Towards FAIR Research Data Management in Experimental Materials Science
A Metadata Pipeline

Filippo Vasone

(Data Steward, University of Bologna)

Supervision: Dr. Michaela Kuepferling (INRiM); Prof. Luca Anselma (UniTo)

Paul Klee, *Hauptweg und Nebenwege*, Public domain, via Wikimedia Commons

Outline

1. Introduction
2. Background
3. Method and Tools
4. Result: Metadata Pipeline
5. Conclusion and Discussion

Outline

1. <u>Introduction</u>
2. Background
3. Method and Tools
4. Result: Metadata Pipeline
5. Conclusion and Discussion

Introduction

Among Open Science practices: research data management according to FAIR principles (Wilkinson et al., 2016).

The Project

Aim: explore and implement a FAIRification process in one of the spintronics laboratories at INRiM (experimental materials science).

Result: metadata pipeline.

Outline

Background

There are several frameworks for FAIRification: general (e.g., Jacobsen et al., 2020; Welter et al., 2023) and specific (e.g., Sinaci et al., 2020)

Two important points (from Welter et al., 2023):

- Choose an **aspect** of the FAIR principles to be implemented - they are principles!
- FAIR-**by-design**.

Background: The General Context

There are many FAIR tools for life sciences (Wilkinson et al., 2016).

Materials science has a lower level of digital maturity:

- There are several resources for computational materials science.

- Experimental materials science has fewer FAIR tools at its disposal (they are being built right now) (Ghiringhelli et al., 2023)
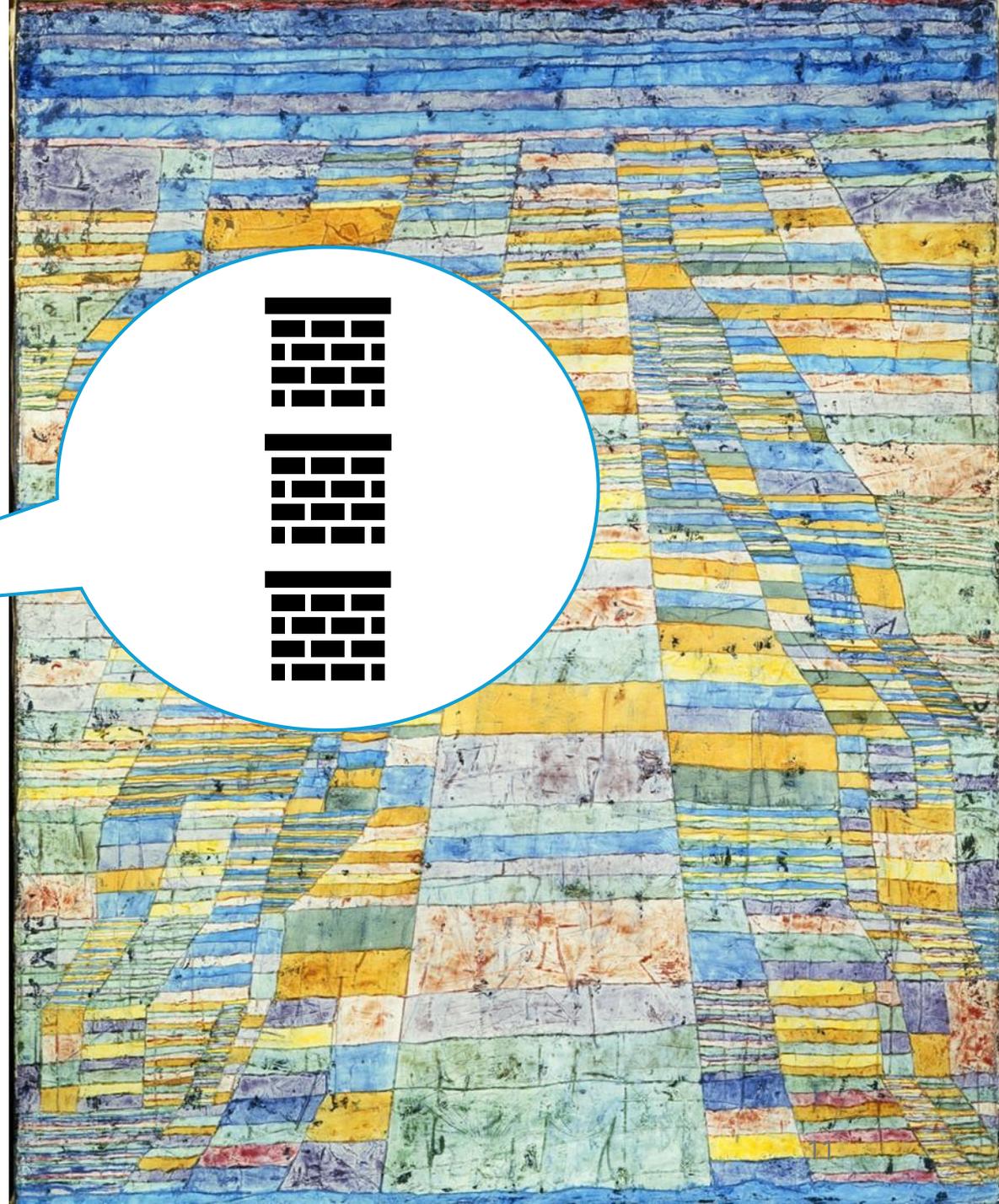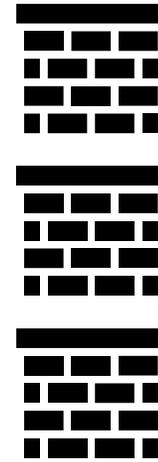
Background: The Experimental Context

