FAIR Data Managemnet

Sandor Brockhauser* and the FAIRmat team

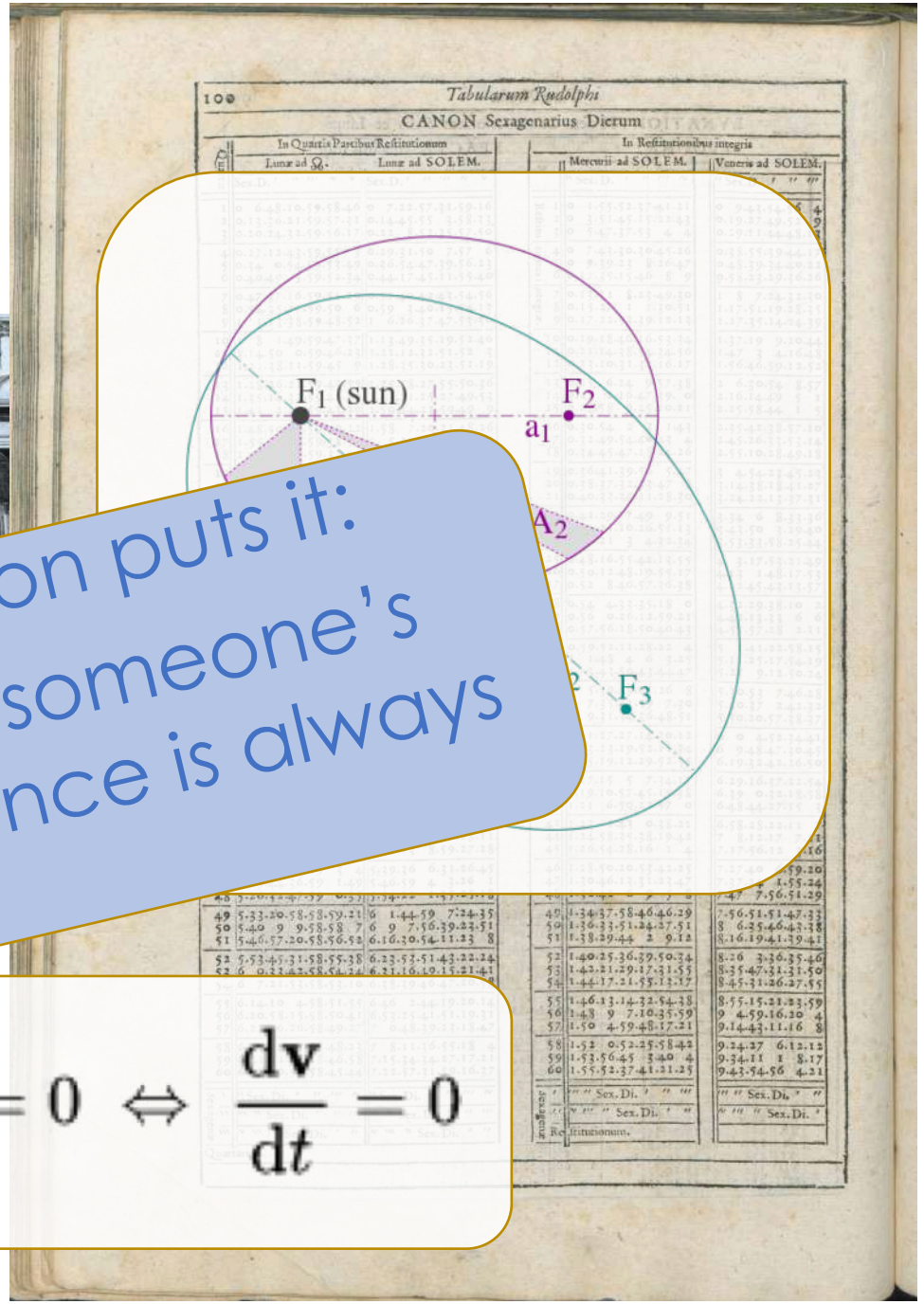
*Center for Materials Science Data,
Humboldt-Universität zu Berlin, Germany

