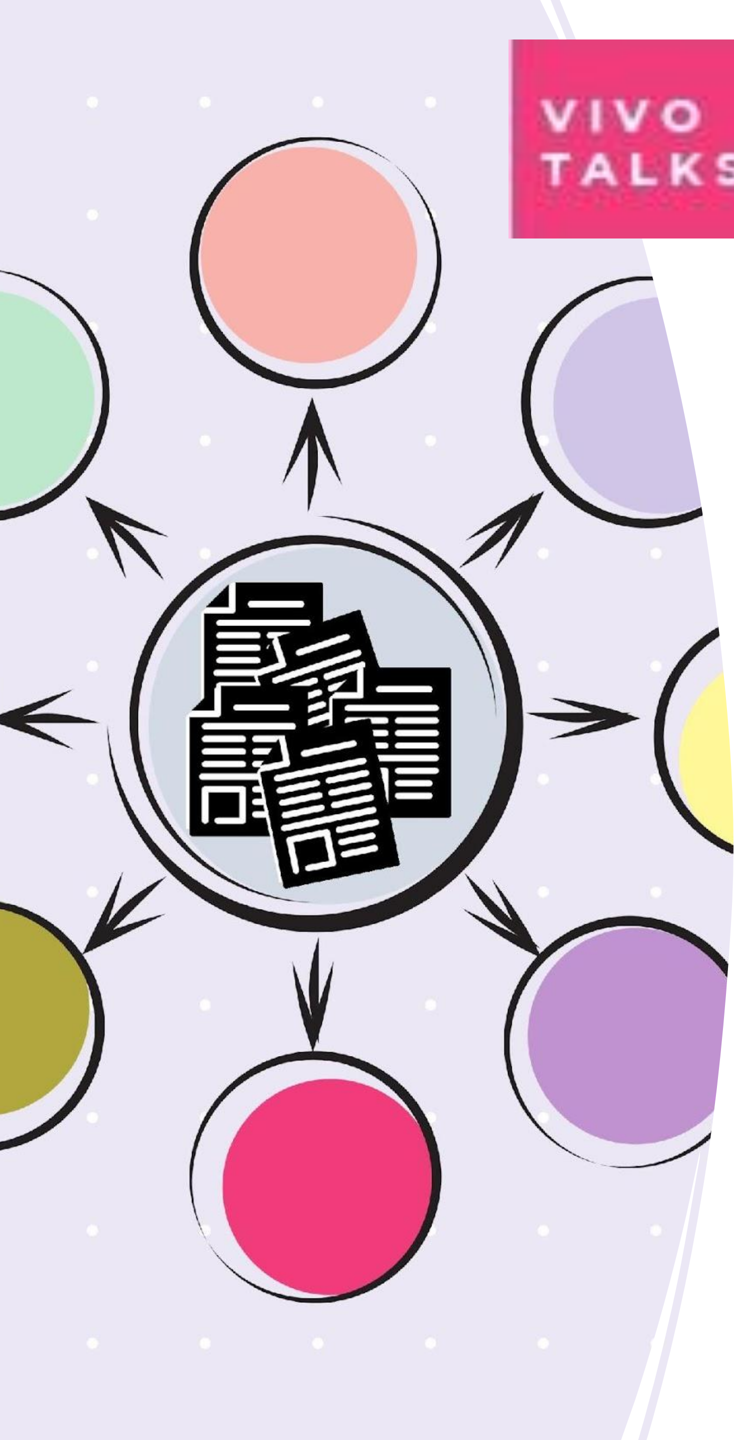


Creating a Visual Research Topic Map for SAMURAI Catalogue and an Introduction to Materials Data Platform (DICE) at NIMS, Japan

Sae Dieb

National Institute for Materials Science (NIMS), Tsukuba, Japan

VIVO Talks, March 23, 2023

Content

- Self Introduction
- NIMS
- Visual Research Topic Map for SAMURAI Catalogue
  - Motivation
  - Method
    - Data collection
    - Data preprocessing
    - Domain knowledge resources
    - Researcher output representation
    - Visualization
    - Topic map construction
  - Validation
    - Effect of authors order
    - Effect of domain knowledge resources
    - Researchers correlation analysis.
- Introduction to DICE
- Conclusion and Future work

Self Introduction

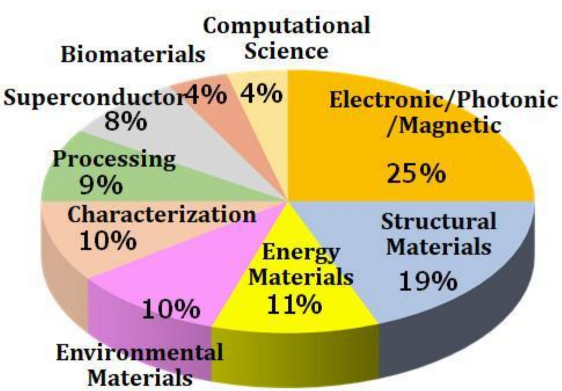
Sae Dieb, PhD.

ディーブ　冴

- *12. 2015: PhD in Information Science and Technology, Hokkaido University, Japan*
- 10. 2011 ~ 3. 2012: *Research assistant, Graduate School of Information Science and Technology, Hokkaido University, Japan*
- *2.2014 Visiting researcher: Department of Physics, University of Southampton, U.K.*
- 2. 2016 ~ 6.2018: *Postdoctoral researcher-machine learning for inverse materials design, Graduate School of Frontier Sciences, The University of Tokyo, Japan*
- *8. 2017 ~ 3.2019: Visiting researcher: RIKEN, Center for Advanced Intelligence Project, Japan*
- *7. 2018 ~ present: Scientific Researcher (Materials Informatics), National Institute for Materials Science, Japan*

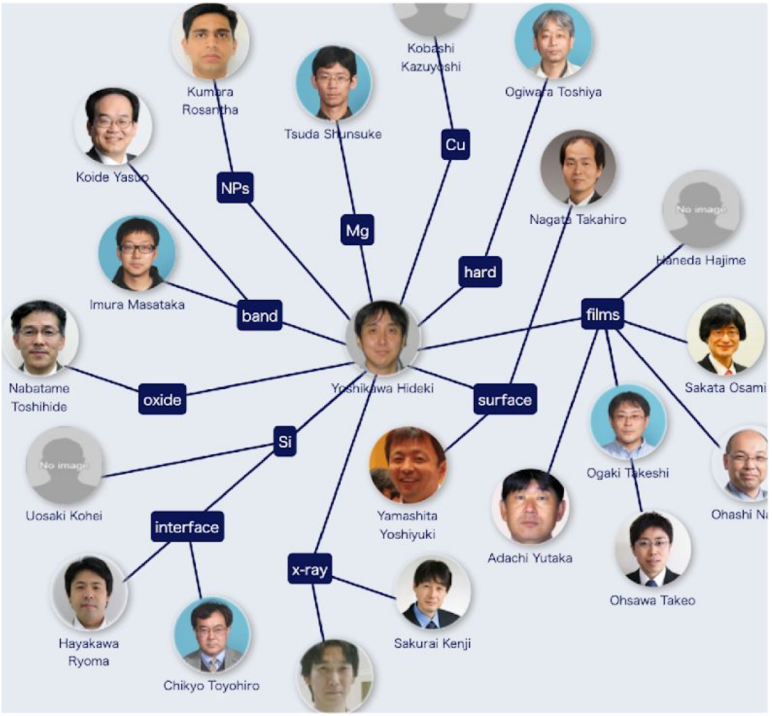
National Institute for Materials Sceince, Japan

