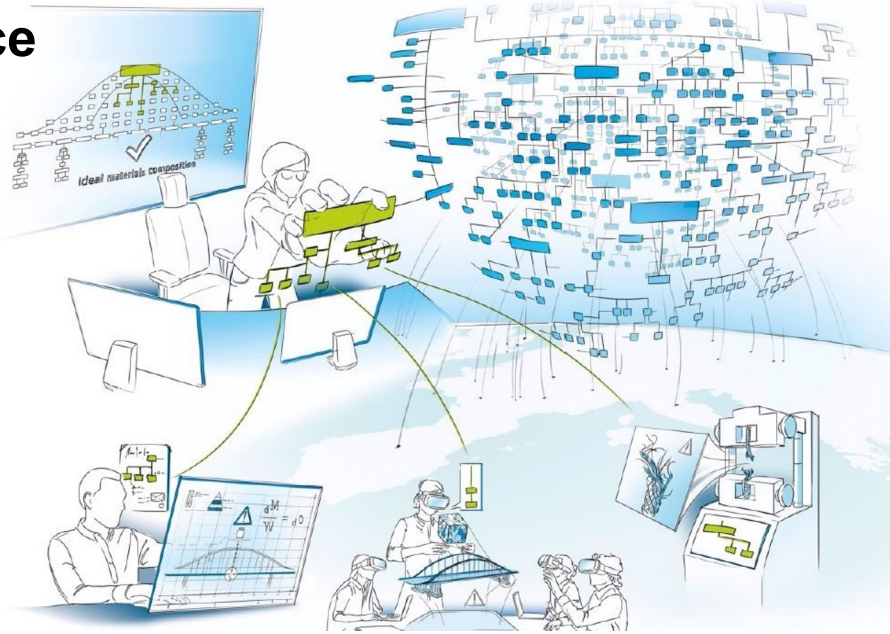
Applying Metadata Management

Extraction, Mapping & Visualisation

NFDI-MatWerk Conference

| Siegburg, 27.-29.06.23



Workshop Team:

Ajay Kirar
Benedikt Heinrichs
Elias Vitali
Gulzaure Abdildina
Reetu Joseph
Rossella Aversa

Agenda

- Introduction
- Demo: Semantic Metadata Extractor
    - Q/A session
- Introduction to FIB/SEM Tomography (PP13)
- Demo: Metadata Extraction & Mapping with the new interface
    - Q/A Session
- Demo: Metadata Visualisation
    - Q/A Session
- Wrap-up

Introduction

Aim of the workshop

To demonstrate how **Metadata Management** can be **applied** in practical scenarios by way of **extraction, mapping and visualisation**, which would ease the task for researchers by **decreasing time and the manual effort** needed.

Marktplatz

- Posters:
  - Metadata Extraction Tool and Schema Mapper for Scanning Electron Microscopy (SEM) images (Elias)
  - Semantic Metadata Extractor (Benedikt)
- Live demos
  - Demonstration of Automatic Metadata Extraction and Mapping for SEM Images (Elias & Ajay)

Demo: Semantic Metadata Extractor

Benedikt Heinrichs, Sirieam Hunke

Introduction

- There is an ongoing task to transform research data to FAIR Digital Objects (FDOs)
- These FAIR Digital Objects contain
  - The Digital Object (research data)
  - Metadata about the Digital Object
  - Some Service Interfaces
  - A Persistent Identifier
- The big point we focus on here today is metadata

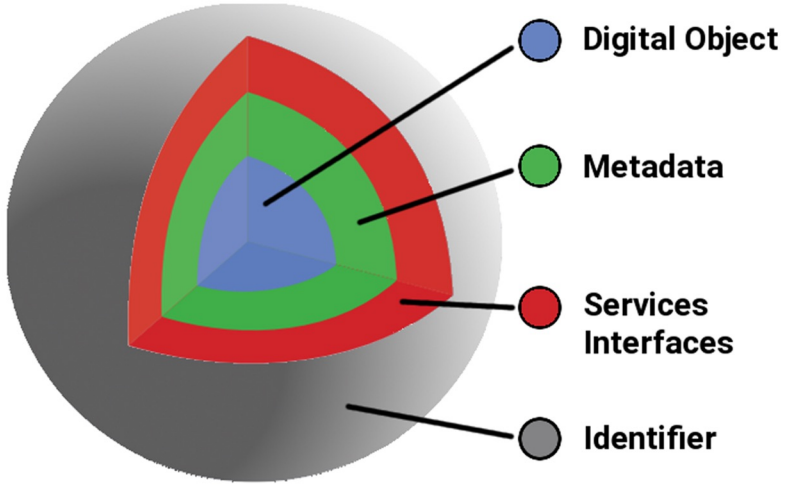
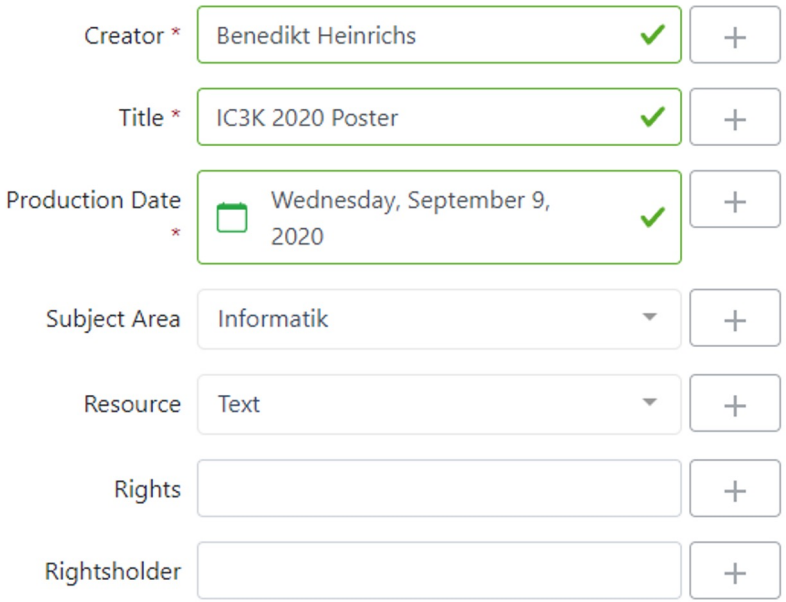


Image from "Digital Objects – FAIR Digital Objects: Which Services Are Required?", located at: https://datascience.codata.org/articles/10.5334/dsj-2020-015/

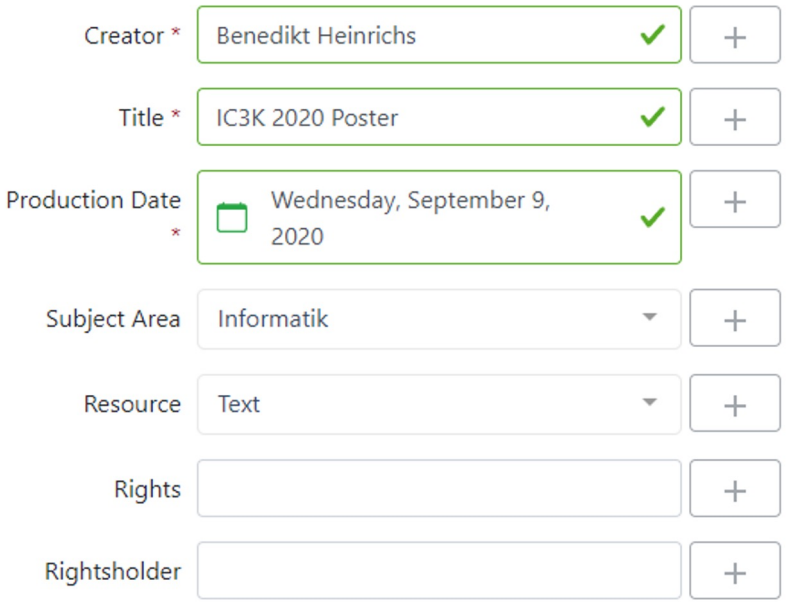
Metadata

- Why Metadata?
  - Describing the information surrounding the generation of a research item
  - Example: Describing a research experiment, the time it took place, etc.
- What Metadata?
  - Administrative: e.g. location or rights
  - Structural: provenance information
  - Descriptive: who, when or what
  - Semantic: Formulated in RDF using ontologies and validated by SHACL (W3C standard)
- How Metadata?
  - Manually input
  - Automatically generated during an experiment



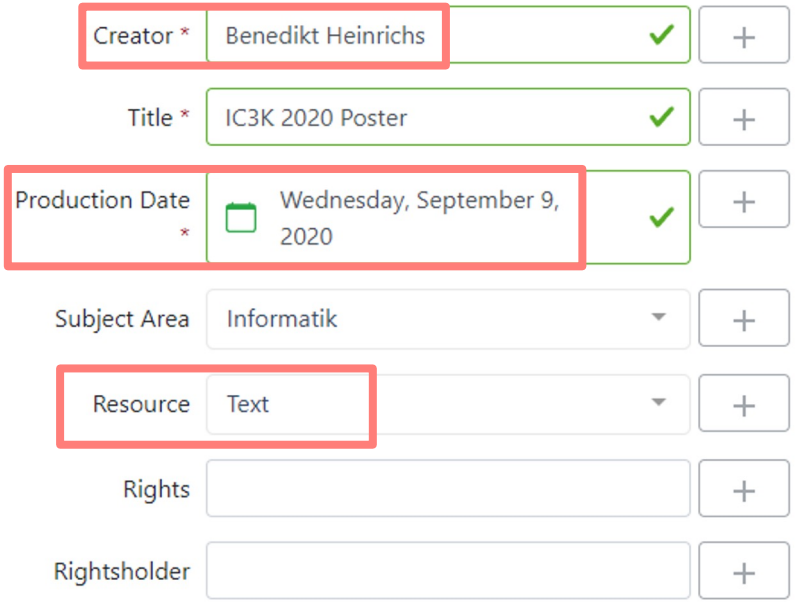
Manual Metadata Input

- Administrative and Structural are usually fairly simple to automatically determine
  - e.g. a platform which manages metadata should be able to be aware of these types of metadata
- Descriptive metadata, however, currently mostly needs to be entered manually
  - This is usually a tedious and time-consuming task
  - A goal of RDM is to make the research process easier and not to create additional hurdles
    - Thankfully, the research data itself a lot of the time brings a subset of the necessary descriptive metadata with itself



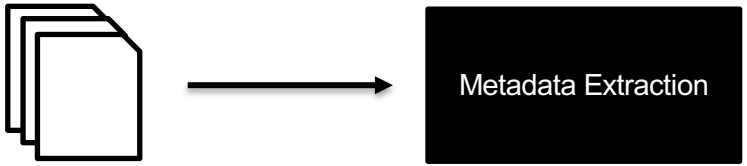
Manual Metadata Input

- Administrative and Structural are usually fairly simple to automatically determine
  - e.g. a platform which manages metadata should be able to be aware of these types of metadata
- Descriptive metadata, however, currently mostly needs to be entered manually
  - This is usually a tedious and time-consuming task
  - A goal of RDM is to make the research process easier and not to create additional hurdles
    - Thankfully, the research data itself a lot of the time brings a subset of the necessary descriptive metadata with itself

