Materials Data Platform overview: metadata, vocabulary, and repository

Mikiko Tanifuji

Materials Data Platform Center

Div. of Materials Data and Integrated System (MaDIS)

National Institute for Materials Science

Who we are

NIMS National Institute for Materials Science

Established in 2001 by merger of two national institutes: (metals + inorganic materials) Now covers materials in general

MaDIS Research & Services Division of Materials Data and Integrated System

Established in 2017 to focus on materials data and integration

DPFC Materials Data Platform Center

Budget 2017 – 2020, 3 billion yen

Who we are: a Facebook

Materials Development

- 高分子材料設計
- 熱制御材料
- 金属系構造材料
- 半導体
- 多孔質材料

Materials Architectures

- 計算シミュレーション：自動化
- SIP-MI：プロセスから一貫予測
- SMILES X：分子人工設計
- 計測インフォマティックス

Data Science

- スパースモデリング（モデル選択）
- 画像解析～パターン認識、深層学習
- 回帰技術～機械学習、ベイズ推定
- 最適化技術～能動学習等
- 自然言語処理

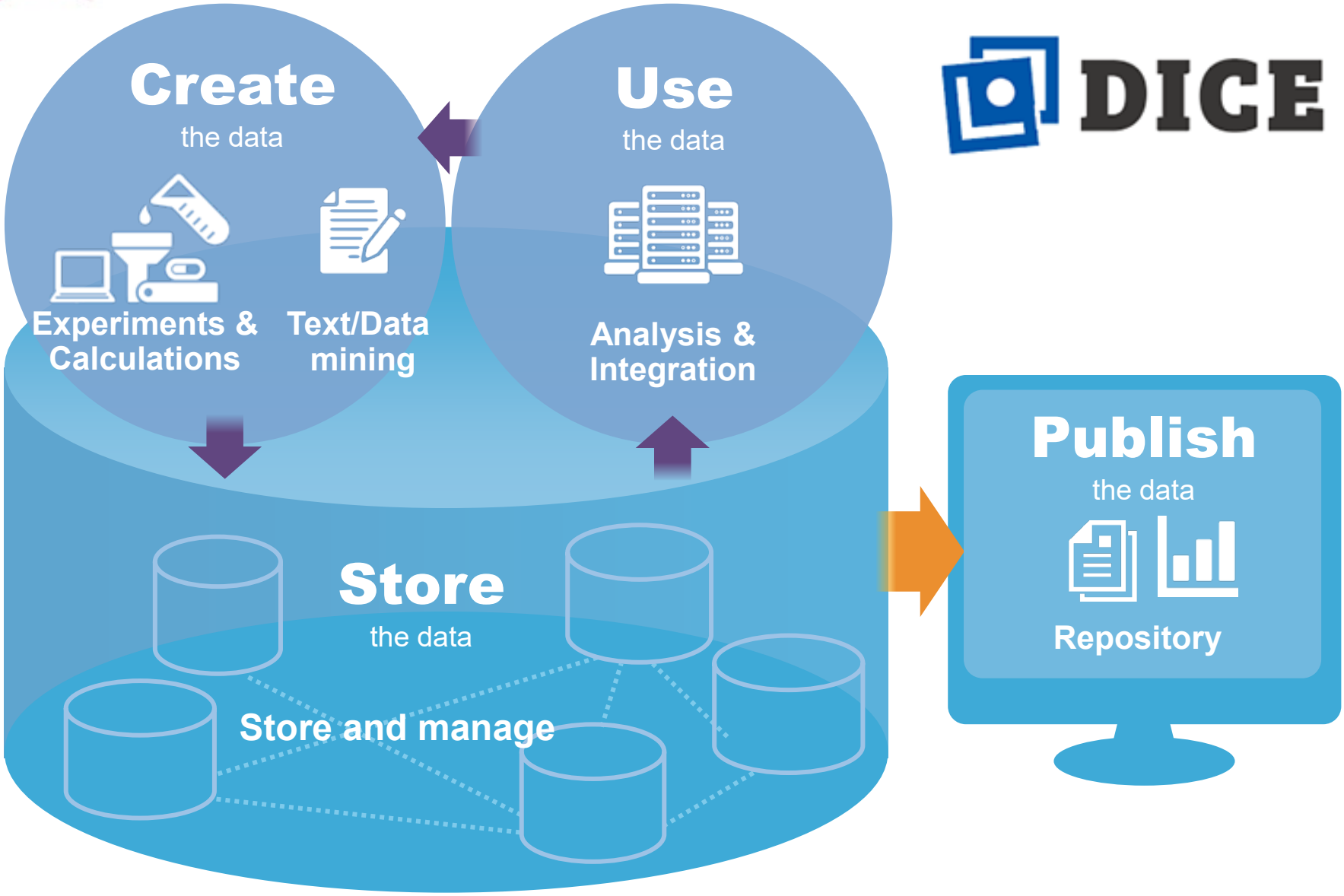
Data infrastructure

- Data structure and modeling
- Data curation
- Data collection and FAIRable
- Data mining from publications
- Data system technology and development

Data-driven people at MaDIS

66 staff at Materials Data Platform Center

Materials Data Platform at NIMS



DICE

Materials Data Platform overview

Building advanced databases

MatNavi Database

PolyInfo

Text and data mining

TDM

meta

Data collection

Data Collection System

IoT data transfer system

M-DaC Data Conv Tools

M-DaC

Large-scale facilities

Other database

Data Provider

Common metadata model

Research Data Management

Materials Vocabulary Wiki

API Framework

Data Management Plan

Data policy

Data Cloud Materials data hub (2021 -)