Spin-Orbit Torque (SOT) Measurement
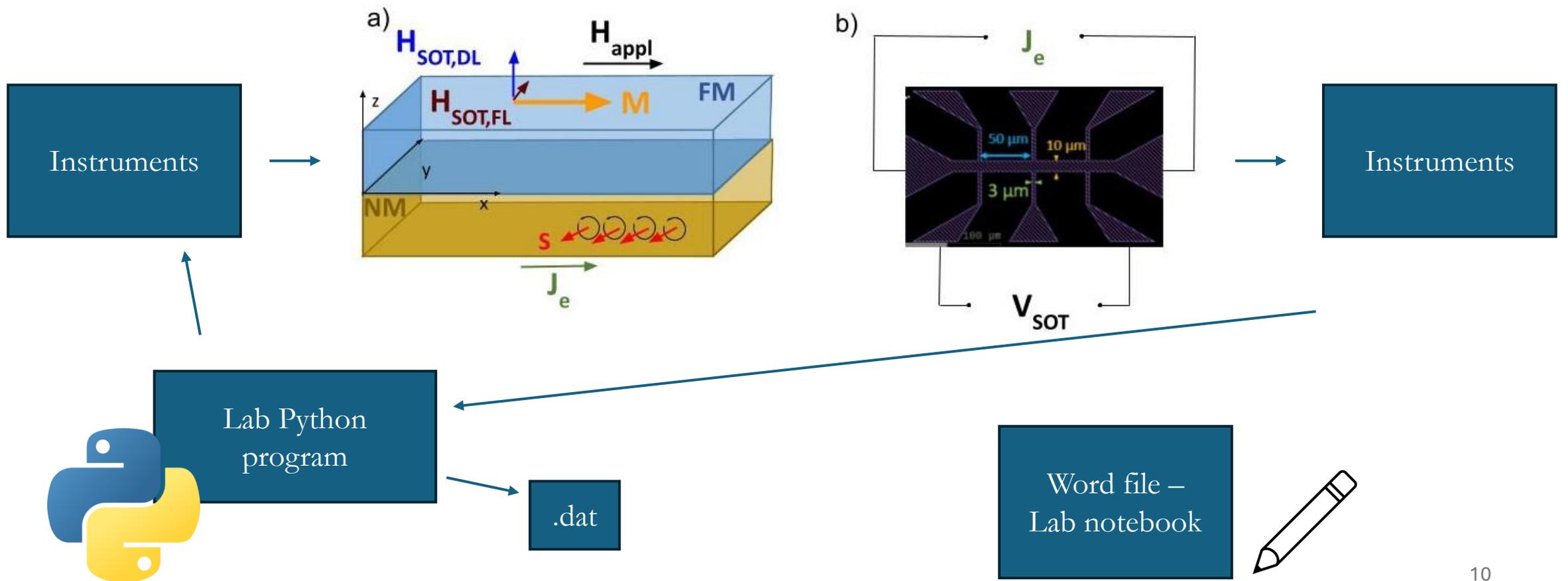
Outline

1. Introduction
2. Background
3. <u>Method and Tools</u>
4. Result: Metadata Pipeline
5. Conclusion and Discussion

Method: Contextual FAIRification

- Fragmented context of FAIRification in experimental materials science
- Usually the 'openness' of Open Science values the pluralism of scientific disciplines (Leonelli, 2023)
- FAIR-by-design

$\Rightarrow$ FAIRification starting from the context

The research and choice of tools was an important part of the project.

(Usually, you would need the support of a data steward)

Focus: Metadata and Interoperability

Metadata: set of key-value pairs that describe a potentially informative object (Pomerantz, 2015)

Interoperability: two or more information systems speak the same language (and refer to the same objects) (Guizzardi, 2020)

(e.g.: 'Temperature': of the environment? of the sample?)

Interoperability (FAIR principles)	'Modest' interoperability
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	Clarity
I2. (meta)data use **vocabularies** that follow FAIR principles	Unambiguity
I3. (meta)data include qualified references to other (meta)data	Qualified references

Tool #1: CHADA

Conceptual metadata schema (PDF, non machine-actionable)

CHADA: 'Characterization Data' (CEN, 2021): CWA 17815:2021

- Controlled vocabulary for material characterization experiments and measurements

- Metadata schema divided into 4 sections
  - User case
  - Experiment
  - Raw data
  - Data processing

Tool #2: eLabFTW

eLabFTW is an electronic laboratory notebook (CARPi et al., 2017).

=> machine-actionable (Higgins et al., 2022)

Two important features:

1) It allows you to record the metadata of an experiment in a custom way as JSON ('extra_fields').

2) Templates for frequently done experiments.

Part of template example on eLabFTW

Tool #3: NOMAD

= FAIR repository and archive for materials science (meta)data (Scheidegen et al., 2023)

The (meta)data are organized into 'sections', which can be default or defined by the user from the NOMAD Metainfo (schema language).

