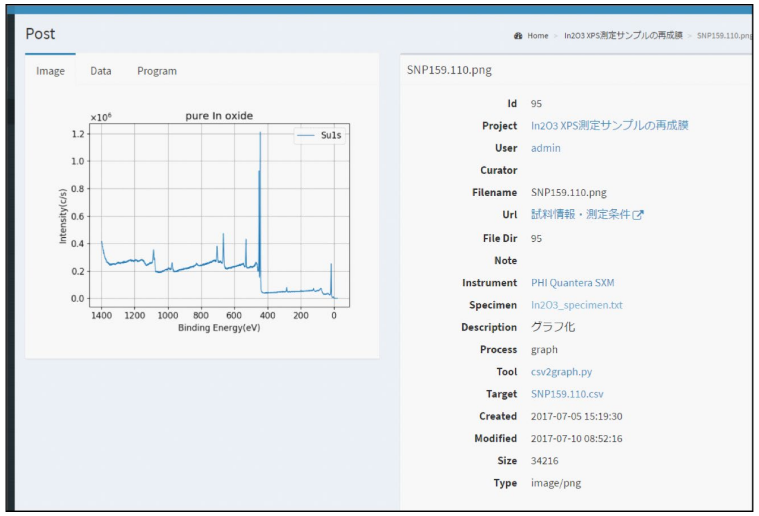
PID

O/C

Data cloud

Data repository

MDR MATERIALS DATA REPOSITORY



Post

Home > In2O3 XPS測定サンプルの再成膜 > SNP159.110.png

Image | Data | Program

SNP159.110.png

Id	95
Project	In2O3 XPS測定サンプルの再成膜
User	admin
Curator	
Filename	SNP159.110.png
Url	試料情報・測定条件
File Dir	95
Note	
Instrument	PHI Quantera SXM
Specimen	In2O3_specimen.txt
Description	グラフ化
Process	graph
Tool	csv2graph.py
Target	SNP159.110.csv
Created	2017-07-05 15:19:30
Modified	2017-07-10 08:52:16
Size	34216
Type	image/png

MDR - a FAIR system of Materials Data Platform

2019 -

NIMS service
DCS
Data Curation System

Public service
MDR MATERIALS DATA REPOSITORY

2020 -

Public service
VocWiki
Vocabulary for Data Management

NIMS service
IoT Data
IoT Data Transferring System

Public service
M-DaC
Materials Data Conversion Tools

NIMS service
RDM
Research Data Management

NIMS service
LabNote
Online Lab Notebooks

NIMS service
TDM NIMS Materials Data Platform for Text and Data Mining

NIMS service
Single Sign-on
A gateway to all data services

NIMS service
Analytics
High performance computer system

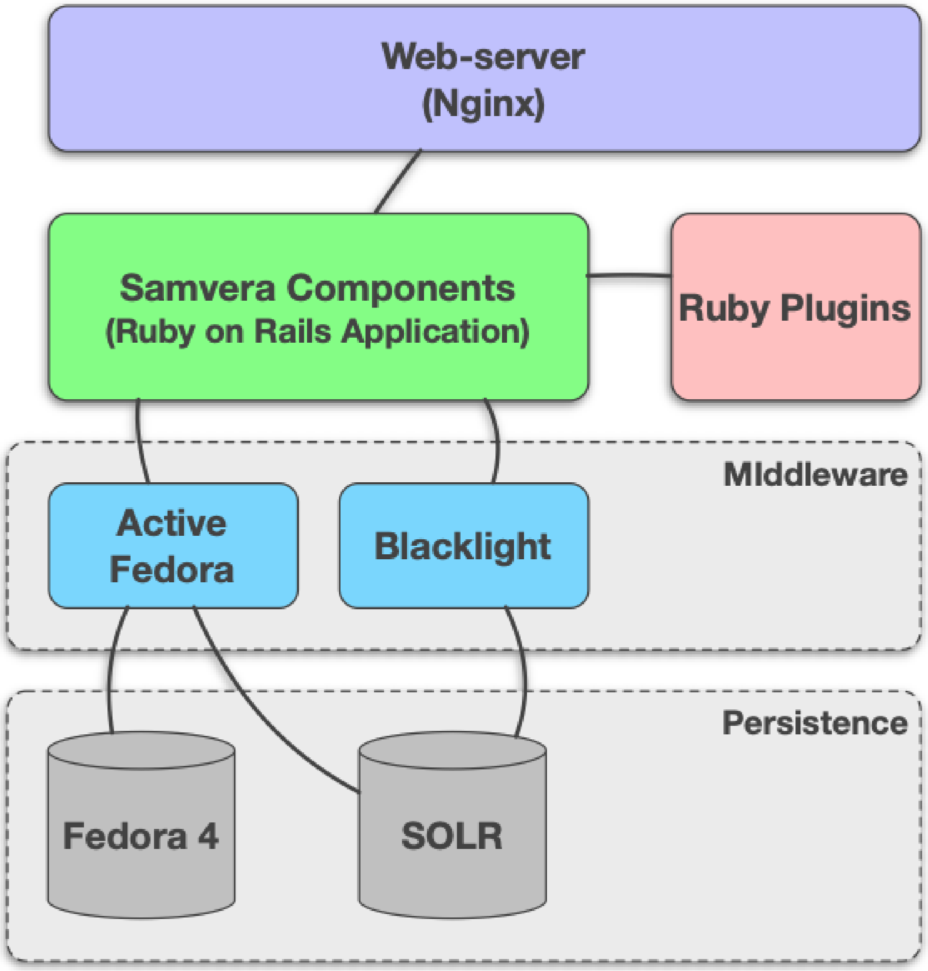
**Data deposit | Data deposit via IoT |
Data search | Data download | Data visualizations |
Data analytics & Informatics**

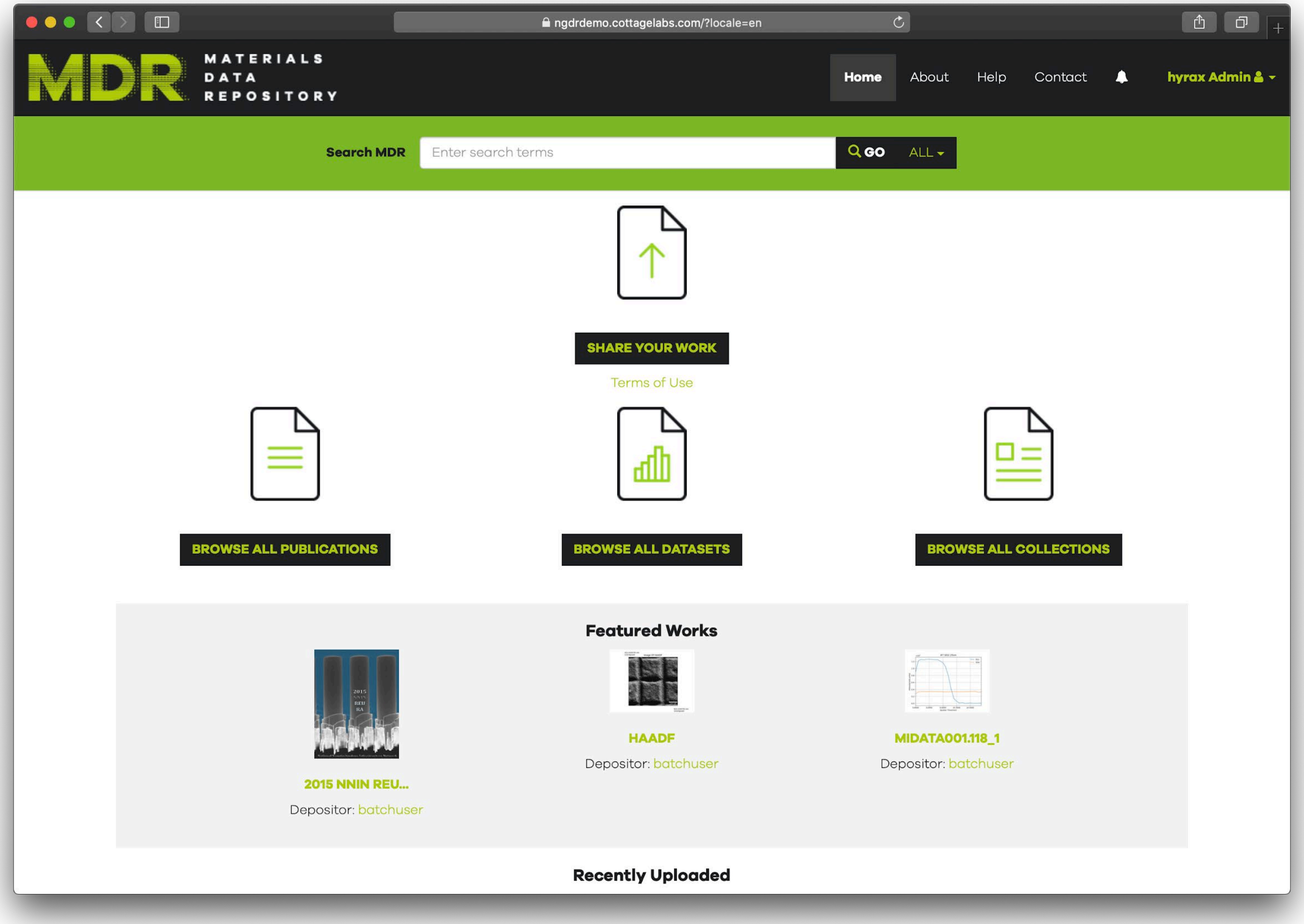
2. The MDR system

Steven Eardley

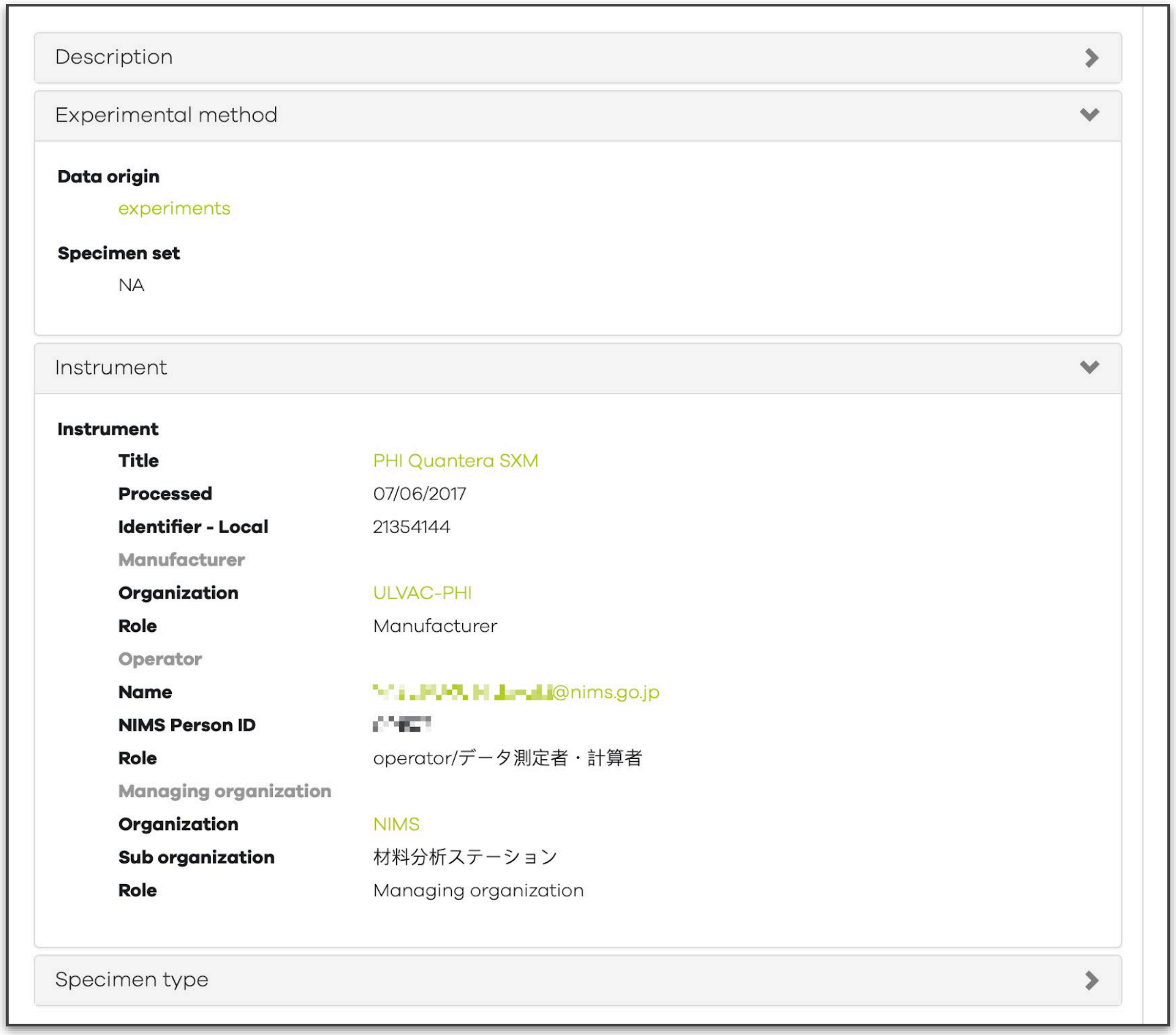
About the Materials Data Repository (MDR)