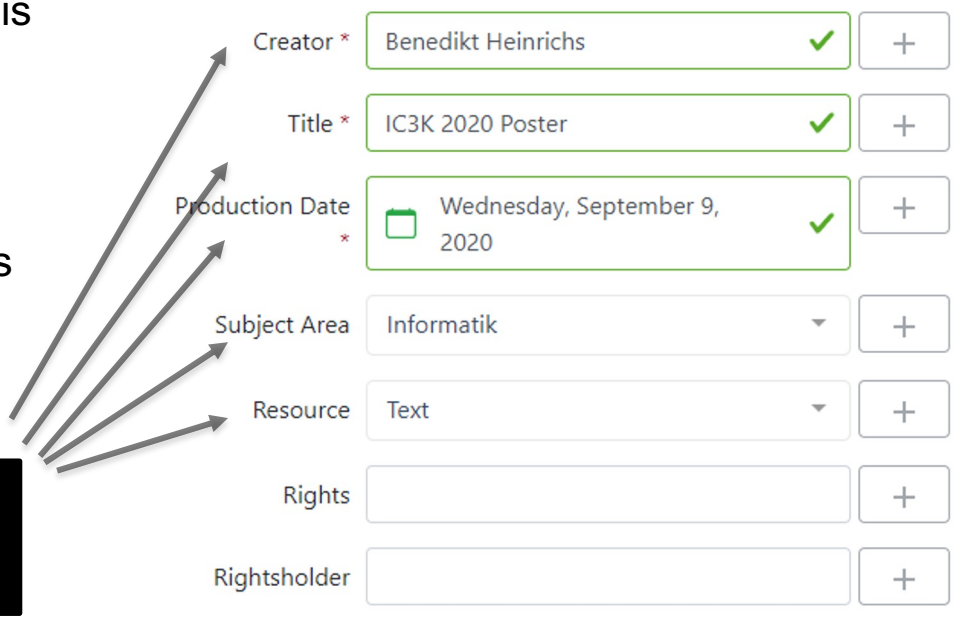


Motivation for Semantic Metadata Extraction

- We want to know what our research data is about
- We want to provide the most detailed information about the content of our research data
- We want to spend only as much time as is necessary to input values into forms
- Proposition:



Research Data



Filled out Metadata

Result Motivation for Semantic Metadata Extraction

```
47        image:mode "RGB" ;
48        ebucore:hasFormat "JPEG" ;
49        ebucore:height "425" ;
50        ebucore:width "640" .
51
52    imageobject:apple rdfs:label "apple" ;
53        imageobject:count "6" .
54
55    imageobject:orange rdfs:label "orange" ;
56        imageobject:count "7" .
```

TestText.txt - Editor

Datei Bearbeiten Ansicht

```
Benedikt and Amin work in the same office.
David is eavesdropping their conversation.
David can also speak chinese and knows the characters '性格'.
```

```
68  <http://pikes.fbk.eu/#Amin> a <http://www.newsreader-project.eu/ontologies/PERSON> .
69
70  <http://pikes.fbk.eu/#Benedikt> a <http://www.newsreader-project.eu/ontologies/PERSON> .
71
72  <http://pikes.fbk.eu/#eavesdropping> rdfs:seeAlso <http://dbpedia.org/resource/Eavesdropping> .
```

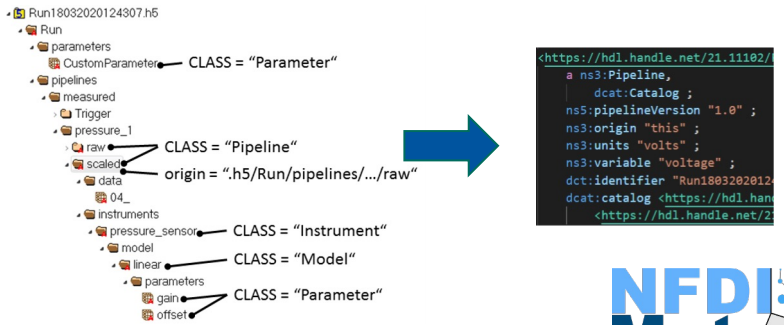
Prepare the deployment for production server

Refer to the product ticket for the description of the ticket.

- [ ] Update AutoSPInstaller.xml for installation
  - [ ] Check if there need to be seperate SP Installer xmls for the Farm join since there is only one farm.
  - [x] Check if we need a separate file. Set the secret keys and replace it with Consul values into the deployment script.
- [x] Detect why the error during installation "Previously installed Office 2019" (PreReqCheck) occurred
- [x] Parameter for the Installation => For Produktiv vs Dev
  - Create a parameter to filter (blacklist) steps
  - Add a json file with default values for config params, also secrets, check if this solves #322
- [ ] Make sure the production deployment is working
- [x] Put CoScInE DB creation into its own step and otherwise throw 3.02 out

Topic/600-productionDeployment

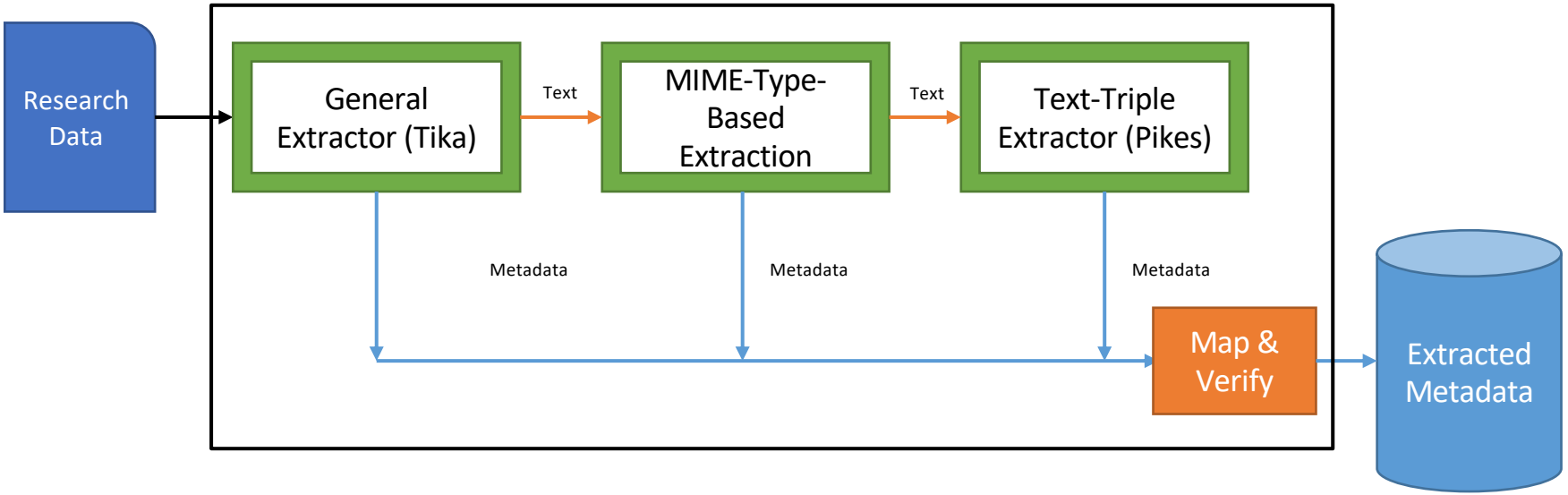
```
87
88    <http://pikes.fbk.eu/#installation> a <http://dbpedia.org/class/yago/Initiation107453195> ;
89        rdfs:seeAlso <http://dbpedia.org/resource/Installation_(computer_programs)> .
90
91    <http://pikes.fbk.eu/#json> rdfs:seeAlso <http://dbpedia.org/resource/JSON> .
92
```



HDF5 File Structure

Metadata Extraction Pipeline

Metadata Extraction



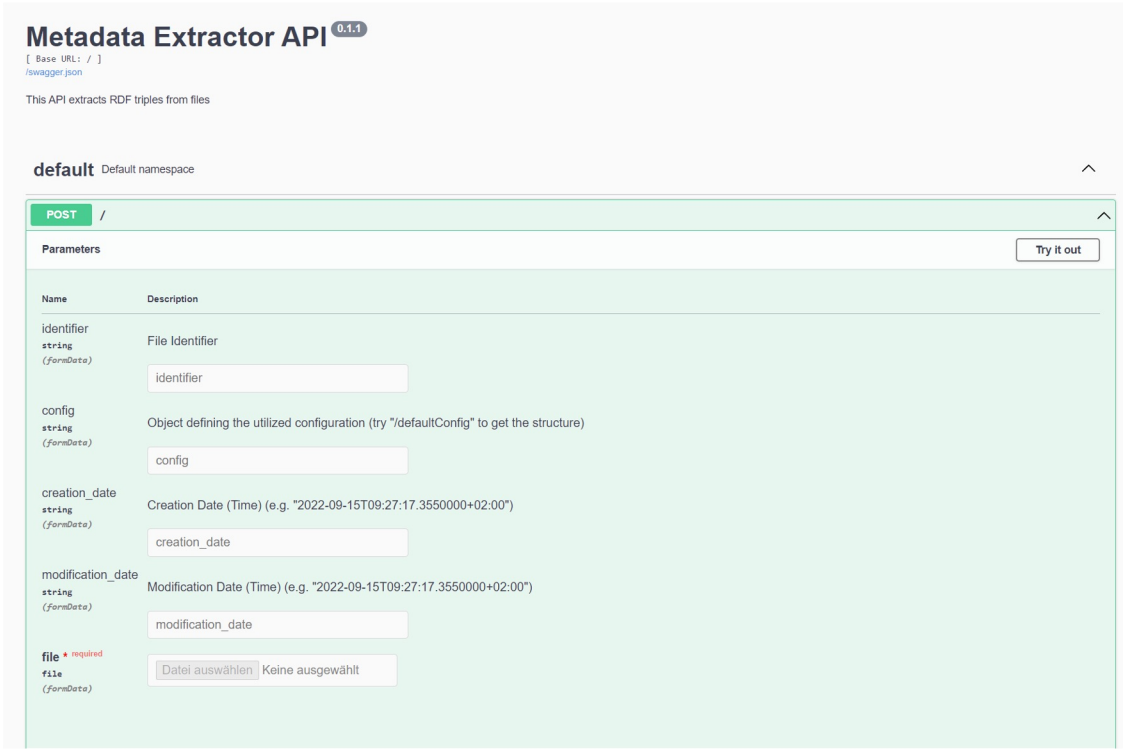
Heinrichs, B. ; Politze, M.
Moving Towards a General Metadata Extraction Solution for Research Data with State-of-the-Art Methods
12th International Conference on Knowledge Discovery and Information Retrieval, KDIR 2020, online, 2 Nov 2020 - 4 Nov 2020