Tool #3: NOMAD

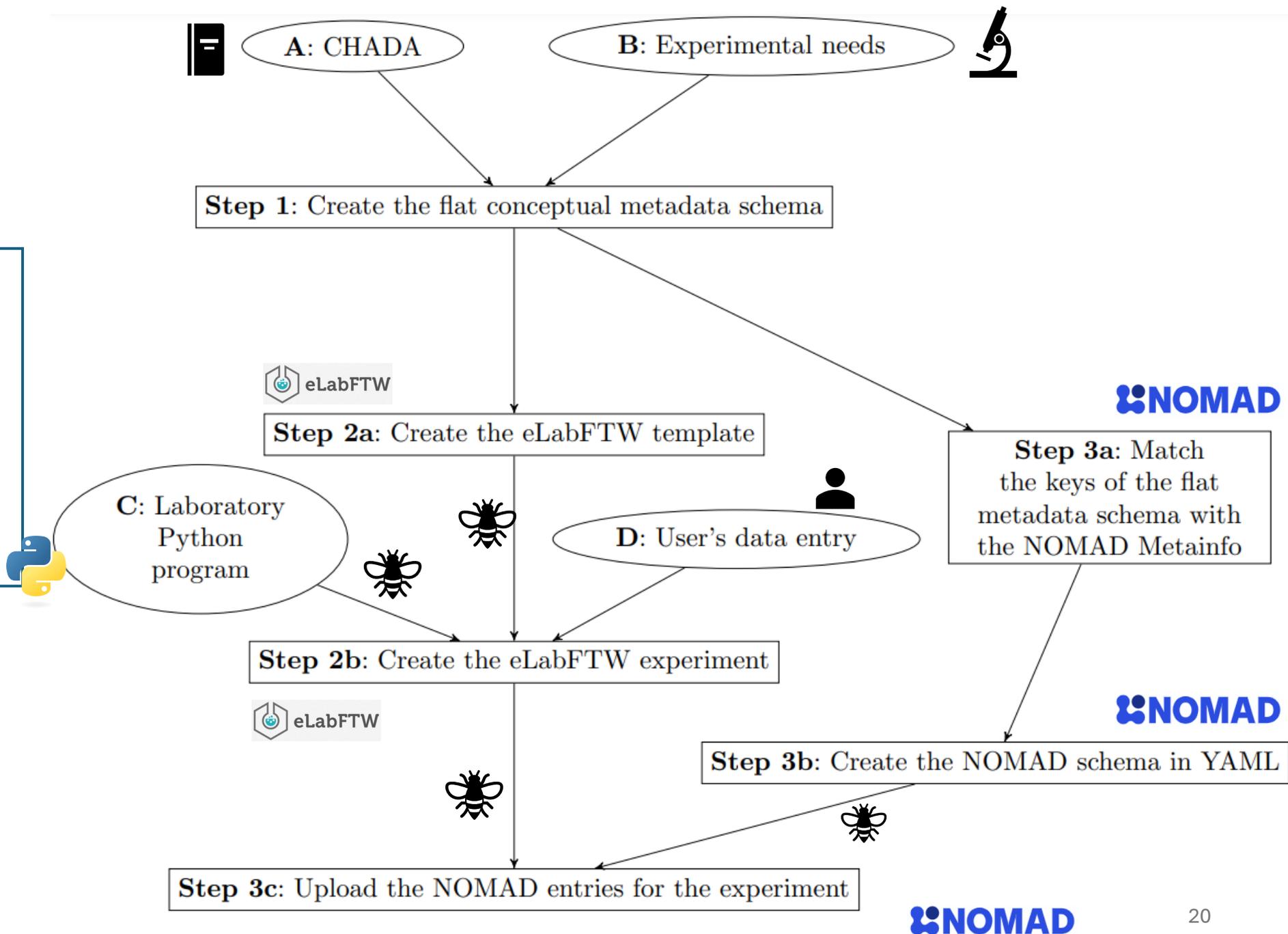Interoperability of NOMAD: example of filter in the search page.

Outline

1. Introduction
2. Background
3. Method and Tools
4. Result: Metadata Pipeline
5. Conclusion and Discussion

Result

Legend

Ellipse: input

Boxes: action/step

Bee: API integration (Python)

Step 1,2,3: temporal order

A: CHADA

B: Experimental needs

Step 1: Create the flat conceptual metadata schema

eLabFTW

Step 2a: Create the eLabFTW template

C: Laboratory Python program

D: User's data entry

Step 3a: Match the keys of the flat metadata schema with the NOMAD Metainfo

Step 2b: Create the eLabFTW experiment

eLabFTW

Step 3b: Create the NOMAD schema in YAML

Step 3c: Upload the NOMAD entries for the experiment

Step 1: flat conceptual metadata schema

Adapt CHADA to the SOT measure (A)

+

Experimental needs (B)

Experiment_RunNumber	Experiment_Detector_detector_1	DataProcessing_data_management
Experiment_DateAndTime	Experiment_Detector_detector_2	UserCase_Sample_device_geometry
UserCase_SamplePreparation	UserCase_Sample_sample_name	Experiment_MeasurementParameters_field_direction
Experiment_InteractionVolume	UserCase_Sample_device_width	Experiment_Detector_detector_1_brand
Experiment_CalibrationProcess	RawData_Host_Location	Experiment_Detector_detector_2_brand
UserCase_CharacterisationEnvironment	Experiment_MeasurementParameters_current_pulse	UserCase_SampleMaterialProperties_layer_1_material
Experiment_Probe&PhysicsOfInteraction_probe_1	UserCase_Sample_device_length	UserCase_SampleMaterialProperties_layer_2_material
Experiment_Probe&PhysicsOfInteraction_probe_2	UserCase_User_operator_name	UserCase_SampleMaterialProperties_layer_3_material
UserCase_Sample_number_of_layers	Experiment_Signal_signal_1_from_current	UserCase_SampleMaterialProperties_layer_4_material
UserCase_Sample_layer_1_thickness	DataProcessing_data_normalisation	UserCase_SampleMaterialProperties_layer_5_material
UserCase_Sample_layer_2_thickness	Experiment_Signal_signal_2_from_field	DataProcessing_main_processed_signals
UserCase_Sample_layer_3_thickness	UserCase_Batch_production_batch_name	Experiment_Probe&PhysicsOfInteraction_physics_of_interaction
UserCase_Sample_layer_4_thickness	DataProcessing_properties_elaborated_data	DataProcessing_data_processing_level_of_expertise
UserCase_Sample_layer_5_thickness	UserCase_User_operator_level_of_expertise	RawData_raw_data_file_column_headers_unit
DataProcessing_processing_reproducibility	DataProcessing_data_analysis_procedure	Experiment_MeasurementParameters_magnetic_field_sequence