Data Driven Science

Kepler's Laws

Tycho de Brahe (1546 – 1601)

Johannes Kepler (1571 – 1630)
Astronomia Nova, 1609
Harmonices Mundi, 161
Tabulae R



As the Royal Society of London puts it:
„Science is about not taking someone's word and so, instead, the science is always in the data"

$$\mathbf{F}_{\text{Net}} = 0 \Leftrightarrow \frac{d\mathbf{v}}{dt} = 0$$

, Kepler (name)/

FAIRmat

Science Data @ EU

FAIRmat

European Open Science Cloud

EOSC is the **Science, Research and Innovation data space** in the European strategy for data. *(European Commission)*

- seamless access
- FAIR (Findability, Accessibility, Interoperability and Reusability) management
- reliable reuse of research data and all other digital objects produced along the research life cycle
  (e.g. methods, software and publications)

Science Data @ EU

European Open Science Cloud

EOSC is the **Science, Research and Innovation data space** in the European strategy for data. *(European Commission)*

320 million €

- 2016-2018 EOSCpilot
- 2018: EOSC Launch
- 2018-2022 INFRAEOSC4 (ESFRI Clusters)
- 2019-2022 INFRAEOSC5 (National Clusters)

FAIRmat

Science Data @ EU

The EU's open science policy

https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en

- Open Data
  FAIR and open data sharing should become the default for the results of EU-funded scientific research.
- European Open Science Cloud
  store, share, process and reuse research digital objects
  (like **publications, data, and software**)
- New generation metrics
  New indicators must be developed to complement the conventional indicators **for research quality and impact**, so as to do **justice to open science** practices.
- Future of scholarly communication
  All peer-reviewed scientific **publications should be freely accessible**
- Research integrity & reproducibility of scientific results
  All publicly funded research in the EU should adhere to **commonly agreed standards** of research integrity.

FAIRmat

wwPDB / IUCr (2008)

– Macromolecular coordinates: mandatory deposition in PDB (created 1971)

– Structure factors: mandatory to accompany coordinates since 2008

11/29/2007

Announcement: Experimental Data Will Be Required for Depositions Starting February 1, 2008

Effective February 1, 2008, structure factor amplitudes/intensities (for crystal structures) and restraints (for NMR structures) will a mandatory requirement for PDB deposition.

These data must be deposited at a member site of the Worldwide Protein Data Bank (www.wwpdb.org): RCSB PDB (www.pdb.org), PDBe (www.ebi.ac.uk/msd), PDBj (www.pdbj.org), or BMRB (www.bmrb.wisc.edu).

Data can be released as soon as they have been processed and approved. There is a one-year limit on the length of time a structure and its experimental data can be put on hold, including structures that are on hold until the associated paper is published (HPUB).

This policy was developed as a result of comments and recommendations from the PDB user community, including the Commission on Biological Macromolecules of the International Union of Crystallography and the NMR Task Force, and has been endorsed by the wwPDB Advisory Committee.

Questions relating to depositions should be sent to info@wwpdb.org.

Science Data @ EU

FAIRmat

Nature Editorials (2009)

Data's shameful neglect. Nature 461, 145 (2009). https://doi.org/10.1038/461145a

Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly.

More and more often these days, a research project's success is measured not just by the publications it produces, but also by the data it makes available to the wider community....

Research funding agencies need to recognize that preservation of and access to digital data are central to their mission, and need to be supported accordingly. ...

“Data management should be woven into every course in science.”...

What is more, funding agencies and researchers alike must ensure that they support not only the hardware needed to store the data, but also the software that will help investigators to do this. One important facet is metadata management software: tools that streamline the tedious process of annotating data with a description of what the bits mean, which instrument collected them, which algorithms have been used to process them and so on — information that is essential if other scientists are to reuse the data effectively.