Analyses and materials integration

Machine learning system

Mi²i

SIP Mint system

MI MATERIALS INTEGRATION

Publishing, linking, open science

Data Repository

MDR MATERIALS DATA REPOSITORY

Closed data

Open data

Collaboration

Industry

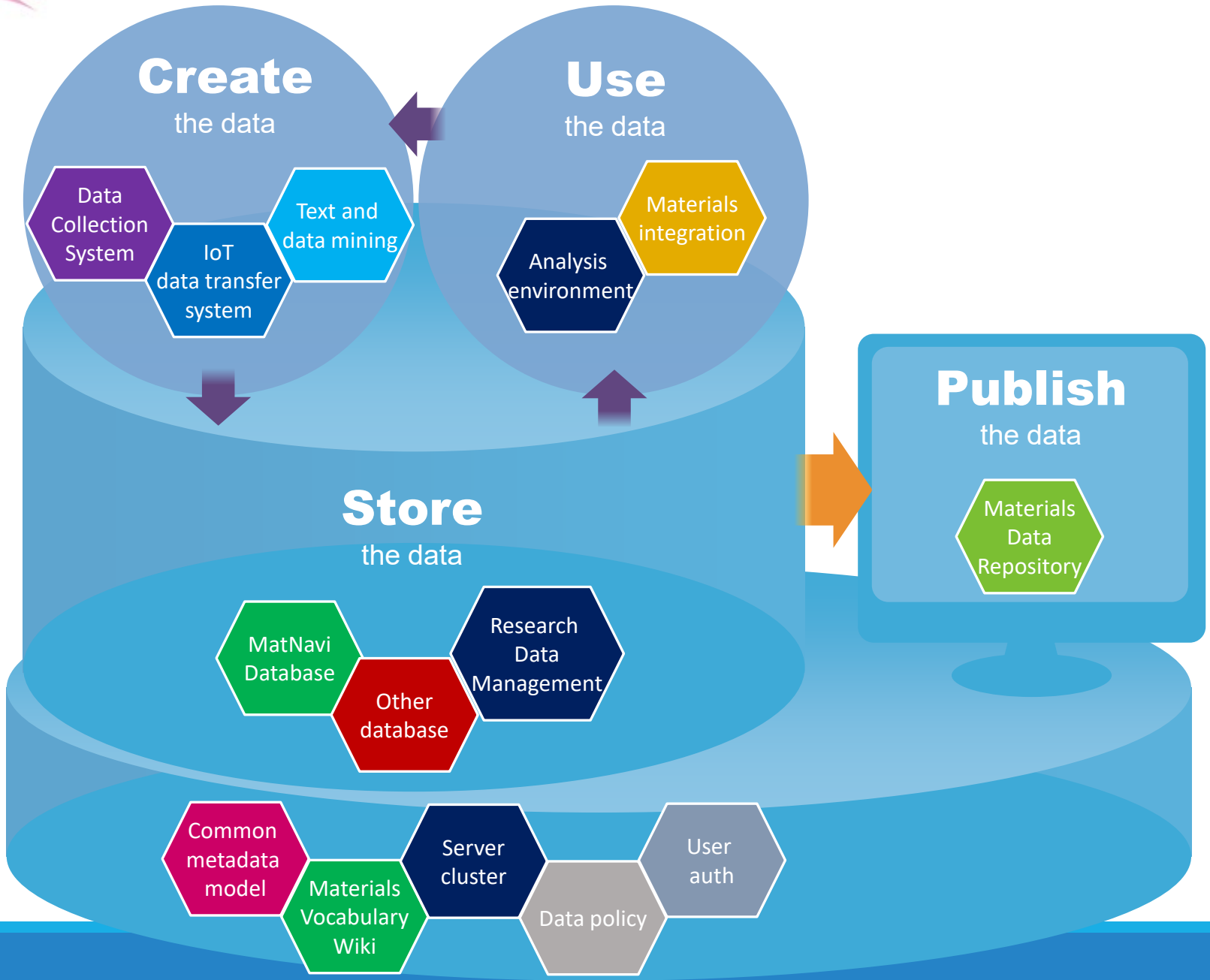
Academia

Publishing, Federation with other repositories

Data Science Academia-Private sectors

Image credit: Koji101, pngimg.com (CC)

Four actions mapped to the platform components



DICE Common Message Format

Common metadata model

METADATA

Mandatory metadata

- Bibliographic metadata
- Administrative metadata
- Subject material

Domain-specific metadata

+

Characterization metadata	Specimen metadata	Property metadata	Synthesis/Process metadata	Calculation metadata
Method, Environment…	Material type, Structure…	Physical properties, Units…	Processed date, Temperature…	Computer software, Version…

Implemented as data model

DATA

+

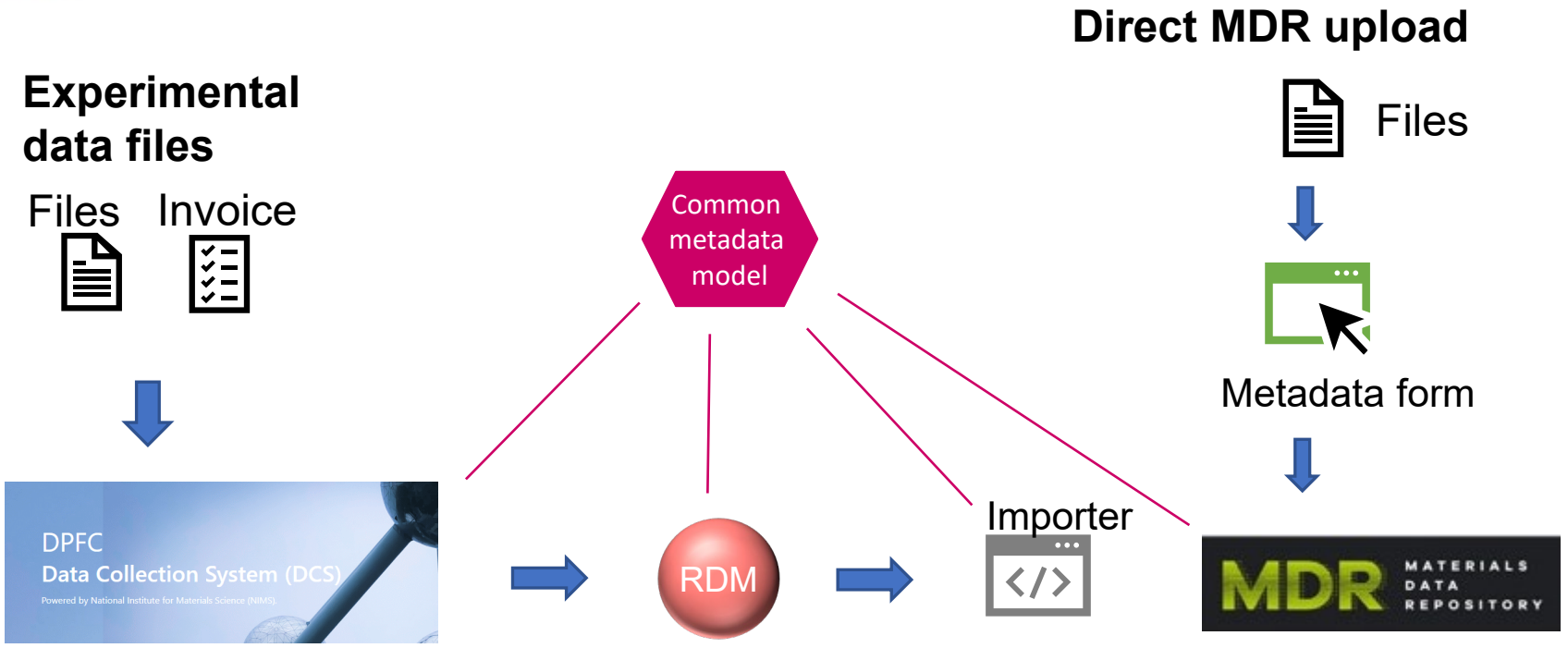
Primary parameters

Characterization primary params	Specimen primary params	Property primary params	Synthesis/Process primary params	Calculation primary params
Data	Data	Data	Data	Data