- Located in Tsukuba science city, 50 KM north east of central Tokyo
- 30% of Japan's national research institute are in Tsukuba Science City
- ***<u>Established in 1956</u>***, *Budget 216 million US$)*
  - ✔ *Staff 1,582 (half research positions)*
  - ✔ *3 campuses in Tsukuba*
- MaDIS:Research & Services Division of Materials Data and Integrated System
  - ✔ *<u>Established in 2017</u> to focus on materials data and integration, and launched* DPFC *for services.*



Researchers by field

Visual Research Topic Map for SAMURAI Catalogue



Mikiko Tanifuji

Masashi Ishii

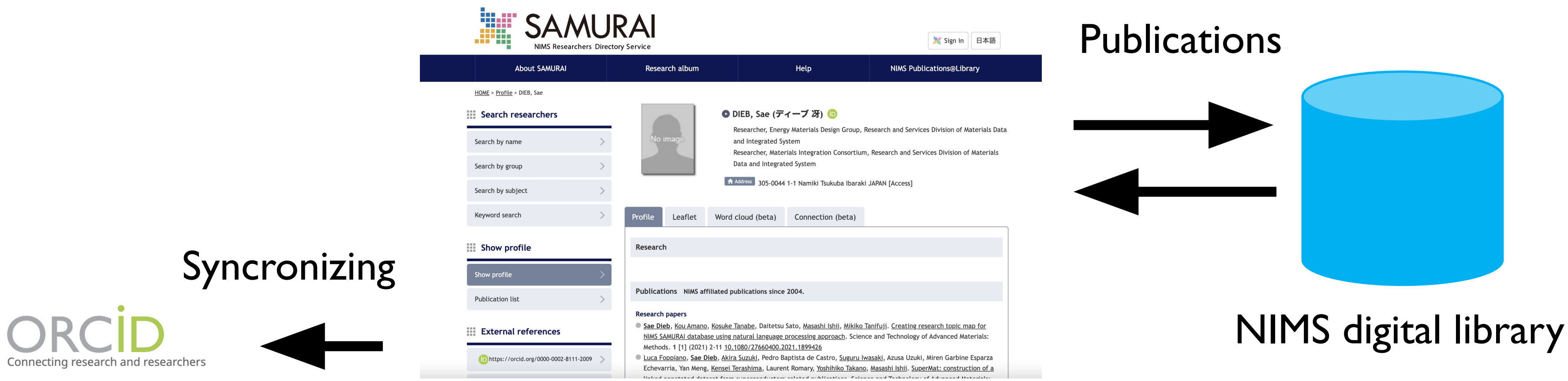
Kosuke Tanabe

Kou Amano

Daitetsu Sato

Collaborators

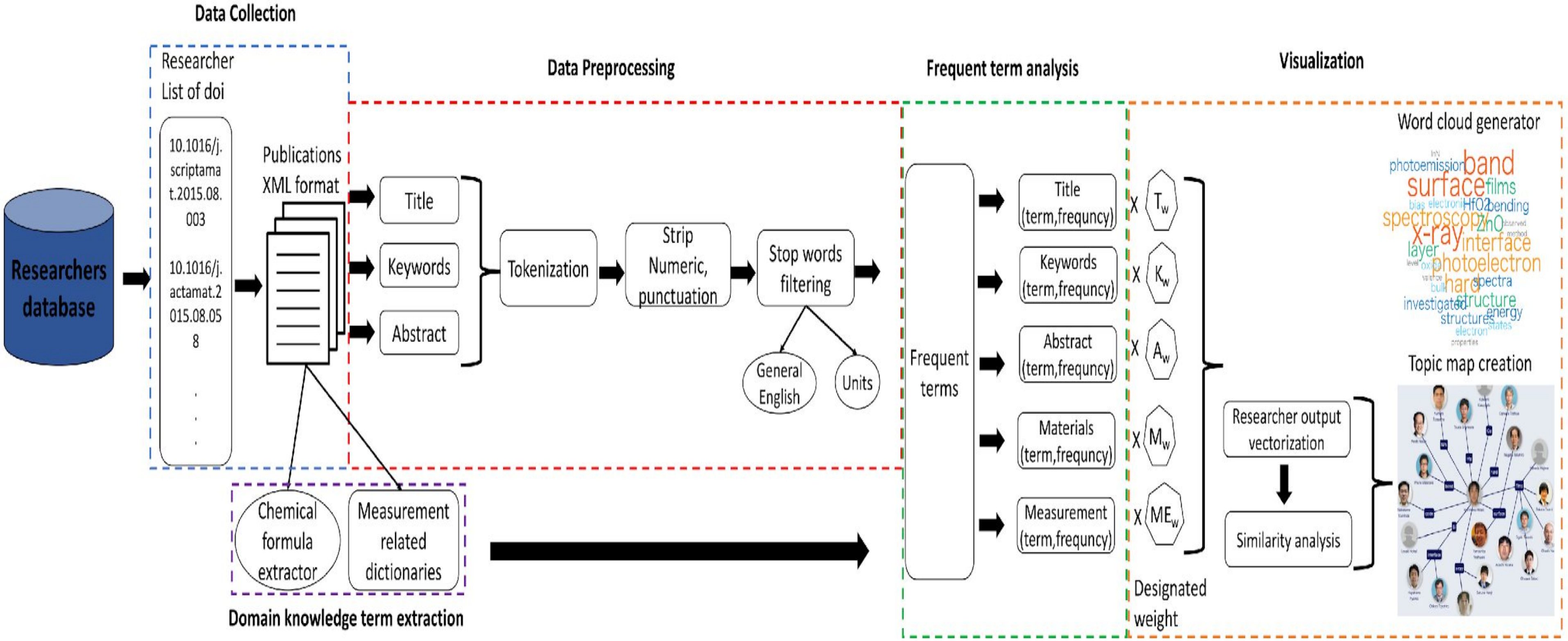
Motivation


Publications
Syncronizing
NIMS digital library
ORCID
Connecting research and researchers
SAMURAI
NIMS Researchers Directory Service
Sign in
日本語
About SAMURAI
Research album
Help
NIMS Publications@Library
Search researchers
Search by name
Search by group
Search by subject
Keyword search
Show profile
Show profile
Publication list
External references
DIEB, Sae (ディーブ 冴)
Profile
Leaflet
Word cloud (beta)
Connection (beta)
Research
Publications NIMS affiliated publications since 2004.
Research papers

Motivation

- Maximizing the information absorbance and intuitively capturing the research characteristics of each materials science researcher.
- Connecting researchers with similar topics aiming to find potential collaborators.