Step 2: eLabFTW

Step 2a: Create an experiment template on eLabFTW from the flat conceptual metadata schema (step 1).

Portion of the JSON metadata attached to the experiment template on eLabFTW.

Step 2: eLabFTW

API Integration

Python library: https://github.com/elabftw/elabapi-python

- GET the experiment template

- POST the experiment

- Editing the metadata coming from the experiment with information from the lab Python program

    Metadata quality considerations

- PATCH the experiment with the metadata coming from the lab Python program **(C)**

Step 2: eLabFTW
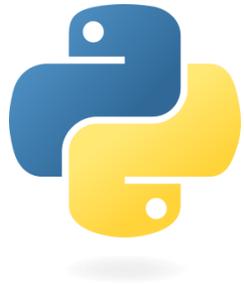
Laboratory Python program

Upload_metadata_to_eLabFTW
_with_API.ipynb

(C)

User's data entry **(D)**

Step 2b: Create the
eLabFTW experiment

=> **Locally FAIR**

Step 3: NOMAD

Step 3a: match the keys of our flat conceptual metadata schema (step 1) with the NOMAD metainfo.

In particular, we are able to model the structure of the sample (the thin film) with the NOMAD Metainfo.

Step 3b: create a NOMAD schema for the definition in YAML.

- 'SOTMeasurement' (experiment entry, in which most of the information converge).

- 'SampleStructure' (thin film).

- 'Substance' (the layers of the thin film).

```yaml
definitions:
  name: SOT
  sections:
    SampleStructure:
      base_sections:
        - nomad.datamodel.metainfo.basesections.CompositeSystem
        - nomad.datamodel.data.EntryData
      quantities:
        UserCase_Sample_device_width:
          type: np.float64
          unit: mm
          m_annotations:
            eln:
              component: NumberEditQuantity
              display:
                unit: mm
      sub_sections:
        layers:
          section:
            quantities:
              layer:
                type: Substance
                m_annotations:
                  eln:
                    component: ReferenceEditQuantity
            repeats: true
    SOTMeasurement:
      base_sections:
        - nomad.datamodel.metainfo.eln.ElnBaseSection
        - nomad.datamodel.data.EntryData
        - nomad.datamodel.metainfo.basesections.CompositeSystemReference
      quantities:
        UserCase_Sample_sample_name:
          type: str
          m_annotations:
            eln:
              component: StringEditQuantity
        Experiment_RunNumber:
          type: int
          m_annotations:
            eln:
              component: NumberEditQuantity
    Substance:
      base_sections:
        - nomad.datamodel.metainfo.eln.Substance
        - nomad.datamodel.data.EntryData
      quantities:
        UserCase_Sample_layer_thickness:
          type: np.float64
          unit: nm
          m_annotations:
            eln:
              component: NumberEditQuantity
              display:
                unit: nm
```

Step 3: NOMAD

Step 3c: the goal is the creation of NOMAD entries for the SOT measurement.

- GET the metadata from the eLabFTW experiment (elabapi-python).
- Creating Python dictionaries for each entry (experiment, thin film, layers of the thin film).
    - Part of the structure was hard-coded on the basis of the structure of NOMAD entries.
    - Dictionaries edited with the metadata coming from eLabFTW.
        - Metadata quality
            - Transforming strings containing numbers into integers and floats.
            - Date format: from Python datetime object into ISO 8601 format (accepted by NOMAD).
- The Python dictionaries are turned into JSON files and, together with the YAML schema, are uploaded onto NOMAD through API.

Step 3: NOMAD

NOMAD parsers create structured and interoperable entries (partial interoperability, due to the (until now) 'low' population of NOMAD).

	Name	Type	Mainfile		Process status ↑	
☐	SOT	Schema	SOT.schema.archive.yaml	📋	SUCCESS	→
☐	Tantalum	Substance	Tantalum.archive.json	📋	SUCCESS	→
☐	CoFeB	Substance	CoFeB.archive.json	📋	SUCCESS	→
☐	Platinum	Substance	Platinum.archive.json	📋	SUCCESS	→
☐	SOT_attempt_API	SOTMeasurement	SOT_attempt_API.archive.json	📋	SUCCESS	→
☐	Silicon dioxide	Substance	Silicon dioxide.archive.json	📋	SUCCESS	→
☐	thin_film	SampleStructure	thin_film.archive.json	📋	SUCCESS	→
☐	Silicon	Substance	Silicon.archive.json	📋	SUCCESS	→

8 entries

28

Step 3: NOMAD

Example of the 'data' section within the 'SOT_attempt_API' entry

Step 3: NOMAD

Example of the 'results' section within the 'Platinum' entry

Step 3: NOMAD

We have achieved partial interoperability.