Save as files

After lots of discussions (still ongoing), schema file published at https://dice.nims.go.jp/

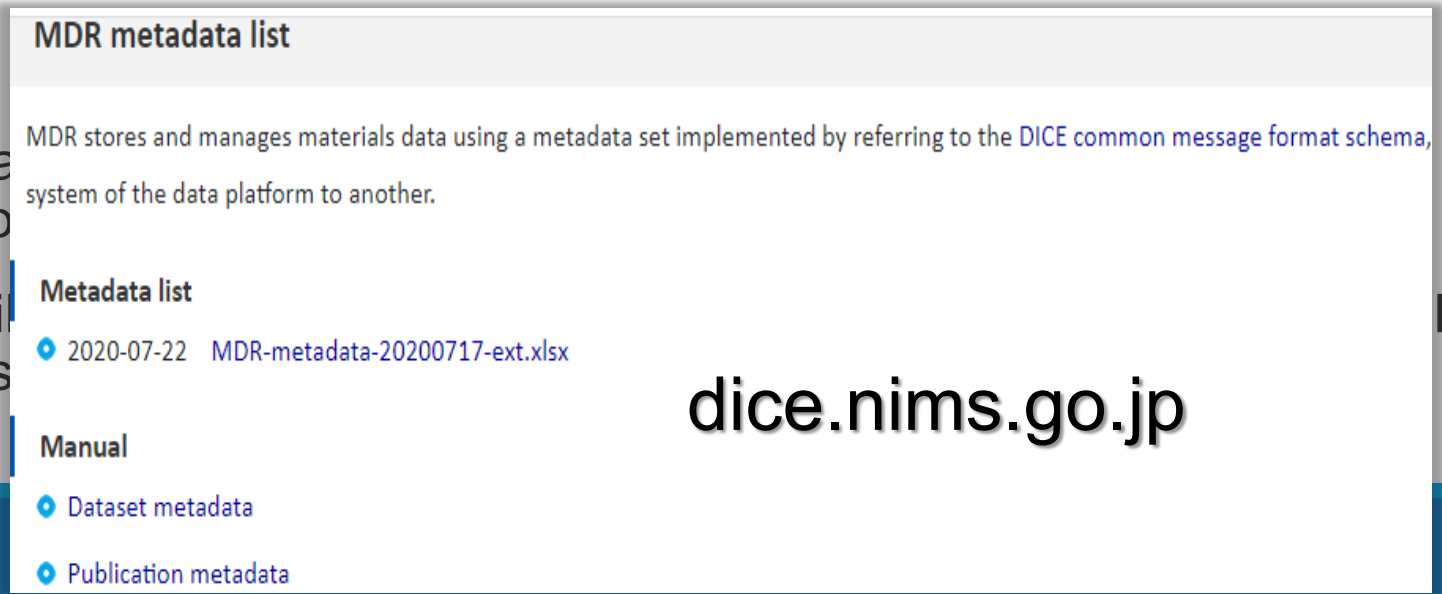
DICE Common Message Format



Data

- Va
  co
- Bi
  us

R



dice.nims.go.jp

Ongoing discussion about “primary parameters”

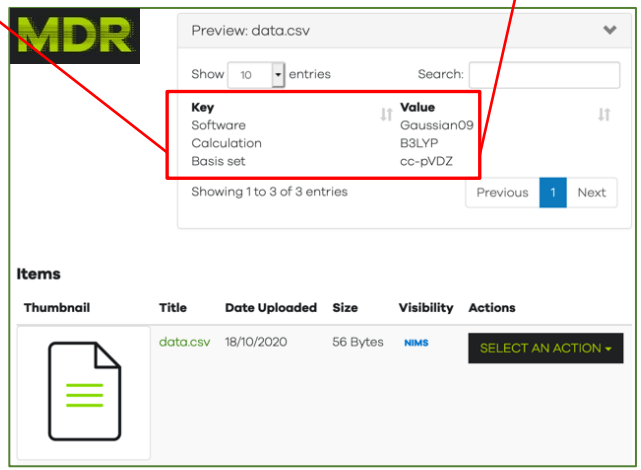
- Highly domain-specific parameters such as
  - Which absorption edge for a XAFS measurement?
  - Which basis set for a Gaussian calculation?

Extremely difficult to agree how to include in the common schema!

➡ At least, save as files?