Method: overview



Data collection

- 1058 SAMURAI researchers
- 8269 published articles (XML format)
- Using the DOI
- Extraction of
  - Title of the paper
  - Keywords" section
  - Abstract of the paper

Data preprocessing

- tokenization
- Noise reduction:
  - Removing numeric values, punctuation marks (for example, \23.5", \!", \?").
  - Filtering general English language stop-words such as \but", \an", \he".
  - Physical units such as \m" (meter) for length measurement, and \K" (Kelvin)

Domain knowledge resources

- General English language-based tokenization schema might have a low matching ratio for domain-specific knowledge
- Extraction of chemical compounds
  - Chemistry-aware tokenization
  - Employed a regular expression tool to recognize chemical formula
- Extraction of measurement related terms
  - Japanese dictionary of physics and chemistry
    - X-ray diffraction microscopy and Maxam-Gilbert method.
    - Simulation category such as the Monte Carlo method.

Extracted terms

Researchers	Terms				
1058		Domain knowledge		Others	Sum
Publications		Chemical compound	Measurement related		
8269	Total	5762	189	98992	104943
	% of sum	5.5 %	0.2 %	94.3 %	100 %
	Unique	1161	40	14262	15463
	% of sum	7.5 %	0.3 %	92.2 %	100 %

Terms extracted for word cloud visualization and topic map creation including domain knowledge terms. Publications are the sum of all retrieved publications for each author.

Most frequent domain knowledge terms

Chemical compound	Measurement terms
In	X-ray photoelectron spectroscopy
Si	Phase
Fe	Monte Carlo
Al	XPS

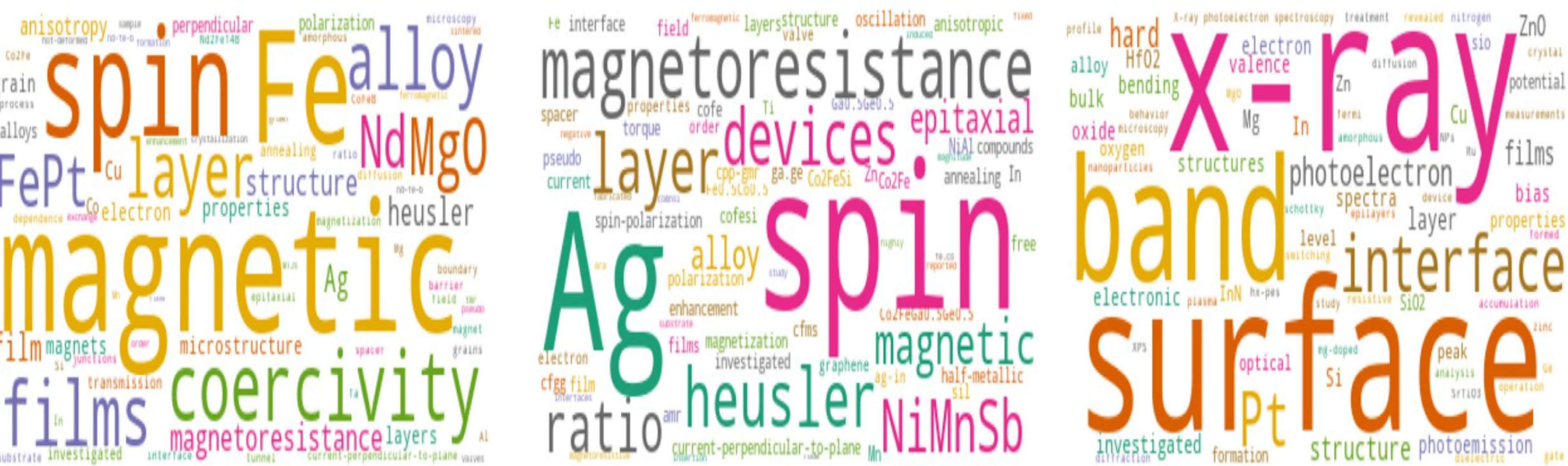
Top domain knowledge terms extracted for word cloud visualization and topic map creation.

Researcher output representation

- Five sets of (term, frequency) were created and normalized based on the length of the extracted sections for each researcher as follows:
  - T_{tf} Title terms extracted from all publications.
  - K_{tf} Keywords terms extracted from all publications.
  - A_{tf} Abstract terms extracted from all publications.
  - M_{tf} Material formulas extracted from all three sections of all research publications.
  - ME_{tf} Measurement related terms extracted from all three sections of all research publications.
  - Each researcher output is then represented with the following equation:
    - $R_{tf} = Tw * T_{tf} \cup Kw * K_{tf} \cup Aw * A_{tf} \cup Mw * M_{tf} \cup MEw * Me_{tf}$
  - Where Tw, Kw, …. Are assigned weights for each section