Metadata Extraction Application – Features

- Dynamic Configuration
  - You can configure every extractor which shall be called and specify certain environment variables
- Registration of custom extractors
  - Once implemented, an extractor will listen to its registration method which can e.g. listen to certain MIME-Types like “image/png”
  - Custom extractors can be excluded from the default configuration, so that specific use cases can be proposed without impacting everything else
- Highly extendable
  - By being open source, this application is easily extendable to different areas
    - https://git.rwth-aachen.de/coscine/research/metadataextractor

Metadata Extraction Service

https://metadataextractor.otc.coscine.dev/



Example Results – Object Detection

```
47          image:mode "RGB" ;
48          ebucore:hasFormat "JPEG" ;
49          ebucore:height "425" ;
50          ebucore:width "640" .
51
52      imageobject:apple rdfs:label "apple" ;
53          imageobject:count "6" .
54
55      imageobject:orange rdfs:label "orange" ;
56          imageobject:count "7" .
```

Example Results – Text

TestText.txt - Editor

Datei Bearbeiten Ansicht

```
Benedikt and Amin work in the same office.
David is eavesdropping their conversation.
David can also speak chinese and knows the characters '性格'.
```

```
68    <http://pikes.fbk.eu/#Amin> a <http://www.newsreader-project.eu/ontologies/PERSON> .
69
70    <http://pikes.fbk.eu/#Benedikt> a <http://www.newsreader-project.eu/ontologies/PERSON> .
71
72    <http://pikes.fbk.eu/#eavesdropping> rdfs:seeAlso <http://dbpedia.org/resource/Eavesdropping> .
```

Example Results – Image to Text

Prepare the deployment for production server

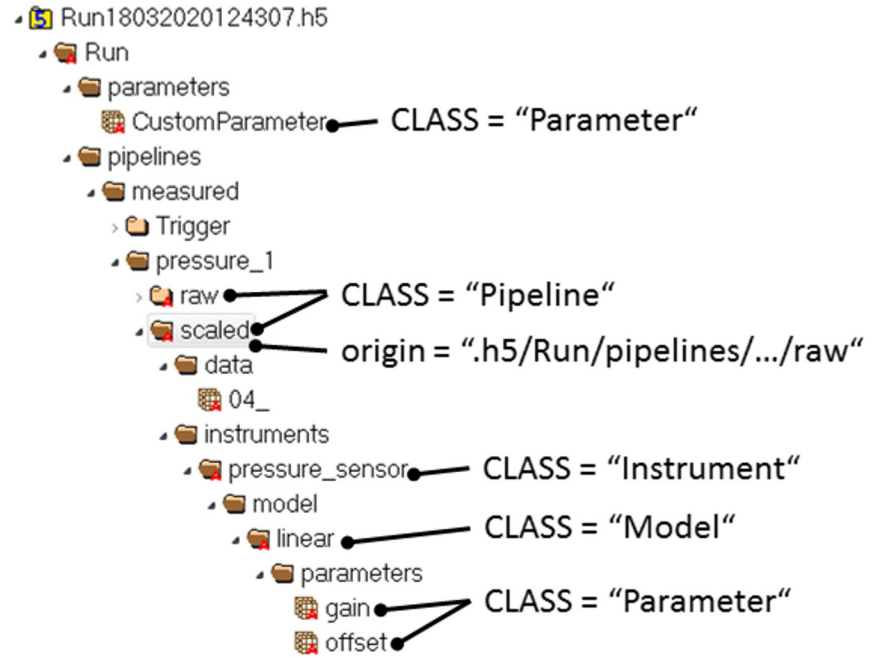
Refer to the product ticket for the description of the ticket.

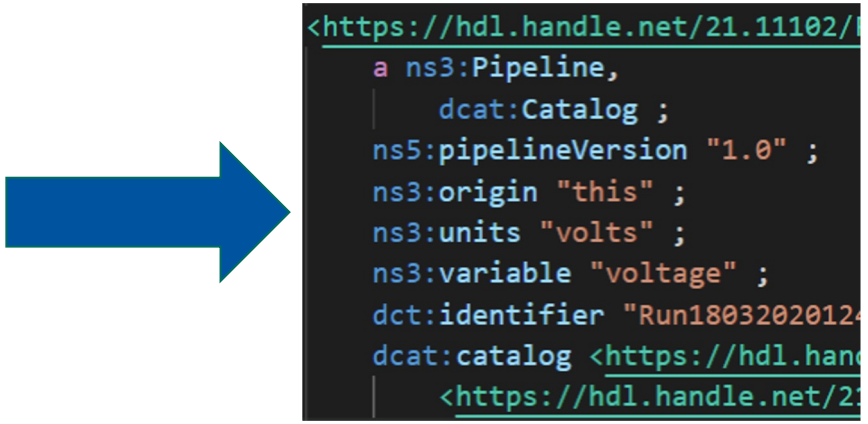
- [ ] Update AutoSPInstaller.xml for installation
  - [ ] Check if there need to be seperate SP Installer xmls for the Farm join since there is only one farm.
  - [x] Check if we need a separate file. Set the secret keys and replace it with Consul values into the deployment script.
- [x] Detect why the error during installation "Previously installed Office 2019" (PreReqCheck) occurred
- [x] Parameter for the Installation => For Produktiv vs Dev
  - Create a parameter to filter (blacklist) steps
  - Add a json file with default values for config params, also secrets, check if this solves #322
- [ ] Make sure the production deployment is working
- [x] Put CoSciE DB creation into its own step and otherwise throw 3.02 out

Topic/600-productionDeployment

```
87
88    <http://pikes.fbk.eu/#installation> a <http://dbpedia.org/class/yago/Initiation107453195> ;
89        rdfs:seeAlso <http://dbpedia.org/resource/Installation_(computer_programs)> .
90
91    <http://pikes.fbk.eu/#json> rdfs:seeAlso <http://dbpedia.org/resource/JSON> .
92
```

Example Results – Real Example with HDF5





HDF5 File Structure

Heinrichs, B. P. A. ; Preuß, N. ; Politze, M. ; Müller, M. S. ; *et al*
Automatic General Metadata Extraction and Mapping in an HDF5 Use-case
Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management

Metadata Extraction Service – Usage

Open Source

- Git Repo (Python Code): https://git.rwth-aachen.de/coscine/research/metadataextractor
- Docker Image: registry.git.rwth-aachen.de/coscine/research/metadataextractor:latest
- Demo: https://metadataextractor.otc.coscine.dev/
- Start adding your own extractor now by using Gitpod: https://gitpod.io/#https://git.rwth-aachen.de/coscine/research/metadataextractor

Metadata Extraction Service – Future Work

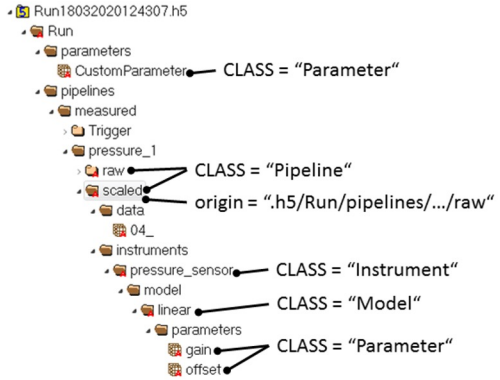
- Inclusion in research data management systems like Coscine
  - Automatically extract the metadata for research data based on a given configuration
  - Fill the metadata form automatically based on a templating engine
  - Make use of the extracted metadata in applications like “search”

- Utilize the extracted metadata to determine the similarity between research data when the MIME-Type is different (e.g. image with text vs. text file)
  - Making use of https://git.rwth-aachen.de/coscine/research/semanticsimilarity

- Improve the performance and make it better scale against big research data

Conclusion

- Today, I showed a look into our proposed solution of semantic metadata extraction
- It is a pipeline which takes research data and tries to describe the content as semantic metadata
- The usage has been demonstrated on different examples with a real-life use case as well



HDF5 File Structure

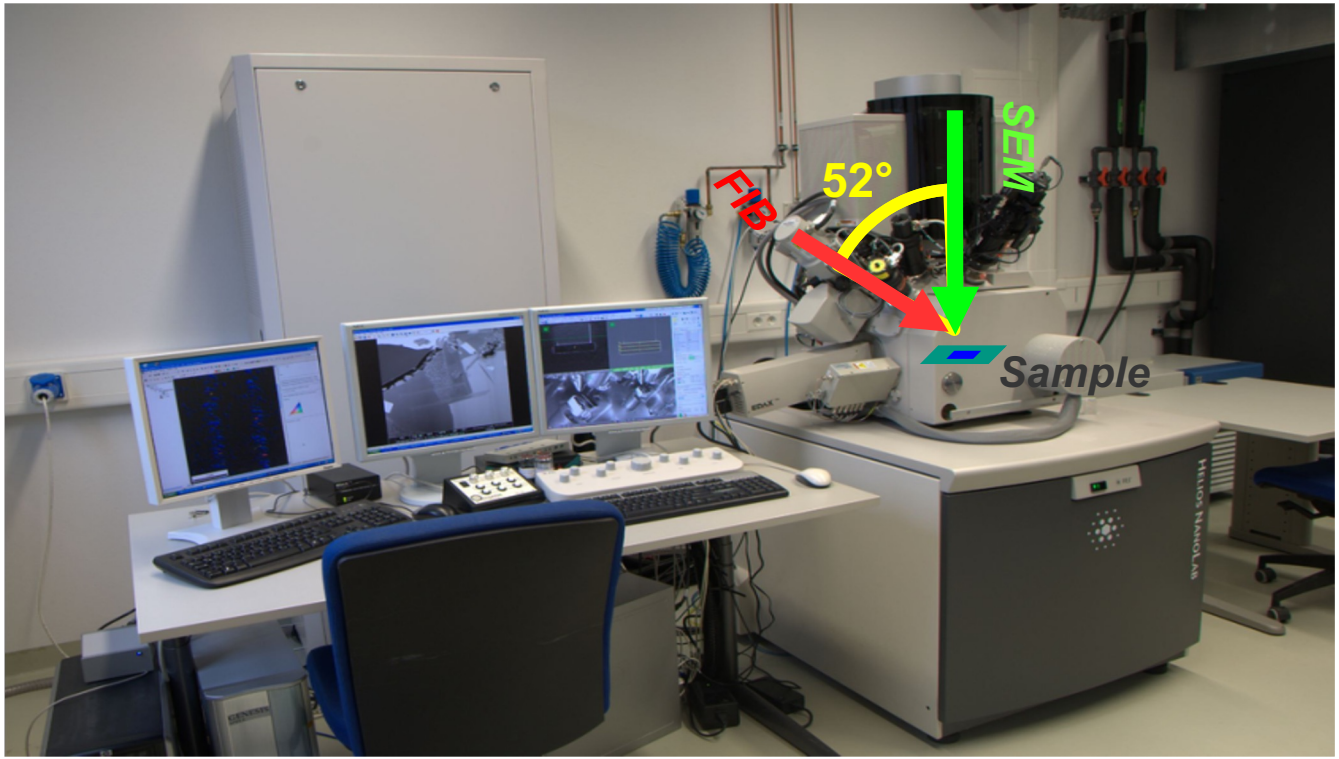
- Future work is being done to utilize this in real life applications