- Hyrax (Samvera)





Nested View



Description

Experimental method

Data origin
experiments

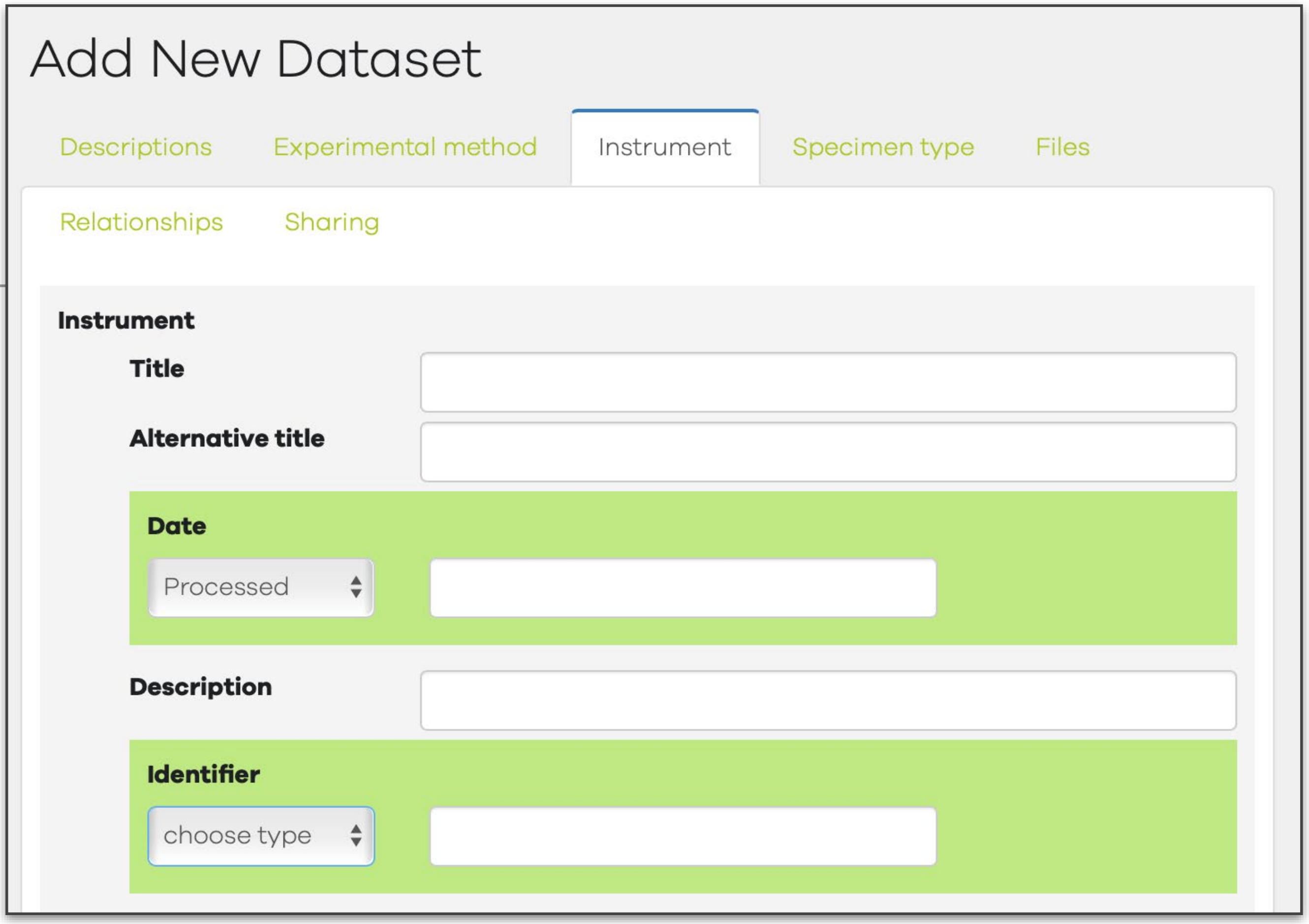
Specimen set
NA

Instrument

Instrument

Title	PHI Quantera SXM
Processed	07/06/2017
Identifier - Local	21354144
Manufacturer	
Organization	ULVAC-PHI
Role	Manufacturer
Operator	
Name	[redacted]@nims.go.jp
NIMS Person ID	[redacted]
Role	operator/データ測定者・計算者
Managing organization	
Organization	NIMS
Sub organization	材料分析ステーション
Role	Managing organization

Specimen type



Add New Dataset

Instrument

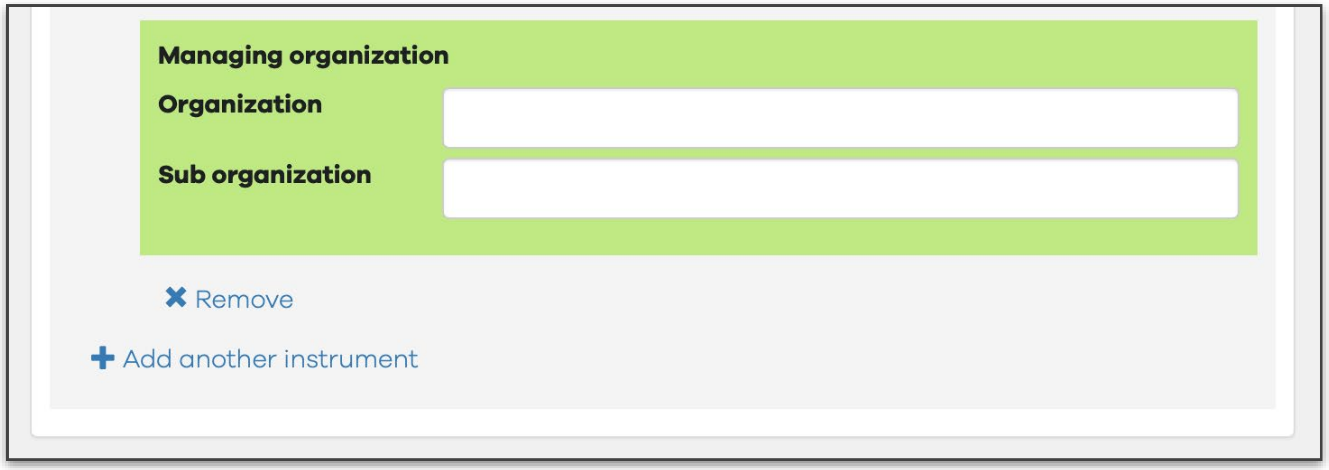
Title

Alternative title

Date
Processed

Description

Identifier
choose type

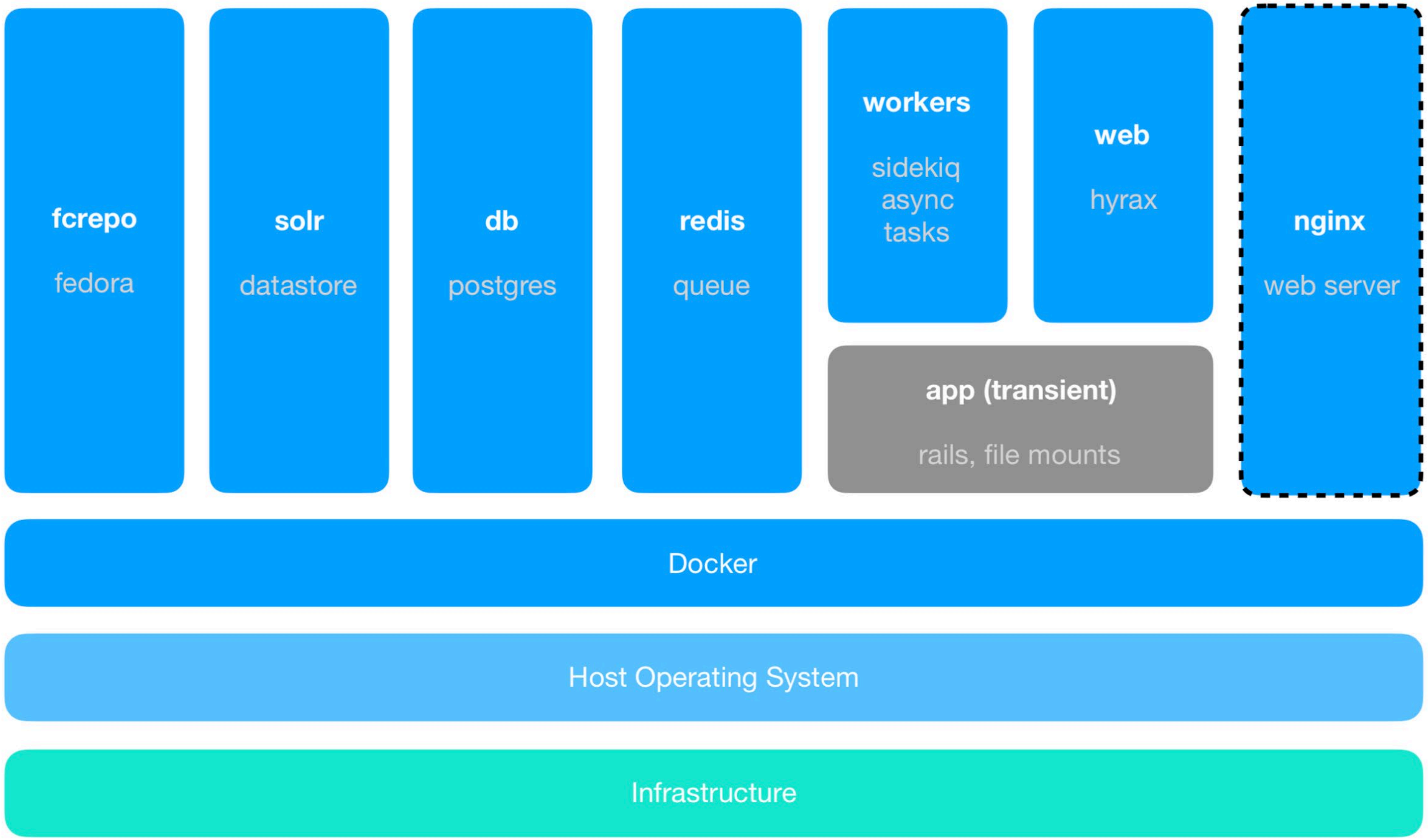


Managing organization

Organization

Sub organization

Containerised Development and Deployment



fcrepo

fedora

solr

datastore

db

postgres

redis

queue

workers

sidekiq
async
tasks

web

hyrax

app (transient)

