Data management and adoption of the FAIR principle; perspective from a research institution

NATIONAL INSTITUTE FOR MATERIALS SCIENCE (NIMS)
MATERIALS DATA PLATFORM CENTER (DPFC)

MIKIKO TANIFUJI, MANAGING DIRECTOR
2021.6.18
Research Data Platform, DICE – Starting 2020

Create
Public Knowledge Data
- Articles
- Experts
- AI

Store
Big scale data processing clusters
10PB/0.5PFLOPS (NIMS, Tsukuba)

Use
R&D
Government institutes, Universities, Industries

Publish
Publish data with DOI via data repository
Share data under research schemes
MDR Shared
Peer-review journal

A sister journal of STAM (IF5.8)

Experimental Data
- Raw data, metadata
- Conversion to Machine-readable, Machine-comprehensive data

Findable Accessible Readable & Reusable Interoperable
Applications & Tools
- AI

Public Knowledge
Data

Experimental Data

AI
ML
ML
研究データをつくる・ためる・つかう、とは？

- 測ったそばから使える形でためる！

成膜装置：新材料の合成
XRF：元素組成の分析
XRD：結晶構造の分析
AFM：表面形状の分析
XPS：化学種の分析

データ構造化・登録システム

様々な実験装置からのデータを一覧

実験データ 説明情報
タグ付

検索やフィルタで絞り込み

実例：
- 3次元アトムプロープのFIB加工条件の記録
- コンピューターテクノロジーによって合成した組成傾斜膜の蛍光X線分析による組成分布画像の記録

論文：
"IoT データ収集システムのデータアーキテクチャ" 情報処理学会論文誌デジタルプラクティス（2021）

For details:
Materials Data Platform - a FAIR System for Data-Driven Materials Science
DOI: 10.1109/IIAI-AAI.2019.00206
DICE Data Services – to use – by FAIRable

Research Data
To collect, To be high valuable

Research data
To store & O/C management

Research data
To share, To publish

Research data
As public knowledge databases

Access monitoring, security systems

Analysis Cluster 100 virtual machines
What we see by CHORUS – capture rates of articles

Research data publishes as part of articles. Therefore, capturing timely and higher rates are desirable!
What we see by CHORUS – Capturing Open Access is not high.

45% of NIMS articles is open access. But capturing rate is not high. Why?
What we see by CHORUS – VOR vs AM

Among top 5 publishers, AM is well distributed APS (No.1) and Elsevier (No.2).

Appreciated publishers who supports authors versions with DOI distributions (so that we can capture via CHORUS).
### What we see by CHORUS – Data sets

<table>
<thead>
<tr>
<th></th>
<th>2021年2月</th>
<th>2021年6月</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIMS articles</td>
<td>5,747</td>
<td>6,176</td>
</tr>
<tr>
<td>No of data sets in articles</td>
<td>97</td>
<td>96</td>
</tr>
<tr>
<td>No of DOI of data sets</td>
<td>242</td>
<td>232</td>
</tr>
<tr>
<td>CCDC &amp; Fiz Karlsruhe</td>
<td>212</td>
<td>207</td>
</tr>
<tr>
<td>Zenodo</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>figshare</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>その他</td>
<td>5 ※1</td>
<td>4 ※2</td>
</tr>
<tr>
<td>Crossref</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

All are the DOIs of articles (not data sets;)

※1：Mendeley Data, Dryad, Citrine Informatics, Apollo (Univ. of Cambridge), CXIDB: Coherent X-ray Imaging Data Bank

※2：Mendeley Data, Dryad, Citrine Informatics, Apollo (Univ. of Cambridge)
What we see by CHORUS – Data sets for scientific use! – The most important reason.

CHORUS: breakdown use of data sets with articles

Reference: Web of Science: Data Citation Index (DCI)
NIMS data sets indexed in databases, repositories

Research data are often indexed in data catalogues and database (ex. CCDC) that are very important FAIR channels as truly useful data resources.
DICE is a data platform for all experts and applications for materials science.

"As a researcher, it is tedious to provide a universal, machine-learned dataset to be used in serendipitous ways."

"As a materials scientist, there is a particular material I am looking to design a model for AI-assisted research."

"As a data scientist, I am looking for materials data that I can use.

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

Acknowledgement:
Many thanks to Ms. Chie Onodera, a Digital Librarian at NIMS Library. Our ideas come from day-to-day wondering, "Why? How?? Still ... But!", And we are still moving on towards Materials DX.