Visualization

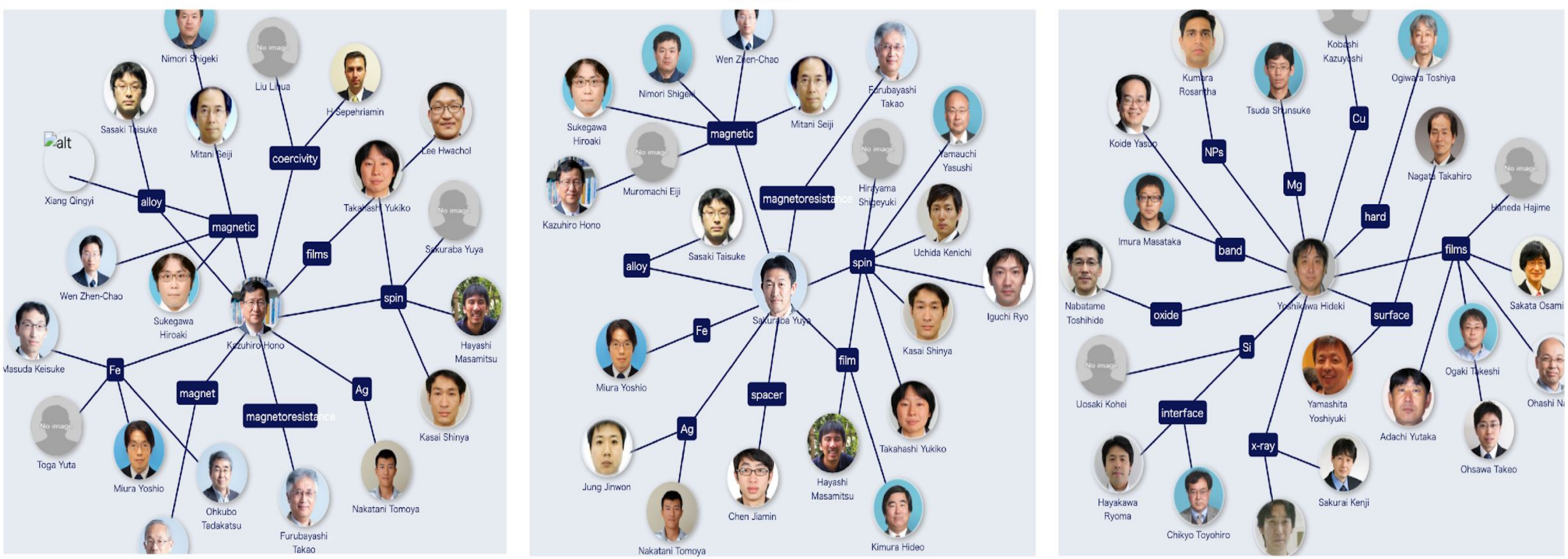
(a) Hono (b) Sakuraba (c) Yoshikawa



Kazuhiro Hono, Research Center for Magnetic and Spintronic Materials.
Yuya Sakuraba, Research Center for Magnetic and Spintronic Materials
Hideki Yoshikawa, Research Center for Advanced Measurement and Characterization.

Topic map construction

- Cosine similarity used to find resarchers with similar topics
  - $simiarity = \cos\theta = \frac{\sum v_i v_j}{\sqrt{\sum {v_i}^2}\sqrt{\sum {v_j}^2}}$

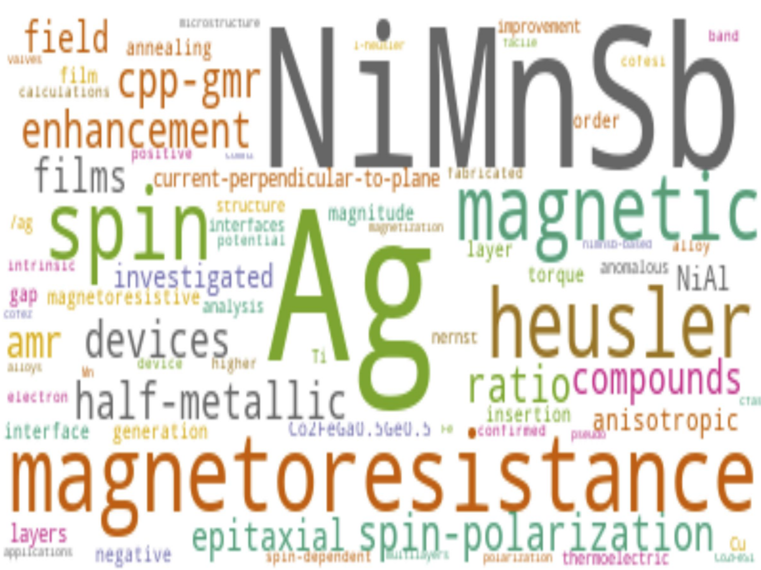


(a) Hono (b) Sakuraba (c) Yoshikawa

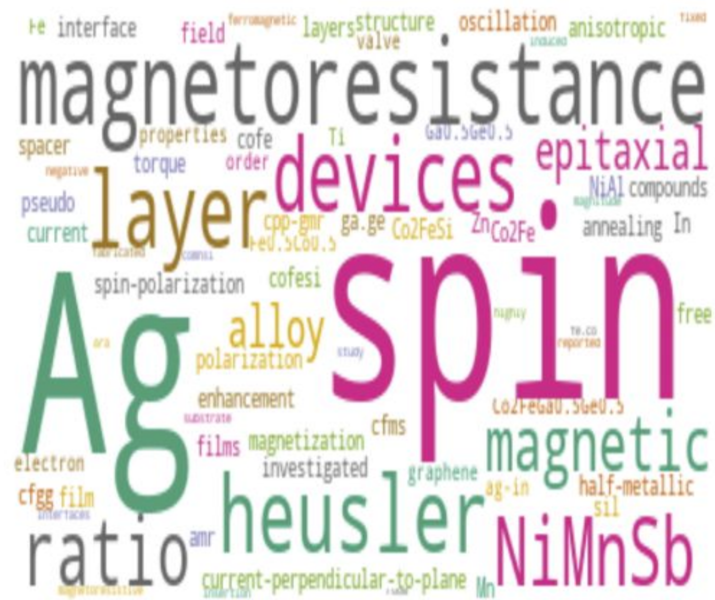
Effect of authors order

- Researcher position in the author list can affect his interest topics (first author, PI, corresponding author).

First or corresponding author



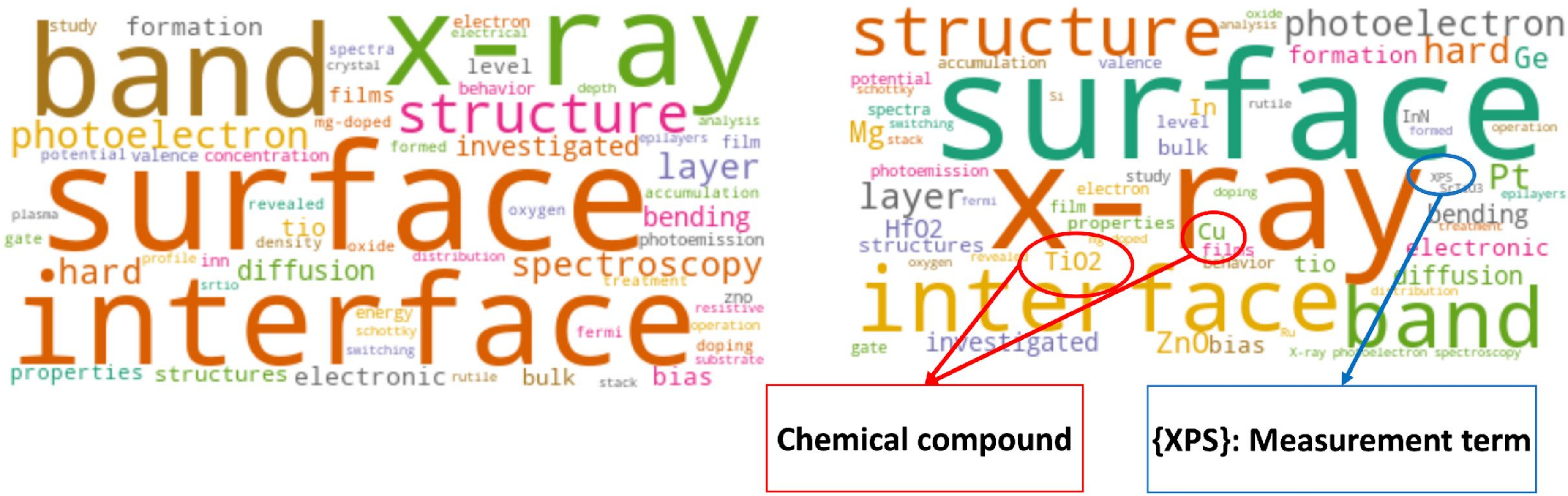
Total publications



Word cloud visualization for the researcher Yuya Sakuraba topics as extracted from publications where he is either the first or the corresponding author.