As a 2 x n key-value CSV

Software	Gaussian09
Calculation	B3LYP
Basis set	cc-pVDZ

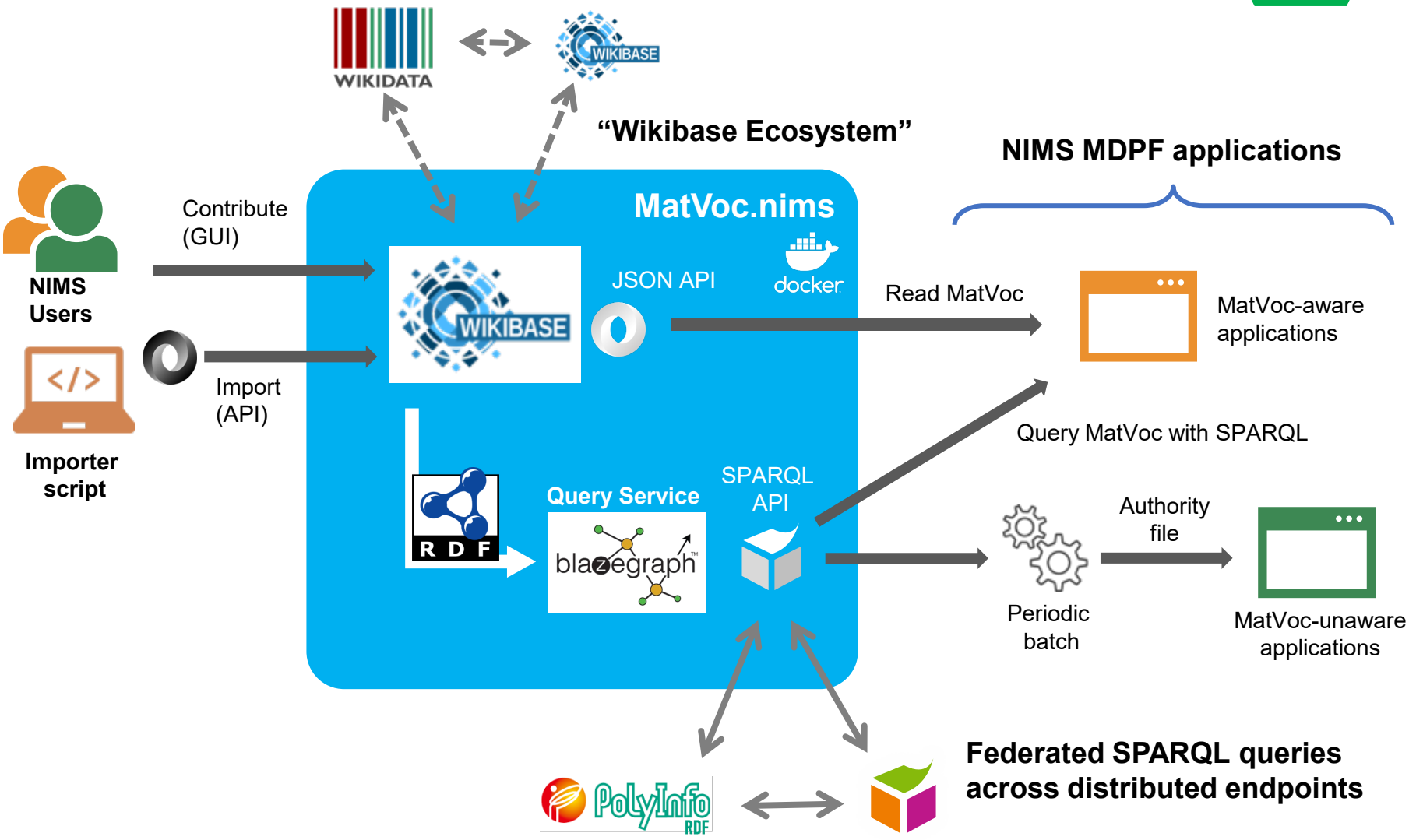


As a Schema.org-like JSON-LD file


```
{ “@graph”: [
  { “@id”: “./”,

    /* Bibliographic metadata */
    “name”: ...,
    “author”: ...,

    /* Scientific metadata */
    “variableMeasured”: ...,

    “hasPart”: {“@id”: “data.dat”}
  }, ......
]}
```

