A Lightweight Introduction to FAIR Digital Objects

Nicolas Blumenröhr – nicolas.blumenroehr@kit.edu
KARLSRUHE INSTITUTE OF TECHNOLOGY
The Struggle of (Meta)data Management

- Scientists waste time with data wrangling

Let’s consider an example for NMR data acquisition:

- **Data collection**: I need NMR spectra data
- **Searching different storage systems**
- **Dealing with different metadata contents**
- **Many steps done manually**

- FAIR Principles provided guidelines for improved stewardship and management
- **But, their implementations are not fully aligned**
Dealing with Storage Systems

• A variety of storage systems exist that implement FAIR principles
• Information retrieval works via access protocols and metadata descriptions
• These are typically diverse

For example, to retrieve license information for data reuse:

In the Zenodo repository:

In the NMRxiv database:

Creative Commons Attribution Share Alike 4.0 International (CC BY-SA 4.0)
Dealing with Metadata

- Metadata Schemas and Standards help to define, organize, and manage metadata
- Enable Machine-readability and interpretability
- Differ between- and within disciplines regarding structure, contents and formats
- Metadata schemas often have different structures and vocabularies

```
"rightsList": [{"lang":"en","rights":"Creative Commons Attribution 4.0 International","rightsUri":"https://creativecommons.org/licenses/by/4.0/legalcode","schemeUri":"https://spdx.org/licenses/","rightsIdentifier":"cc-by-4.0","rightsIdentifierScheme":"SPDX"}]
```

```
"license":{"title":"Creative Commons Attribution 4.0 International (CC BY 4.0)","slug":"cc-by-4.0","spdx_id":"CC-BY-4.0","url":"https:\/\/creativecommons.org\/licenses\/by\/4.0\/...
```
Machine-actionability

- Requires machine-readability and interpretability
- Automated systems act on digital resources and their metadata
- No, or less human intervention is required
How to Tackle?

• High-level information should be harmonized
• Underlying systems and standards must not be changed

Harmonization of essential information

Additional, but uniform representation format

License example:
• Name: License
• Description: A URL referring to a license that defines the scope of use for a digital resource
• A unique PID
• Typing --> is a URL for a existing license from an enumeration list
The FAIR Digital Objects Concept

- Representation of digital resources in a uniform way (Digital Object)
- Integrates the essential elements for FAIRness
Information is Reduced and Standardized

- Each FAIR Digital Object (FDO) is based on the same structure
- Information at this level is unified and can be treated equally

Harmonization of essential information

- Readable and interpretable for humans and machines

PID record:

... License(PID):
Name: License
Value: https://.../licenses/by/4.0/..

... License(PID):
Name: License
Value: https://.../licenses/by-sa/4.0/..
This initiative has received funding from the EU’s H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

Original Handle Record Example

Handle Value for: 21.11152/865d3383-55a4-4620-b4ef-e806382e7e09

<table>
<thead>
<tr>
<th>Index</th>
<th>Type</th>
<th>Timestamp</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>21.T11148/6ae999952a0d2dca14d6</td>
<td>2024-05-31 19:42:07Z</td>
<td>Aminolevulinic Acid</td>
</tr>
<tr>
<td>4</td>
<td>21.T11148/aafdf5fb4e7222e2d950a</td>
<td>2024-05-31 19:42:07Z</td>
<td>2004-09-16T00:00:00.000000Z</td>
</tr>
<tr>
<td>5</td>
<td>21.T11148/397d831aa3a9d1e8eb52c</td>
<td>2024-05-31 19:42:07Z</td>
<td>2024-05-25T00:00:00.000000Z</td>
</tr>
<tr>
<td>6</td>
<td>21.T11148/2f3f14c8fe5fb6a0063a8</td>
<td>2024-05-31 19:42:07Z</td>
<td><a href="https://creativecommons.org/licenses/by/4.0/deed.en">https://creativecommons.org/licenses/by/4.0/deed.en</a></td>
</tr>
<tr>
<td>9</td>
<td>21.T11148/e83481d4bf467110e7e9</td>
<td>2024-05-31 19:42:07Z</td>
<td>application/zip</td>
</tr>
<tr>
<td>10</td>
<td>21.T11148/82e2503c34920ee987740</td>
<td>2024-05-31 19:42:07Z</td>
<td>{ &quot;sha256sum&quot;: &quot;5174fd6992c4a6c1f718711a19d2c6314d6908402488:</td>
</tr>
<tr>
<td>13</td>
<td>21.T11148/60d1c34a6ab5d67049f</td>
<td>2024-05-31 19:42:07Z</td>
<td>C5H9NO3</td>
</tr>
<tr>
<td>100</td>
<td>HS_ADMIN</td>
<td>2024-05-31 19:42:07Z</td>
<td>handle=21.11152/USER01; index=300; [create hdle,delete hdle,create del admin,del admin,add admin,list]</td>
</tr>
</tbody>
</table>