- Description of the sample structure in an interoperable way.
- Adherence to the idea of modest interoperability (clarity, unambiguity, qualified references).

Esempio dei riferimenti nell'entry 'thin_film'

Outline

1. Introduction
2. Background
3. Method and Tools
4. Result: Metadata Pipeline
5. <u>Conclusion and Discussion</u>

Conclusion and discussion
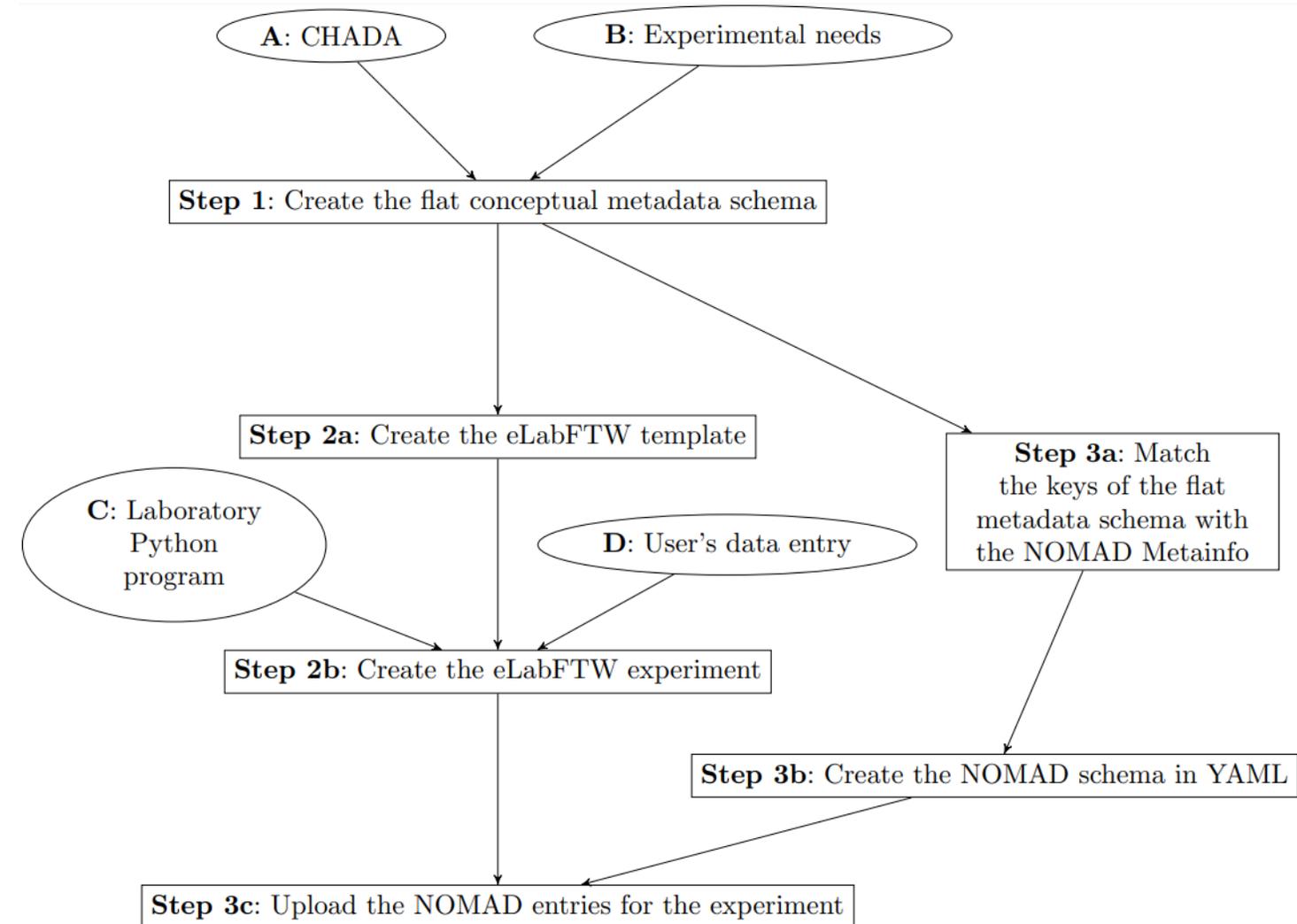
- The project achieved a semi-automation and FAIRification of an initially manual and non machine-actionable process.

- This technological infrastructure allows possible future data analysis, even with ML tools.

- Electronic laboratory notebooks have great potential for carrying out FAIRification projects.

Conclusion and discussion

Limitations

- This pipeline has two 'bosses': CHADA and NOMAD.

- eLabFTW has limited interoperability.