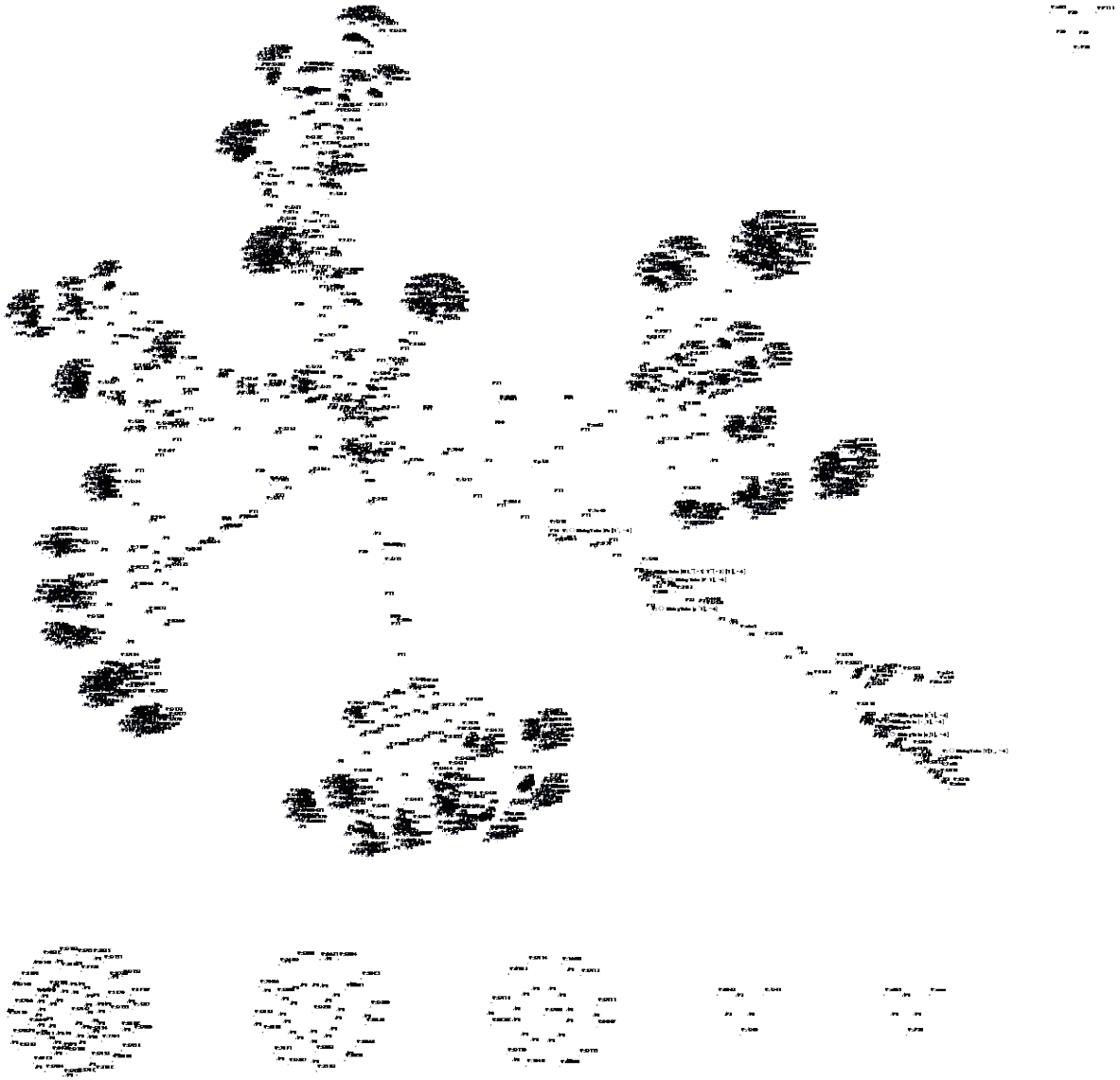
Vocabulary system overview

Materials Vocabulary Wiki



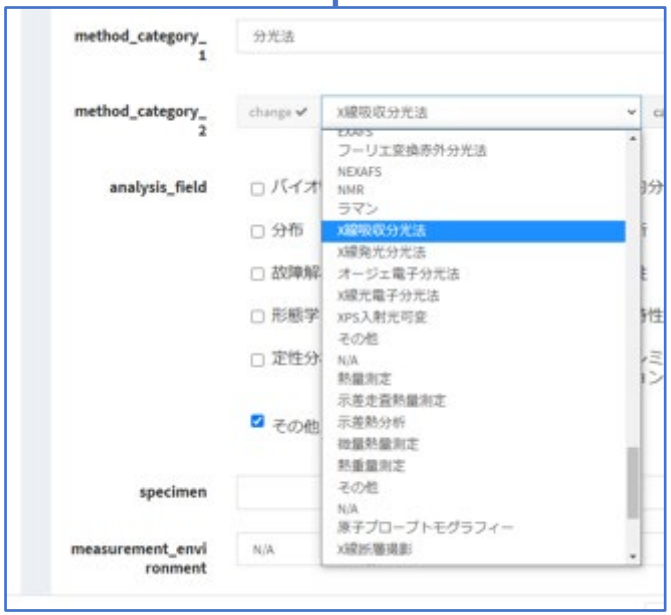
Bird's eye view of our vocabulary now

Materials Vocabulary Wiki



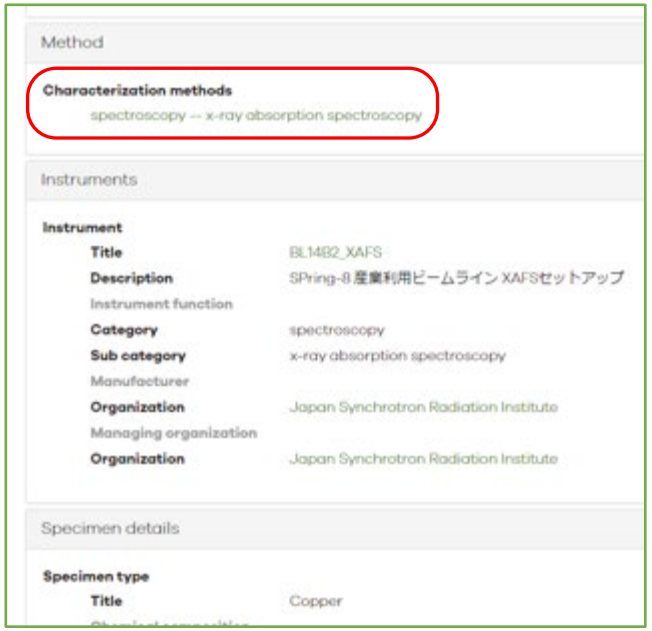
Vocabulary-based metadata transfer between systems