rails, file mounts

nginx

web server

Docker

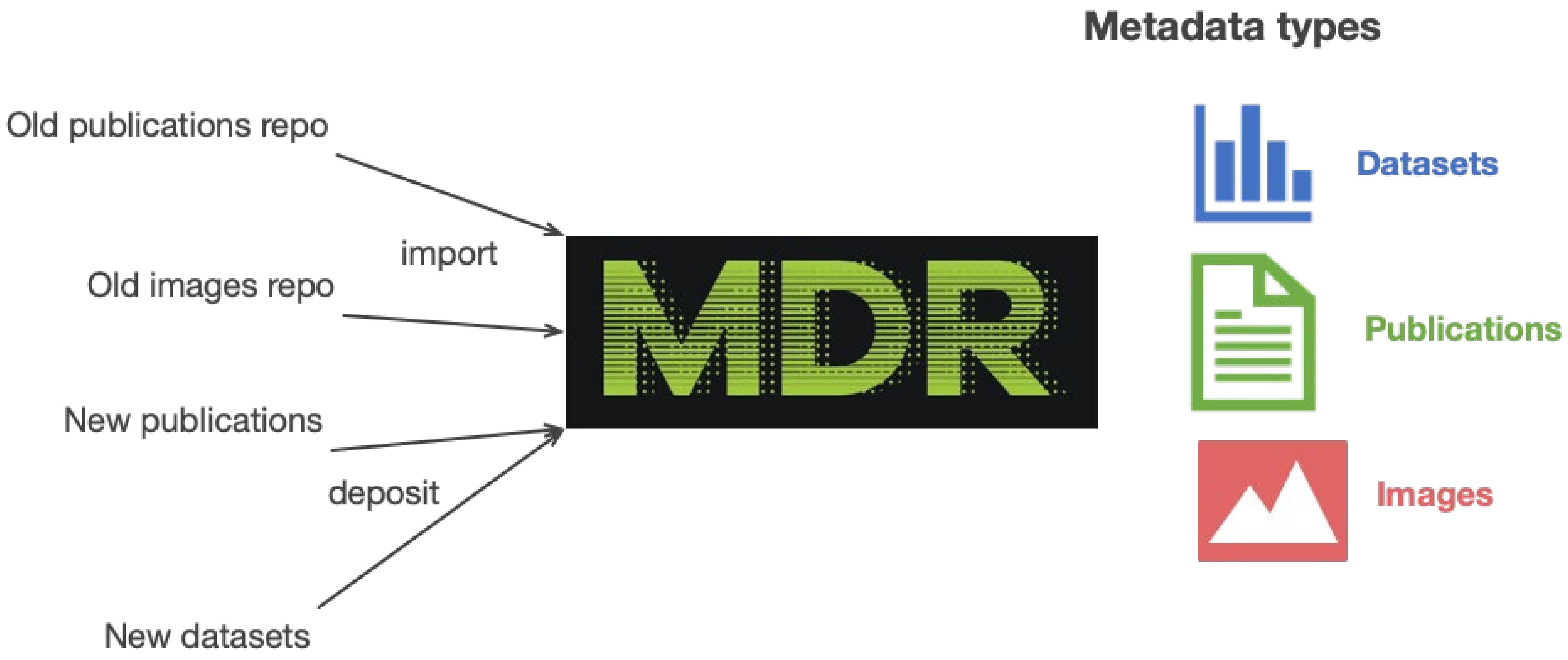
Host Operating System

Infrastructure

3. A focus on metadata

Asahiko Matsuda

Datasets, publications, & images coexisting in MDR



Old publications repo

Old images repo

import

New publications

deposit

New datasets

Metadata types

Datasets

Publications

Images

Metadata for...

Publications

- Title
- Authors
- Publication
- Issue
- Date
- ...

Datasets

- Method
- Specimen
- Facility
- Temperature
- Acceleration energy
- ...

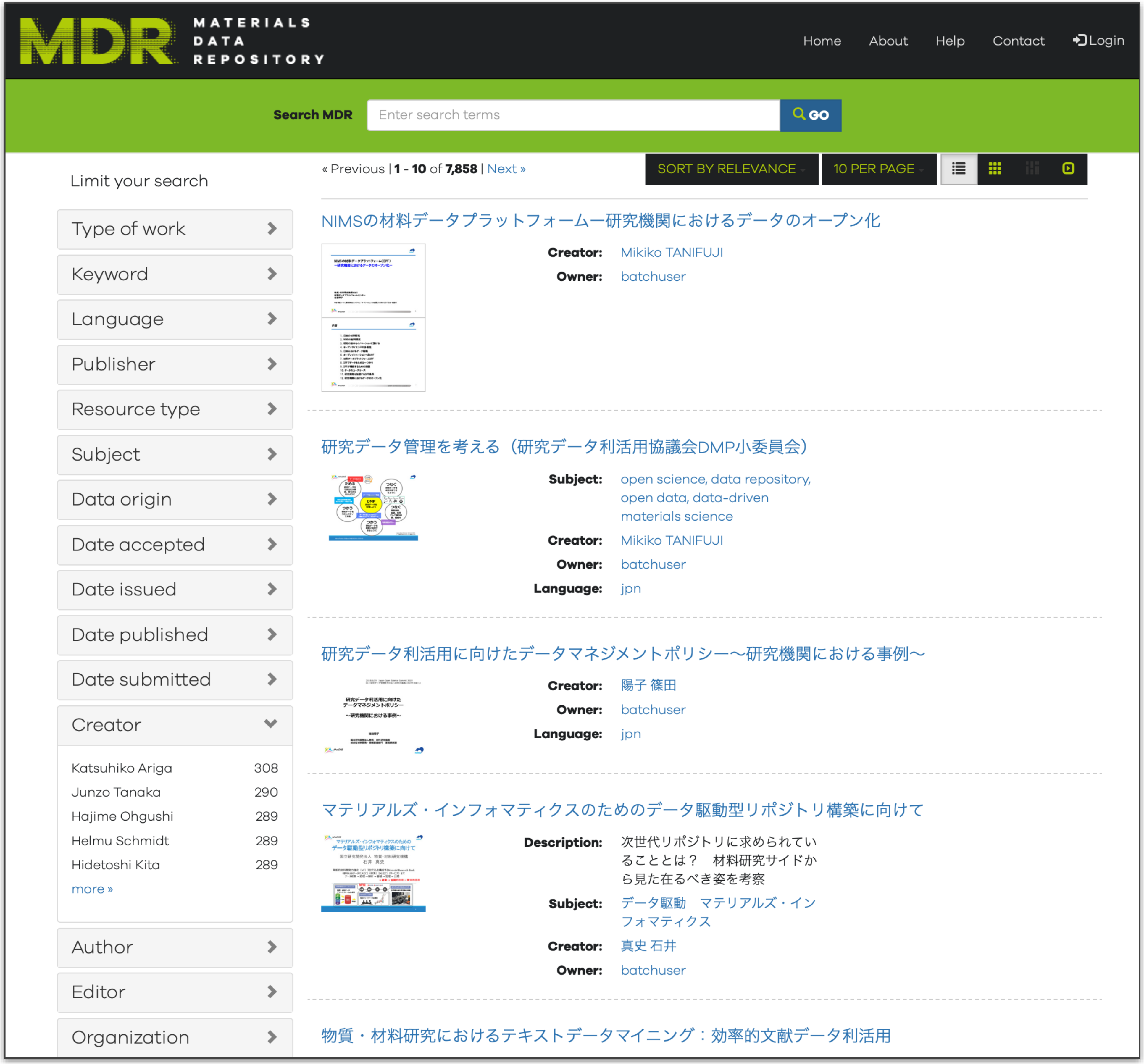
Extremely domain-specific !
How can we model this ?

Tiered and nested metadata model for datasets

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

Metadata view and deposit form also reflect this model

Metadata used for faceted browsing & searching



MATERIALS DATA REPOSITORY

Home About Help Contact Login

Search MDR | Enter search terms | GO

« Previous | **1** - **10** of **7,858** | Next »

SORT BY RELEVANCE | 10 PER PAGE

Limit your search

- Type of work
- Keyword
- Language
- Publisher
- Resource type
- Subject
- Data origin
- Date accepted
- Date issued
- Date published
- Date submitted
- Creator

Katsuhiko Ariga	308
Junzo Tanaka	290
Hajime Ohgushi	289
Helmu Schmidt	289
Hidetoshi Kita	289

more »

- Author
- Editor
- Organization

NIMSの材料データプラットフォーム―研究機関におけるデータのオープン化

Creator: Mikiko TANIFUJI
Owner: batchuser

研究データ管理を考える（研究データ利活用協議会DMP小委員会）

Subject: open science, data repository, open data, data-driven materials science
Creator: Mikiko TANIFUJI
Owner: batchuser
Language: jpn

