HELMHOLTZ Metadata Collaboration

Karlsruhe Institute of Technology

Metadata Management: From Key Essentials to Practice
Helmholtz Metadata Collaboration (HMC)

Sabrine Chelbi

Scientific Computing Center (SCC) – Karlsruhe Institute of Technology (KIT)

Learning outcomes

- Understand the importance and added value of metadata.
- Get an overview of JSON Schema.
- Learn how to create, edit, and manage metadata.
- Search for data from existing metadata.

Outline

First part (Theory Part):

- Motivation
- Recap of FAIR principles
- Basics of metadata
- Basics of JSON Schema and metadata documents
- Try the Game

Second part (Practical Part):

- Use Case Introduction
- Intro to the Demo Frontend
  - Mapping Service
  - Metadata Repository
  - Site Search
- Metadata Editor
- Survey

Data-related research use cases

You might want to:

- Compare your results with similar ones in the literature
- Reproduce/reuse results available in the literature
- Take delivery of the project handed over by a student/colleague who left
- Exchange data with your colleagues to collaborate on a research project
- Allow others to reproduce/reuse your results to be cited

Data-related research questions

- Which data do support the results in this paper?
- How can I reproduce the data?
- How were these measures achieved/performed?
- How can I search for data with specific features?
- How can I publish my data in such a way that others can reuse it and cite it?

The FAIR Guiding Principles

https://www.go-fair.org/fair-principles/

The FAIR Guiding Principles

Findable

(Meta)data should be easy to find for both humans and computers

PIDs

- F1: (Meta)data are assigned globally unique and persistent identifiers (PIDs)
- F3: Metadata clearly and explicitly include the identifier of the data they describe

The FAIR Guiding Principles

Accessible

It should be known how (meta)data can be accessed

Metadata Repositories

- A2: **Metadata should be accessible** even when the data is no longer available
- A1.2: The protocol allows for an **authentication and authorization (AAI)** procedure where necessary.

The FAIR Guiding Principles

Interoperable

Data should be exchanged and interpreted by humans and computers

Structured metadata

- I1: (Meta)data use a formal, accessible, shared, and broadly applicable **language** for knowledge representation

The FAIR Guiding Principles

Reusable

It should be clear how data can be reused and/or replicated

Standards, Licence

- R1: Metadata should richly describe the data with a plurality of **accurate and relevant attributes**
- R1.1: (Meta)data are released with a clear and accessible data **usage licence**

Basics of metadata: concepts

Data vs Metadata

Structured Metadata

Metadata Standards

Metadata Repositories

PIDs

Licences

Data vs Metadata

Metadata: data describing data.

https://imgflip.com/i/92po3j

Data vs Metadata

“Data is stuff. It is raw, unprocessed, possibly even untouched by human hands, unviewed by human eyes, un-thought-about by human minds”.

J. Pomerantz (2015). Metadata. The MIT Press.

“Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource”.

National Information Standards Organization (2004) from “Big Data, Little Data, No Data”, C. L. Borgman (2015)

Data vs Metadata

	Data	Metadata
Nature and Content	**Raw facts**, measurements, observations	Information **providing context** and attributes
Usage	Analysis, decision making, **research**	Data management, discovery, **interpretation**
Representation	Mostly **unstructured**	**Structured**
Relationship	Independent and **stand-alone**	**Linked** to data/other metadata
Purpose and Function	Primary **source of information**	**Supporting framework** for organization, management, interpretation

Jejkal T., How to Manage Metadata. The MetaRepo: Introduction (2024). DOI: 10.5445/IR/1000172638

Basics of metadata: concepts

Data vs Metadata

Structured Metadata

Metadata Standards

Metadata Repositories

PIDs

Licences

Structured metadata

Format: the language for knowledge representation and exchange

Semantics: the (agreed) terminology to describe the attributes

Schema: the template which specifies the expected attributes and how they should be structured

Metadata format: XML

- eXtensible Markup Language
- Main purpose: transfer and storage of arbitrary data on the World Wide Web
- Human- and machine-readable
- Hierarchical (tree-like) structure
- Elements are wrapped in start <...> and end </...> “tags”

```
<example>
    <title>This is the example title</title>
    <description>A simple XML example</description>
    <wordCount>1</wordCount>
</example>
```

```
<person>
  <firstName>John</firstName>
  <lastName>Doe</lastName>
</person>
```

Gerlich, S., Strupp, A., Hofmann, V., Sandfeld, S. (2023). Fundamentals of Scientific Metadata. The Carpentries Incubator. DOI: 10.5281/zenodo.10091708

Metadata format: JSON

- JavaScript Object Notation
- Main purpose: transfer and storage of data
- Human- and machine-readable
- Hierarchical structure
- Elements are defined in key:value pairs
- Elements are wrapped as {objects} or [arrays]

```
{
    "key":"value",
    "aString":"string",
    "anInteger":5,
    "aFloat":0.5,
    "aBoolean":true,
    "anArray": ["item1", "item2", "item3"],
    "anObject": {
        "key1":"value1",
        "key2":"value2",
        "key3":"value3"
    }
}
```

Gerlich, S., Strupp, A., Hofmann, V., Sandfeld, S. (2023). Fundamentals of Scientific Metadata. The Carpentries Incubator. DOI: 10.5281/zenodo.10091708

Semantics

- **Vocabulary:** set of terms pertaining to a particular domain + definitions. Useful to ensure that the data is described consistently

- **Taxonomy:** hierarchical structure of the terms. Useful to organize data into categories which are meaningful in a particular domain

- **Ontology:** formal description of the terms, their properties and their relationships within a particular domain. Useful to consistently represent the knowledge about a domain

Metadata Schema

- Template which specifies the expected elements and how they are structured:
  - Names (vocabulary)
  - Value types
  - Rules
  - Mandatory/optional
- XML Schema Definition (XSD): less frequently used
- JSON Schema: uses the JSON Schema Vocabulary https://json-schema.org to specify & syntactically validate the structure

Metadata Schema vs Metadata Document

Metadata Schema

Photo by M. Coghlan on Flickr (licence CC-BY-SA 2.0)

Metadata Document

Photo by M. Carrati on Unsplash

Recchia A., How to edit metadata. The Metadata Editor (2024)

Metadata Schema vs Metadata Document

Metadata Schema

Template which specifies the expected elements and how they are structured

```
"givenName": {
    "type": "string",
    "description": "(Optional) - Given name of the user"
},
"familyName": {
    "type": "string",
    "description": "(Optional) - Family name of the user"
},
"age": {
    "type": "number",
    "description": "(Optional) - Age of the user"
},
```

Metadata Document

An instance of a Metadata Schema which describes a given resource and conforms to the specified definitions

```
"givenName": "John",
"familyName": "Doe",
"age": 20
```

Benefits of structured metadata

- Same parameters for all data, more harmonized description
- Structured metadata can be more easily interpreted (also by machines)
- Results can be more easily reproduced/reused
- Data can be more easily compared/exchanged
- Data can be found based on their attributes (search, filter)
- Metadata can be validated.
    - Note: schema validation only checks for syntactical validity (required property, corresponding value, and whether the value conforms with the expected data type)

Basics of metadata: concepts

Data vs Metadata

Structured Metadata

Metadata Standards

Metadata Repositories

PIDs

Licences

Metadata Standards

- Metadata schemas which are well-established, endorsed, and widely accepted by the user community

General purpose

DublinCore http://dublincore.org/schemas/

DataCite http://schema.datacite.org

Schema.org https://schema.org

Neutron, x-ray, muon

NeXus http://www.nexusformat.org

```
▼<xs:sequence>
  ▼<xs:choice minOccurs="0" maxOccurs="unbounded">
     <xs:element ref="any"/>
   </xs:choice>
 </xs:sequence>
```