Q & A

Introduction to FIB/SEM Serial Sectioning Tomography (PP13)

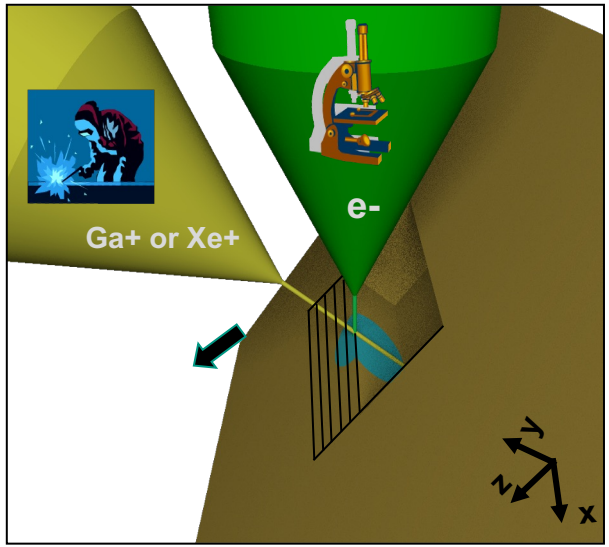
Reetu Joseph, Rossella Aversa, Christoph Pauly

FIB/SEM Serial Sectioning Tomography Apparatus

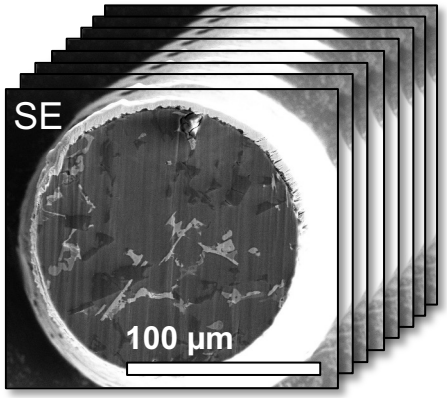


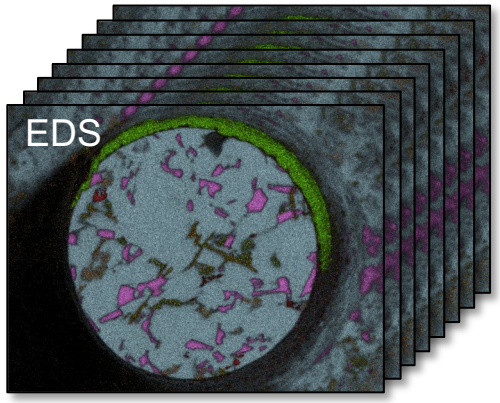
Courtesy: Christoph Pauly & Michael Engstler, Saarland University

I. Serial sectioning and multimodal imaging

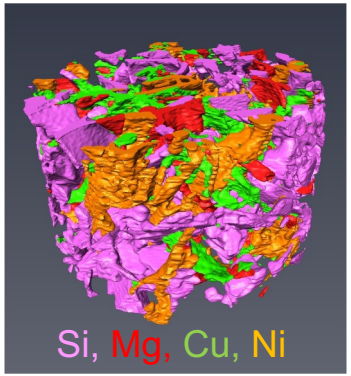


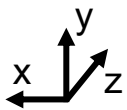
II. Image Processing and 3D reconstruction





2D image stack
(x-y-plane)

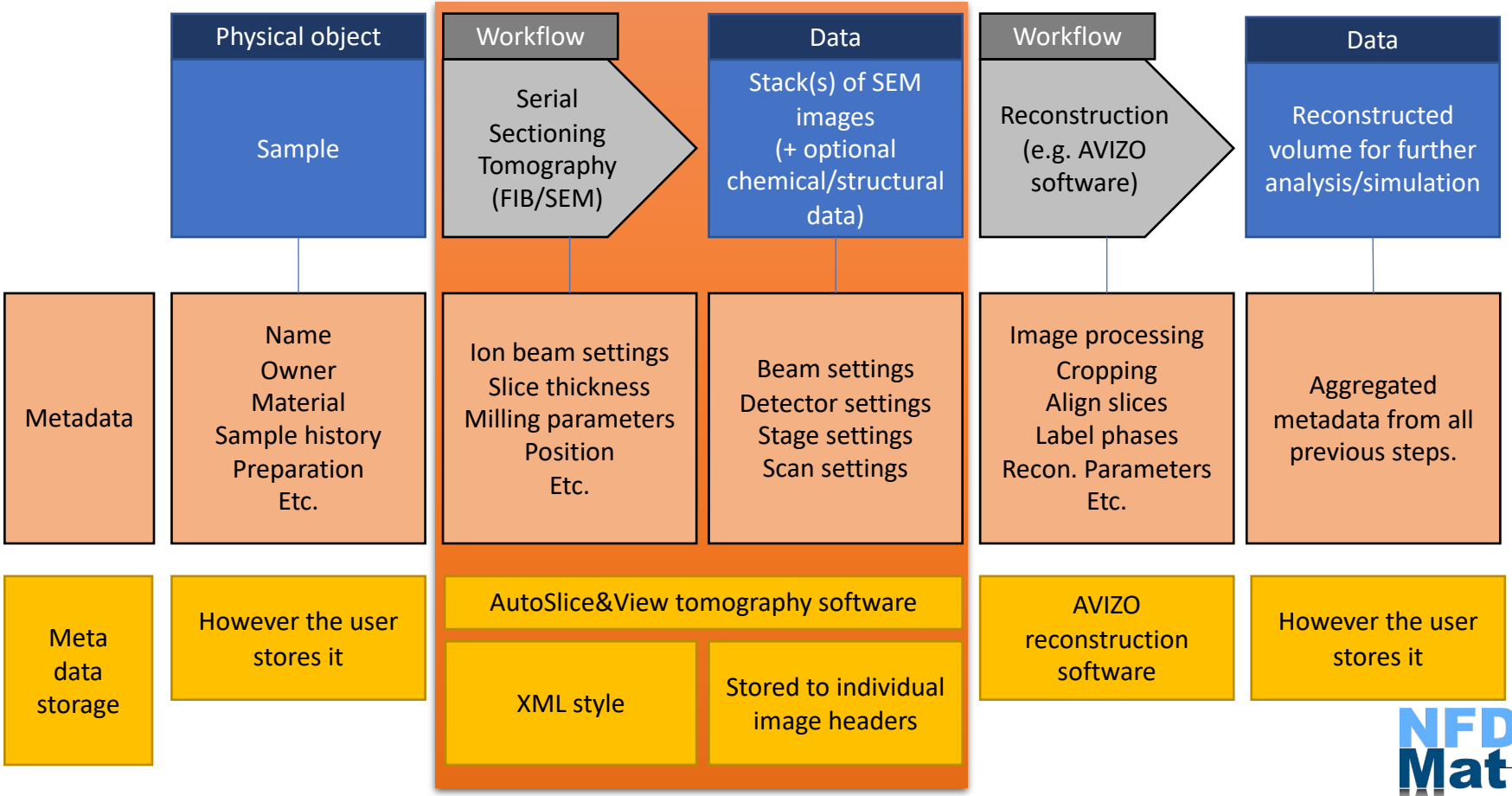




3D volume based
on EDS data (x-y-z)

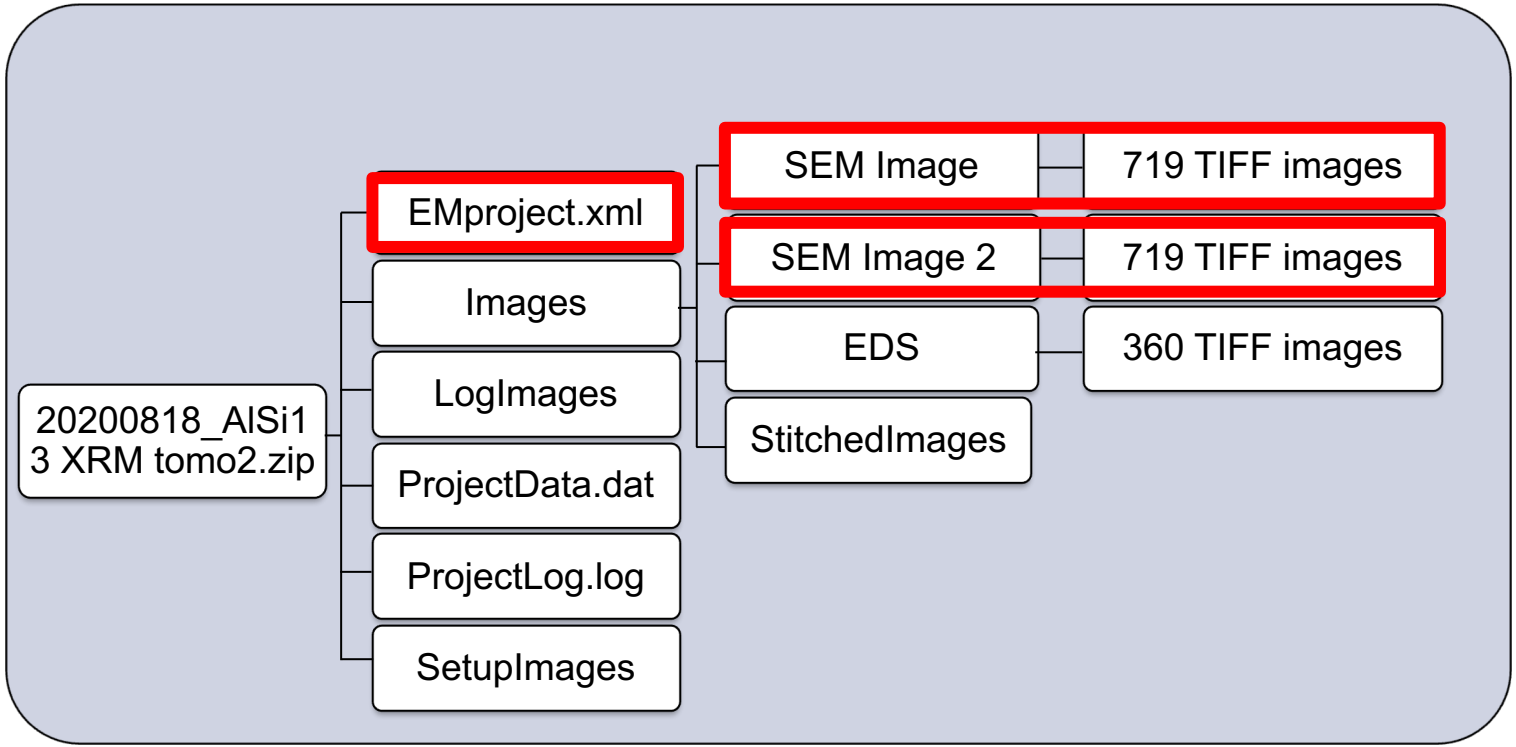
NFDI MatWerk

Courtesy: Christoph Pauly & Michael Engstler, Saarland University



Metadata Handling

Acquistion Data



Demo + Presentation: Metadata Extraction & Mapping – SEM/FIB Serial Sectioning Tomography