研究データ利活用に向けたデータマネジメントポリシー～研究機関における事例～

Creator: 陽子 篠田
Owner: batchuser
Language: jpn

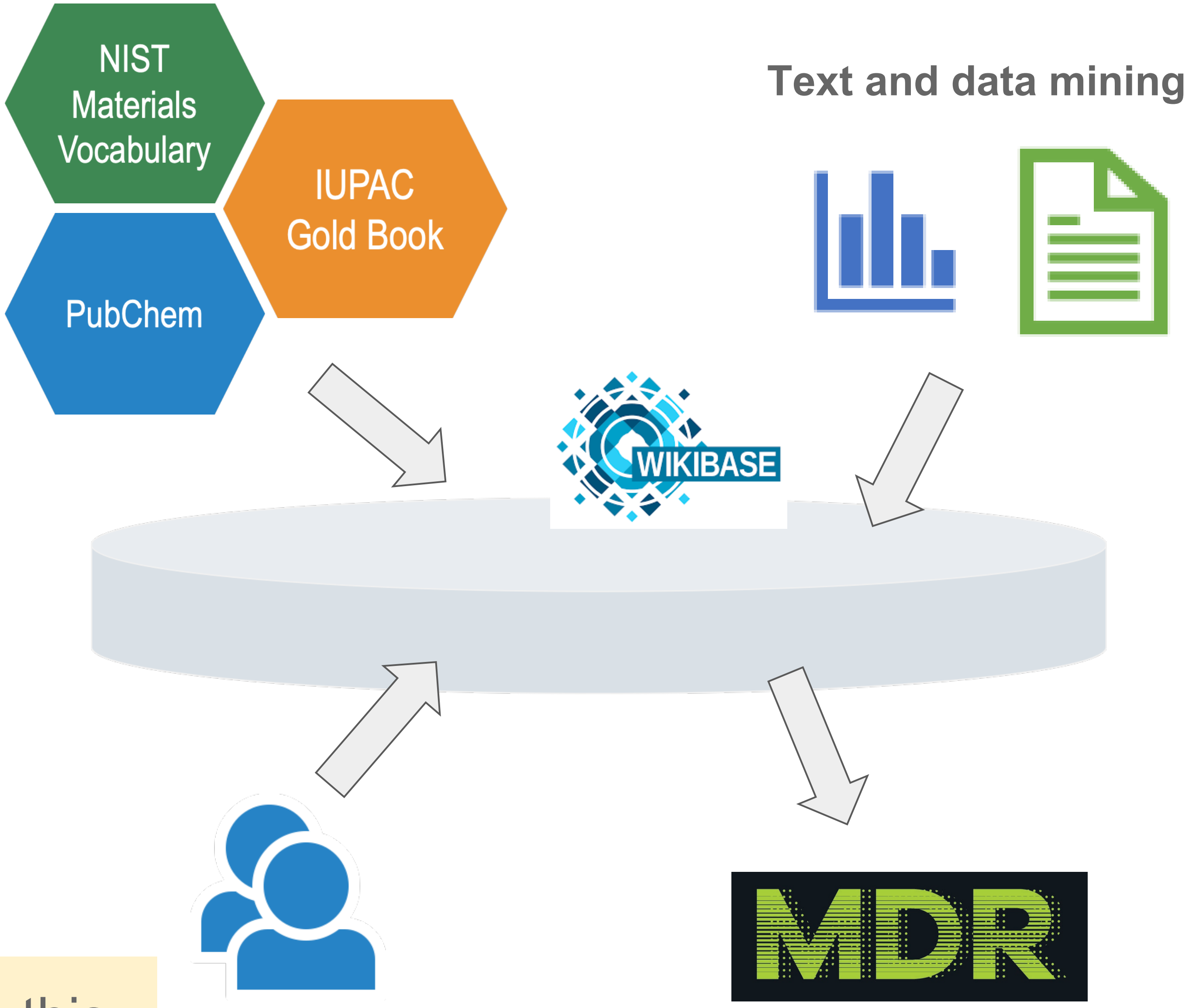
マテリアルズ・インフォマティクスのためのデータ駆動型リポジトリ構築に向けて

Description: 次世代リポジトリに求められていることとは？　材料研究サイドから見た在るべき姿を考察
Subject: データ駆動　マテリアルズ・インフォマティクス
Creator: 真史 石井
Owner: batchuser

物質・材料研究におけるテキストデータマイニング：効率的文献データ利活用

Enriching metadata with vocabularies

- 3 sources of vocabulary terms:
  1. Controlled vocabularies
     - Community governed
  2. Machine-generated
     - Terms extracted by text/data-mining
  3. Crowd-sourced
     - User-generated terms
     - From NIMS research community
     - "Folksonomy"

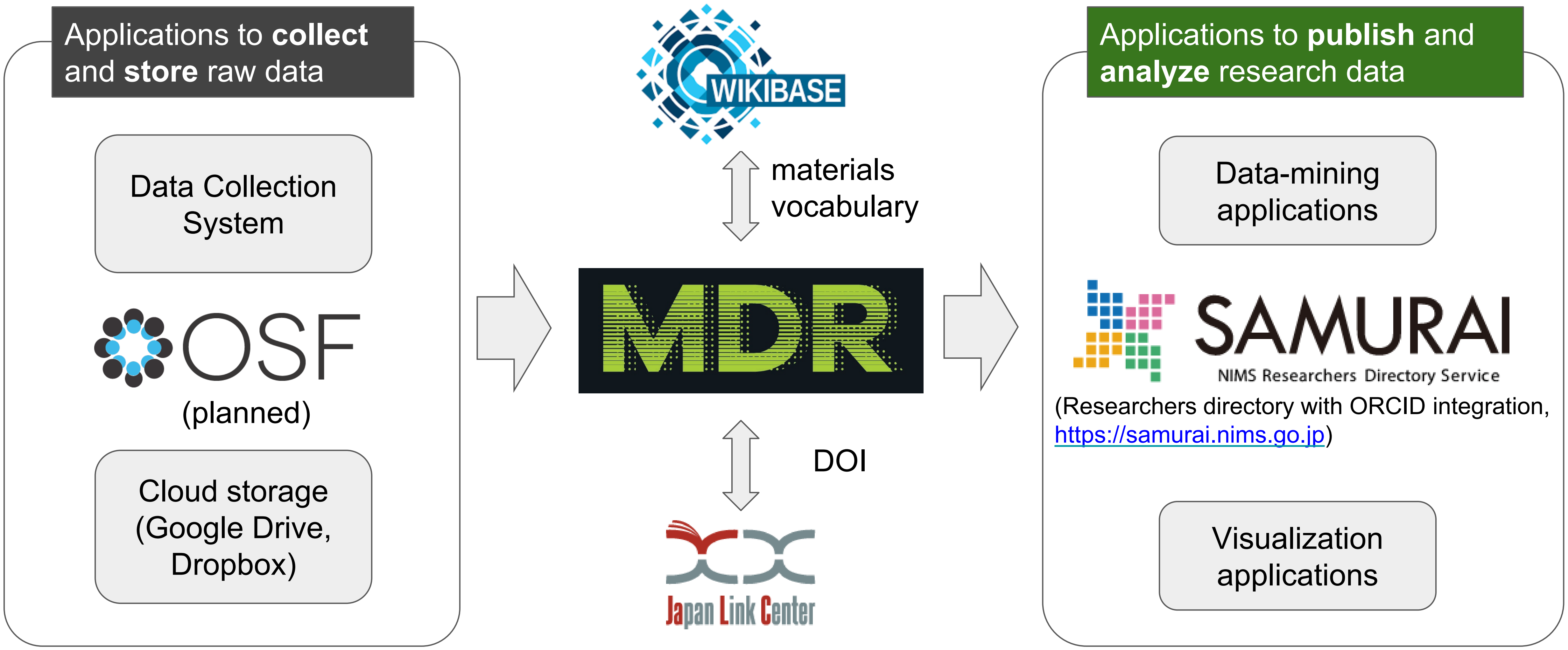


We have a separate poster focusing on this.

4. Integration

Kosuke Tanabe

Overview of integrations



Applications to **collect** and **store** raw data

Data Collection System

OSF (planned)

Cloud storage (Google Drive, Dropbox)

WIKIBASE

materials vocabulary

MDR

DOI

Japan Link Center

Applications to **publish** and **analyze** research data

Data-mining applications

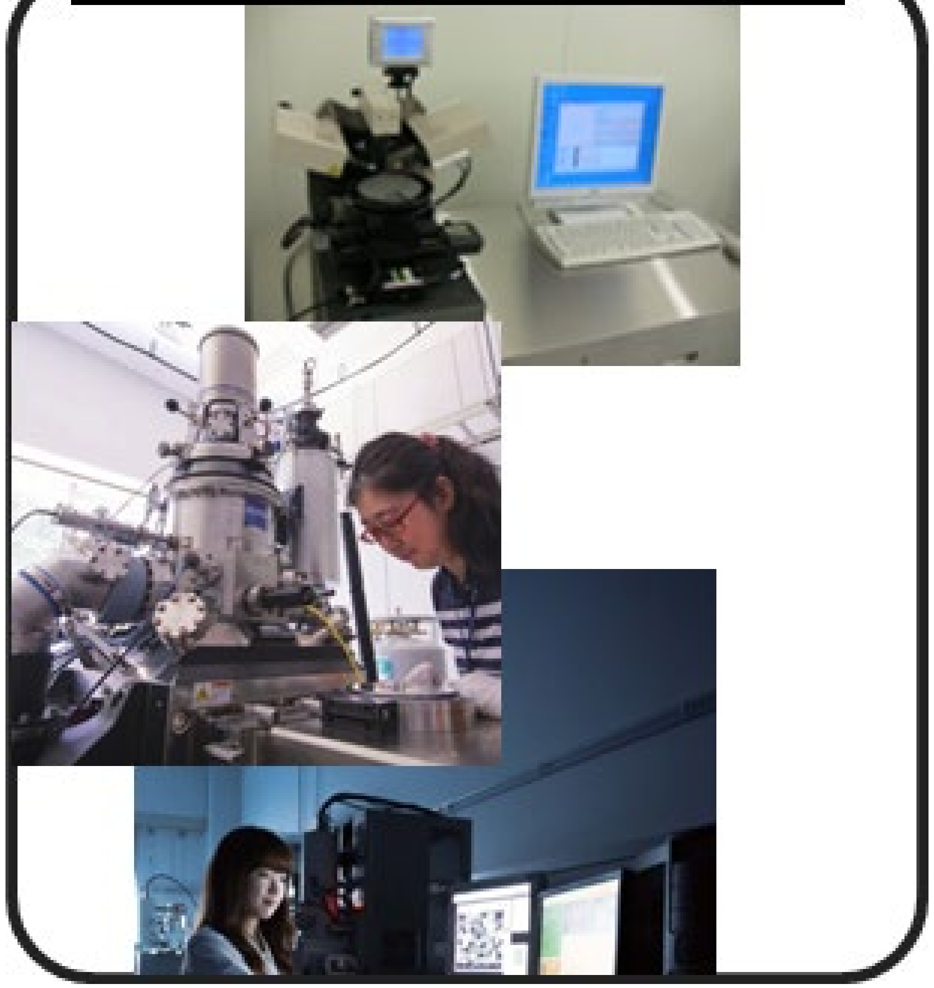
SAMURAI NIMS Researchers Directory Service

(Researchers directory with ORCID integration, https://samurai.nims.go.jp)

Visualization applications

Use case for depositing experimental data

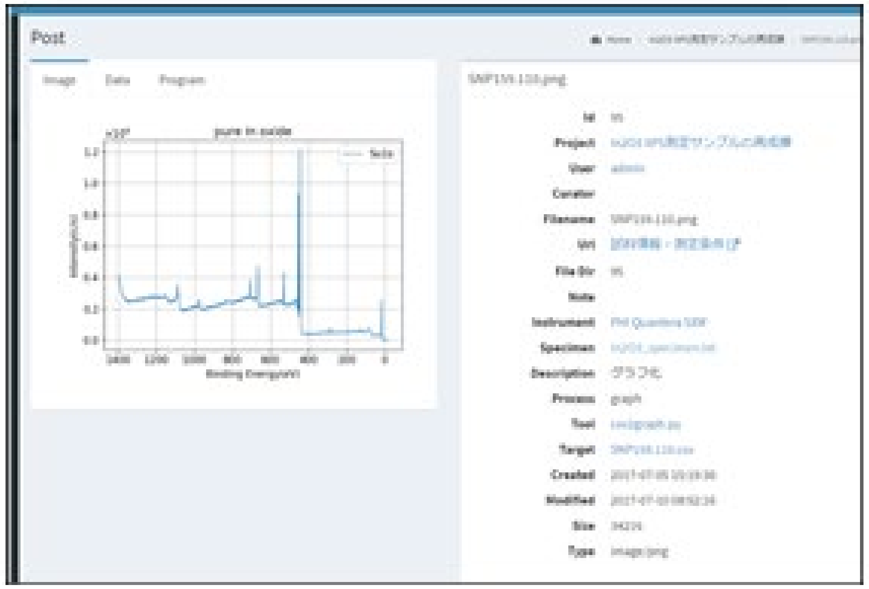
Experimental facilities



Deposit

Data repository

MDR MATERIALS DATA REPOSITORY



Post

Home > [illegible] > [illegible]