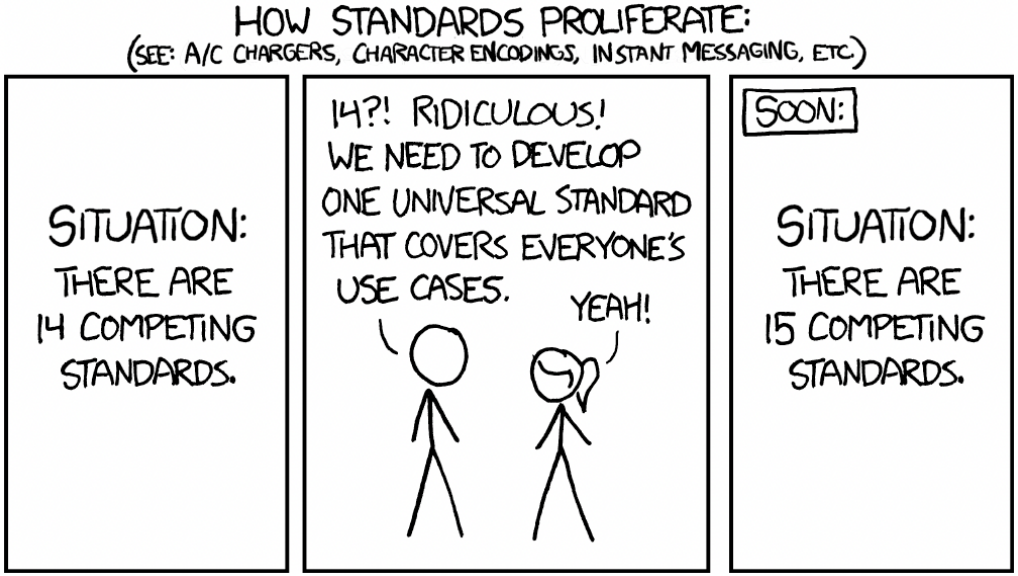
```
<xs:element name="givenName" minOccurs="0"/>
<xs:element name="familyName" minOccurs="0"/>
```

```
"colleague": [
  "http://www.xyz.edu/students/alicejones.html",
  "http://www.xyz.edu/students/bobsmith.html"
],
```

```
entry:NXentry
   raw:NXsubentry
      definition="NXsas"
   reduced:NXsubentry
      definition="NXcanSAS"
   fluo:NXsubentry
      definition="NXfluo"
```

Take-home message

- Before describing your data on your own, you should look for existing metadata schemas or standards.



https://xkcd.com/927/

Questions?

What if a standard does not exist?

Where do I publish my metadata?

How do I link my data to metadata?

How can my (meta)data be reused and cited?

Basics of metadata: concepts

Data vs Metadata

Structured Metadata

Metadata Standards

Metadata Repositories

PIDs

Licences

Metadata Repository

- “Information system used to store, manage and provide access to Metadata, following a policy or a set of rules that define storage and access norms.” Aversa R., et al. (2024) DOI: 10.5281/zenodo.10663833
- Register/find metadata schemas
- Register/find metadata documents
- Validate metadata documents against the schema
- Versioning
- Access control management
- User authentication

Basics of metadata: concepts

Data vs Metadata

Structured Metadata

Metadata Standards

Metadata Repositories

PIDs

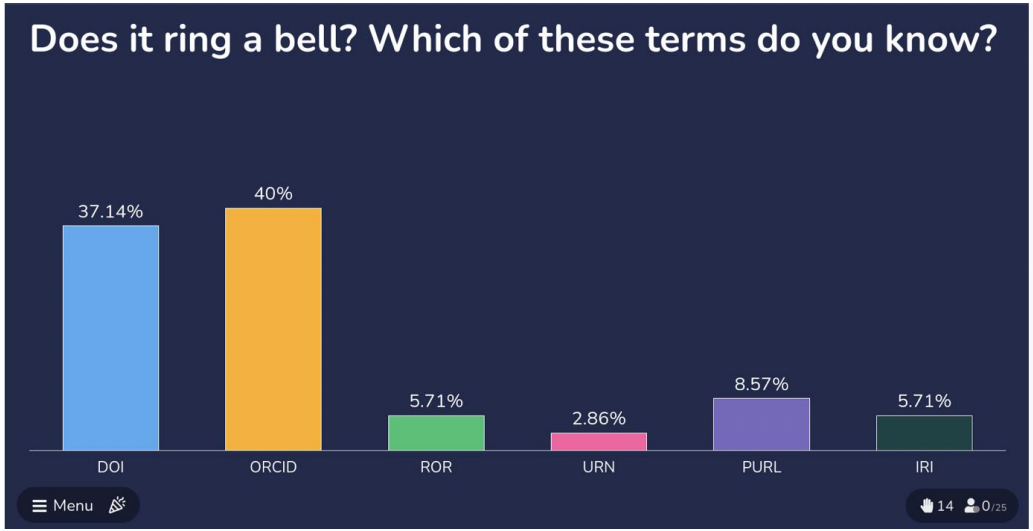
Licences

Persistent Identifiers (PIDs)

- **Persistent identifier:** long-lasting reference to locate and identify a resource, even if it changes over time
- Globally unique and persistent over time (until the PID provider is maintained!)
- Connected to a set of metadata describing a resource rather than to the resource itself
- **Benefit of PIDs:** allow different platforms to exchange information consistently, e.g. to track citations and reuse.

PID services

Mint, manage and resolve PIDs.



- DOI: Digital Object Identifier
- ORCID: Open Researcher and Contributor ID
- ROR: Research Organization Registry
- URN: Uniform Resource Name
- PURL: Persistent Uniform Resource Locator
- IRI: Internationalized Resource Identifier

Gerlich, S., Strupp, A., Hofmann, V., Sandfeld, S. (2023). Fundamentals of Scientific Metadata. The Carpentries Incubator. DOI: 10.5281/zenodo.10091708

Take-home message

- PIDs are largely used to identify researchers, institutions, research articles, data resources, metadata records, code, software, …
- Ready to publish?

Basics of metadata: concepts

Data vs Metadata

Structured Metadata

Metadata Standards

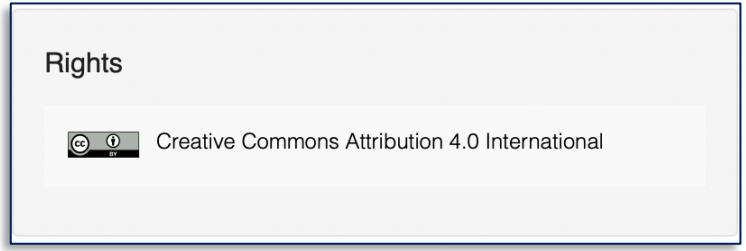
Metadata Repositories

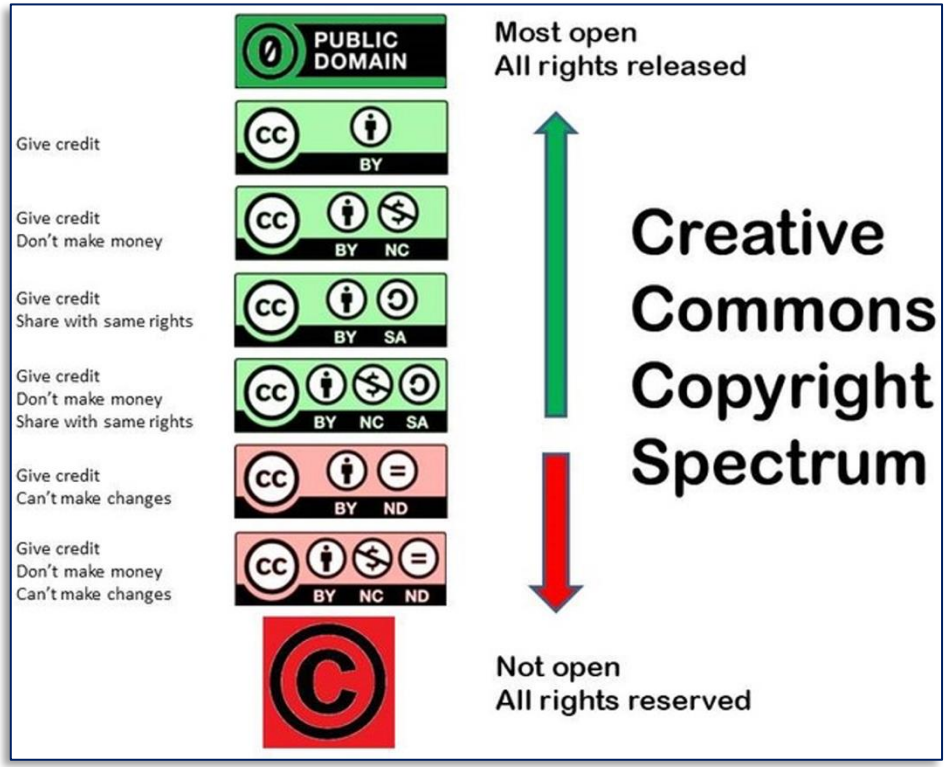
PIDs

Licences

Data Licence

Legal arrangement between the creator of the data and the end-user, or the place where the data will be deposited, specifying what users can do with the data.