Science Data @ EU

FAIRmat

Naure Editorials & Science Editorial (2014)

Journals unite for reproducibility. Nature 515, 7 (2014). https://doi.org/10.1038/515007a

a group of editors representing more than 30 major journals; representatives from funding agencies; and scientific leaders assembled at the American Association for the Advancement of Science's headquarters in June 2014 to discuss principles and guidelines for preclinical biomedical research. The gathering was convened by the US National Institutes of Health, Nature and Science (see Science 346, 679; 2014)....

Reproducibility, rigour, transparency and independent verification are cornerstones of the scientific method...

Journals should recommend deposition of data in public repositories, where available, and link data bidirectionally when the paper is published. Journals should strongly encourage, as appropriate, that all materials used in the experiment be shared with those who wish to replicate the experiment....

The hope is that these guidelines will be viewed not as onerous, but as part of the quality control that justifies the public trust in science.

Science Data @ EU

FAIRmat

Acta Crystallographica D, IUCr (2019)

Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography, Acta Cryst D, Struct. Biology
https://doi.org/10.1107/S2059798319004844

(i) Authors should provide a permanent and prominent link from their article to the raw data sets which underpin their journal publication and associated database deposition of processed diffraction data (e.g. structure factor amplitudes and intensities) and coordinates, and which should obey the `FAIR' principles that their raw diffraction data sets should be Findable, Accessible, Interoperable and Re-usable (https://www.force11.org/group/fairgroup/fairprinciples).

(ii) A registered Digital Object Identifier (doi) should be the persistent identifier of choice (rather than a Uniform Resource Locator, url) as the most sustainable way to identify and locate a raw diffraction data set.

In 2018, the IUCr Commission on Biological Macromolecules (CBM) and the IUCr Committee on Data submitted a memorandum to the IUCr Executive Committee and proposed a mechanism for making diffraction experiments publicly available. The goal of ensuring better reproducibility of scientific discoveries in structural biology would be achieved, in part, by:

(1) Allowing the scientific community to identify and re-use the original diffraction image data from a diffraction experiment, which is the primary source of information used to determine a particular macromolecular structure.

(2) Facilitating structure re-determination using those original diffraction image data.

(3) Providing researchers with a straightforward mechanism that will permit assessing the correctness of the structure determination process.

(4) Providing a mechanism to ensure that the structures in the PDB and the publications derived from them are of the highest possible quality.

IUCr Journals are now taking the lead by encouraging authors to provide a doi for their deposited original raw diffraction data when they submit an article describing a new structure or a new method tested on unpublished diffraction data