Image Data Program

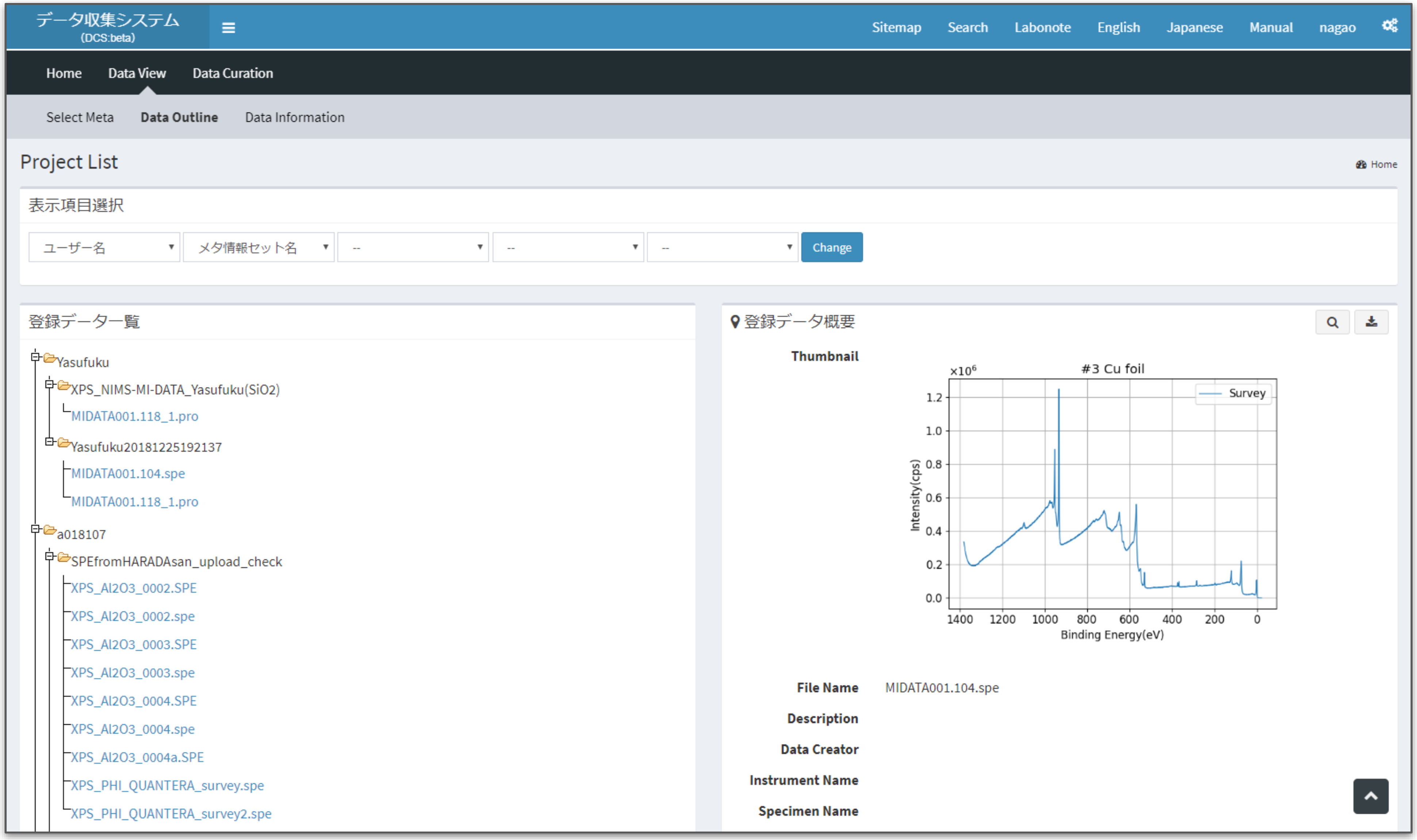
SNP159-110.png

Id	95
Project	[illegible]
User	admin
Curator	
Filename	SNP159-110.png
Url	[illegible]
File Dir	95
Note	
Instrument	PHI Quantera SXM
Specimen	[illegible]
Description	グラフ化
Process	graph
Tool	[illegible]
Target	SNP159-110.csv
Created	2017-07-05 [illegible]
Modified	2017-07-10 [illegible]
Size	[illegible]
Type	image/png

Data Collection System (DCS)

- A system to convert raw measurement data, assign metadata, draw a graph, and hand them over to MDR

- NIMS researchers' home-grown application



Project List

表示項目選択

ユーザー名 ▾ | メタ情報セット名 ▾ | -- ▾ | -- ▾ | -- ▾

登録データ一覧

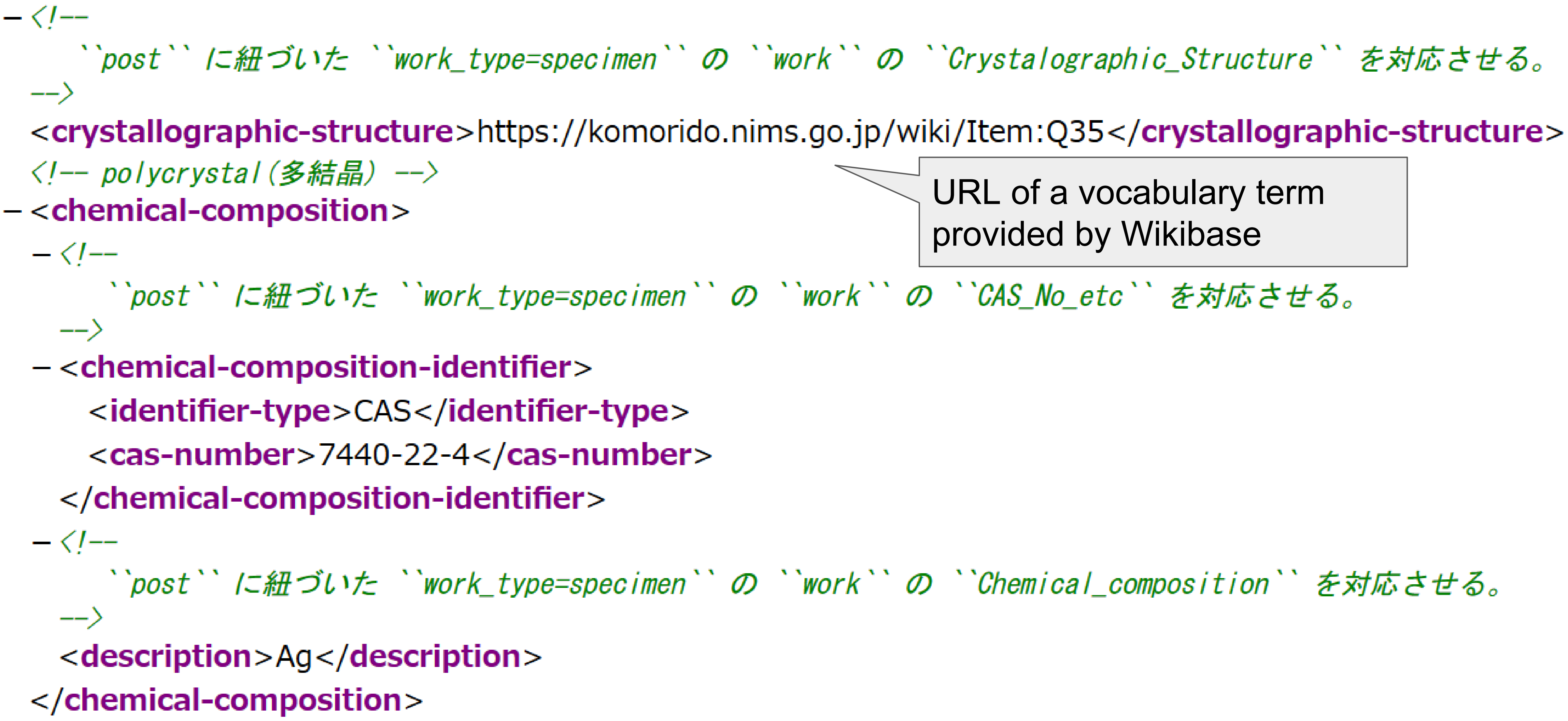
- Yasufuku
  - XPS_NIMS-MI-DATA_Yasufuku(SiO2)
    - MIDATA001.118_1.pro
  - Yasufuku20181225192137
    - MIDATA001.104.spe
    - MIDATA001.118_1.pro
- a018107
  - SPEfromHARADAsan_upload_check
    - XPS_Al2O3_0002.SPE
    - XPS_Al2O3_0002.spe
    - XPS_Al2O3_0003.SPE
    - XPS_Al2O3_0003.spe
    - XPS_Al2O3_0004.SPE
    - XPS_Al2O3_0004.spe
    - XPS_Al2O3_0004a.SPE
    - XPS_PHI_QUANTERA_survey.spe
    - XPS_PHI_QUANTERA_survey2.spe

登録データ概要

Thumbnail

File Name: MIDATA001.104.spe
Description:
Data Creator:
Instrument Name:
Specimen Name:

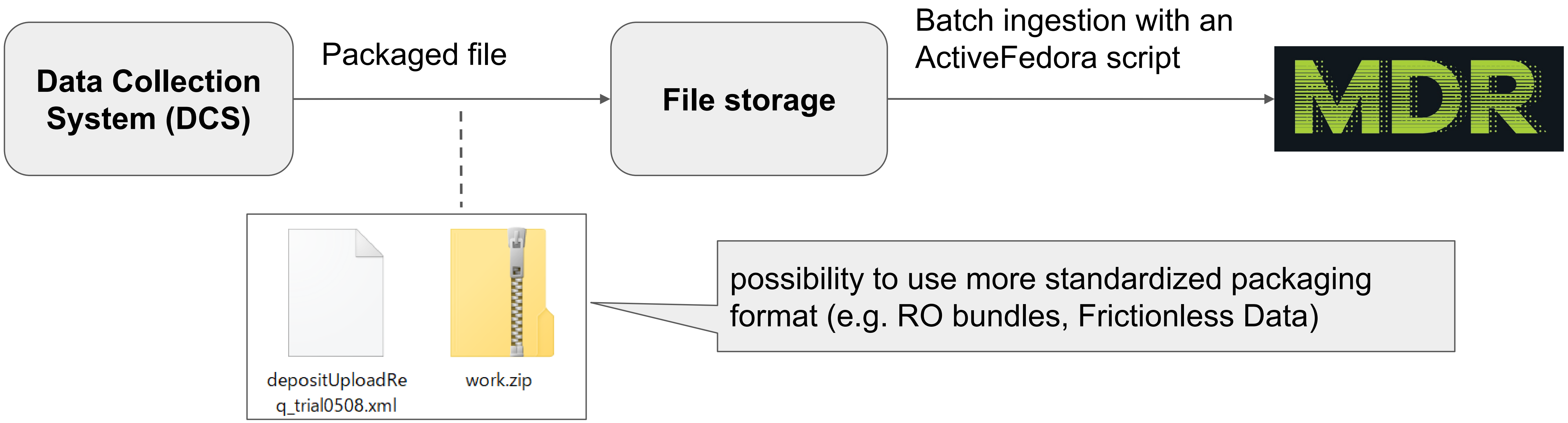
Metadata from DCS to MDR



```
– <!--
    ``post`` に紐づいた ``work_type=specimen`` の ``work`` の ``Crystalographic_Structure`` を対応させる。
  -->
  <crystallographic-structure>https://komorido.nims.go.jp/wiki/Item:Q35</crystallographic-structure>
  <!-- polycrystal(多結晶) -->
– <chemical-composition>
  – <!--
      ``post`` に紐づいた ``work_type=specimen`` の ``work`` の ``CAS_No_etc`` を対応させる。
    -->
  – <chemical-composition-identifier>
      <identifier-type>CAS</identifier-type>
      <cas-number>7440-22-4</cas-number>
    </chemical-composition-identifier>
  – <!--
      ``post`` に紐づいた ``work_type=specimen`` の ``work`` の ``Chemical_composition`` を対応させる。
    -->
    <description>Ag</description>
  </chemical-composition>
```