https://creativecommons.org/share-your-work/cclicenses/

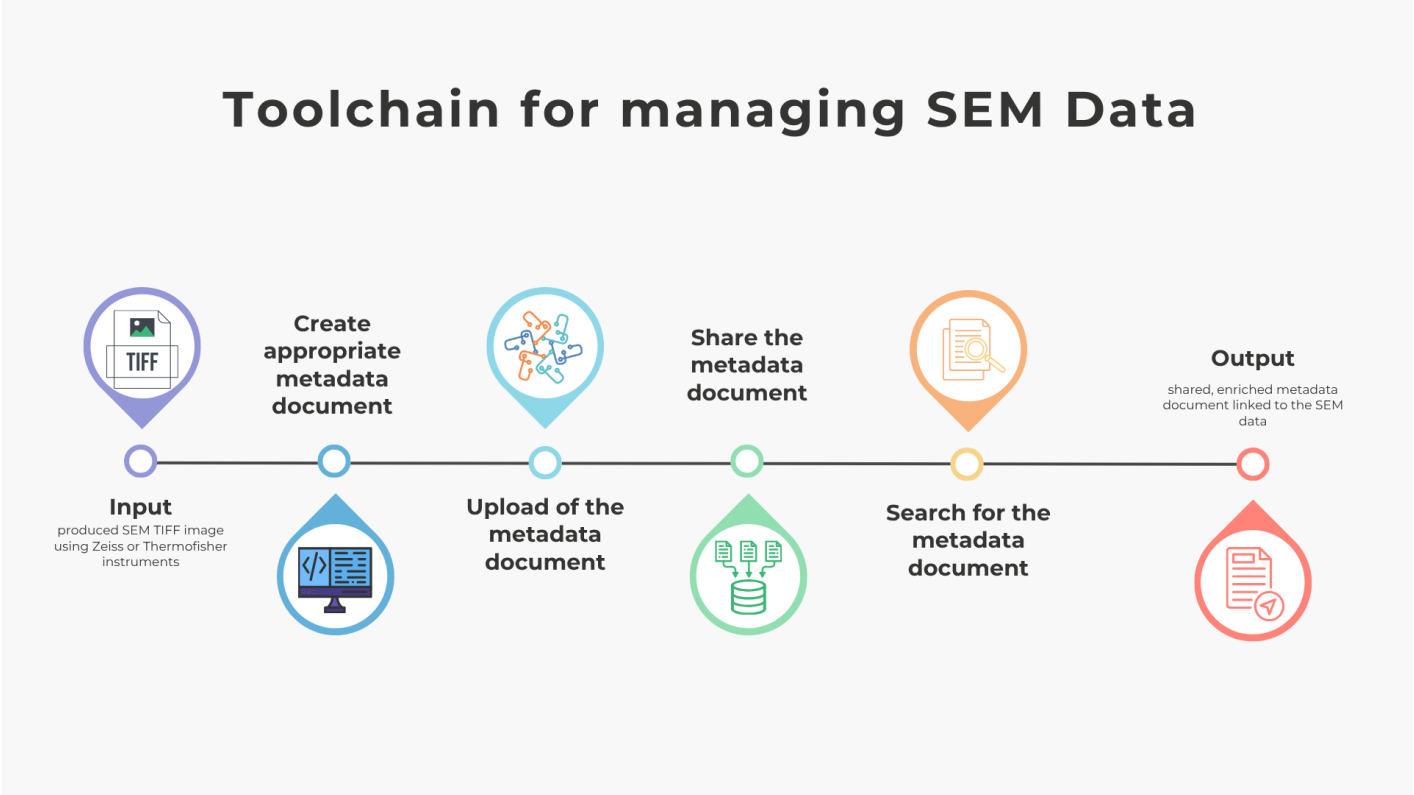
Conclusions

- FAIR Principles are guidelines for (meta)data management
- Structured metadata are helpful for better data interpretation, exchange, reuse
- Adoption of existing standards or community best practices avoids proliferation of descriptions
- Persistent and globally unique identifiers allows you to link resources and cite them
- Licences are keys when you publish or reuse results