- Interoperability is partial due to the 'low' population of the NOMAD Metainfo.
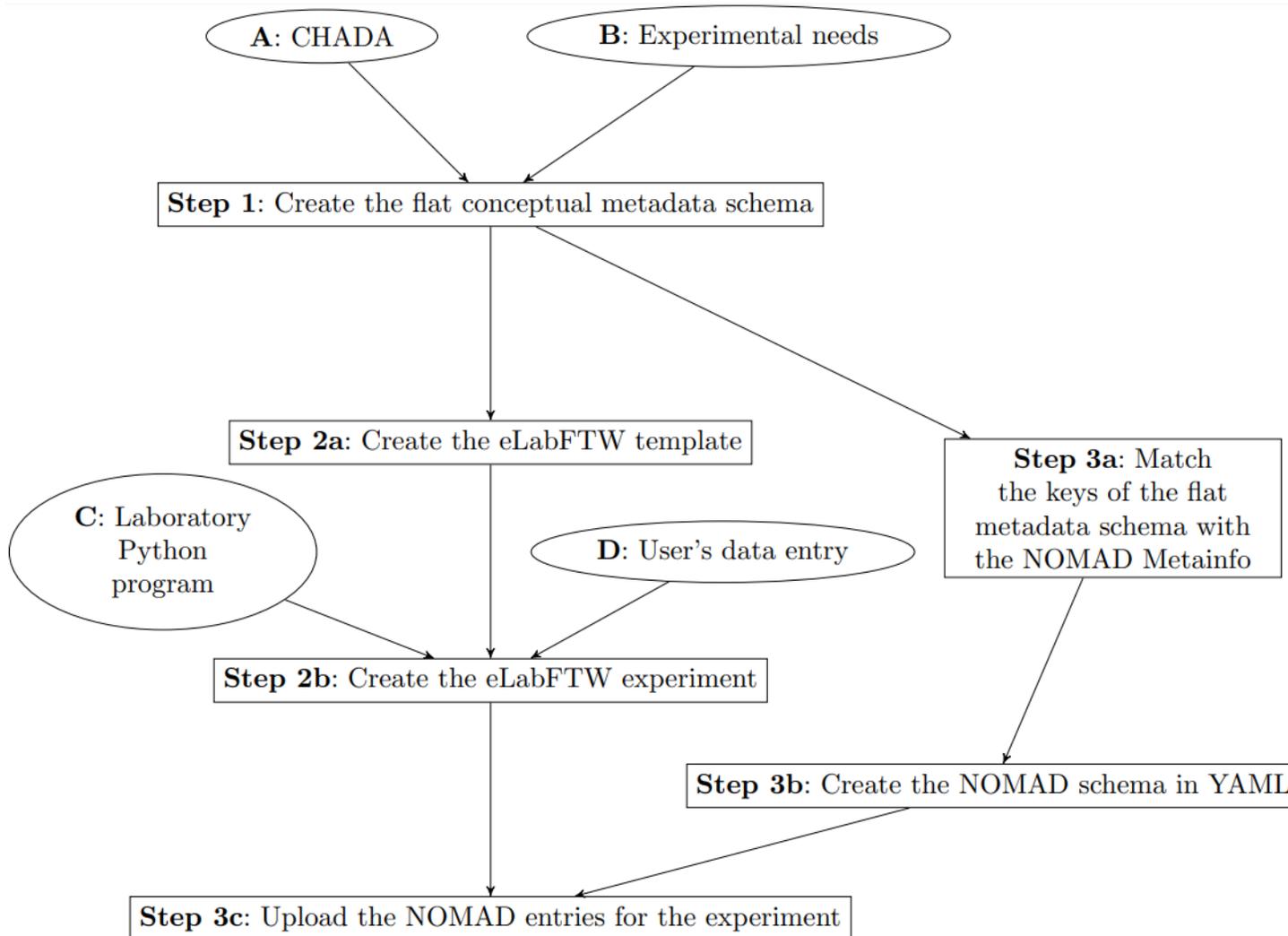
Conclusion and discussion
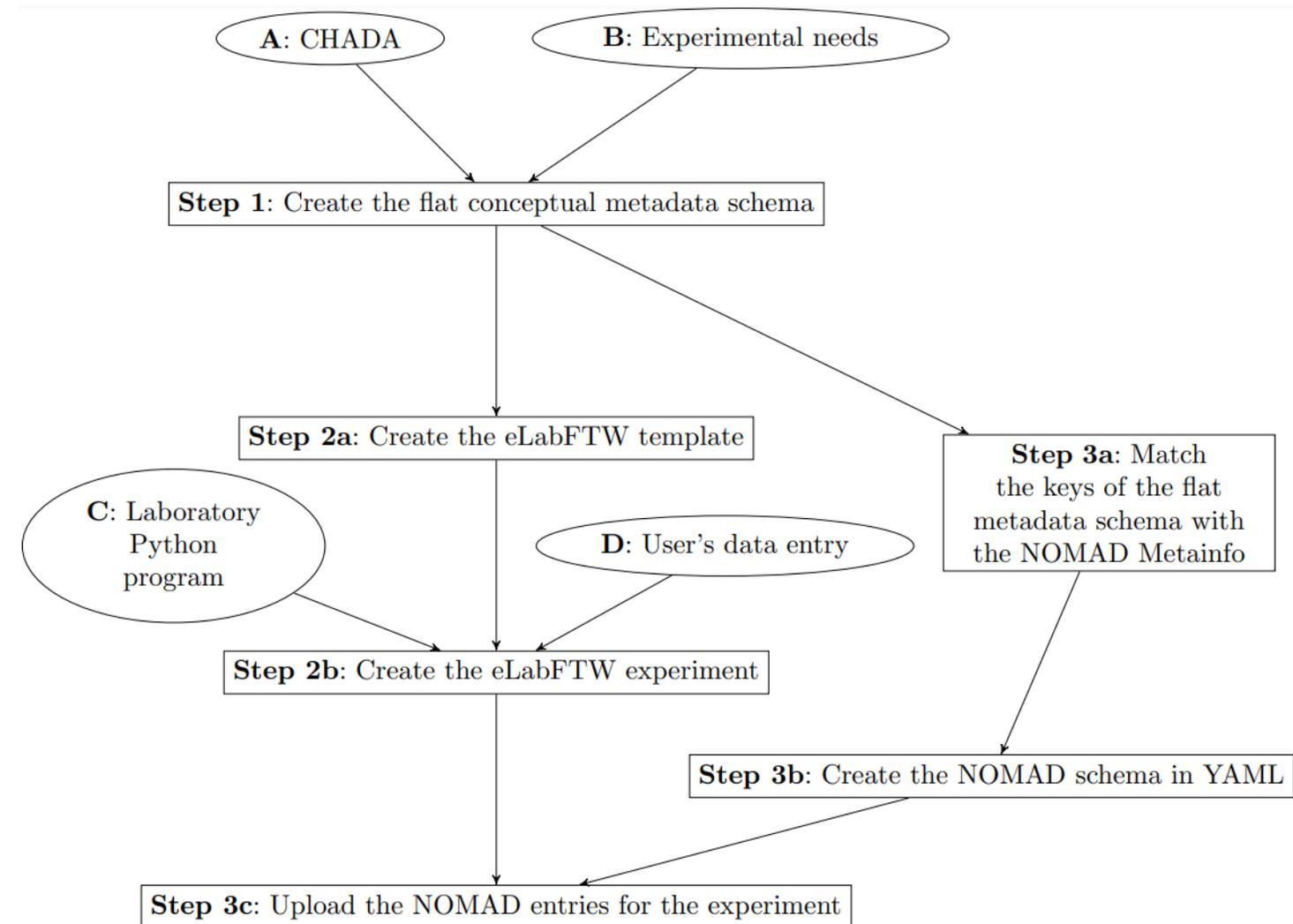
Future developments