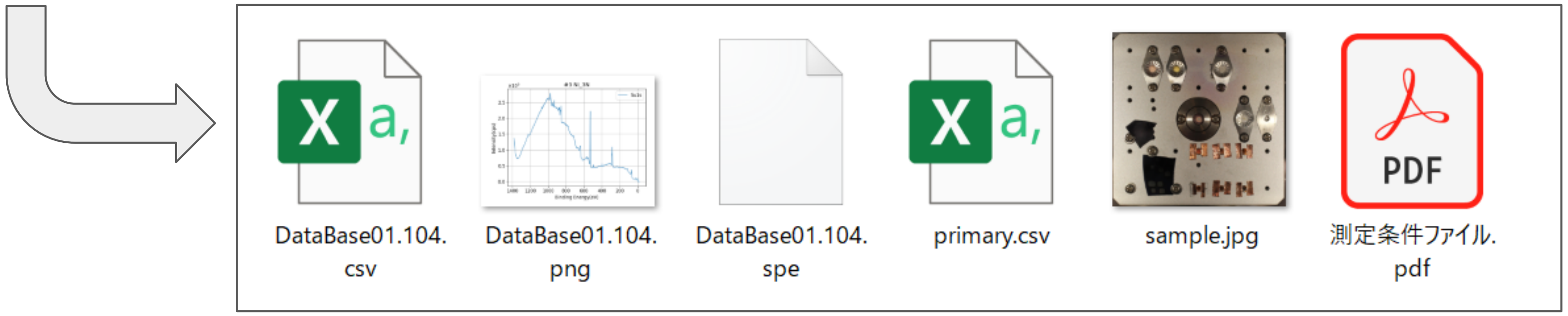
URL of a vocabulary term provided by Wikibase

Dataflow between DCS and MDR


Data Collection System (DCS)
Packaged file
File storage
Batch ingestion with an ActiveFedora script
MDR
depositUploadReq_trial0508.xml
work.zip

possibility to use more standardized packaging format (e.g. RO bundles, Frictionless Data)

- XML metadata file
- Zipped data file



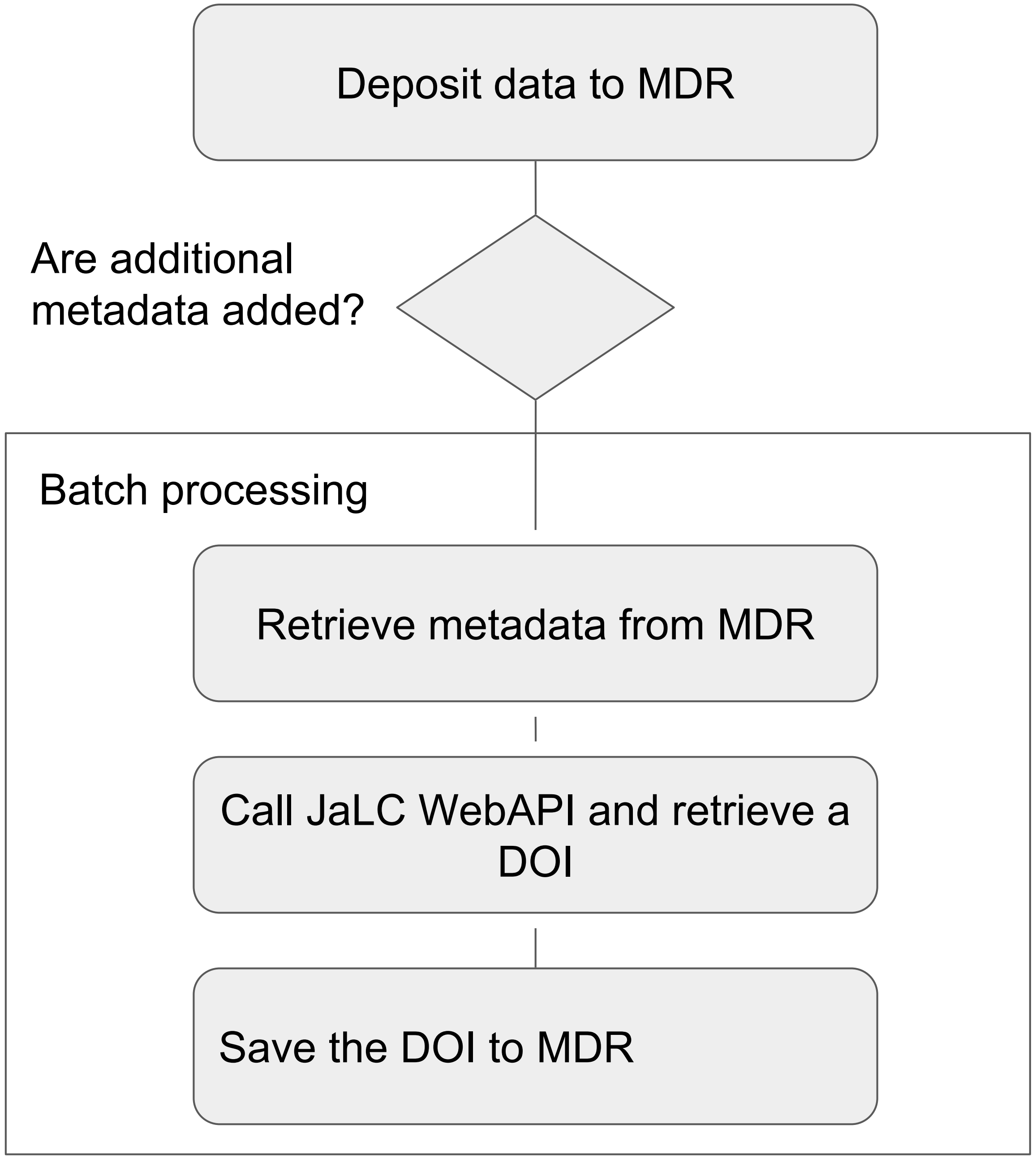
Integration with DOI Registration System

- MDR supports JaLC DOI

https://japanlinkcenter.org/
(DOI RA in Japan)

- Only datasets with both mandatory and domain-specific metadata will be minted DOIs

- The DOI minting is processed by a batch script invoked by MDR

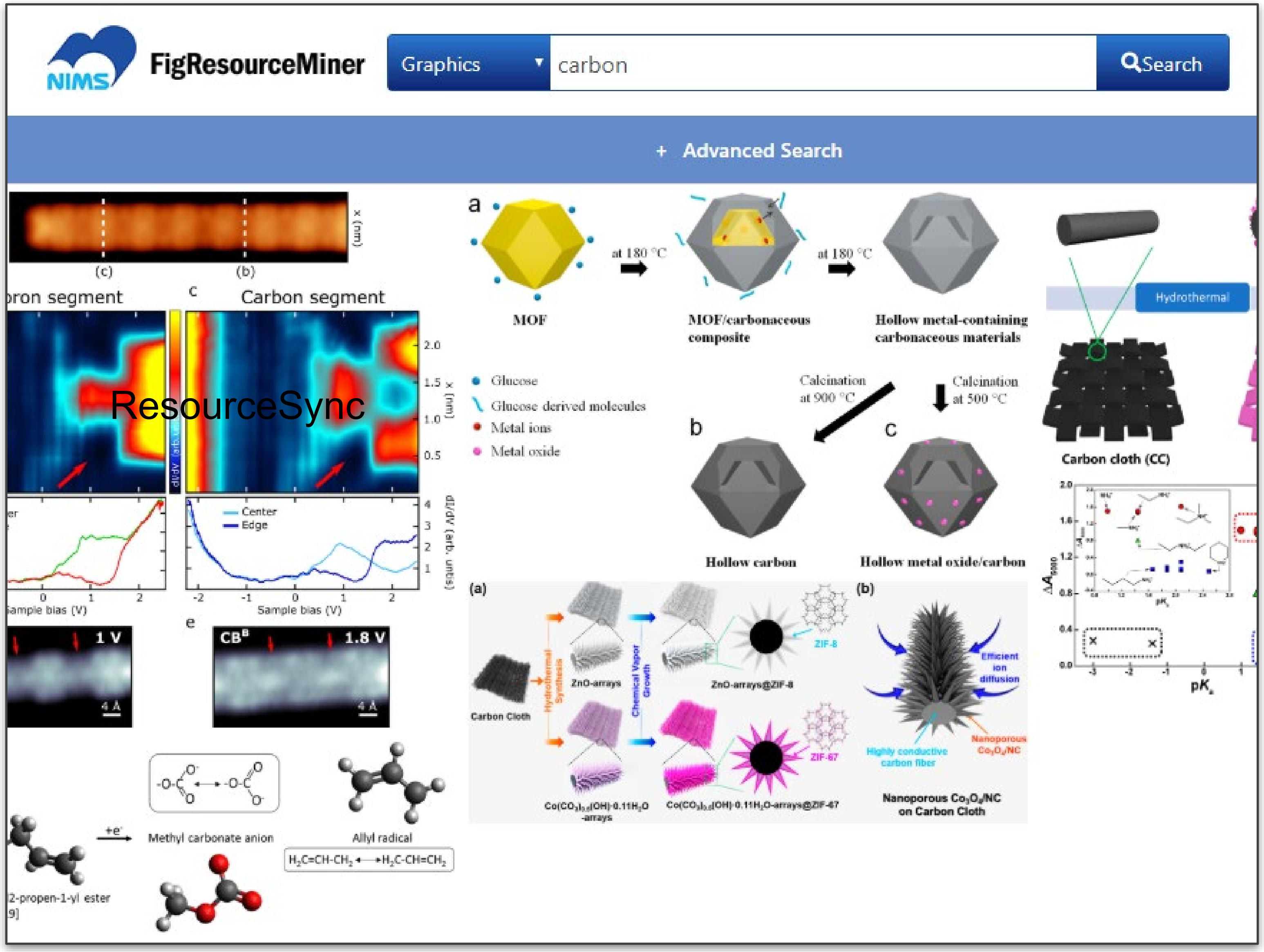

Deposit data to MDR
Are additional metadata added?
Batch processing
Retrieve metadata from MDR
Call JaLC WebAPI and retrieve a DOI
Save the DOI to MDR

Application using data on MDR: FigResourceMiner

- Data mining service
- Extract text information from figures and images in articles and datasets
- FigResourceMiner harvests files from MDR

MDR

ResourceSync



ResourceSync

Challenge in integration

- Depositing huge data from collaborators outside NIMS network
  - Sometimes over 4TB
  - Collaborators are expected to deposit those data to their local repository, then we can harvest metadata for search
  - Don’t we need actual data (not just metadata) for data mining?