Elias Vitali, Ajay Kirar, Reetu Joseph, Rossella Aversa

Acquistion Data

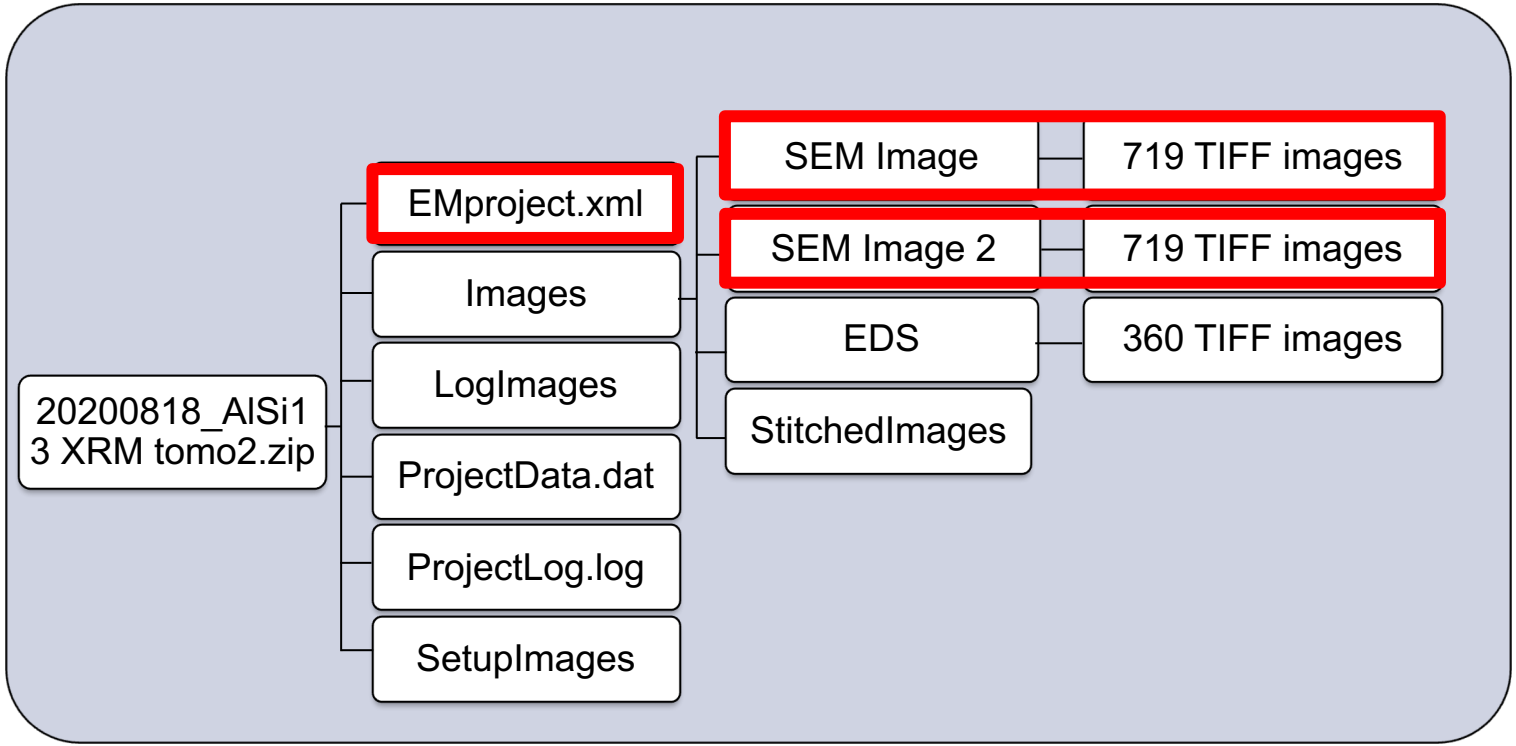
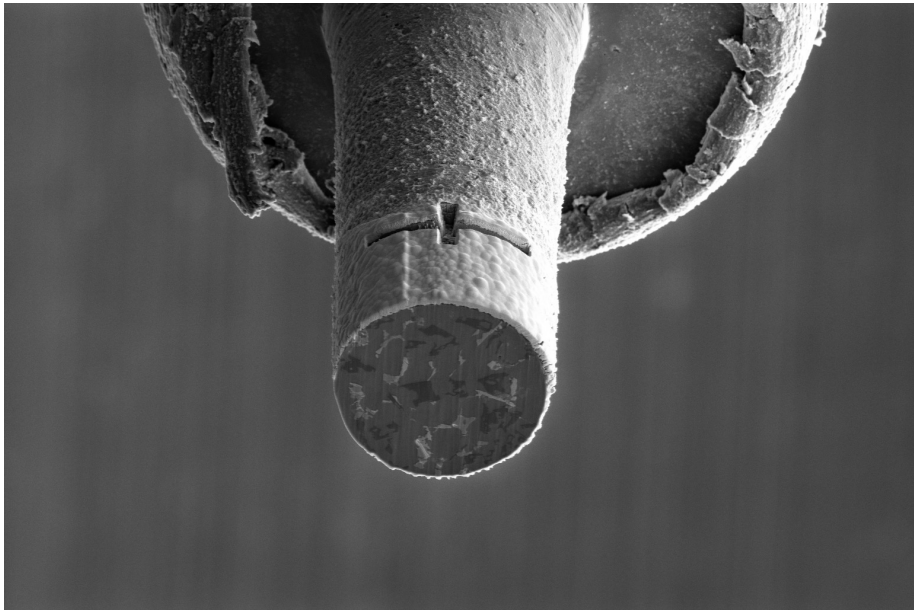


Image Metadata



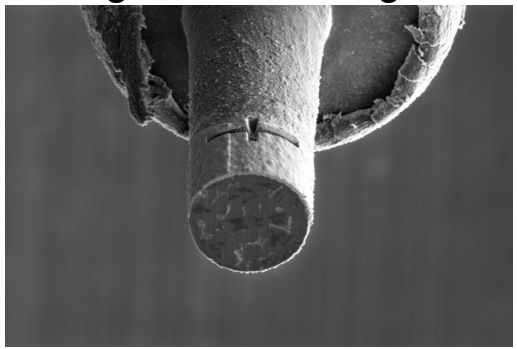
- Hidden in this image is a ton of metadata
- Need an efficient method to extract this metadata

Desired Result

- We want a metadata document that describes:
  - the acquisition;
    - each dataset;
      - each image.
- A schema dictates the format which this metadata document should take
  - This lets us ”map” the metadata

Process at a glance

1. Original TIFF image with all metadata



```
HV=15000
HFW=0.000592
VFW=0.000394667
WD=0.00402349
BeamCurrent=1.6e-009
```

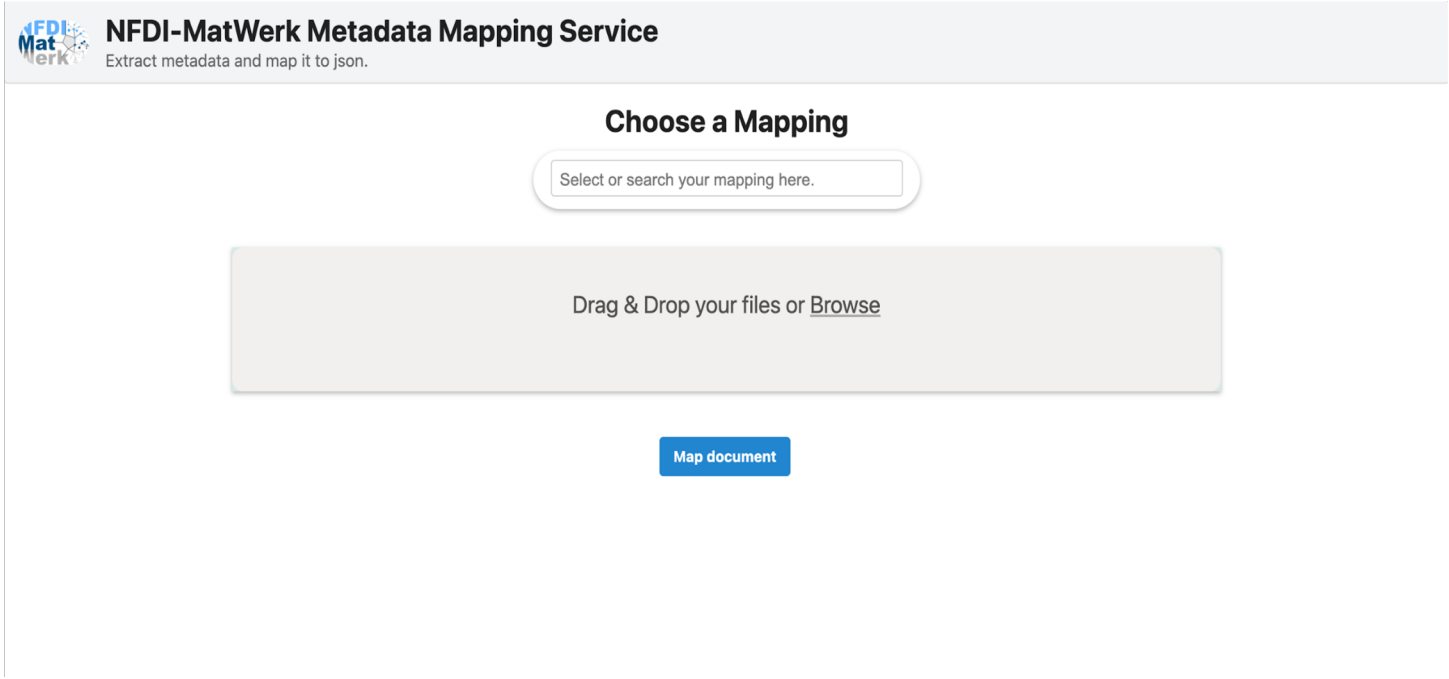
```
"acquisition.dataset.instrument.eBeam.accelerationVoltage.value":"Images.SEM Image.SliceImage.EBeam.hv",
"acquisition.dataset.instrument.eBeam.beamCurrent.value":"Images.SEM Image.SliceImage.EBeam.beamcurrent",
"acquisition.dataset.instrument.eBeam.scanRotation.value":"Images.SEM Image.SliceImage.EBeam.scanrotation",
```