Try the Game

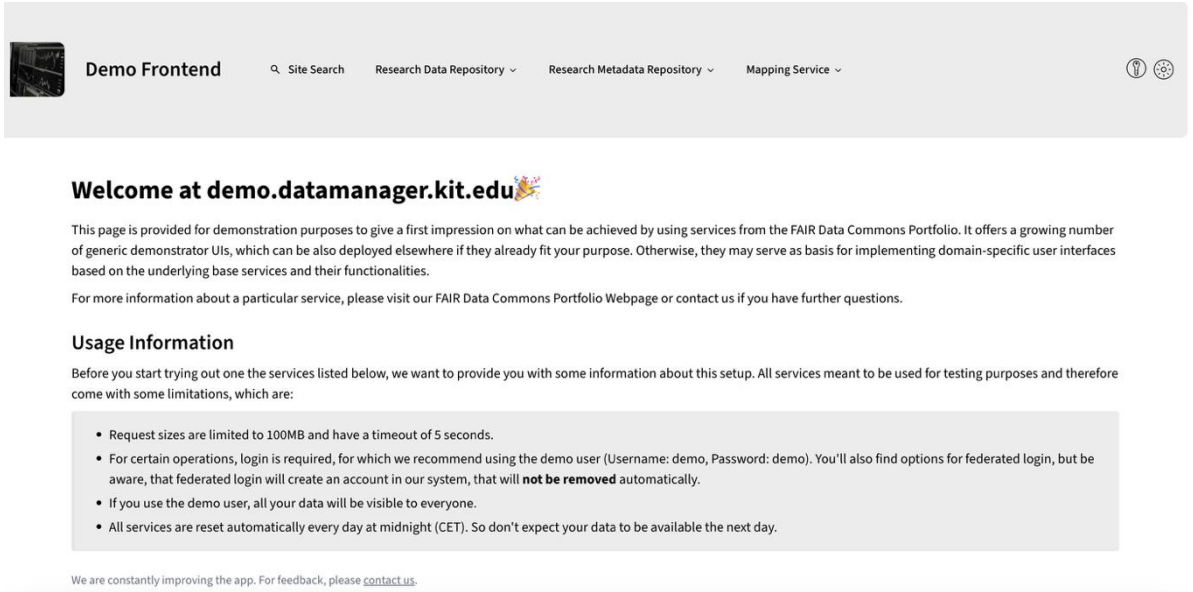
https://t1p.de/metadata-game

Use Case Introduction

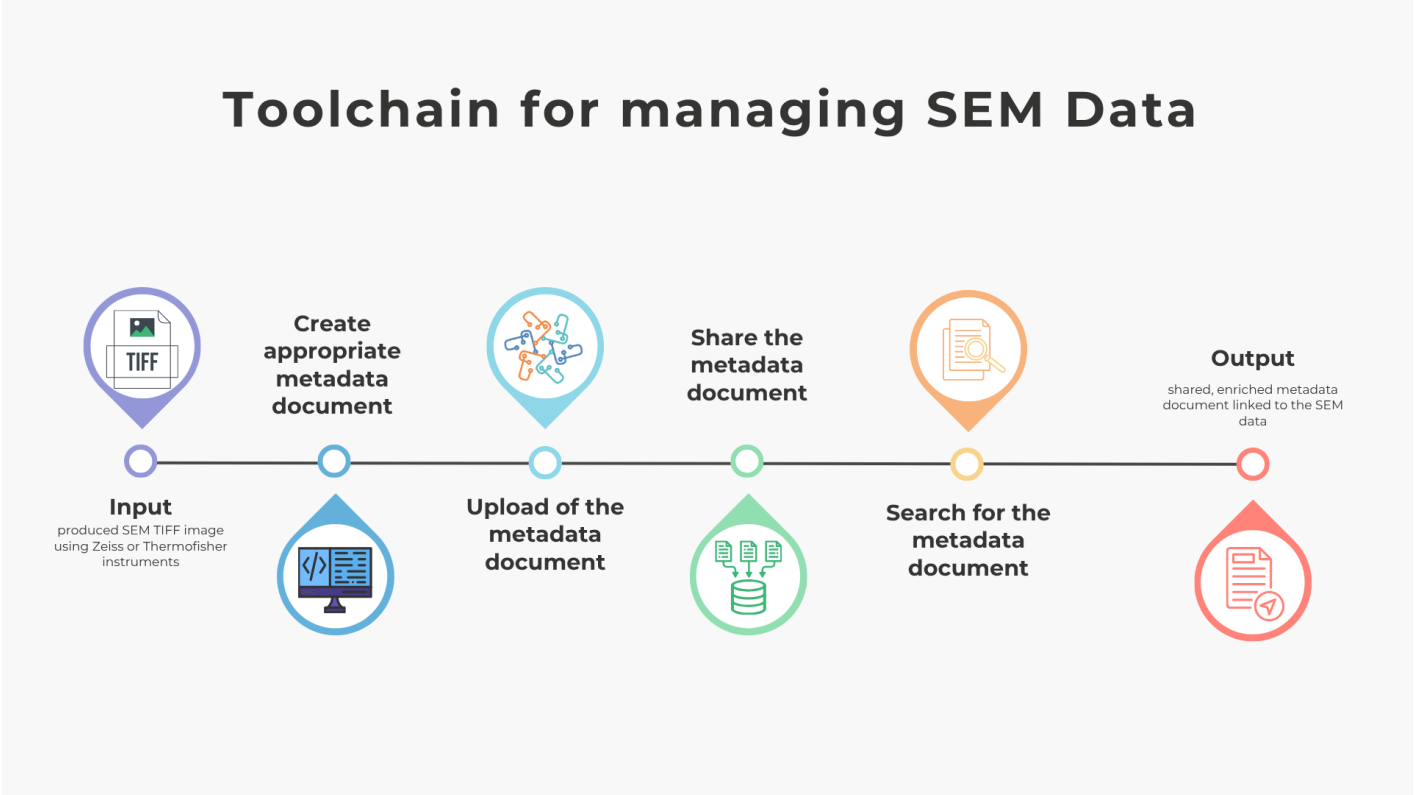


Intro to the Demo Frontend

- Provided services for demonstration purposes
  - Site Search
  - Data Repository
  - Metadata Repository
  - Mapping Service
- https://demo.datamanager.kit.edu/

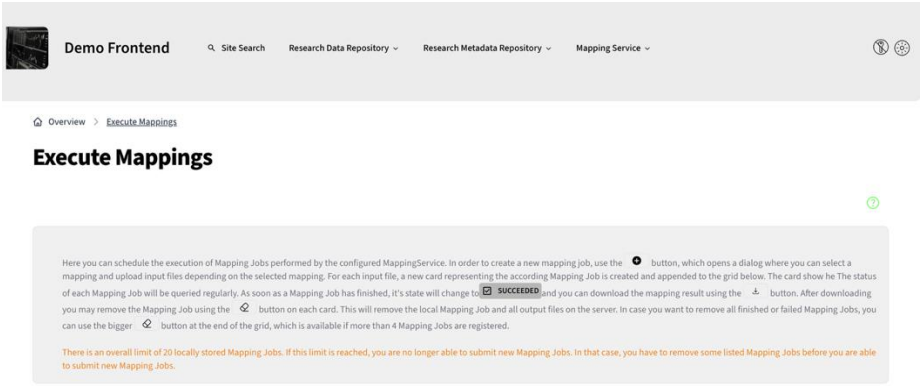


Use Case: Metadata Extraction



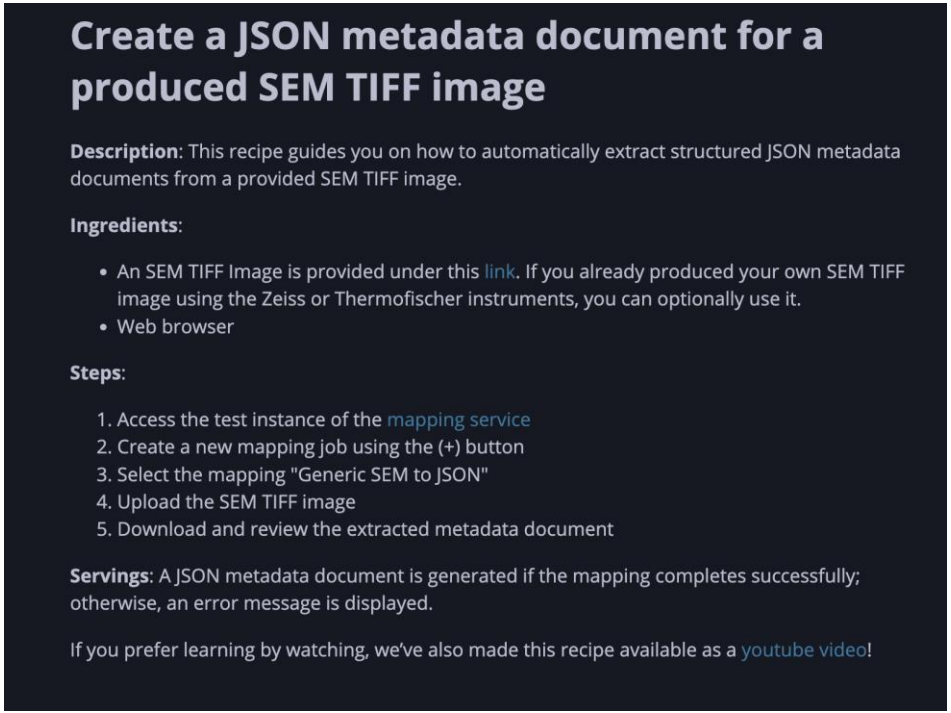
Metadata Extraction: Mapping Service

- Tool enabling the extraction of metadata from different data files and mapping the extracted metadata to published metadata schemas.
- Different mappings available:
  - Zeiss/Thermofischer/Tescan/Jeol SEM to JSON
  - MRI to JSON
  - APE-HE NeXus to JSON
- More information: https://t1p.de/mapping-service-docu
- Link to the service: https://demo.datamanager.kit.edu/mapping/



Metadata Extraction: Mapping Service

- SEM images available:
  - https://matwerk.datamanager.kit.edu/workshop/zeiss_sem/
- Link to the tutorial
  - https://t1p.de/metadata-extraction
- Link to the video
  - https://www.youtube.com/watch?v=0a-tA42eHuM



Create a JSON metadata document for a produced SEM TIFF image

Description: This recipe guides you on how to automatically extract structured JSON metadata documents from a provided SEM TIFF image.

Ingredients:

- An SEM TIFF Image is provided under this link. If you already produced your own SEM TIFF image using the Zeiss or Thermofischer instruments, you can optionally use it.
- Web browser

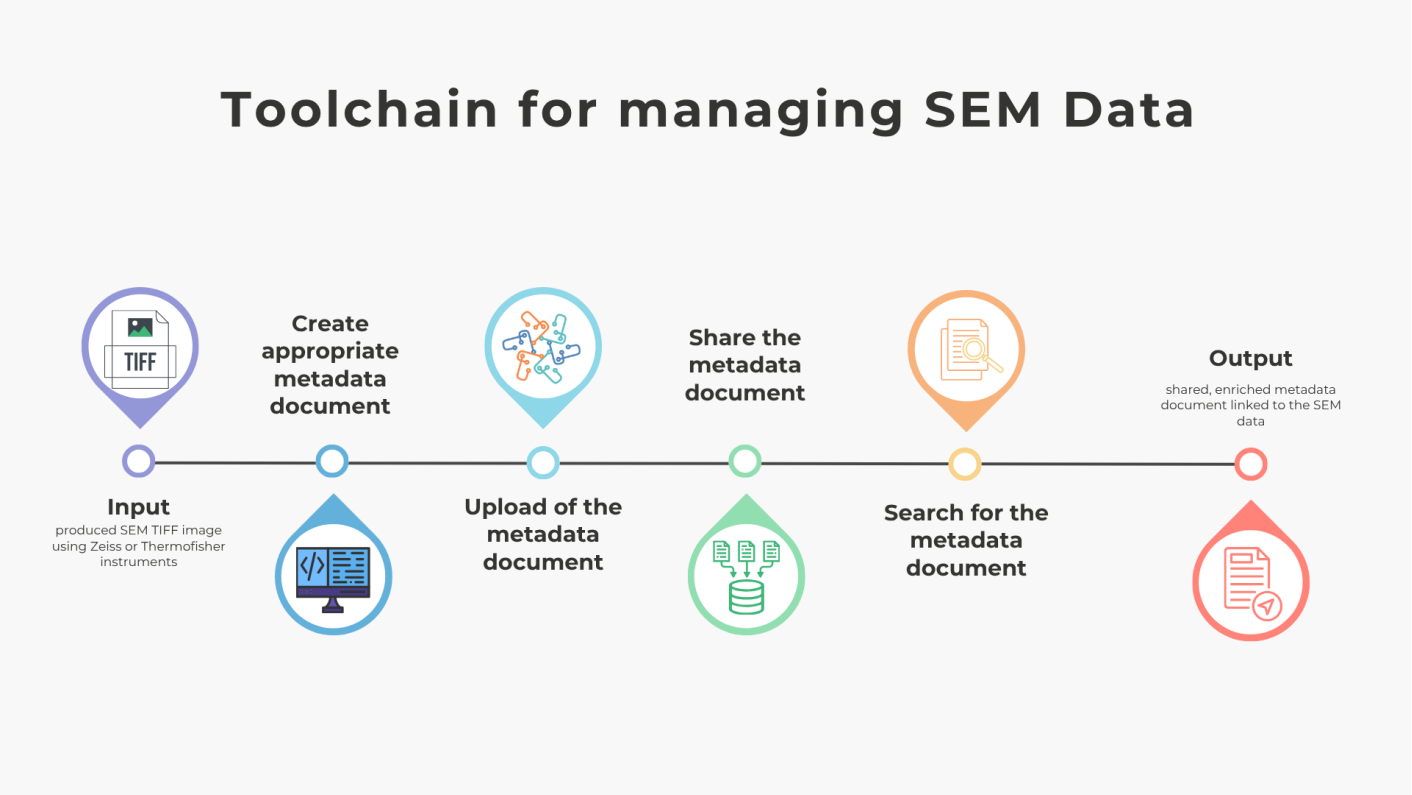
Steps:

1. Access the test instance of the mapping service
2. Create a new mapping job using the (+) button
3. Select the mapping "Generic SEM to JSON"
4. Upload the SEM TIFF image
5. Download and review the extracted metadata document

Servings: A JSON metadata document is generated if the mapping completes successfully; otherwise, an error message is displayed.

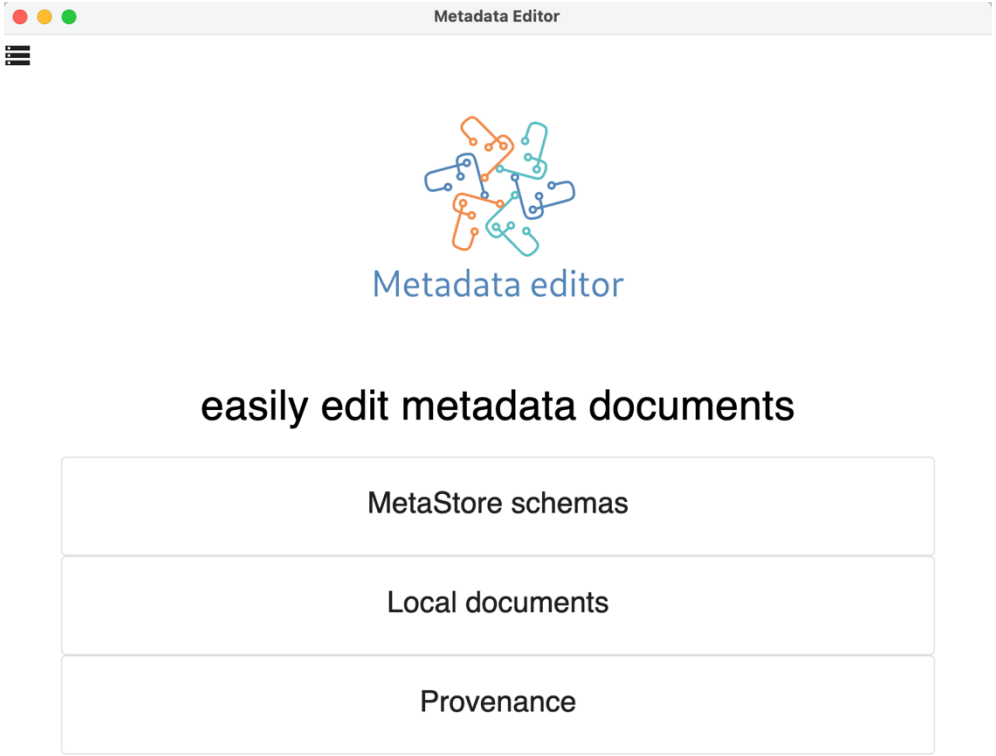
If you prefer learning by watching, we've also made this recipe available as a youtube video!

Use Case: Metadata Editing



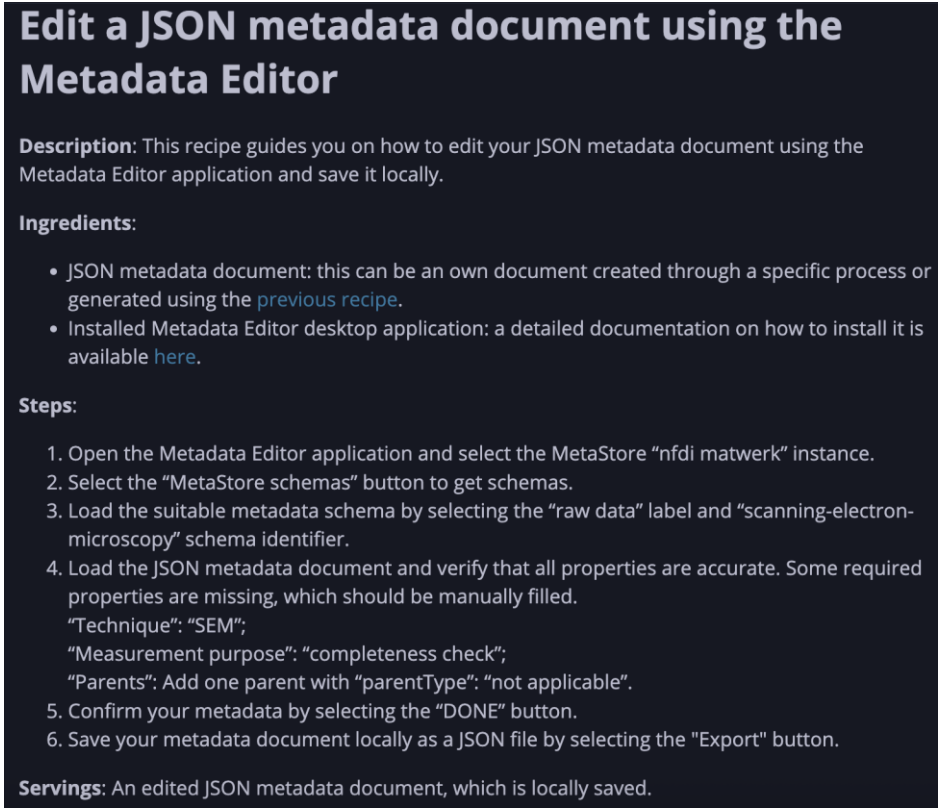
Metadata Editing: Metadata Editor

- desktop application developed and maintained by eXact lab
- intuitive user interface to support users while ingesting and editing their metadata documents
- retrieves schemas and uploads metadata documents from different metadata repository instances
- https://metadata-editor.gitlab.io/documentation/



Metadata Editing: Metadata Editor

- Link to the tutorial
  - https://t1p.de/metadata-editor
- Link to the video
  - https://www.youtube.com/watch?v=w4VlrsNK24I



Edit a JSON metadata document using the Metadata Editor

Description: This recipe guides you on how to edit your JSON metadata document using the Metadata Editor application and save it locally.