Data Collection:
Research Data Express



Common message format
depositUploadReq.json

Materials Data Repository



RDE local dictionary

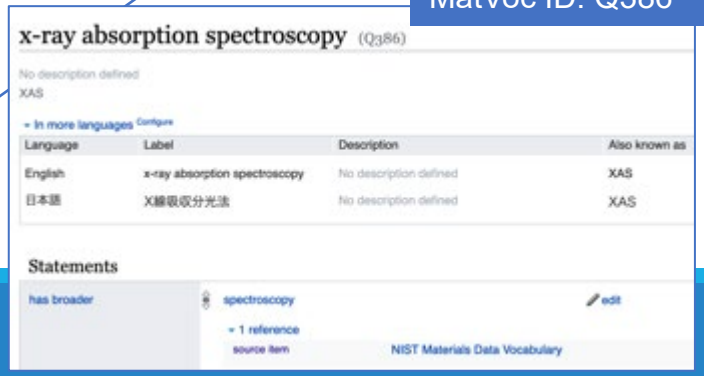
API query

refer

synced

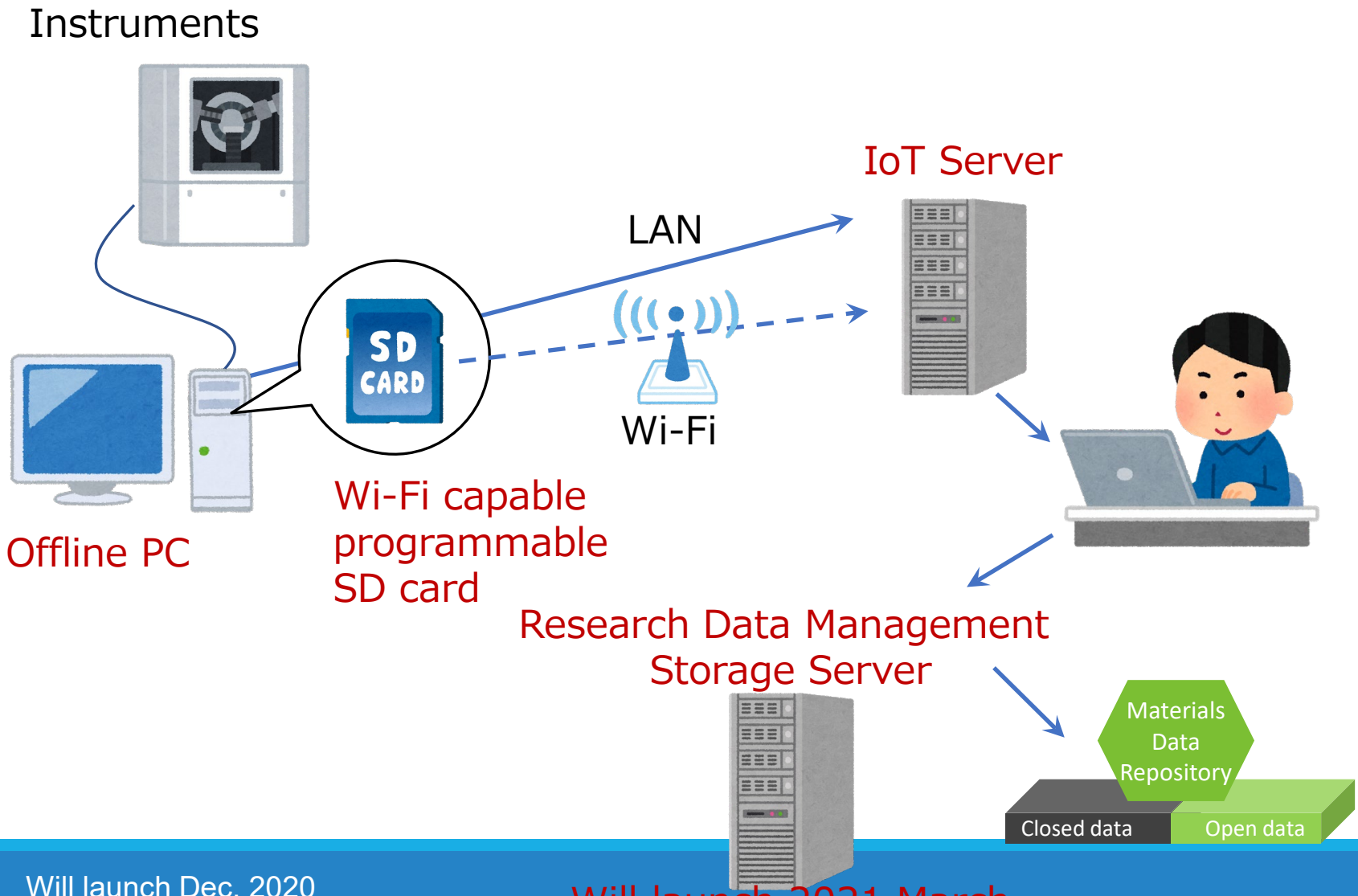
MatVoc

MatVoc ID: Q386



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

Automatic data collection using WiFi-SD cards towards FAIR data



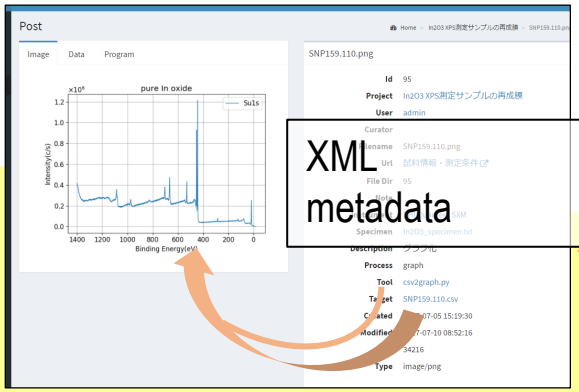
Will launch Dec, 2020

Will launch 2021 March

Data Collection System for efficient measurement data collection and automatic conversion