Effect of domain knowledge resources

(a) Without domain knowledge resources **(b) With domain knowledge resources**



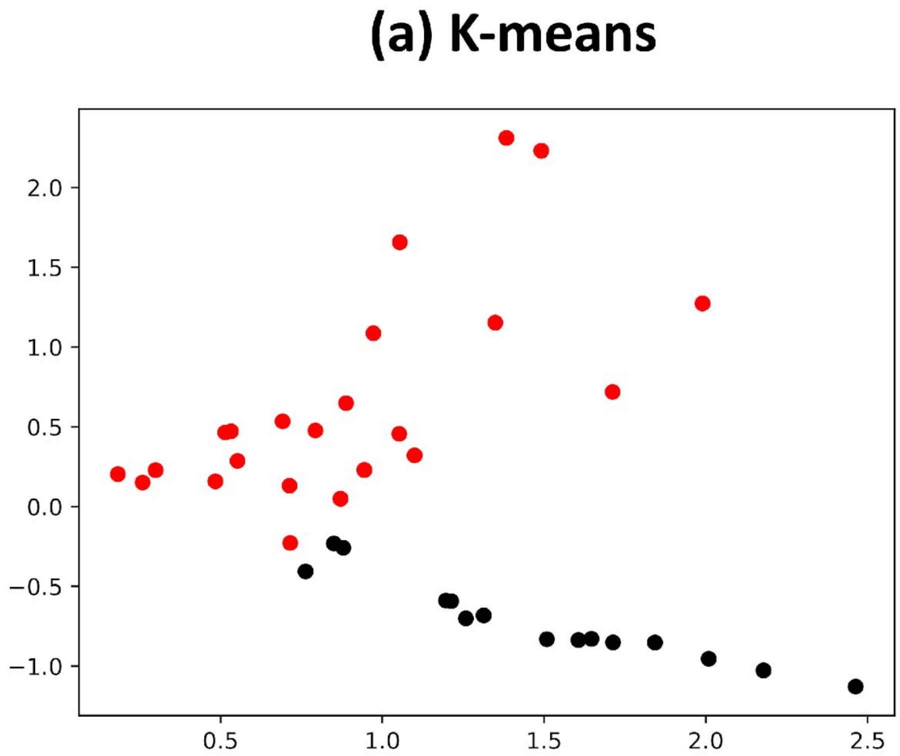
A researcher affiliated with both the Electric and Electronic Materials Field, Nano Electronics Device Materials Group, Research Center for Functional Materials, and the Quantum Beam Field, Synchrotron X-ray Group, Research Center for Advanced Measurement and Characterization.

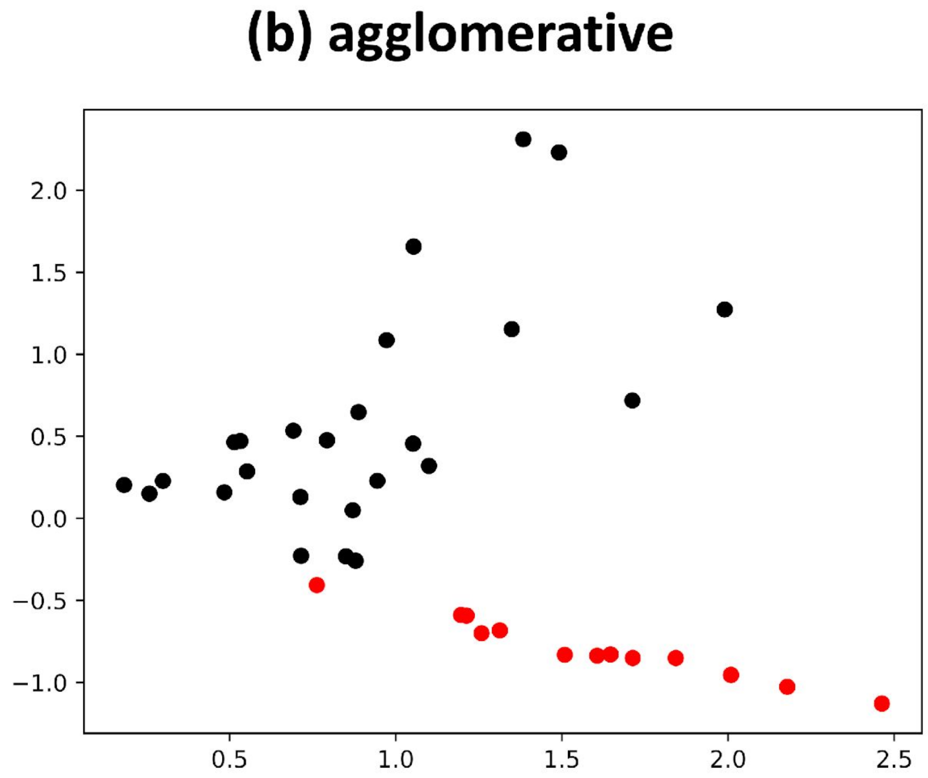
Researcher correlation analysis

Validation:
Researchers who belong to the same center are expected to have a stronger correlation with each other and a weaker correlation with researchers in different centers.
Experimenting on 2 research centers
Clustering researchers output who belong to different research centers:

- K means
- Agglomerative clustering

Researcher correlation analysis





Clustering results for researchers with two different affiliations in NIMS based on their word cloud.
Data points are projected on 2D using truncated singular value decomposition.

Researcher correlation analysis

Clustering method	Internal evaluation	External evaluation	
	Davies–Bouldin	Purity	Adjusted Rand
K-means	2.75	**0.92**	**0.70**
Agglomerative	**2.54**	0.87	0.53

Clustering evaluation metrics: internal and external for 2 groups of researchers in NIMS.