Ingredients:

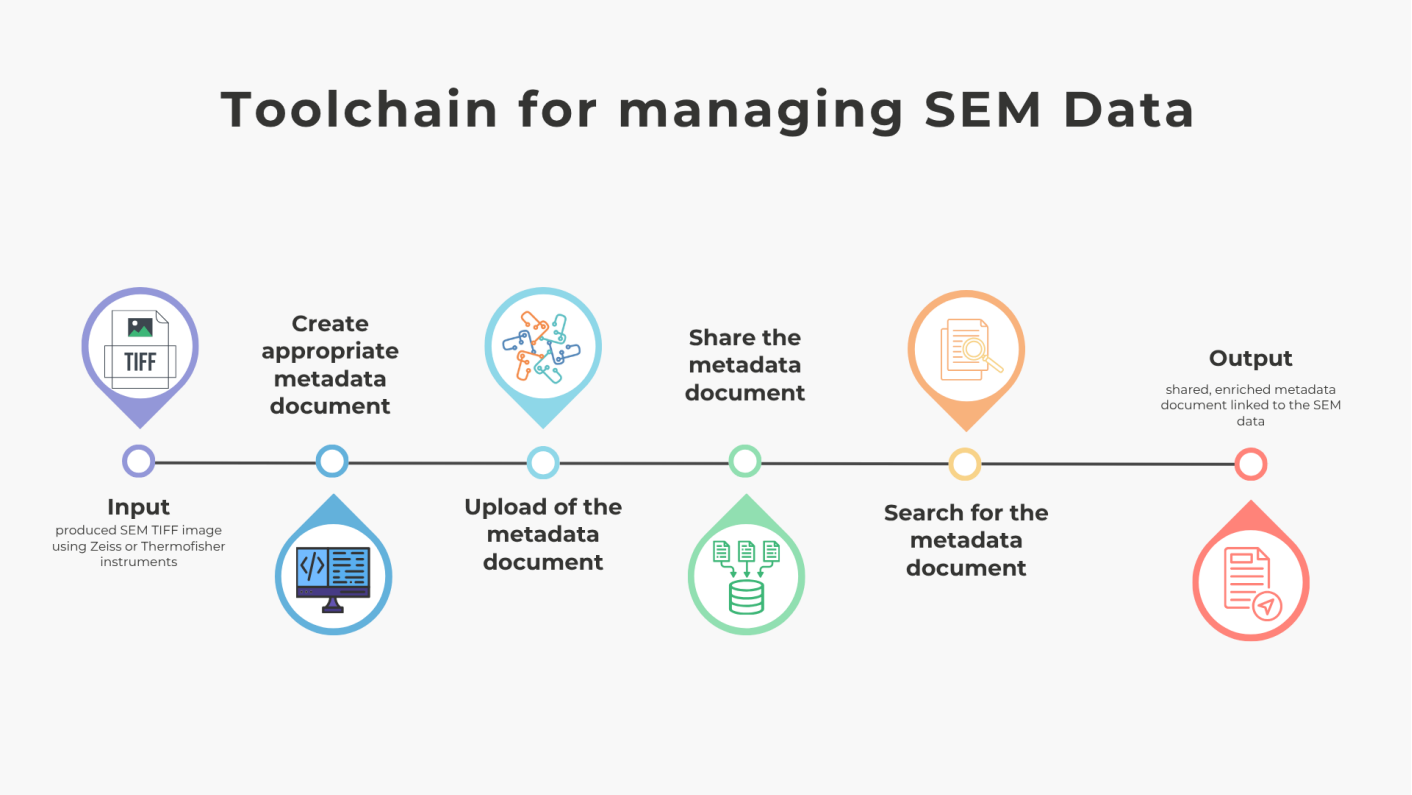
- JSON metadata document: this can be an own document created through a specific process or generated using the previous recipe.
- Installed Metadata Editor desktop application: a detailed documentation on how to install it is available here.

Steps:

1. Open the Metadata Editor application and select the MetaStore “nfdi matwerk” instance.
2. Select the “MetaStore schemas” button to get schemas.
3. Load the suitable metadata schema by selecting the “raw data” label and “scanning-electron-microscopy” schema identifier.
4. Load the JSON metadata document and verify that all properties are accurate. Some required properties are missing, which should be manually filled.
   “Technique”: “SEM”;
   “Measurement purpose”: “completeness check”;
   “Parents”: Add one parent with “parentType”: “not applicable”.
5. Confirm your metadata by selecting the “DONE” button.
6. Save your metadata document locally as a JSON file by selecting the "Export" button.

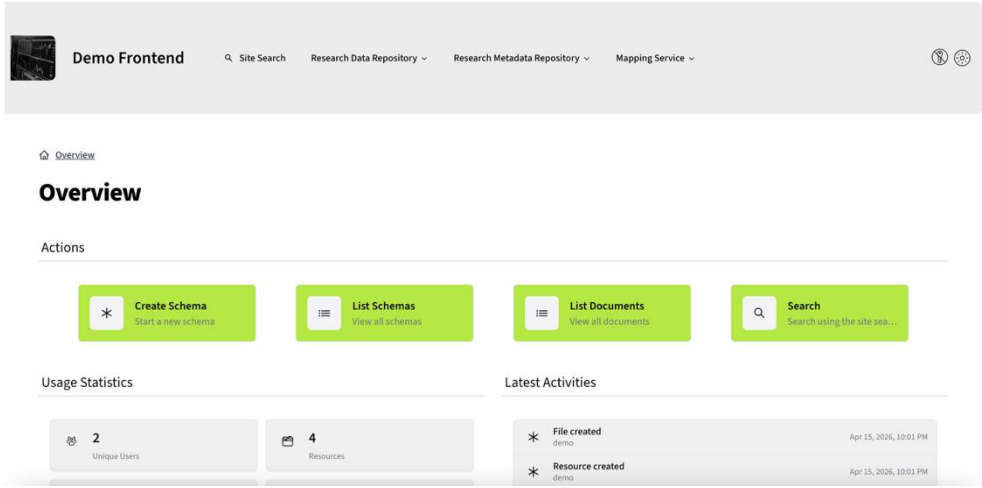
Servings: An edited JSON metadata document, which is locally saved.

Use Case: Metadata Upload



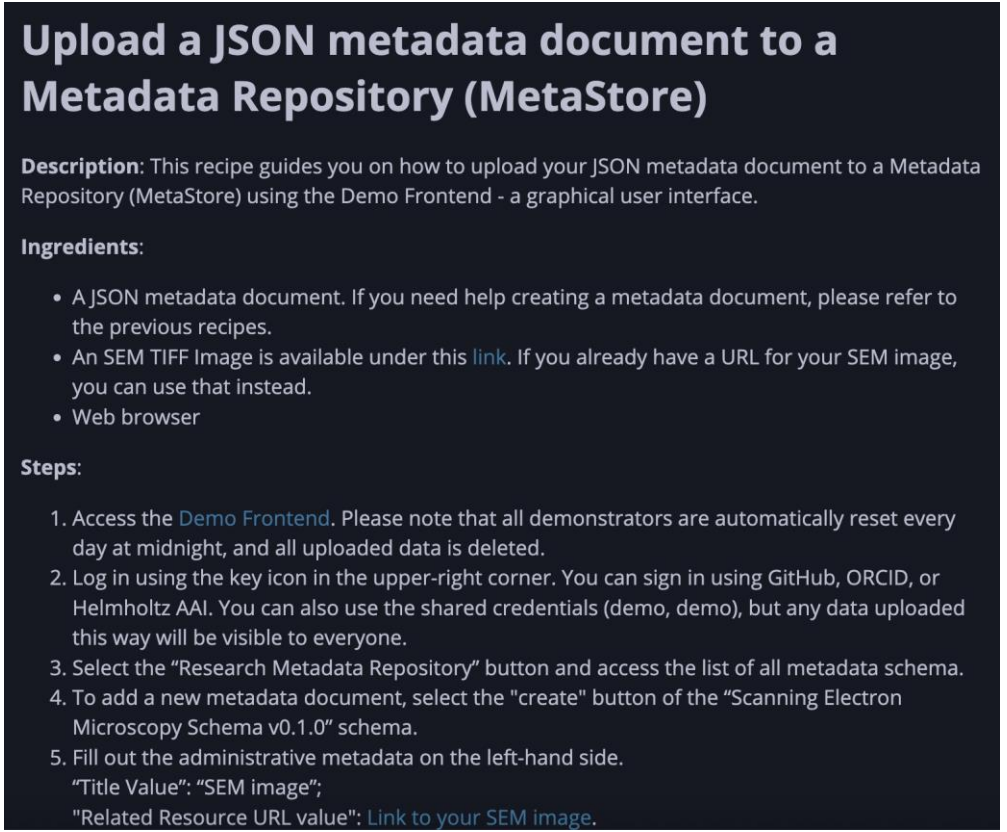
Metadata Upload: Metadata Repository (MetaStore)

- metadata repository framework for storing schemas and metadata documents and supports both formats XML and JSON. MetaStore allows the user to:
- register and later update XML/JSON schemas
- add metadata documents written according to a registered schema and later update them
- validate metadata against a registered schema
- https://demo.datamanager.kit.edu/metastore/



Metadata Upload: Metadata Repository (MetaStore)

- Link to the tutorial
  - https://t1p.de/metadata-upload



Upload a JSON metadata document to a Metadata Repository (MetaStore)

Description: This recipe guides you on how to upload your JSON metadata document to a Metadata Repository (MetaStore) using the Demo Frontend - a graphical user interface.