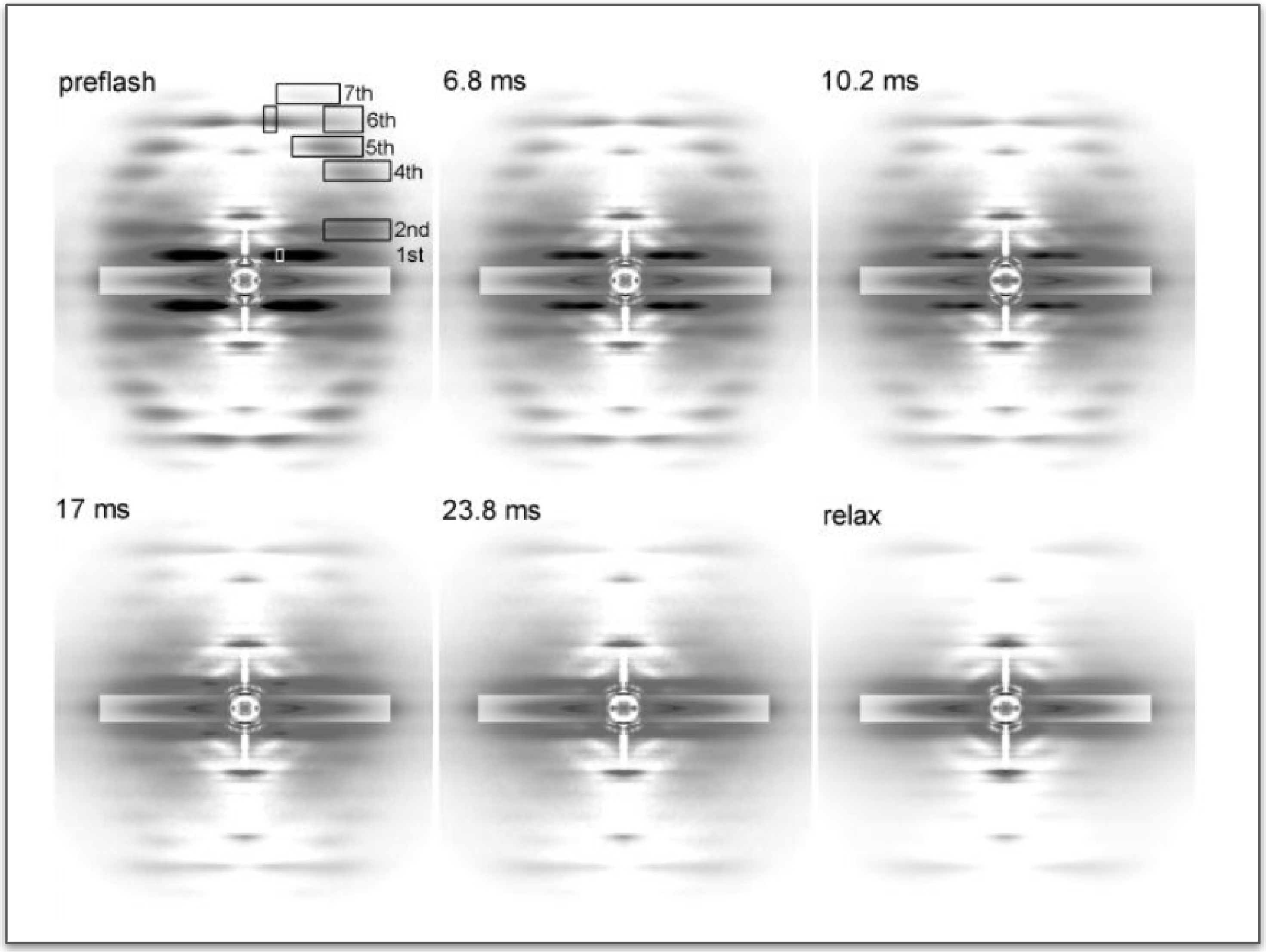


Image data files generated by the X-ray beamline in SPring-8, located outside NIMS

http://www.spring8.or.jp/wkg/BL40XU/solution/lang/SOL-0000001622

5. Supporting discovery

Paul Walk

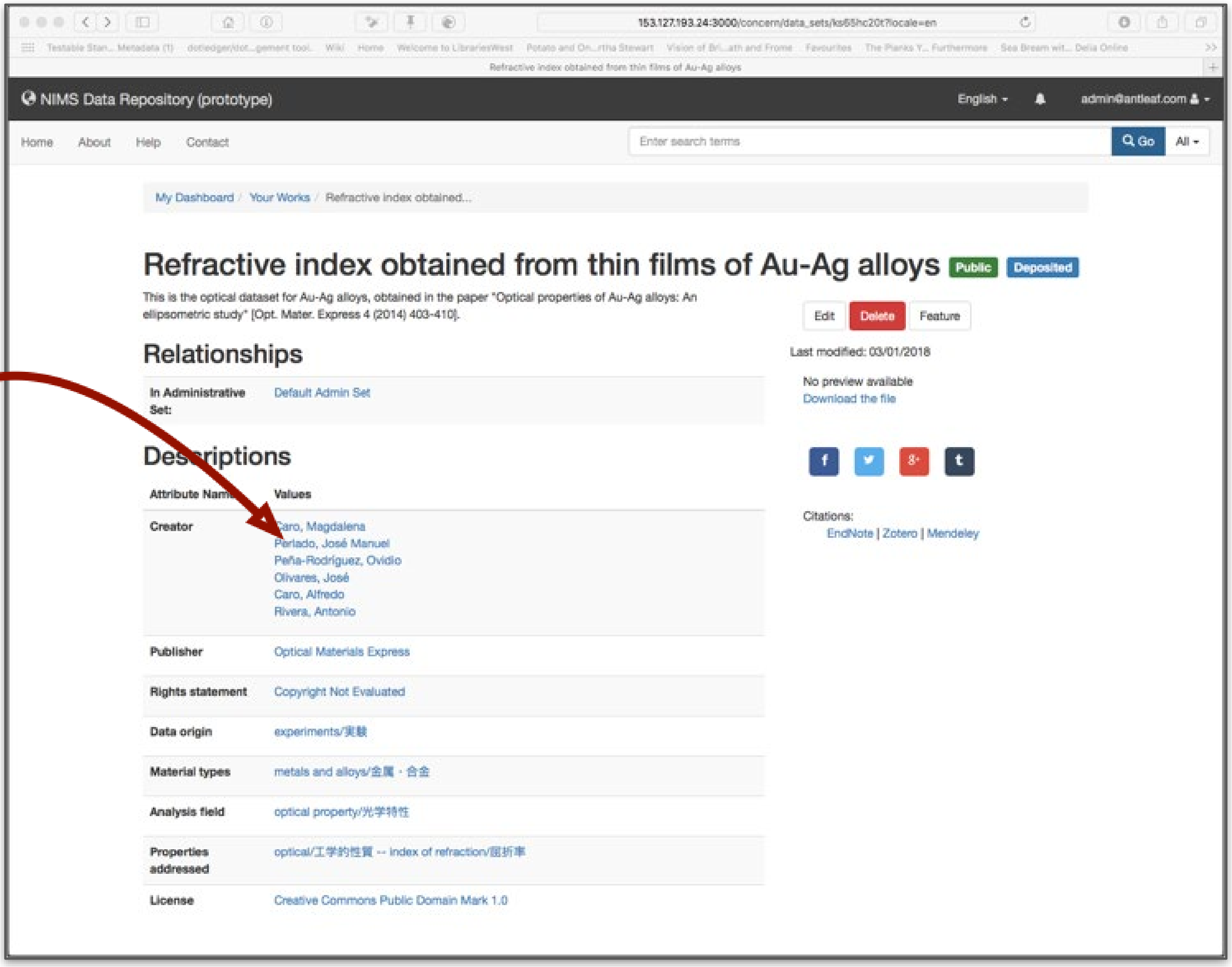
COAR and *Next Generation Repositories*

- Defined "behaviours":
    - Exposing Identifiers
    - Declaring Licenses at the Resource Level
    - **Discovery Through Navigation**
    - Interacting with Resources (Annotation, Commentary, and Review)
    - Resource Transfer
    - **Batch Discovery**
    - Collecting and Exposing Activities
    - Identification of Users
    - Authentication of Users
    - Exposing Standardized Usage Metrics
    - Preserving Resources

NEXT GENERATION REPOSITORIES

Discovery Through Navigation (for humans)

- Faceted browsing and searching
- Using vocabulary terms derived from:
    - Controlled vocabularies
    - Terms extracted algorithmically
    - Crowd-sourced keywords


153.127.193.24:3000/concern/data_sets/ks65hc20t?locale=en
Testable Stan... Metadata (1)
dotledger/dot...gement tool.
Wiki
Home
Welcome to LibrariesWest
Potato and On...rtha Stewart
Vision of Bri...ath and Frome
Favourites
The Planks Y... Furthermore
Sea Bream wit... Delia Online
Refractive index obtained from thin films of Au-Ag alloys
NIMS Data Repository (prototype)
English
admin@antleaf.com
Home
About
Help
Contact
Enter search terms
Go
All
My Dashboard / Your Works / Refractive index obtained...
Public
Deposited
Edit
Delete
Feature
f
t

Refractive index obtained from thin films of Au-Ag alloys

This is the optical dataset for Au-Ag alloys, obtained in the paper "Optical properties of Au-Ag alloys: An ellipsometric study" [Opt. Mater. Express 4 (2014) 403-410].