- NOMAD plugin for CHADA.
- Thin film modeling in the NOMAD Metainfo.
- UI for the pipeline.
- Integration of the Metainfo NOMAD with materials science ontologies or foundational ontologies.

Conclusion and discussion
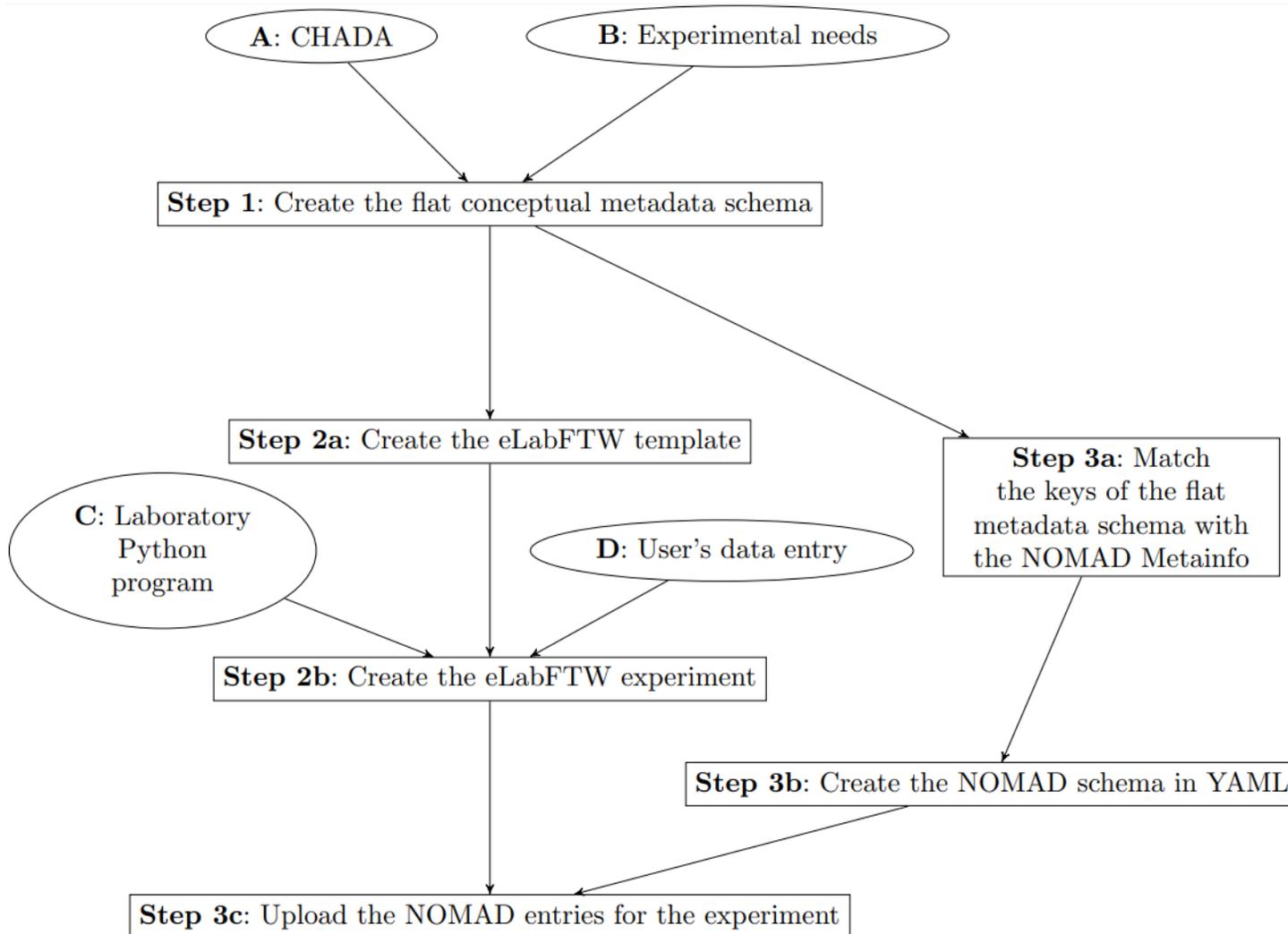
The code is available on GitHub:

https://github.com/filippo-vasone/FAIR-metadata-pipeline-for-experimental-matsci

The thesis is available on Zenodo:

https://zenodo.org/records/15226191

Take-home message #1

FAIRification is not a yes/no matter but it is a **gradual process** of concretization of the FAIR principles to your case.