Davies-Bouldin Index:

$$\mathrm{DBI} = \frac{1}{n}\sum_{1}^{n} \max(j \neq i)\left(\frac{\sigma_i+\sigma_j}{\mathrm{distance}(c_i, c_j)}\right)$$

Purity:

$$\mathrm{Purity} = \frac{1}{N}\sum_{m \in M} \max_{d \in D}\left|m \cap d\right|$$

Adjusted Rand index:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{Expected}_{\mathrm{RI}}}{\max(\mathrm{RI}) - \mathrm{Expected}_{\mathrm{RI}}}$$

$$\mathrm{RI} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}$$

This work was published in Science and Technology of Advanced Materials: Methods
https://www.tandfonline.com/doi/full/10.1080/27660400.2021.1899426

Conclusion

- WE PRESENTED AN APPROACH TO CREATE A TOPIC MAP FOR THE MATERIALS SCIENCE RESEARCHERS USING NATURAL LANGUAGE PROCESSING (NLP).
- WE AIM TO MAXIMIZE INFORMATION ABSORBANCE AND FIND LINKS BETWEEN RESEARCHERS WITH SIMILAR TOPICS TO ENCOURAGE COLLABORATION
- DOMAIN KNOWLEDGE RESOURCES WERE UTILIZED.

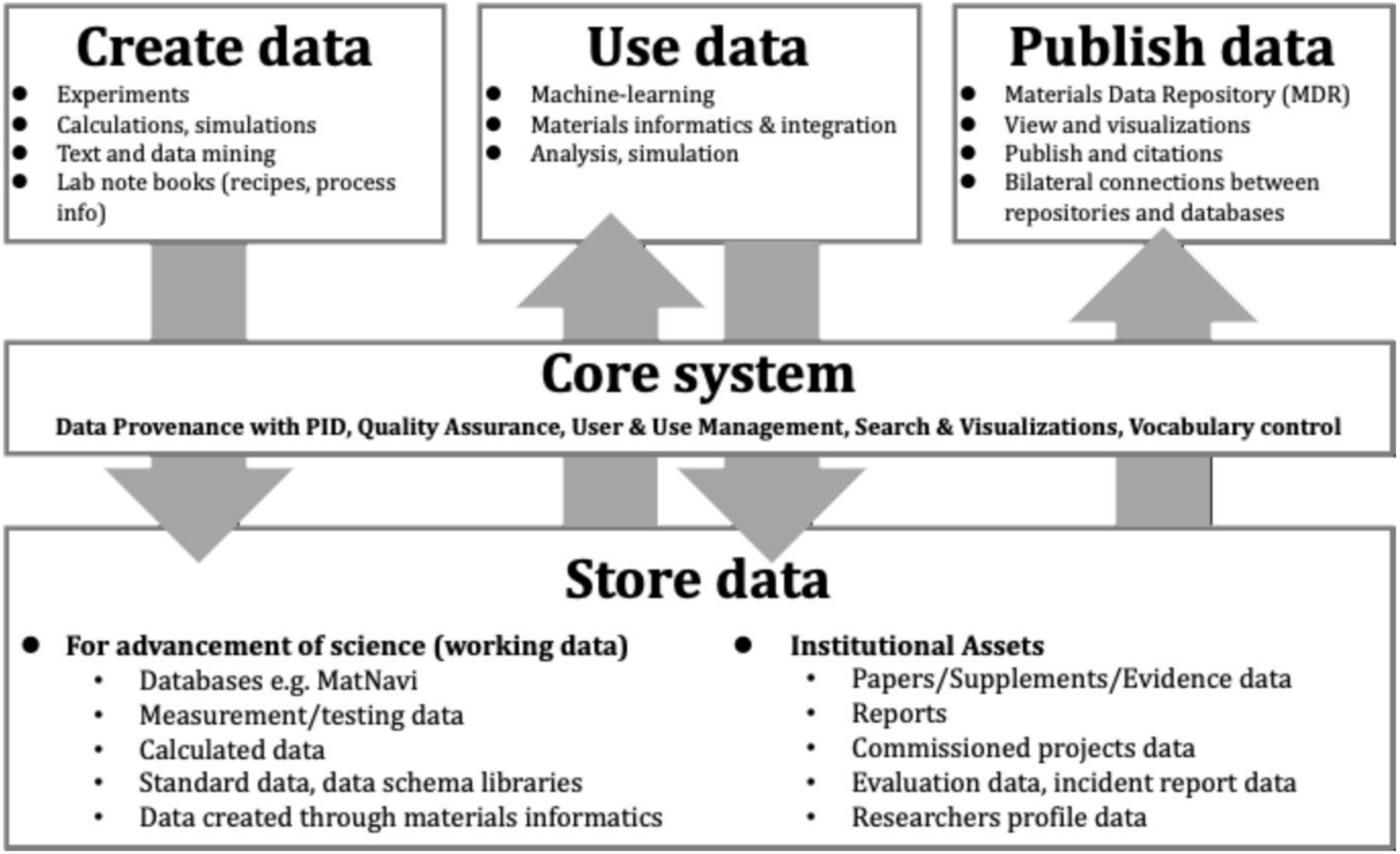
Future work

- Using of language models to detect topics
- conduct morphological analysis to improve tokenization efficiency.
- a weighting factor for each researcher based on his position in the author list.
- collect and analyze other types of resources such as research notes.
- interactive evaluation and adjustment system.

DICE

- DICE is a data platform for all experts offering quality data and applications for materials science.
- will be expanded to Japan-wide service in materials data-driven science
- https://dice.nims.go.jp/en/