Last modified: 03/01/2018

No preview available
Download the file

Citations:
EndNote | Zotero | Mendeley

Relationships

In Administrative Set:	Default Admin Set

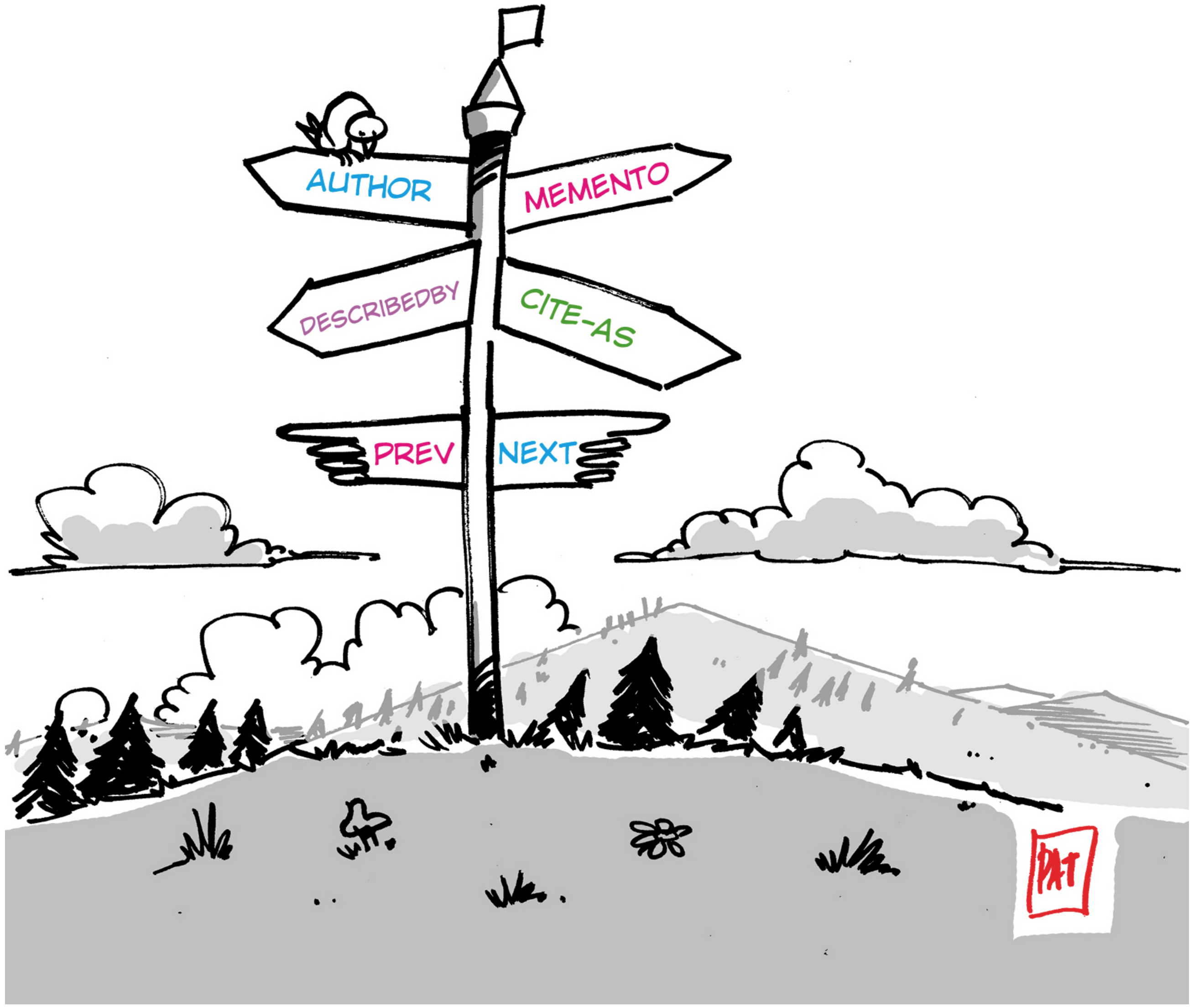
Descriptions

Attribute Name	Values
Creator	Caro, Magdalena Perlado, José Manuel Peña-Rodríguez, Ovidio Olivares, José Caro, Alfredo Rivera, Antonio
Publisher	Optical Materials Express
Rights statement	Copyright Not Evaluated
Data origin	experiments/実験
Material types	metals and alloys/金属・合金
Analysis field	optical property/光学特性
Properties addressed	optical/工学的性質 -- index of refraction/屈折率
License	Creative Commons Public Domain Mark 1.0

Discovery Through Navigation (for machines)

- *Signposting* has defined patterns relating to bibliographic resources:
  - Author
  - Bibliographic Metadata
  - Identifier
  - Publication Boundary
  - Resource Type
- It does define a "dataset" resource type…. but...
- **How do we navigate heterogeneous & complex datasets (multiple files)?**

"Signposting the Scholarly Web"

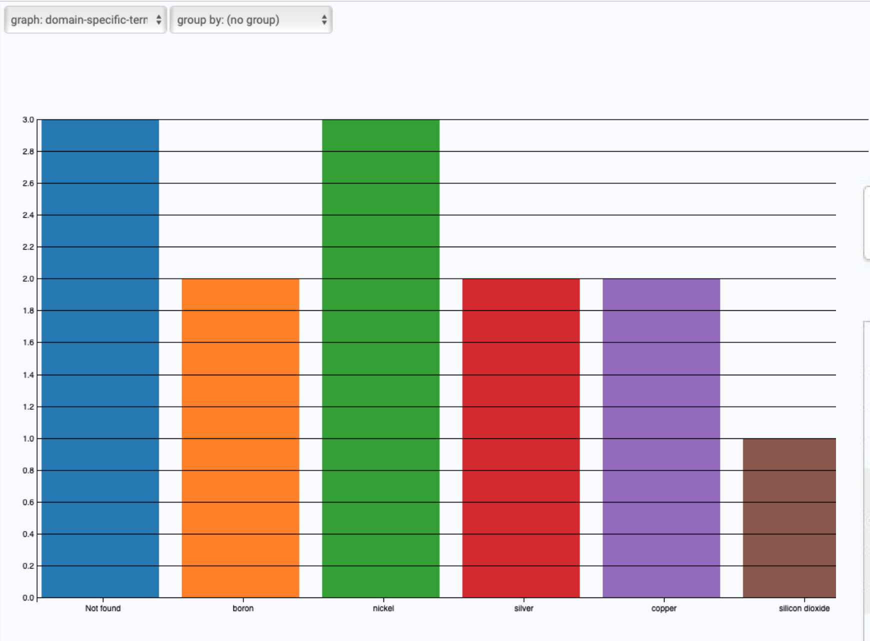


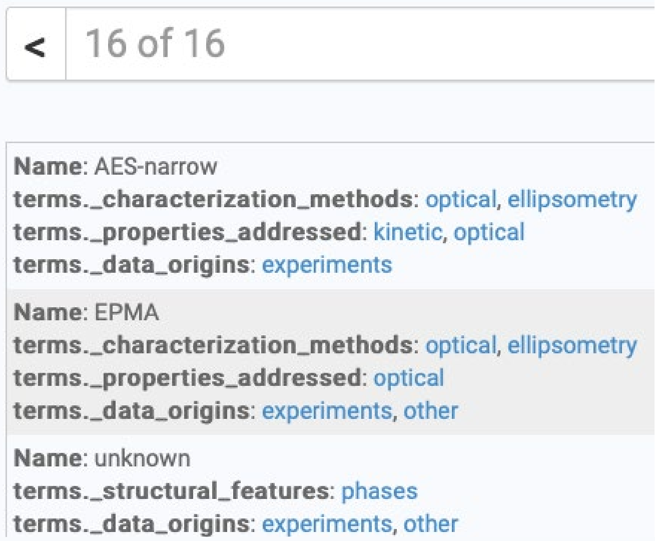
Batch Discovery (1)

- Aggregation is still an important tactic in the "knowledge commons"
  - mitigates network latency and facilitates processing at scale
- Many conceivable services built on research data will require the data to be harvested and aggregated
- OAI-PMH does not support the harvesting of content
- ResourceSync is an important technology for this
- Implemented in the MDR, about to be tested in collaboration with the Open University Core service

Batch Discovery (2)

- Once the data is enabled for batch discovery, many new interfaces, tools etc are possible….

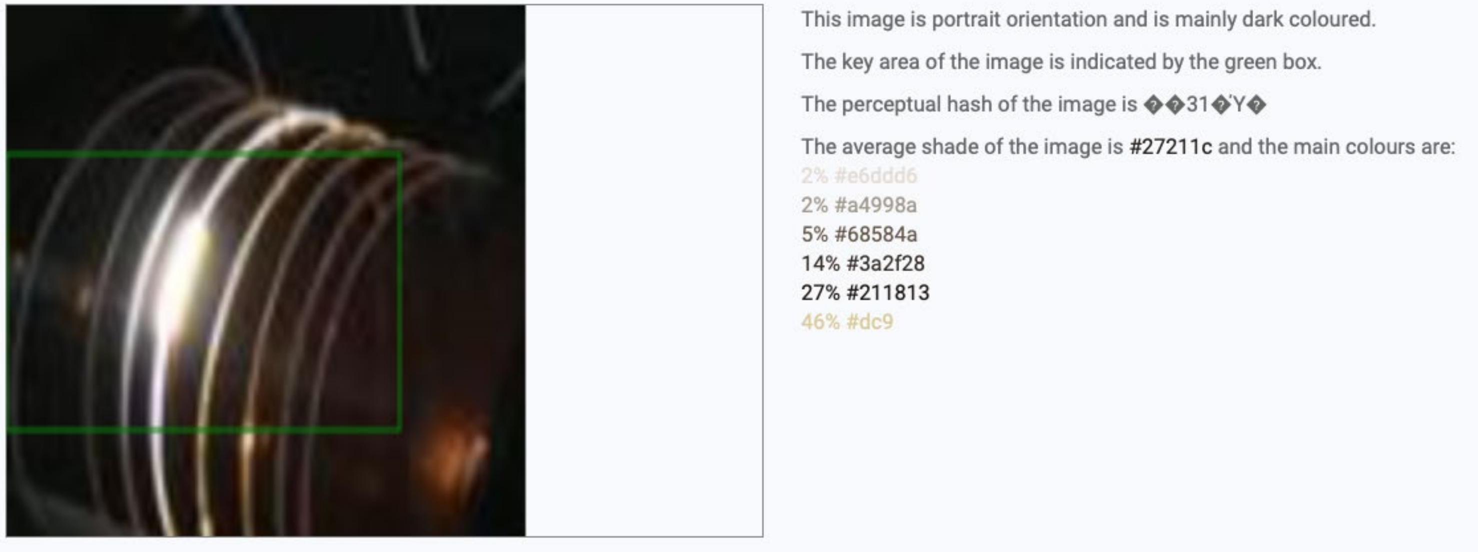




Name: AES-narrow
terms._characterization_methods: optical, ellipsometry
terms._properties_addressed: kinetic, optical
terms._data_origins: experiments

Name: EPMA
terms._characterization_methods: optical, ellipsometry
terms._properties_addressed: optical
terms._data_origins: experiments, other

Name: unknown
terms._structural_features: phases
terms._data_origins: experiments, other

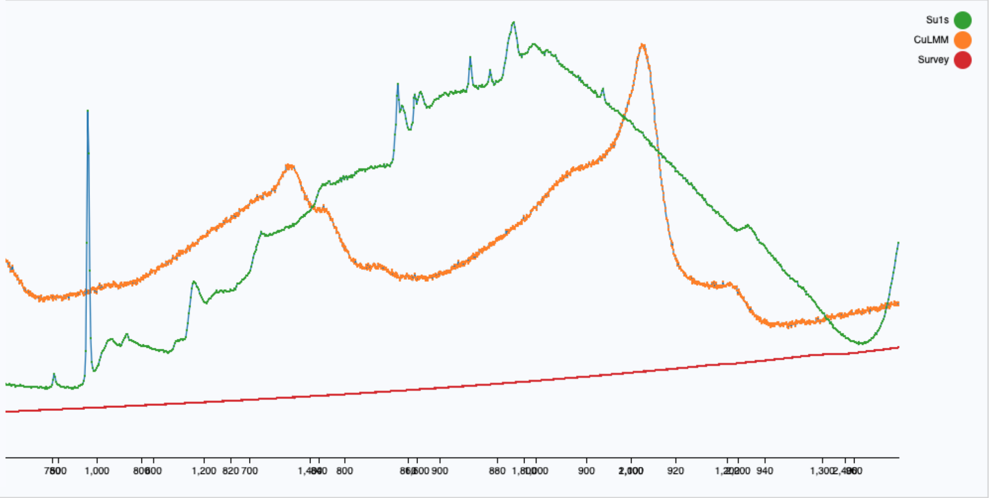


This image is portrait orientation and is mainly dark coloured.

The key area of the image is indicated by the green box.

The perceptual hash of the image is ��31�'Y�

The average shade of the image is **#27211c** and the main colours are:
2% #e6ddd6
2% #a4998a
5% #68584a
14% #3a2f28
27% #211813
46% #dc9



Conclusions

- By September 2019, we will have launched the Materials Data Repository, which:
  - Is a platform to collect and showcase the work of NIMS's researchers
    - Shows some of COAR's Next Generation Repository behaviours
    - Is integrated with a number of other NIMS systems
    - Is playing its part as a significant 'node' in the global knowledge commons
  - **By April, 2020 April, MDR is scheduled to be opened to public**
    - a publicly accessible platform for R&D of materials

ありがとうございました

Arigatō

Danke schön!

Thank you!