Enables machine-readability and interpretability
Tooling for Content Assessment

**PID Information Record**

<table>
<thead>
<tr>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>kernelInformationProfile</td>
<td>21.T11148/b9b76f887845e32d29f7</td>
</tr>
<tr>
<td>dateModified</td>
<td>2023-08-01T00:00:00+00:00</td>
</tr>
<tr>
<td>checksum</td>
<td>&quot;sha512sum&quot;: &quot;f0a6e42dc67335e6857b6&quot;</td>
</tr>
<tr>
<td>dateCreated</td>
<td>2023-02-07T00:00:00+00:00</td>
</tr>
</tbody>
</table>

**FAIR DO Graph**

https://kit-data-manager.github.io/fairdoscope/?pid=21.11152/b0b5de04-6e11-480b-ab66-2d4a5f42ea9e
Machine-Actionability as the Final Goal

- What do we need?
  - An entry point for the user
  - An infrastructure to implement the concept
  - Software to work with the components
  - A mechanism for machines to act on the components
NEP Virtual Access Services

- NMR Graph for retrieval of NMR spectra resources
  - A service to provide a unified search interface of NMR spectra data
- MRI Prediction tool for the prediction of Magnetic Resonance Image data (DICOM format)
  - A service that uses AI software to estimate experimental outcomes by existing results
NMR Graph Service - Motivation

https://www.sciencedirect.com/topics/chemistry/1h-nmr-spectrum

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PCA</td>
<td>H</td>
<td>H</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PCA</td>
<td>HA</td>
<td>H</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PCA</td>
<td>HB3</td>
<td>H</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PCA</td>
<td>HB2</td>
<td>H</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PCA</td>
<td>HG3</td>
<td>H</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PCA</td>
<td>HG2</td>
<td>H</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>GLN</td>
<td>H</td>
<td>H</td>
<td>1</td>
<td>8.23</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>GLN</td>
<td>HA</td>
<td>H</td>
<td>1</td>
<td>4.26</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>GLN</td>
<td>HB2</td>
<td>H</td>
<td>1</td>
<td>1.75</td>
</tr>
</tbody>
</table>

• NMR spectra are provided in different formats, are distributed over different storage systems and described using different metadata schemas

• Relations to related digital resources like publications, metadata documents or software are not easily assessable

• A common representation of various NMR spectra enables a unified search interface at this level

• Discovery, evaluation and retrieval is facilitated
What is the Baseline?

• Certain types of metadata exists for all NMR spectra resources and is typically used for information retrieval.

• This information can be unified and transferred into a machine-readable and interpretable format.

• We used the concept of FAIR Digital Objects (FDOs) to describe various NMR spectra this way.
Graph Format for Extended Usability

- FDOs are entities that contain reusable, interconnected elements
- A graph representation enables the assessment of contents these FDOs describe
- Assessment by graph queries using SPARQL

```sparql
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX fdo: <http://anonymized-namespace/FDO-Graph>

SELECT ?profile ?operation ?expectedOutput ?fdo
WHERE {
  VALUES ?profileName {'Profile1' 'Profile2' 
  FILTER(?profileName IN (?profile))
  ?fdo a fdo:FDO ; fdo:hasProfile ?profile ;
  rdfs:label ?fdo .
  ?operation a fdo:Operation ; fdo:isOperationFor ?fdo ;
  rdfs:label ?operation .
  ?attribute a fdo:Attribute ; rdfs:label
  ?expectedOutput .
}
```

- Search interface for users (via the GUI) and query endpoint for machines
- Try it out:
  - Visit: [https://metarepo.nffa.eu/start_query](https://metarepo.nffa.eu/start_query)
MRI Prediction Tool - Motivation

• Magnetic Resonance Imaging is a measurement technique mostly known from medical imaging – also applied in the materials science field
• Measurements take long time
• Often, many measurement sequences are required for analytics
• The tissue contrast (T1, T2, PD) must be optimized

• Can be reduced by digital acquisition of estimated measurement results

https://mriquestions.com/image-contrast-trte.html
The Approach

- Two main parameters that need to be adjusted – TE and TR
- Instead of measuring each parameter setting, a minimum of required experimental data is collected and applied to an AI model
- Model predicts the image of an alternative parameter setting

CuSO₄ in Millimolar (mM)

T1- weighted - TE: 5 ms
TR: 100 ms

T2- weighted - TE: 25 ms
TR: 5000 ms
What is Possible?

• Currently, the model is specialized for a particular sample type
• Perspectives: prediction of images for a more versatile sample set

• Try it out:
  • Download a test file (DICOM format)
  • Visit: [https://metarepo.nffa.eu/prediction](https://metarepo.nffa.eu/prediction)