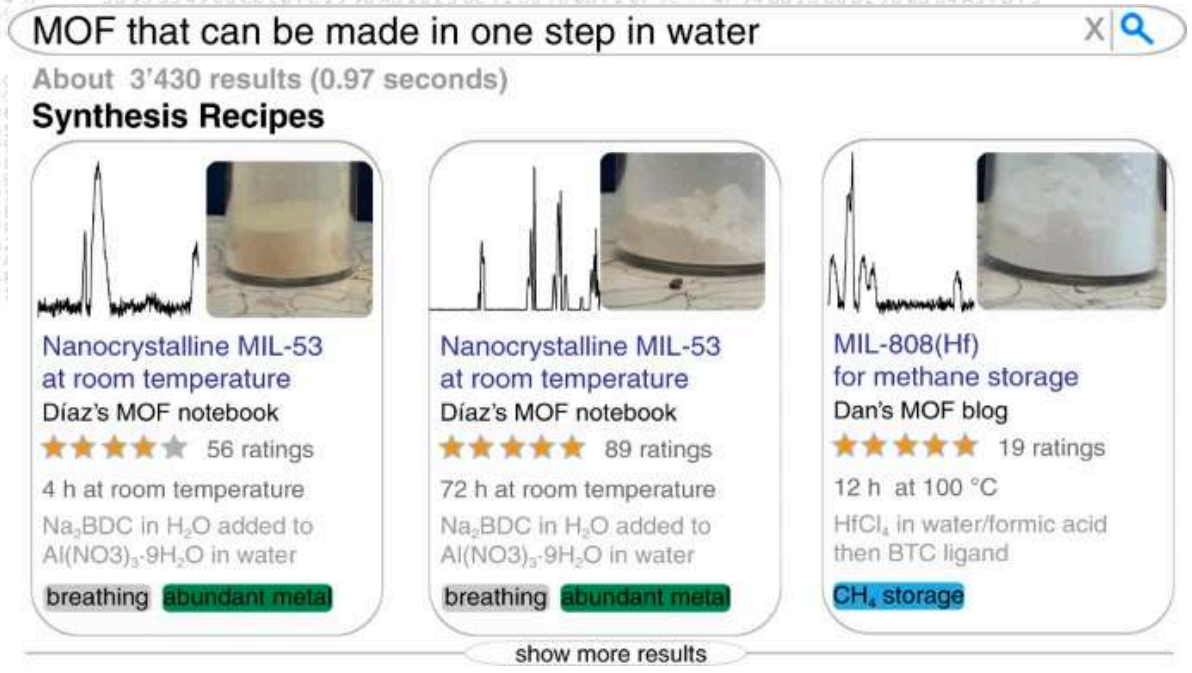
Science Data @ EU

FAIRmat

Nature Chemistry (April 2022)

Making the collective knowledge of chemistry open and machine actionable

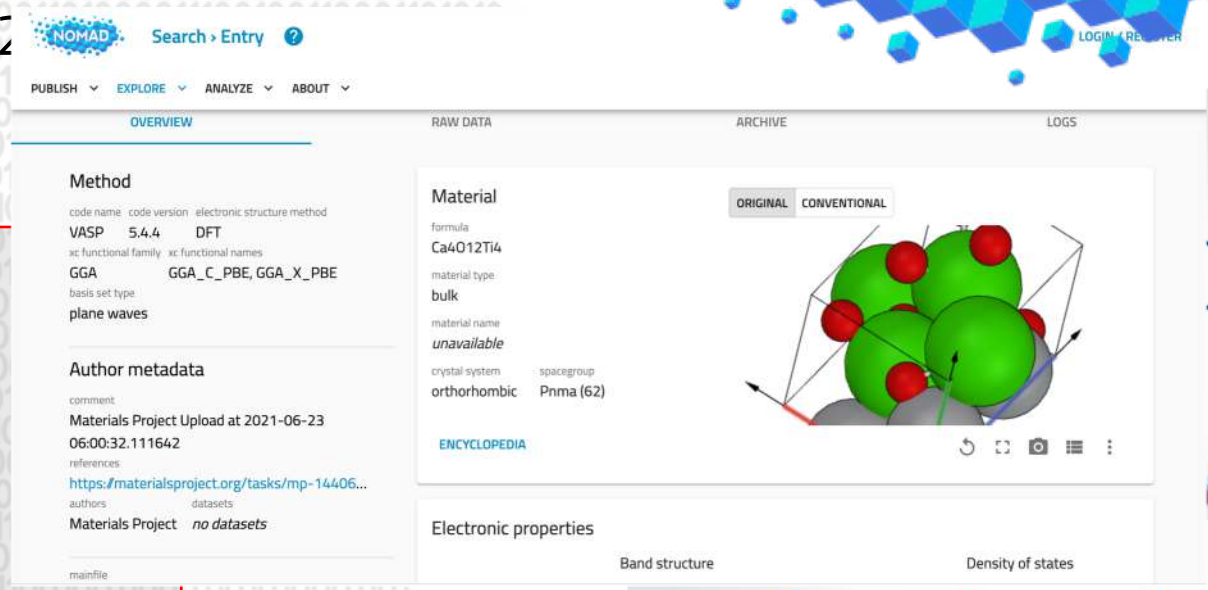
Jablonka, K.M., Patiny, L. & Smit, B. Making the collective knowledge of chemistry open and machine actionable. Nat. Chem. 14, 365–376 (2022). https://doi.org/10.1038/s41557-022-00910-7