Take-home message #2:
The advantages of ELN

- An ELN facilitates **collaboration within the laboratory** – even with **staff turnover.**

- An ELN promotes **collaboration between laboratories**: standardization **fosters interoperability** in the collection and production of data and metadata.

- An ELN improves **data integrity**: it prevents data loss and enables **version control** and **time-stamping**.

- An ELN promotes the **findability and accessibility** of data, metadata, and protocols. For example, it allows you to reconstruct laboratory activities in the **past**.

Take-home message #3: Consideration for implementation

- Consider the **goal** (and therefore the desiderata and constraints) with which the ELN is adopted.

- Some ELNs need to be installed on a **local server**, others rely on the **cloud**: evaluate the best option.

- Try the **demos**, if available.

- Consider **backup and data security** measures.

- If you are adopting an **open source ELN**, rely on software that has been backed by a strong community of developers for a long time.

- The ELN should allow the export of (meta)data in **interoperable formats** (e.g., JSON, CSV, ELN).

References

CARPi, N., Minges, A., & Piel, M. (2017). eLabFTW: An open source laboratory notebook for research labs. *The Journal of Open Source Software*, *2*(12), 146. https://doi.org/10.21105/joss.00146

Comité Européen de Normalisation. (2021). *CEN Workshop Agreement 17815: 2021*. Retrieved on 27/02/2025 from https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/ICT/cwa17815.pdf.

Ghiringhelli, L. M., Baldauf, C., Bereau, T., Brockhauser, S., Carbogno, C., Chamanara, J., Cozzini, S., Curtarolo, S., Draxl, C., Dwaraknath, S., Fekete, Á., Kermode, J., Koch, C. T., Kühbach, M., Ladines, A. N., Lambrix, P., Himmer, M.-O., Levchenko, S. V., Oliveira, M., … Scheffler, M. (2023). Shared metadata for data-centric materials science. *Scientific Data*, *10*(1), 626. https://doi.org/10.1038/s41597-023-02501-8

Guizzardi, G. (2020). Ontology, Ontologies and the "I" of FAIR. *Data Intelligence*, *2*(1–2), 181–191. https://doi.org/10.1162/dint_a_00040

Higgins, S. G., Nogiwa-Valdez, A. A., & Stevens, M. M. (2022). Considerations for implementing electronic laboratory notebooks in an academic research environment. *Nature Protocols*, *17*(2), 179–189. https://doi.org/10.1038/s41596-021-00645-8

Jacobsen, A., Kaliyaperumal, R., Da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A Generic Workflow for the Data FAIRification Process. *Data Intelligence*, *2*(1–2), 56–65. https://doi.org/10.1162/dint_a_00028

Leonelli, S. (2023). *Philosophy of open science*. Cambridge University Press. https://doi.org/10.1017/9781009416368

References

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

Scheidgen, M., Himanen, L., Ladines, A. N., Sikter, D., Nakhaee, M., Fekete, Á., Chang, T., Golparvar, A., Márquez, J. A., Brockhauser, S., Brückner, S., Ghiringhelli, L. M., Dietrich, F., Lehmberg, D., Denell, T., Albino, A., Näsström, H., Shabih, S., Dobener, F., … Draxl, C. (2023). NOMAD: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, *8*(90), 5388. https://doi.org/10.21105/joss.05388

Sinaci, A. A., Núñez-Benjumea, F. J., Gencturk, M., Jauer, M.-L., Deserno, T., Chronaki, C., Cangioli, G., Cavero-Barca, C., Rodríguez-Pérez, J. M., Pérez-Pérez, M. M., Laleci Erturkmen, G. B., Hernández-Pérez, T., Méndez-Rodríguez, E., & Parra-Calderón, C. L. (2020). From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. *Methods of Information in Medicine*, *59*(S 01), e21–e32. https://doi.org/10.1055/s-0040-1713684

Welter, D., Juty, N., Rocca-Serra, P., Xu, F., Henderson, D., Gu, W., Strubel, J., Giessmann, R. T., Emam, I., Gadiya, Y., Abbassi-Daloii, T., Alharbi, E., Gray, A. J. G., Courtot, M., Gribbon, P., Ioannidis, V., Reilly, D. S., Lynch, N., Boiten, J.-W., … Burdett, T. (2023). FAIR in action—A flexible framework to guide FAIRification. *Scientific Data*, *10*(1), 291. https://doi.org/10.1038/s41597-023-02167-2.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Further Reading on ELNs

https://www.kuleuven.be/rdm/en/tools/eln;

https://book.the-turing-way.org/reproducible-research/rdm/rdm-elns;

Thank you for your attention!

Contacts:
filippo.vasone@unibo.it