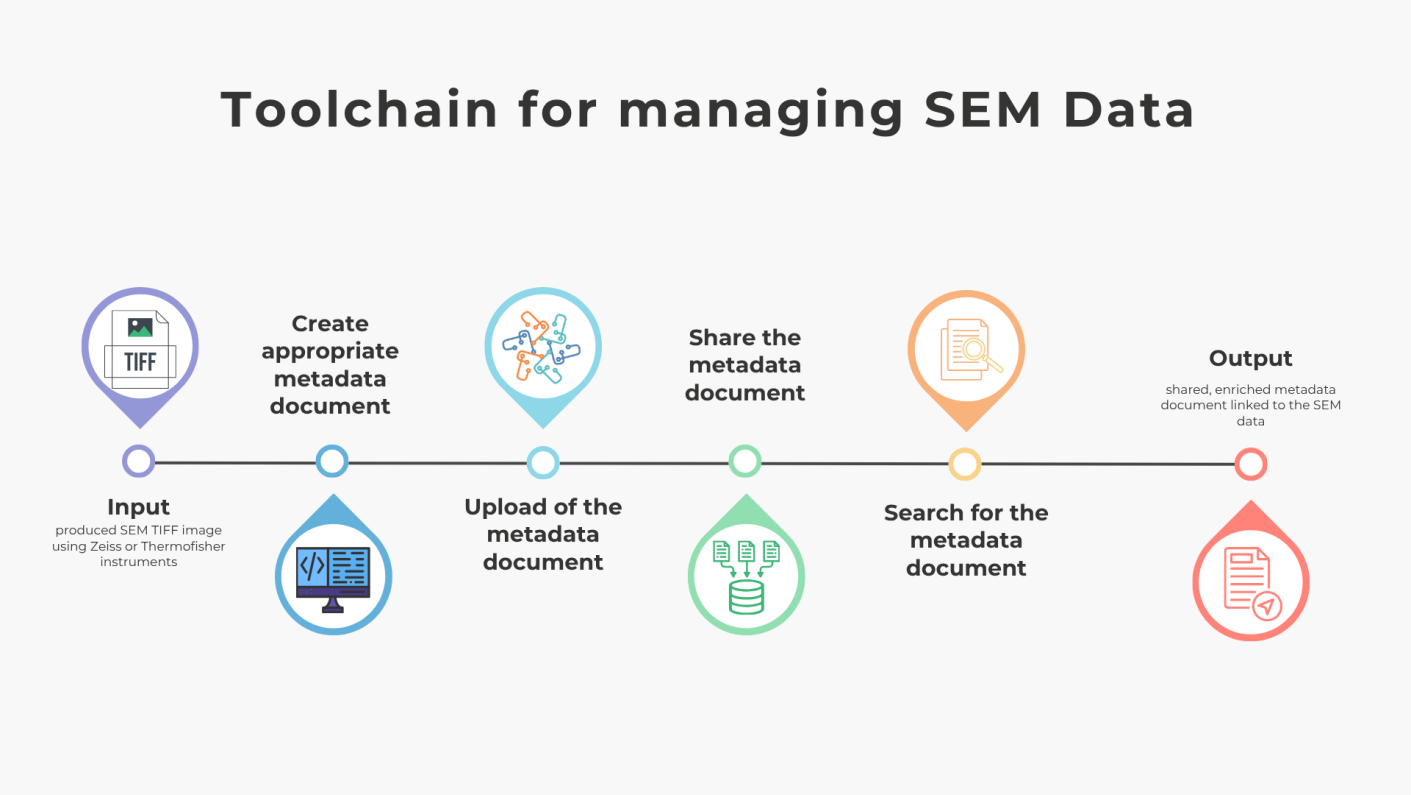
Ingredients:

- A JSON metadata document. If you need help creating a metadata document, please refer to the previous recipes.
- An SEM TIFF Image is available under this link. If you already have a URL for your SEM image, you can use that instead.
- Web browser

Steps:

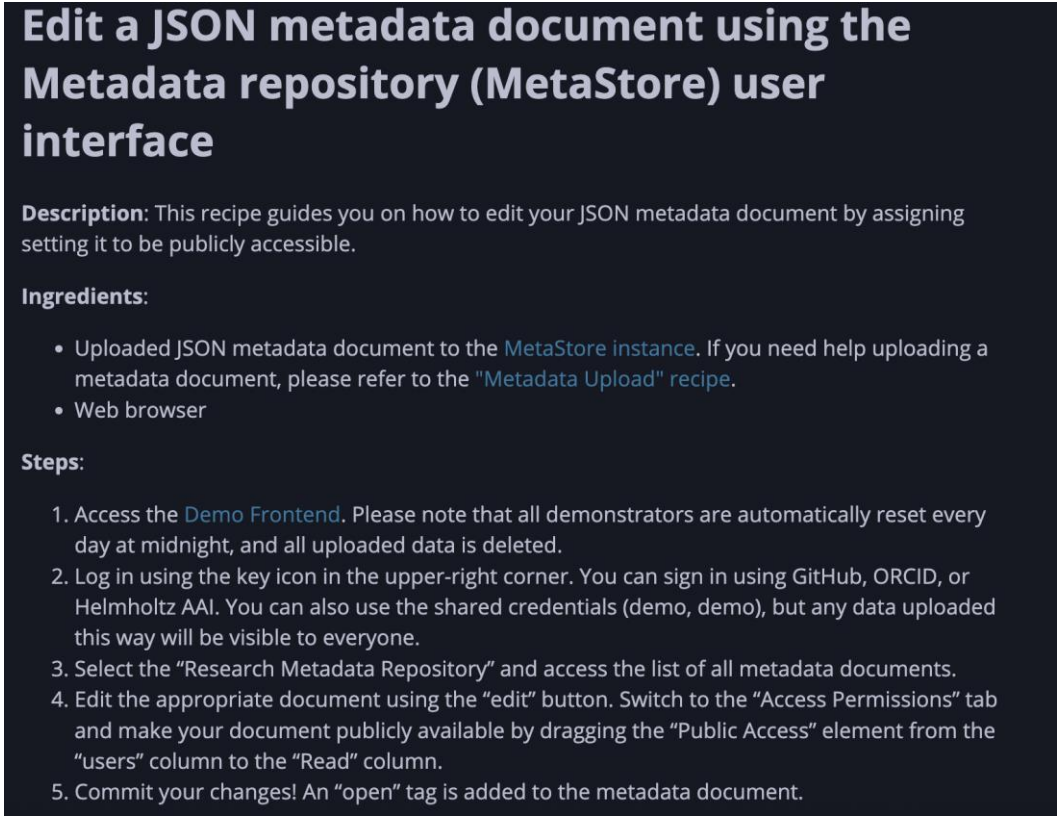
1. Access the Demo Frontend. Please note that all demonstrators are automatically reset every day at midnight, and all uploaded data is deleted.
2. Log in using the key icon in the upper-right corner. You can sign in using GitHub, ORCID, or Helmholtz AAI. You can also use the shared credentials (demo, demo), but any data uploaded this way will be visible to everyone.
3. Select the “Research Metadata Repository” button and access the list of all metadata schema.
4. To add a new metadata document, select the "create" button of the “Scanning Electron Microscopy Schema v0.1.0” schema.
5. Fill out the administrative metadata on the left-hand side.
   “Title Value”: “SEM image”;
   "Related Resource URL value": Link to your SEM image.

Use Case: Metadata Editing



Metadata Editing: Metadata Repository(MetaStore) User Interface

- Link to the tutorial
  - https://t1p.de/metastore-ui



Edit a JSON metadata document using the Metadata repository (MetaStore) user interface

Description: This recipe guides you on how to edit your JSON metadata document by assigning setting it to be publicly accessible.