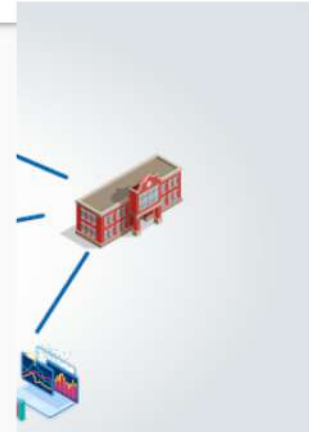


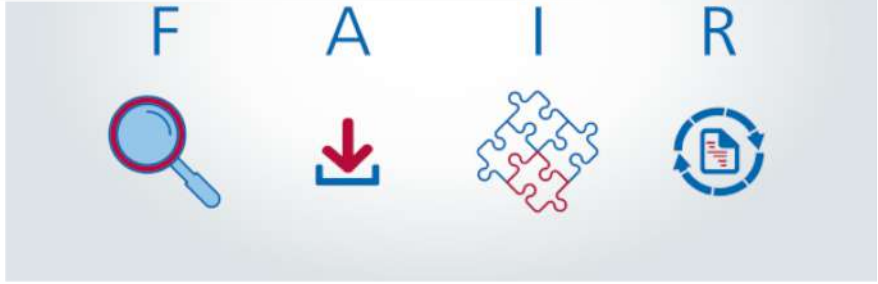
FAIRmat @ NFDI

German National Research Data Infrastructure, NFDI

9 projects started in 2020
10 projects started in 2





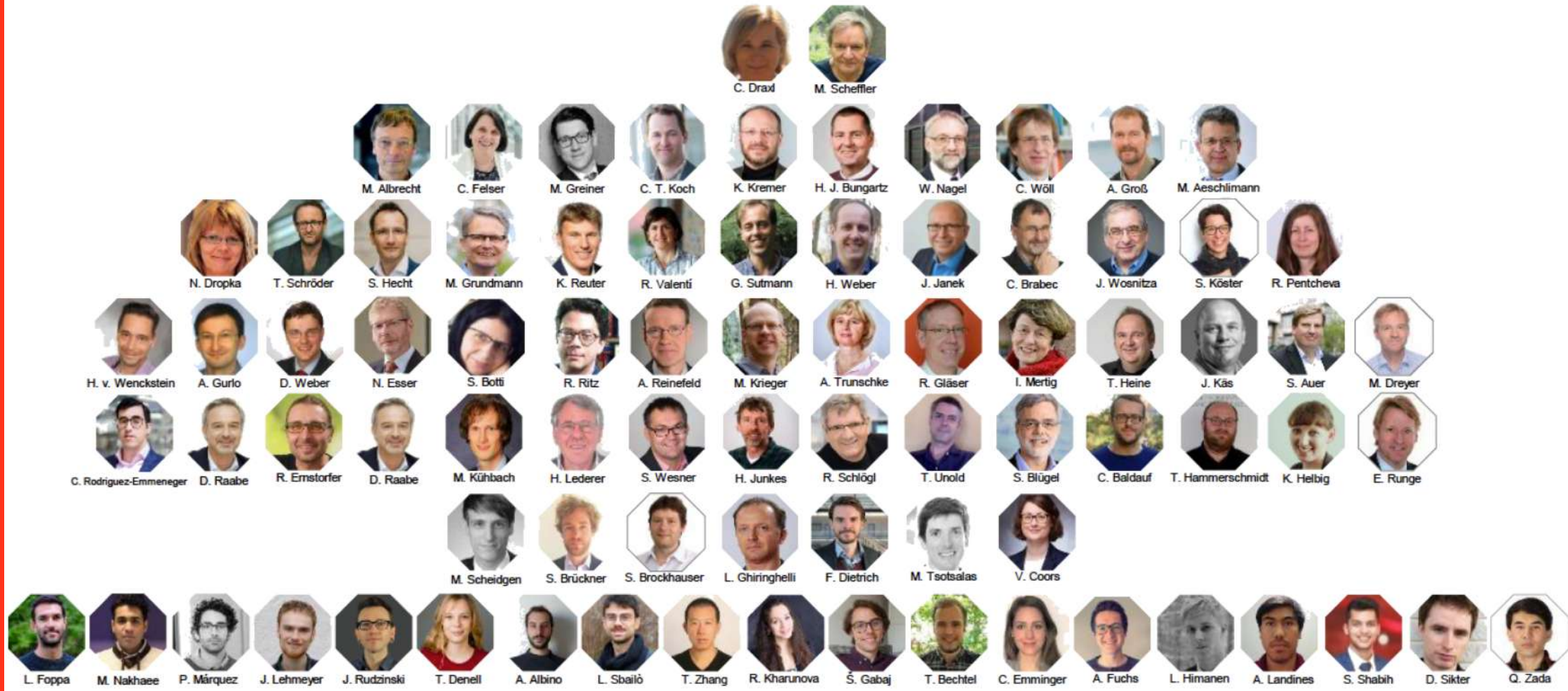


FAIRmat project
5 years
9 institutes