2. Map file which translates original keys to schema keys

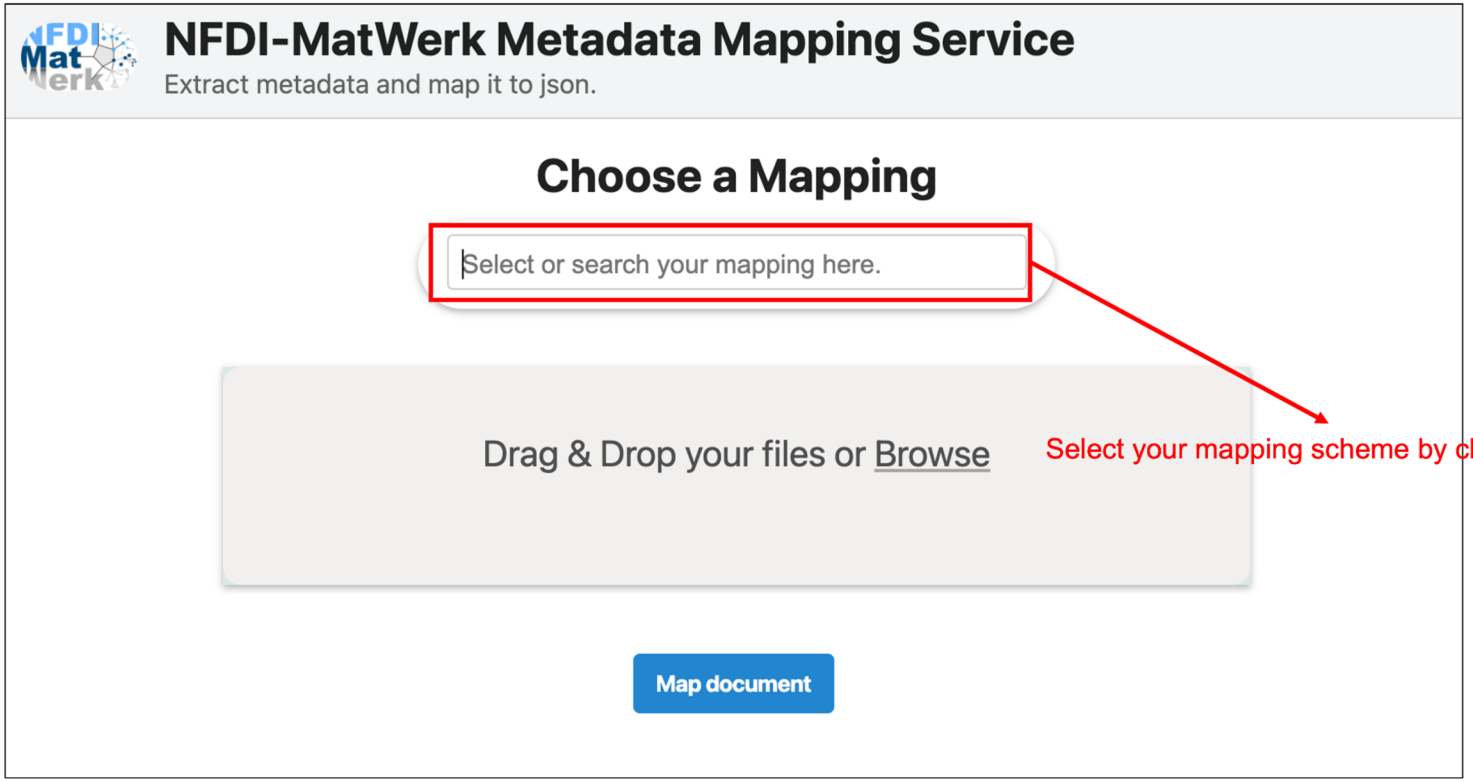
3. Formatted metadata document

```
{
    "acquisition": {
        "genericMetadata": {…
        },
        "dataset": [
            {
                "rows": "1",
                "columns": "1",
                "tileColumn": "0",
                "user": {…
                },
                "program": {…
                },
                "instrument": {
                    "beamType": "EBeam",
                    "spot": "1",
                    "eBeam": {
                        "accelerationVoltage": {
                            "value": "15000"
                        },
                        "beamCurrent": {
                            "value": "1.6e-009"
                        },
```