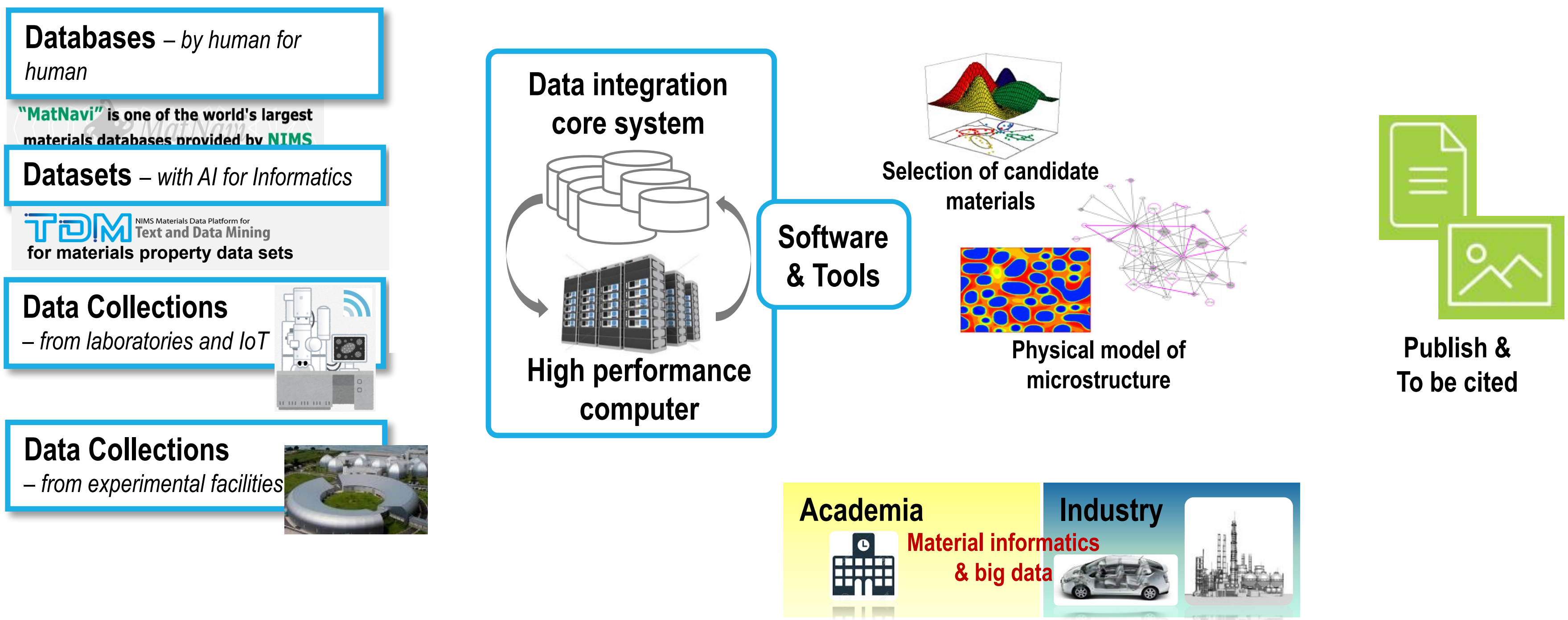
DICE



Key functionalities of the materials data platform, DICE.

Materials Data Bank Project 2017 – 2021: DICE

Materials Data Platform for integration knowledge and informatics



Importance of informatics is drastically increased in materials science !

Credit : Mikiko Tanifuji

DICE: Key Concepts

1. Quality of the data

- Identify who/what/when/how in the metadata
- Integrity of the data (hashes)

2. Accessibility

- URI / DOI / PID - based management
- Lab RDM with DMP MDR analysis
- MDR ⇔ other repos and databases

3. Usability of the data

- Licensing (CC, CC-BY-NC, MIT, etc.)
- Machine-readability of datasets and its metadata

4. Safe environment

- Single-sign-on
- 10-year preservation of data
- User policies (for depositors, downloaders)

5. Research aiding functionalities

- Vocabulary for TDM and materials informatics
- Data analysis environments
- API to connect platform services

PID: persistent identifier
RDM: research data management
DMP: data management plan
MDR: materials data repository
TDM: text & data mining

Credit : Mikiko Tanifuji

MatNavi

DATABASE

Basic Properties

PoLyInfo

The Polymer Database (PoLyInfo) contains information on polymer names, structures, properties, measurement conditions, polymerization methods, molding methods, etc. extracted from scientific literature.

> MORE

> LOGIN

AtomWork

The Inorganic Materials Database (AtomWork) contains information on crystal structures, X-ray diffraction, properties, and phase diagrams extracted from scientific literature.

> MORE

> LOGIN

CPDDB

The Computational Phase Diagram Database (CPDDB) contains information on the Gibbs energy functions of the phases are accumulated in a form of TDB (Thermodynamic DataBase) files, which are obtained from the CALPHAD-type thermodynamic assessments.

> MORE

> LOGIN

CompES-X

The Computational Electronic Structure Database (CompES-X) contains information on predicted by the first-principles calculations (VASP) for crystalline inorganic compounds.

> MORE

> LOGIN

Kakusan Diffusion Database

The Diffusion Database (Kakusan) contains information on the basic diffusion data (diffusion atom, diffusion coefficient, activation energy, etc.) of metallic and inorganic materials from scientific literature.

> MORE

> LOGIN

Thermophysical property

The Thermophysical Property Database contains information on density, surface tension and viscosity coefficient of metals, alloys and ceramics obtained by floating experiments using an electrostatic levitation furnace.

> MORE

> LOGIN

MatNavi

Engineering

Kinzoku

The Metallic Material Database (Kinzoku) contains information on tensile properties, creep properties, creep rupture properties and fatigue properties of NIMS Structural Materials Data Sheets.

> MORE

> LOGIN

CCTD

The CCT Diagram Database (CCTD) contains information on CCT diagrams for welding various steel materials and related data.

> MORE

> LOGIN

NIMS Structural Materials Data Sheet Online

Creep Data Sheet

The Creep Data Sheet (CDS) contains PDF documents of the Creep Data Sheet published by NIMS.

> MORE

> LOGIN

Fatigue Data Sheet

Fatigue Data Sheet (FDS) contains PDF documents of the Fatigue Data Sheet published by NIMS.

> MORE

> LOGIN

Corrosion Data Sheet

The Corrosion Data Sheet (CoDS) contains PDF documents of the Corrosion Data Sheet published by NIMS.

> MORE

> LOGIN

Space Use Materials Strength Data Sheet

The Space Use Materials Strength Data Sheet (SDS) contains PDF documents of the Space Use Materials Strength Data Sheet published by NIMS.

> MORE

> LOGIN

Applications

CompoTherm

The Composite Design & Property Prediction System (CompoTherm) is an advanced simulation tool for searching for candidate composite materials with optimal thermo-physical properties and structures.

> MORE

SurfSeg

The Metal Segregation Prediction System (SurfSeg) is a system that predict of surface segregation between two annealed metals.

> MORE

InterChemBond

Interface Bonding Prediction System (InerChemBond) is a system that predict of interface chemical bonding between metal oxide (AxOy) and metal (M) or alloy (MB dissolved in MA).