52 research groups
14 M€

FAIRmat

FAIRmat @ NFDI

FAIRmat for Condensed-matter and chemical-physics



https://www.fairmat-nfdi.eu/fairmat/fairmat_/fairmatteam

FAIRmat

FAIRmat @ NFDI

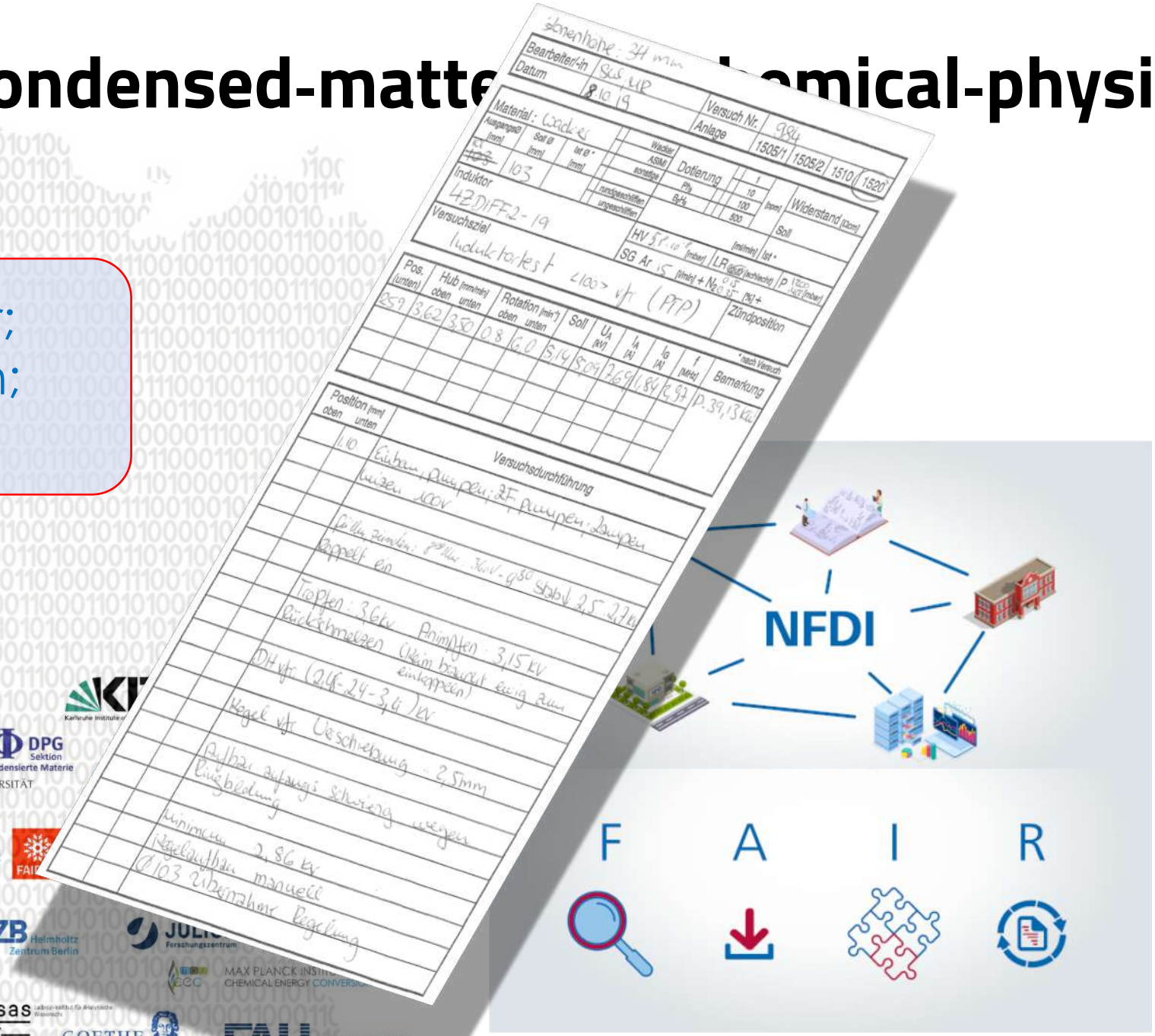
FAIRmat for Condensed-matter and chemical-physics