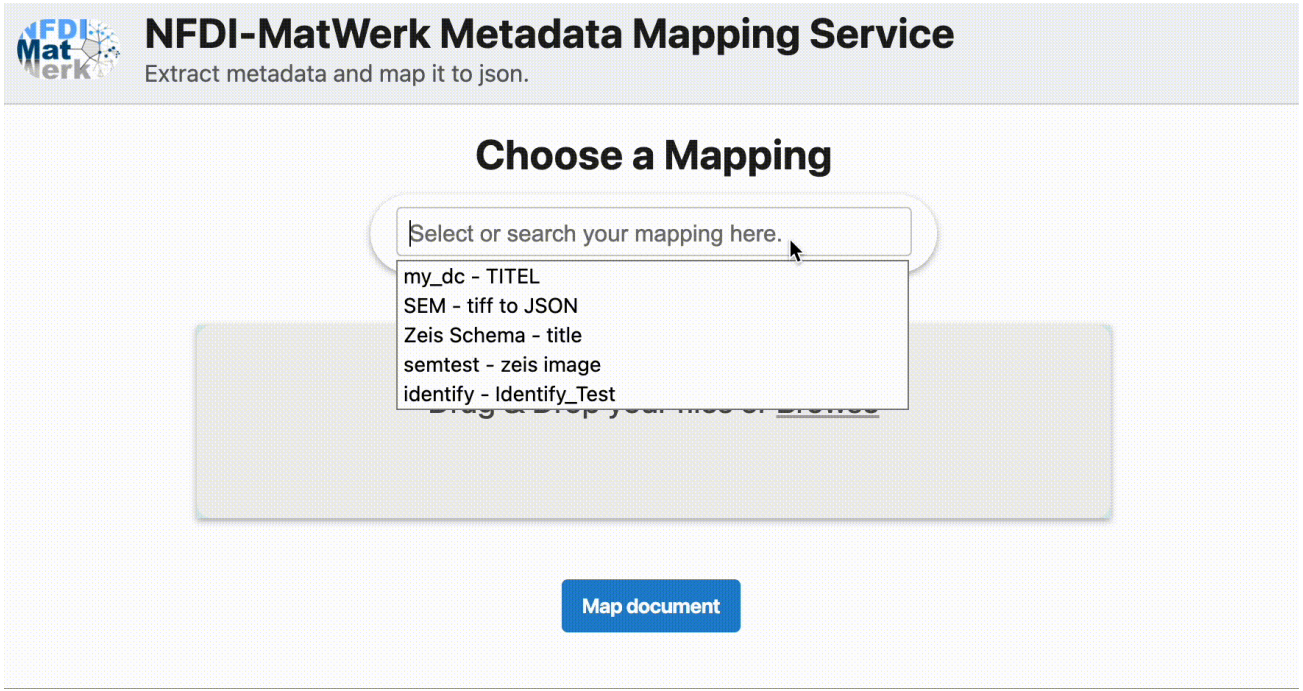
The Mapping Service



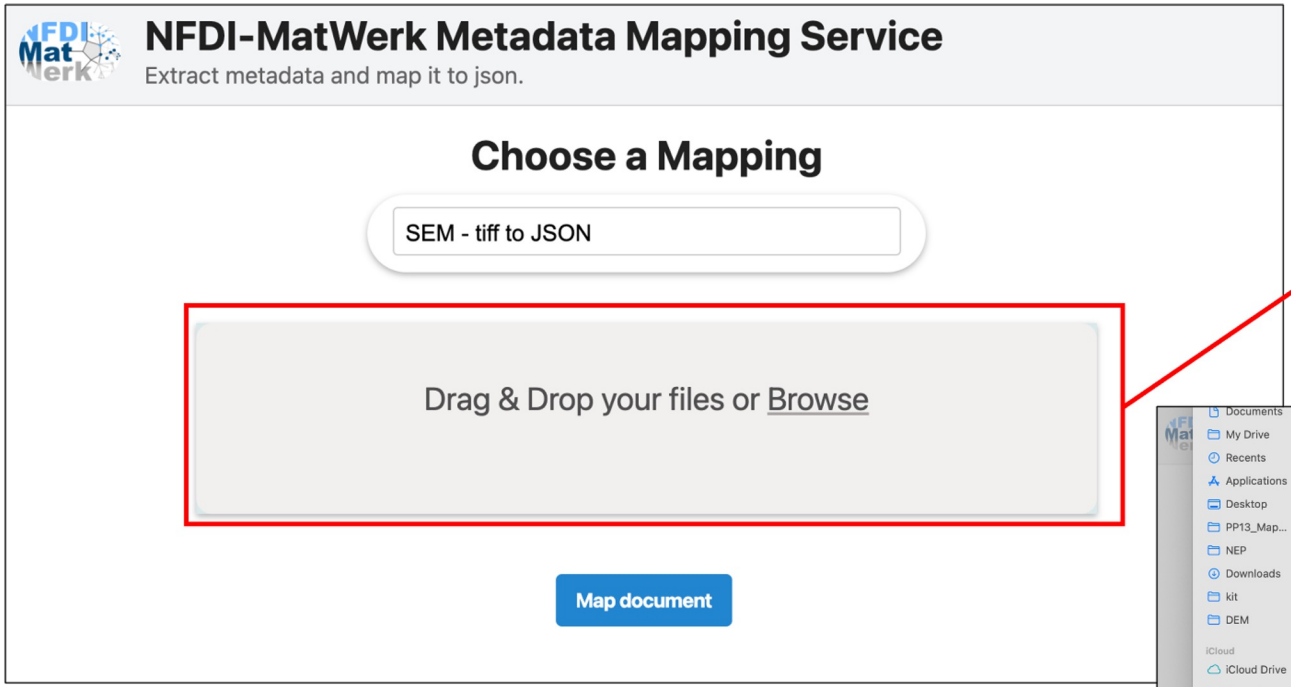
1. Choose the mapping scheme



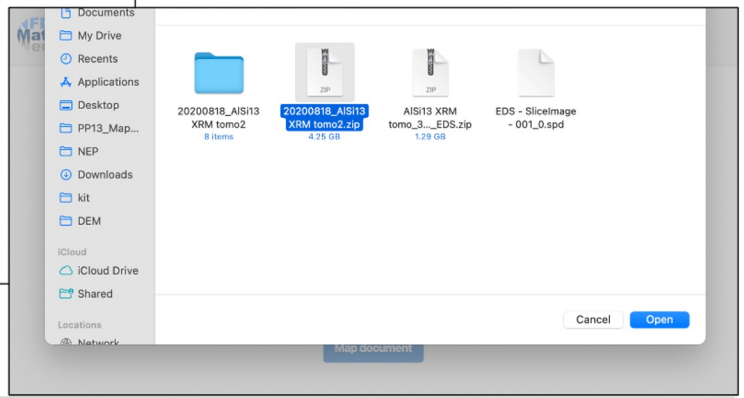
1. Choose the mapping scheme



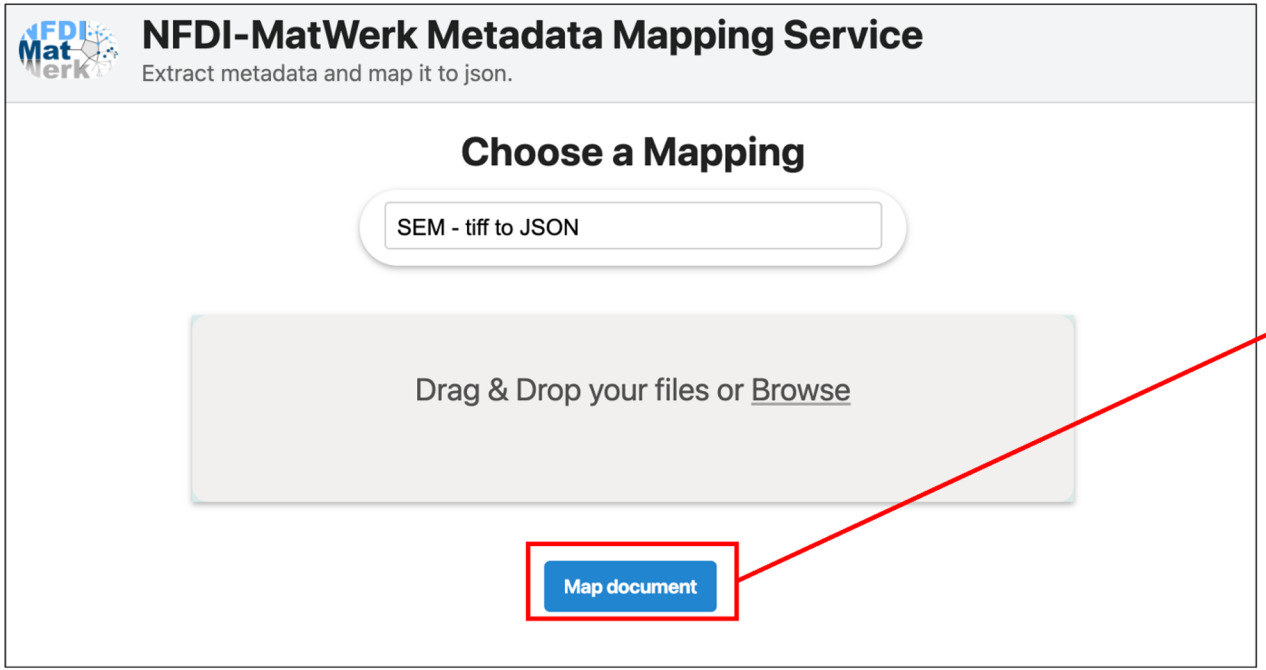
2. Uploading the dataset



Click on “Choose File” and upload the desired dataset



3. Map and download



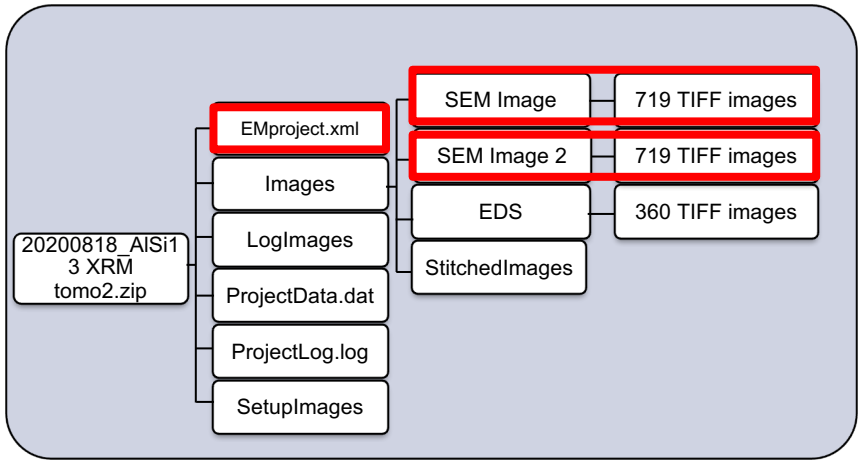
Click “Map document”. A download of the results file should automatically start, if it doesn’t, you can also click on the ”Download result” button which will appear after mapping.

4. Examine the results


```
{
    "acquisition": {
        "genericMetadata": {
            "program": {
                "programName": "Auto Slice & View 4",
                "programVersion": "4.2.1.1982"
            },
            "applicationId": {
                "identifierValue": "ASV"
            },
            "fileVersion": "1.2",
            "projectName": "20200818_AlSi13 XRM tomo2",
            "zCutSpacing": {
                "value": "2.0000000000000002E-07"
            },
            "numberOfCuts": "719",
            "pump": "TMP",
            "column": "Elstar",
            "source": "FEG",
            "eucentricWorkingDistance": {
                "value": "0.004"
            },
```

Summary

- Automatically extract data from hundreds of images in one swift process
- Reorganized metadata in a structured format according to a published schema




```
{
    "acquisition": {
        "genericMetadata": {
            "program": {
                "programName": "Auto Slice & View 4",
                "programVersion": "4.2.1.1982"
            },
            "applicationId": {
                "identifierValue": "ASV"
            },
            "fileVersion": "1.2",
            "projectName": "20200818_AlSi13 XRM tomo2",
            "zCutSpacing": {
                "value": "2.0000000000000002E-07"
            },
            "numberOfCuts": "719",
            "pump": "TMP",
            "column": "Elstar",
```

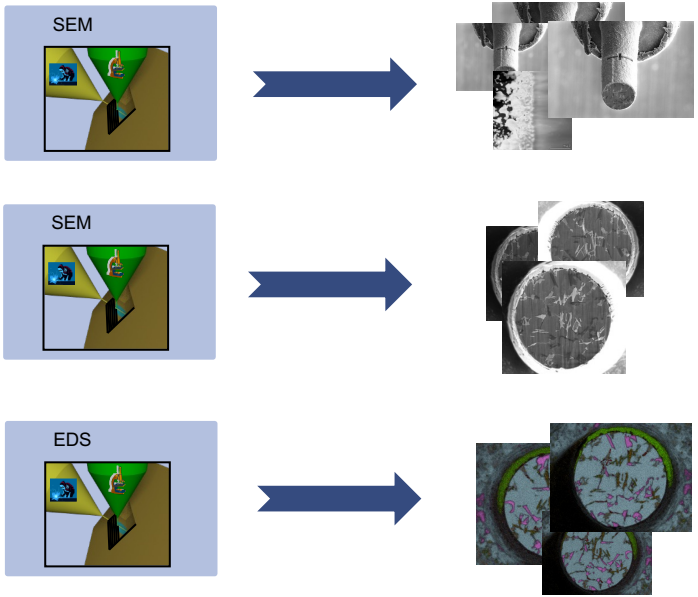
Q & A

Visual Analysis of complex metadata of SEM/FIB Tomography TIFF images

Gulzaure Abdildina, Rossella Aversa

Problem

Each experiment produces a large amout of data



Scientists need to:

- Analyze a large amount of experimental results.
- Monitor the experiment's progress for expected outcomes.
- Compare results from multiple experiments.

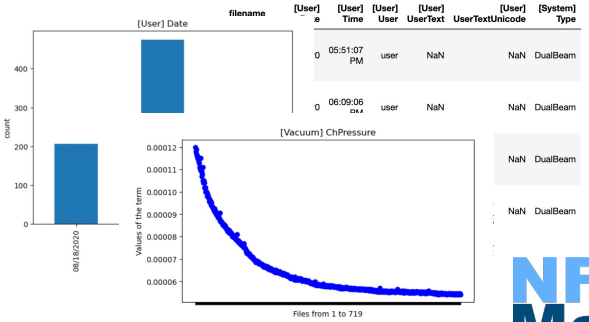
How can these processes be automated to facilitate the decision-making process?

Visual Analyzer

Python-based Jupyter Notebook which allows to:

- quickly overview metadata of SEM/FIB Tomography TIFF images
- visualize metadata for the analysis of experiment results -> makes complex data more accessible and understandable
- support scientists in accelerating decision-making processes

applying Visual Analyzer



Quick overview of the metadata

- Overall view of the metadata: statistics and common trends

	filename	user Date	user userName	program programName	spot	beamType	scan scanType	eBeam apertureSetting Size	accelerati
count	719	719	719	719	719	719	719	719	
unique	719	3	1	1	1	1	1	1	
top	SEM Image - SliceImage - 085.tif	08/19/2020	user	14.5.1.432	1	EBeam	EScan	4.53e-005	
freq	1	474	719	719	719	719	719	719	

4 rows × 75 columns

Quick overview of the metadata

Analysis of the metadata:

- Total number of metadata terms - 73
- Terms with constant values - 65
- Terms with changing values - 8 :
  - according to schema changing values can be 16

Quick overview of the metadata with changing values

Metadata with changing values

```
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   filename                719 non-null    object
 1   stage StageX            719 non-null    float64
 2   stage StageY            719 non-null    float64
 3   stage StageTa           719 non-null    float64
 4   stage workingDistance   719 non-null    float64
 5   imaging driftcorrected  719 non-null    object
 6   vacuum chamberPressure  719 non-null    float64
 7   image specimenCurrent   719 non-null    float64
 8   images creationTime     719 non-null    object
```

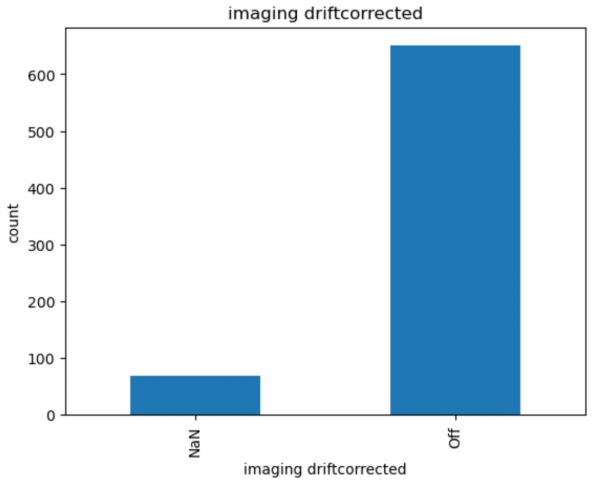
Overview of the values

filename	user Date	stage StageX	stage StageY	stage StageTa	stage workingDistance	driftc
SEM Image - SliceImage - 001.tif	08/18/2020	0.000225271	-0.00467317	0.336851	0.00402349	
SEM Image - SliceImage - 002.tif	08/18/2020	0.000225271	-0.00467317	0.336851	0.00402365	
SEM Image - SliceImage - 003.tif	08/18/2020	0.000225271	-0.00467317	0.33685	0.00402381	
SEM Image - SliceImage - 004.tif	08/18/2020	0.000225271	-0.00467317	0.336851	0.00402396	