> MORE

DICE: Lab data (closed)

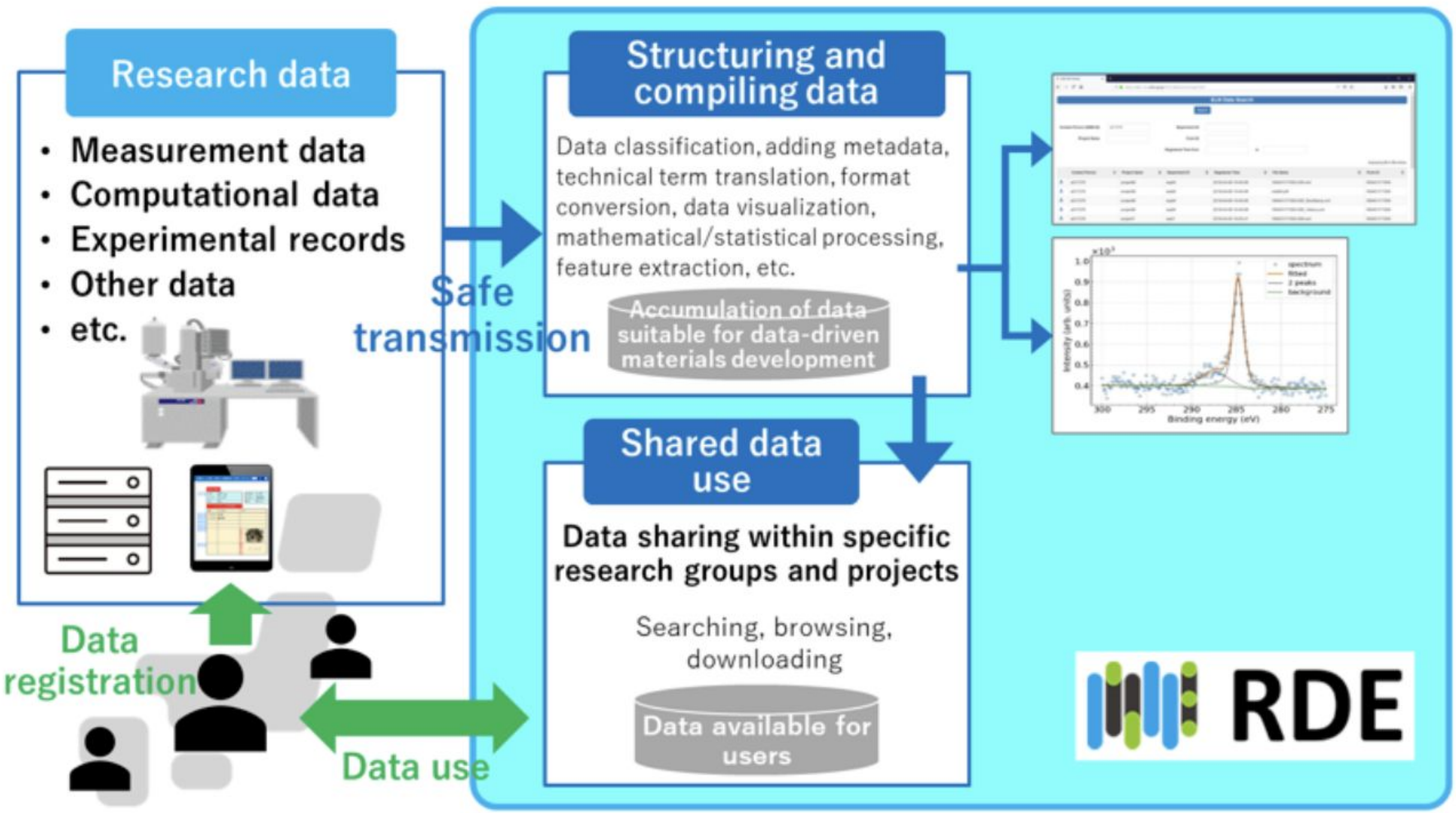
- from experimental facilities and IoT
- to data repository

Credit : Mikiko Tanifuji

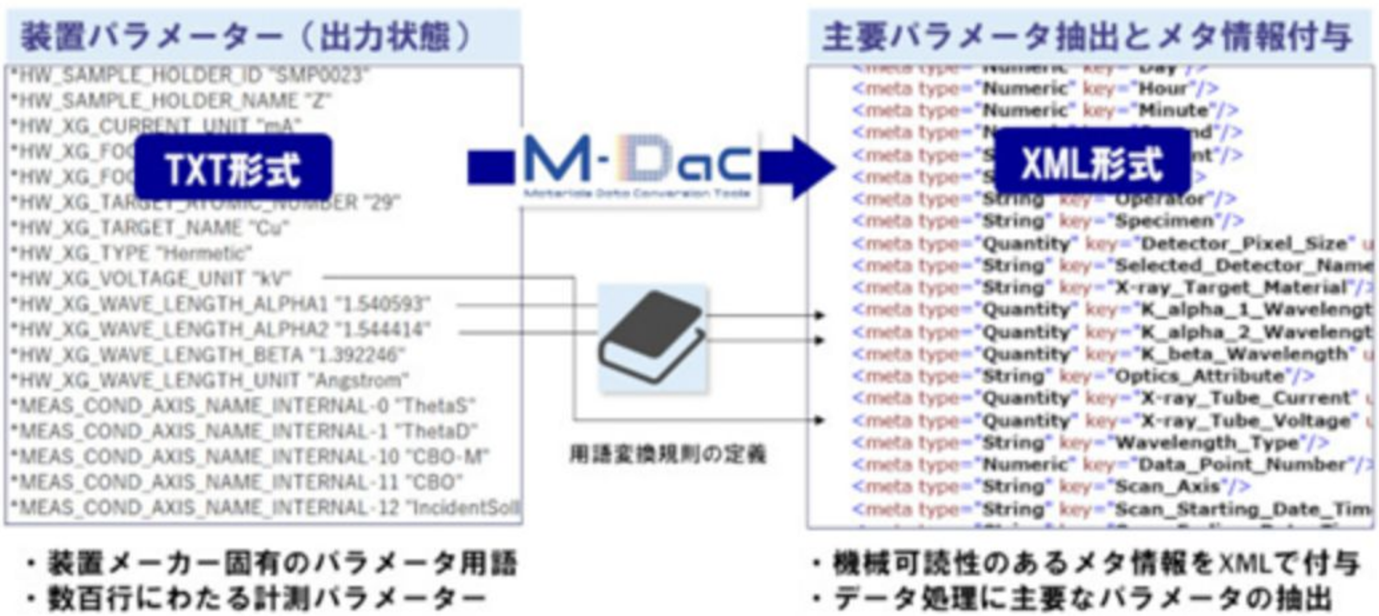
MDR MATERIALS DATA REPOSITORY

The Materials Data Repository (MDR) is a repository to collect and host not only papers and presentations, but also materials data, providing them for use in further materials research and materials informatics. Users can discover publications and datasets using metadata tailored for materials or by a full-text search, and can view and download them.

RDE



Research data
• Measurement data
• Computational data
• Experimental records
• Other data
• etc.
Safe transmission
Structuring and compiling data
Data classification, adding metadata, technical term translation, format conversion, data visualization, mathematical/statistical processing, feature extraction, etc.
Accumulation of data suitable for data-driven materials development
Shared data use
Data sharing within specific research groups and projects
Searching, browsing, downloading
Data available for users
Data registration
Data use
Binding energy (eV)
RDE



M-DaC is a set of software tools that extract meta-information such as measurement conditions and specimen information from raw data generated by measurement instruments and it converts them into highly machine-readable XML files.

Thank you for listening

- Comments or questions

  DIEB.Sae@nims.go.jp