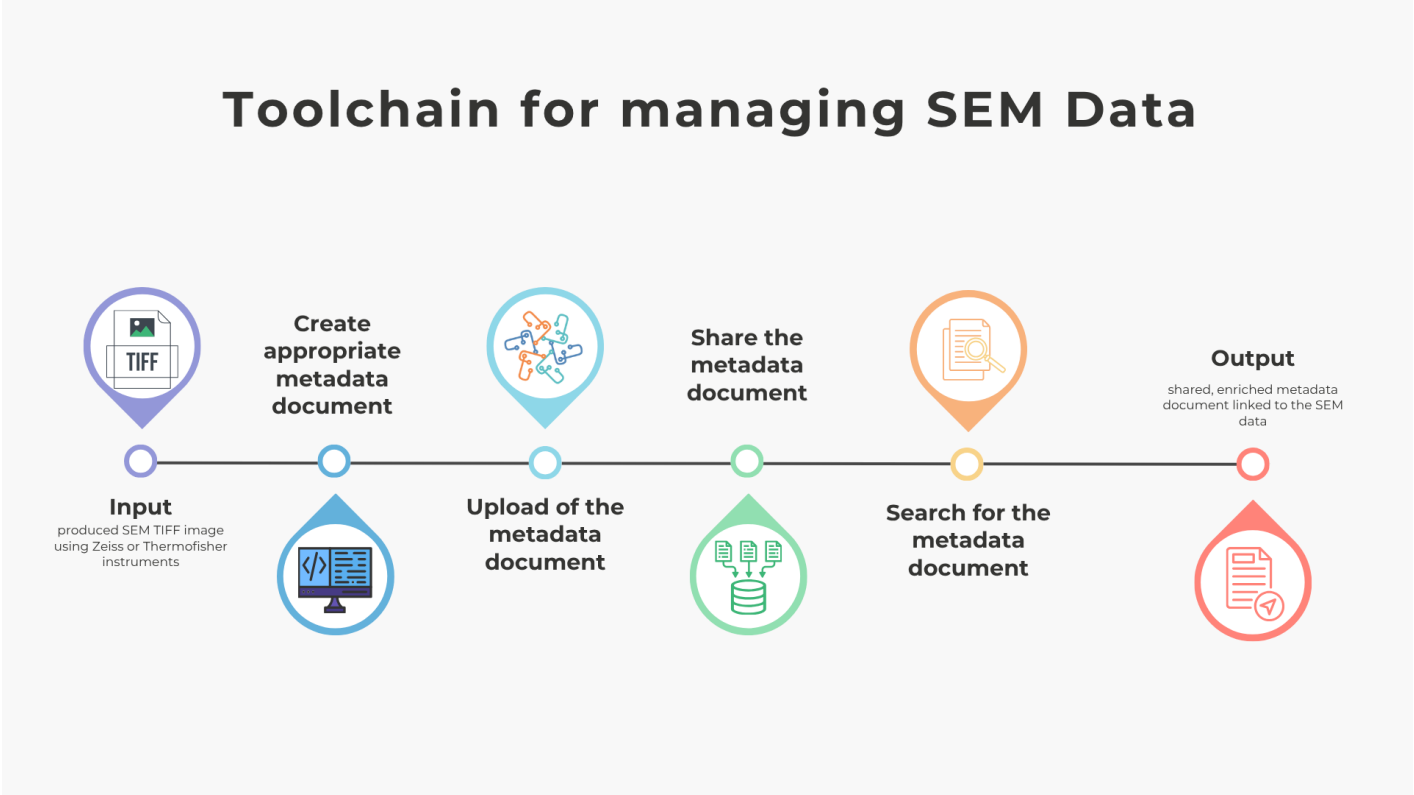
Ingredients:

- Uploaded JSON metadata document to the MetaStore instance. If you need help uploading a metadata document, please refer to the "Metadata Upload" recipe.
- Web browser

Steps:

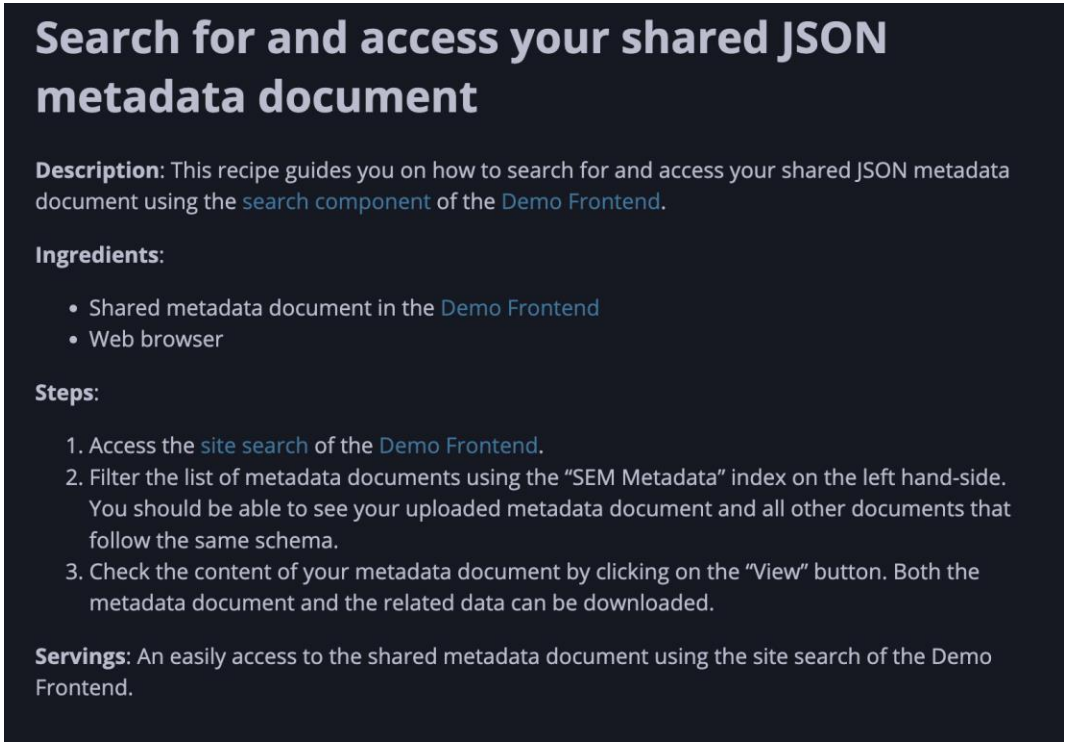
1. Access the Demo Frontend. Please note that all demonstrators are automatically reset every day at midnight, and all uploaded data is deleted.
2. Log in using the key icon in the upper-right corner. You can sign in using GitHub, ORCID, or Helmholtz AAI. You can also use the shared credentials (demo, demo), but any data uploaded this way will be visible to everyone.
3. Select the “Research Metadata Repository” and access the list of all metadata documents.
4. Edit the appropriate document using the “edit” button. Switch to the “Access Permissions” tab and make your document publicly available by dragging the “Public Access” element from the “users” column to the “Read” column.
5. Commit your changes! An “open” tag is added to the metadata document.

Use Case: Metadata Access



Metadata Access: Site Search

- Now you can access your publicly available metadata document.
- Link to the tutorial
  - https://t1p.de/metadata-access



Search for and access your shared JSON metadata document

Description: This recipe guides you on how to search for and access your shared JSON metadata document using the search component of the Demo Frontend.

Ingredients:

- Shared metadata document in the Demo Frontend
- Web browser

Steps:

1. Access the site search of the Demo Frontend.
2. Filter the list of metadata documents using the "SEM Metadata" index on the left hand-side. You should be able to see your uploaded metadata document and all other documents that follow the same schema.
3. Check the content of your metadata document by clicking on the "View" button. Both the metadata document and the related data can be downloaded.

Servings: An easily access to the shared metadata document using the site search of the Demo Frontend.

Materials

- Metadata Management Cookbook
  - https://kit-data-manager.github.io/metadata-management-cookbook/intro.html
- Youtube Channel
  - https://www.youtube.com/@ScientificComputingCenter
- Demo Frontend
  - https://demo.datamanager.kit.edu/
- Email:
  - sabrine.chelbi@kit.edu / training@scc.kit.edu

Acknowledgements

Contributions:
Rossella Aversa, Thomas Jejkal, Andrea Recchia

Used material:
Aversa, A. (2024). The journey towards Metadata Management. DOI: 10.5445/IR/1000174166
Gerlich, S., Strupp, A., Hofmann, V., Sandfeld, S. (2023). Fundamentals of Scientific Metadata. The Carpentries Incubator. DOI: 10.5281/zenodo.10091708

HELMHOLTZ
Metadata Collaboration

Thank you for your attention

www.helmholtz-metadaten.de