FAIRmat

FAIRmat @ NFDI

FAIRmat for Condensed-matte[illegible]mical-physics

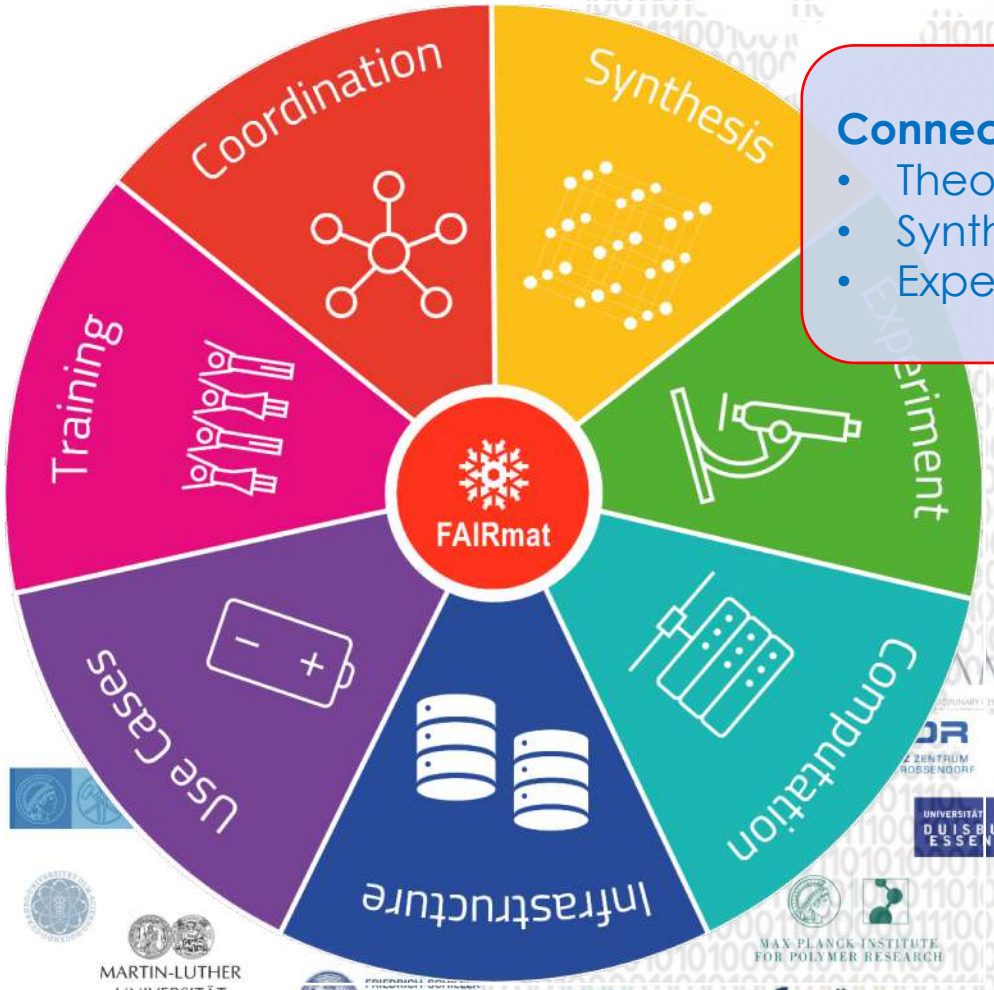
Data Center;
Digitialisation;
ELNs; LIMS