binary raw data

Text conversion program with cooperation of the instrument manufacturers

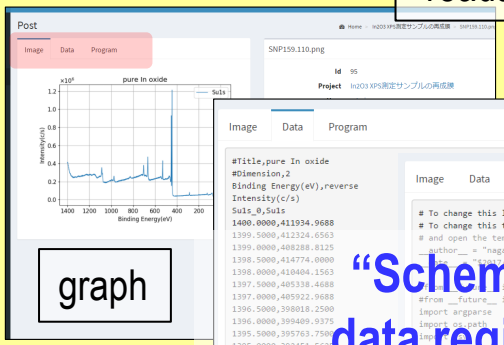


XML metadata

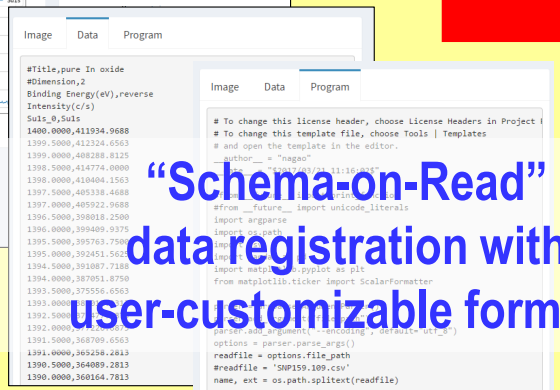
text file

Program developed in-house, for format conversion and controlled metadata

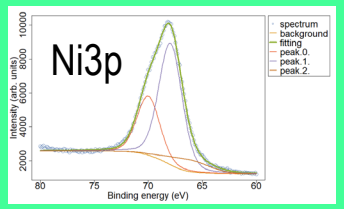
“readable” data



graph



“Schema-on-Read” data registration with user-customizable formats

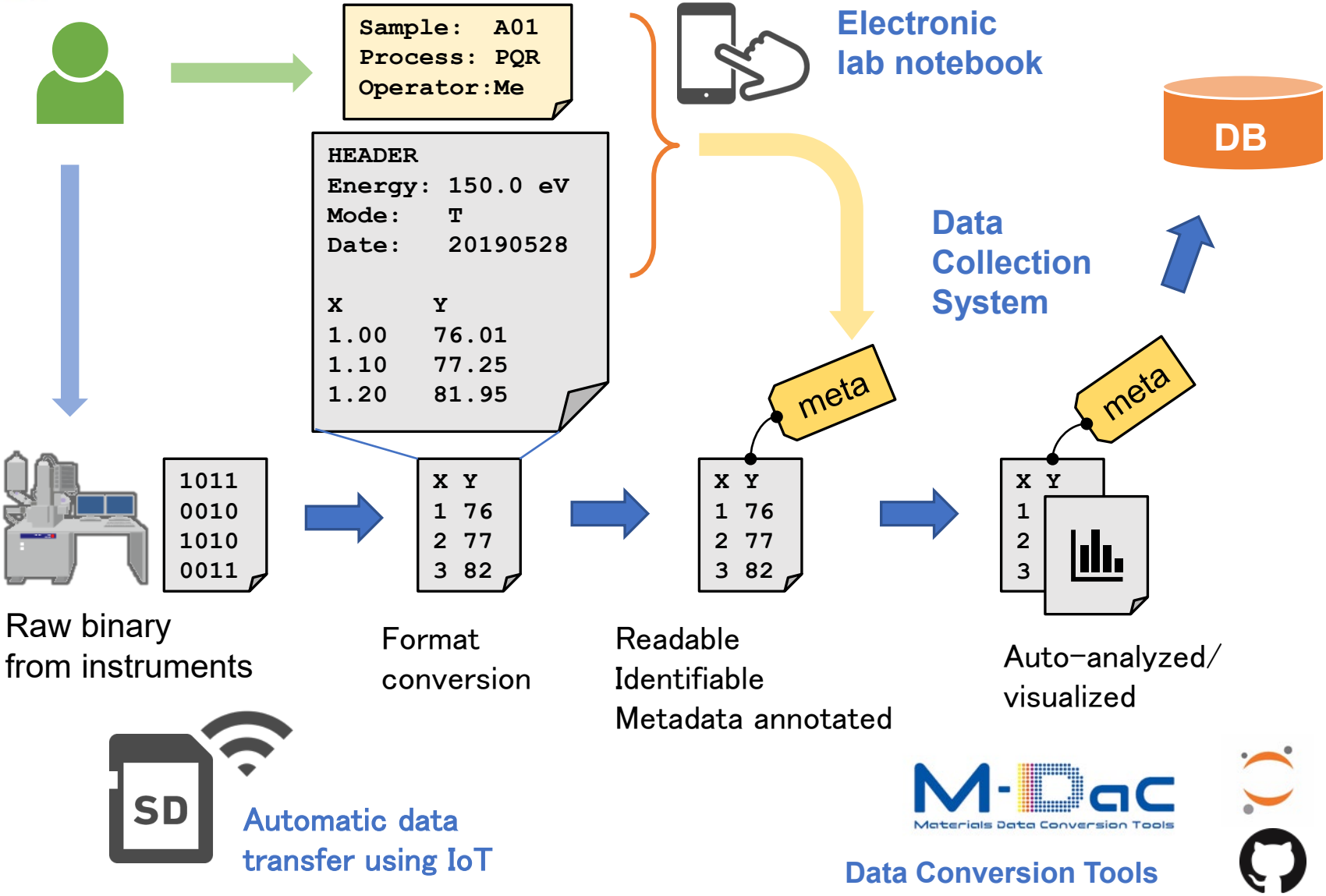


Spectral data automatically sparse-modeled

Converted numeric data (easier for humans)

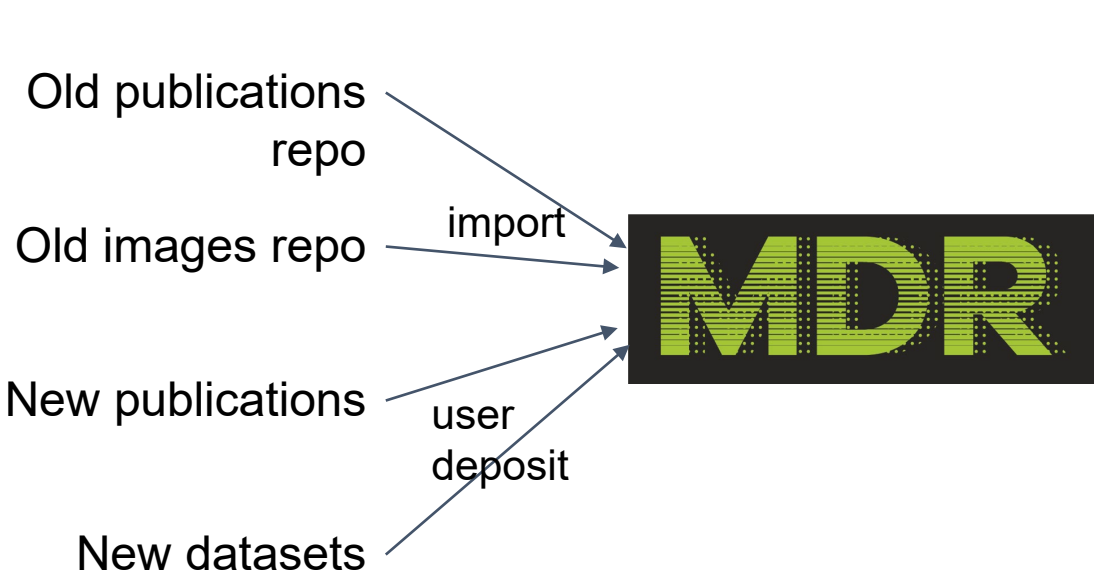
Python program (parser) that interprets the converted numeric data and visualizes them

Collecting experimental data



Materials Data Repository – transition stage 1

Datasets, publications, and images published on the Materials Data Repository



Metadata types:

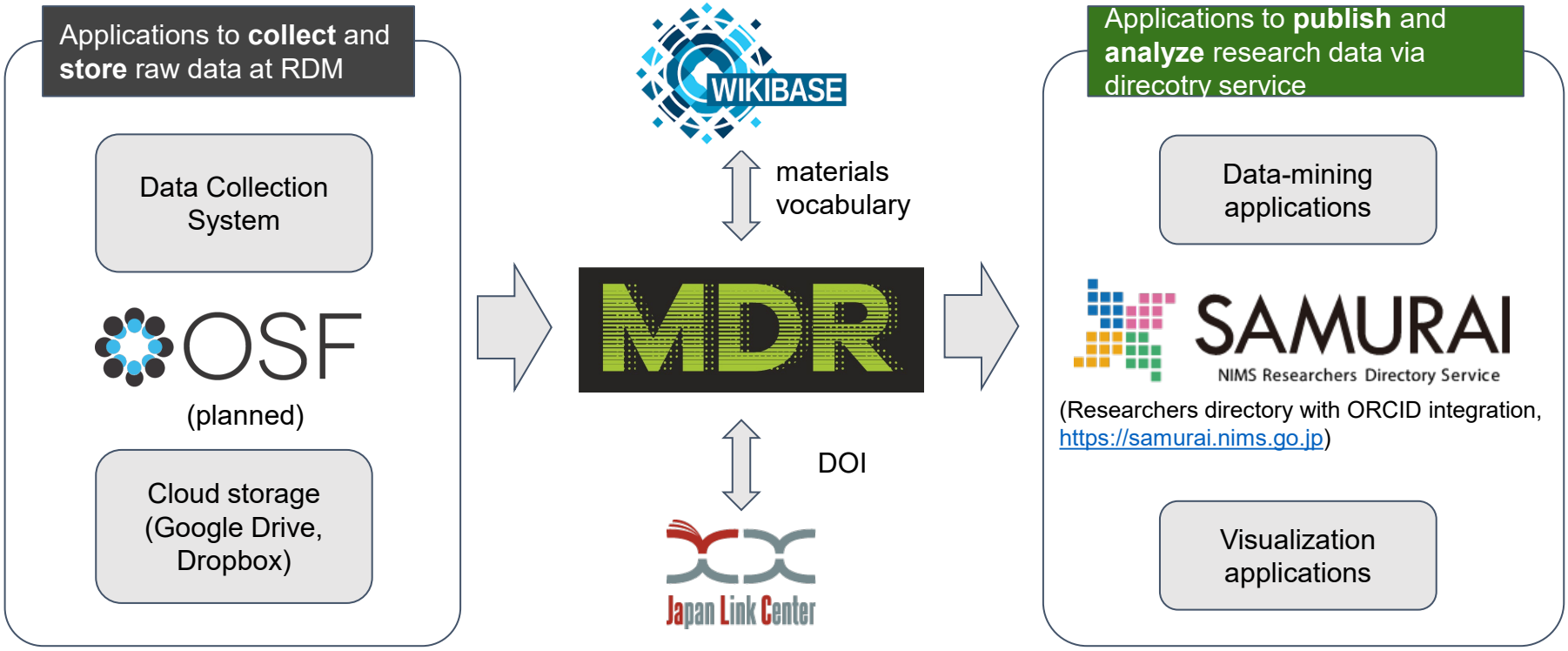
Datasets

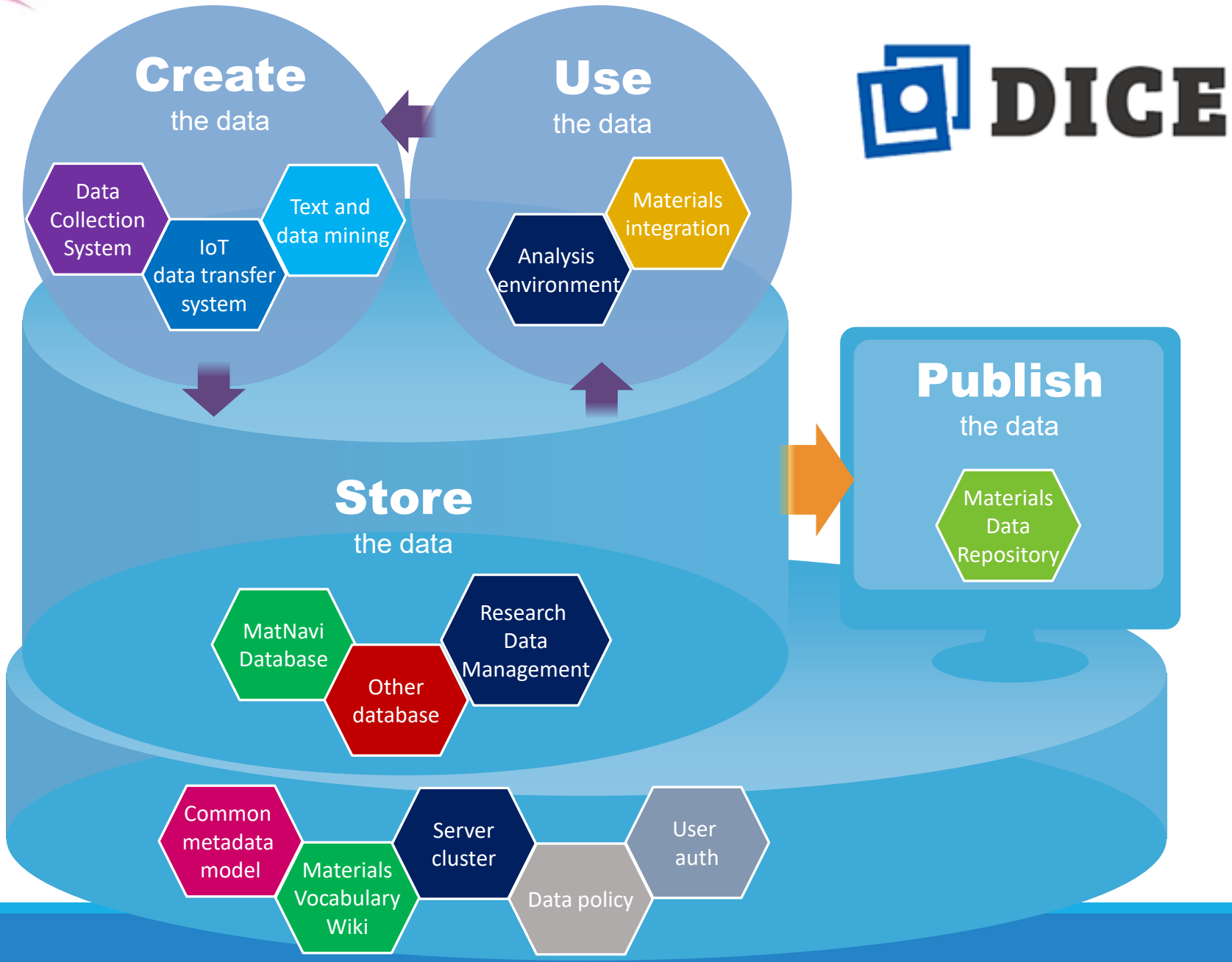
Publications

Images

Materials Data Repository – transition stage 2

Integration <–> FAIRable <–> Accumulation

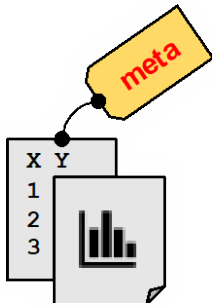



Create
the data
Data Collection System
IoT data transfer system
Text and data mining
Use
the data
Analysis environment
Materials integration
Publish
the data
Materials Data Repository
Store
the data
MatNavi Database
Other database
Research Data Management
Common metadata model
Materials Vocabulary Wiki
Server cluster
Data policy
User auth

DICE

Summary

- Materials Data Platform, DICE as FAIRable platform
- Public and internal services will be launched 2020 - 2021
- DICE as R&D platform for academia and industries
- DICE aims to Japan-wide data hub with:
  - library of data models and meta data schemas for target materials
  - data quality guideline (recommendation) in respond to what data scientists need between what materials scientists can cope with.



DICE

JP / EN

Services ▾ About Guide News Link Contact ▾

Store
Analyze

Publish
Discover

Apply
Create use

DICE is a data platform for all experts offering quality data and applications for materials science

"As a researcher, it is tedious to provide a universal, machine-comprehensible metadata to allow my data to be used in serendipitous ways."

"As a materials scientist, there is a particular material I am looking for, but I don't know how to design a model for AI-assisted research."

"As a data scientist, I am looking for materials data that I can use for my data-driven operations."

DICE provides three things: high-quality data, applications, and knowledge -- bridging between the domain experts to aid rapid advancement in materials science.

Special Thanks to

Dr. Takuya Kadohira

Dr. Kosuke Tanabe

Dr. Asahiko Matsuda

And Dr. Hideki Yoshikawa