Visualizing Metadata

Grouping images by *driftcorrected*:

- Metadata *driftcorrected* has two values
- We can also investigate how many and which images has missing values

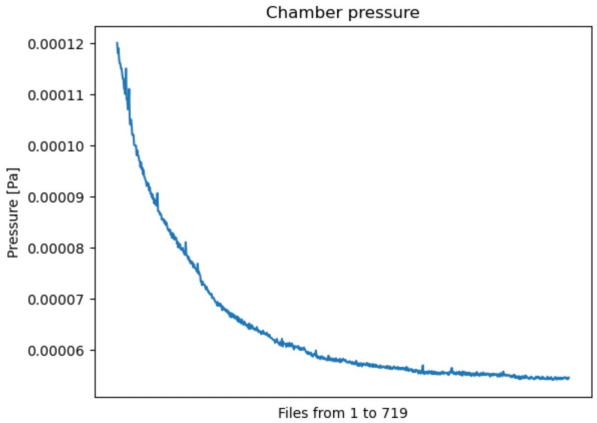


Images with missing *driftcorrected* values

	filename	imaging driftcorrected
673	SEM Image - SliceImage - 048.tif	NaN
101	SEM Image - SliceImage - 050.tif	NaN
5	SEM Image - SliceImage - 052.tif	NaN
107	SEM Image - SliceImage - 054.tif	NaN
205	SEM Image - SliceImage - 056.tif	NaN
717	SEM Image - SliceImage - 063.tif	NaN
629	SEM Image - SliceImage - 065.tif	NaN
551	SEM Image - SliceImage - 073.tif	NaN
701	SEM Image - SliceImage - 103.tif	NaN

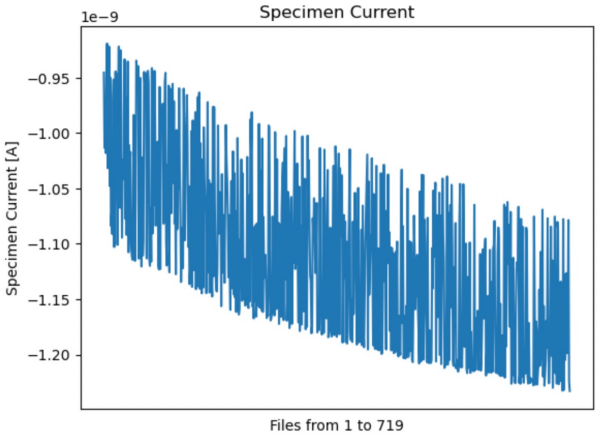
Visualizing Metadata

- Visualizing the evolution of chamber pressure values during the experiment
- Chamber pressure affects milling accuracy
- Expected to decrease over the course of the experiment



Visualizing Metadata

- Visualizing the changes of specimen current values during the experiment
- Specimen current affects the brightness of the SEM image and Energy-Dispersive X-ray Spectroscopy (EDS) signal



Conclusions

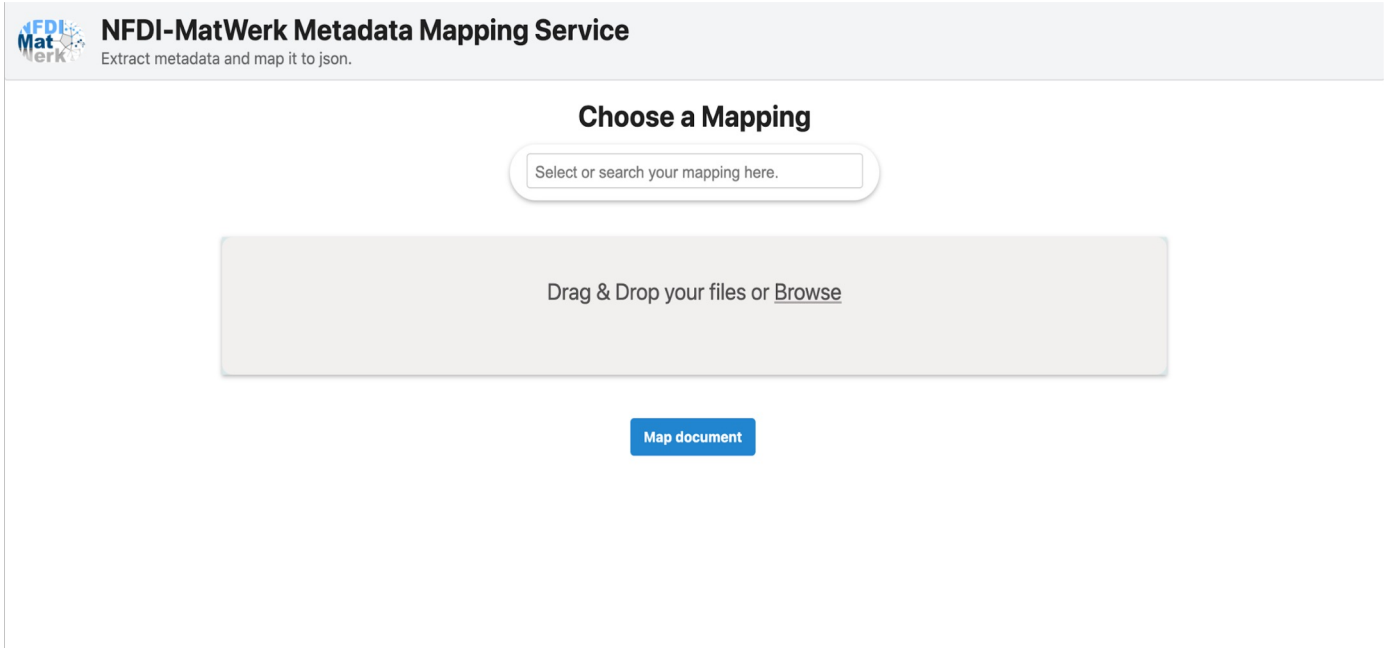
- The Visual Analyzer:
  - allows for a quick overview of the experiment as a whole
  - facilitates the comparison of experiment results
  - automates and optimizes the analysis of metadata, expediting the decision-making process
- It can be expanded and adapted according to the needs of the scientists
- Open source:
  - https://bwsyncandshare.kit.edu/s/nemrjdMGKnLfmt2

Q & A

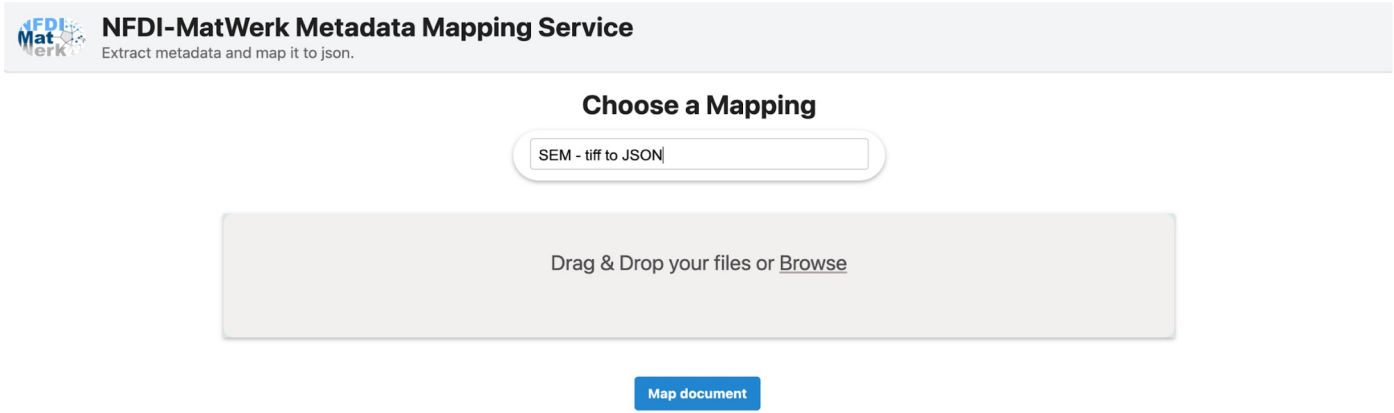
Thank You

Back-up

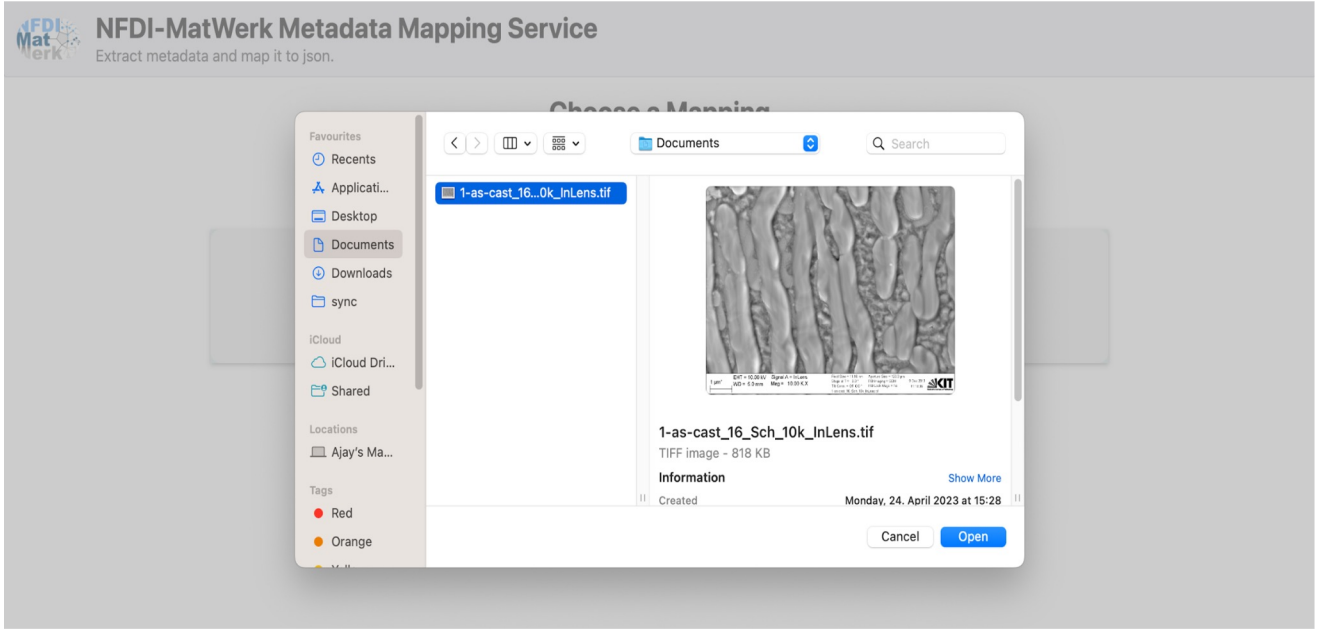
Interface for Mapping Service



Interface for Mapping Service



Interface for Mapping Service



Interface for Mapping Service