FAIRmat

FAIRmat @ NFDI

FAIRmat for Condensed-matter and chemical-physics



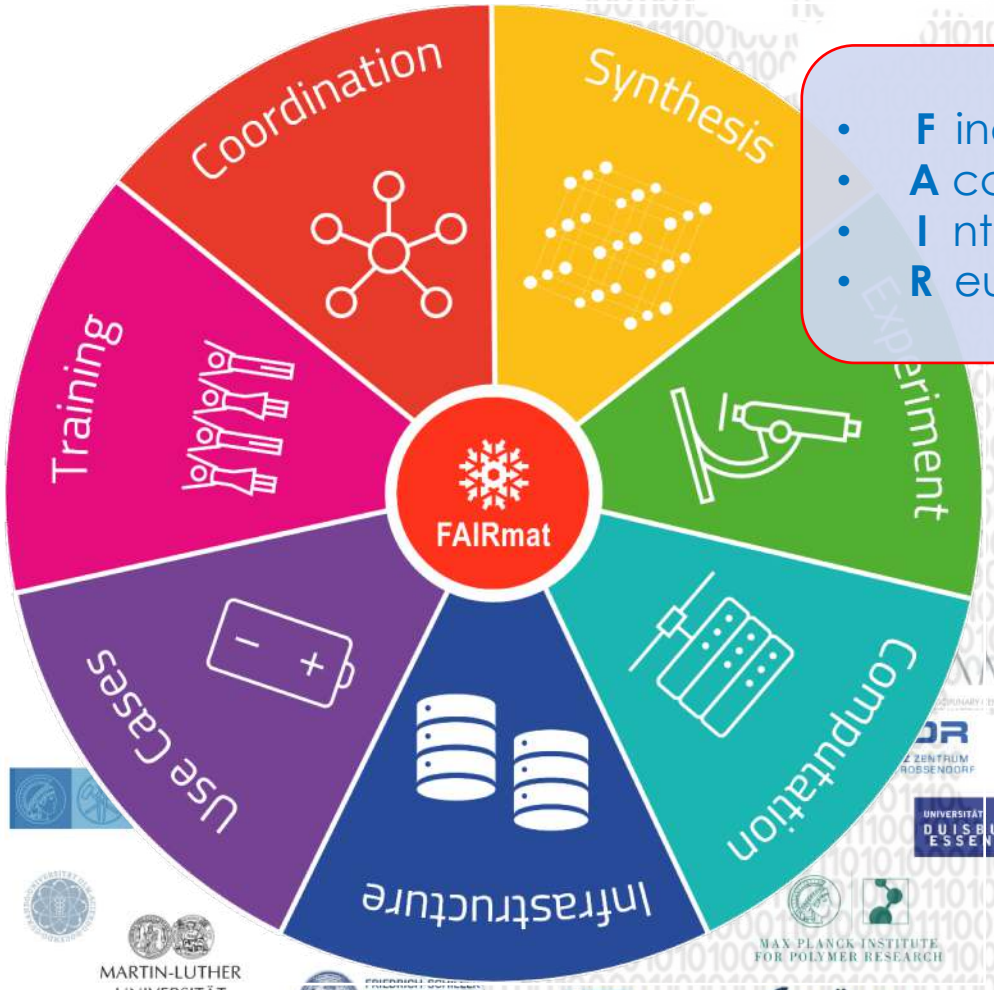
Connect

- Theory & Computation
- Synthesis
- Experiments

FAIRmat

FAIR Principles

FAIRmat

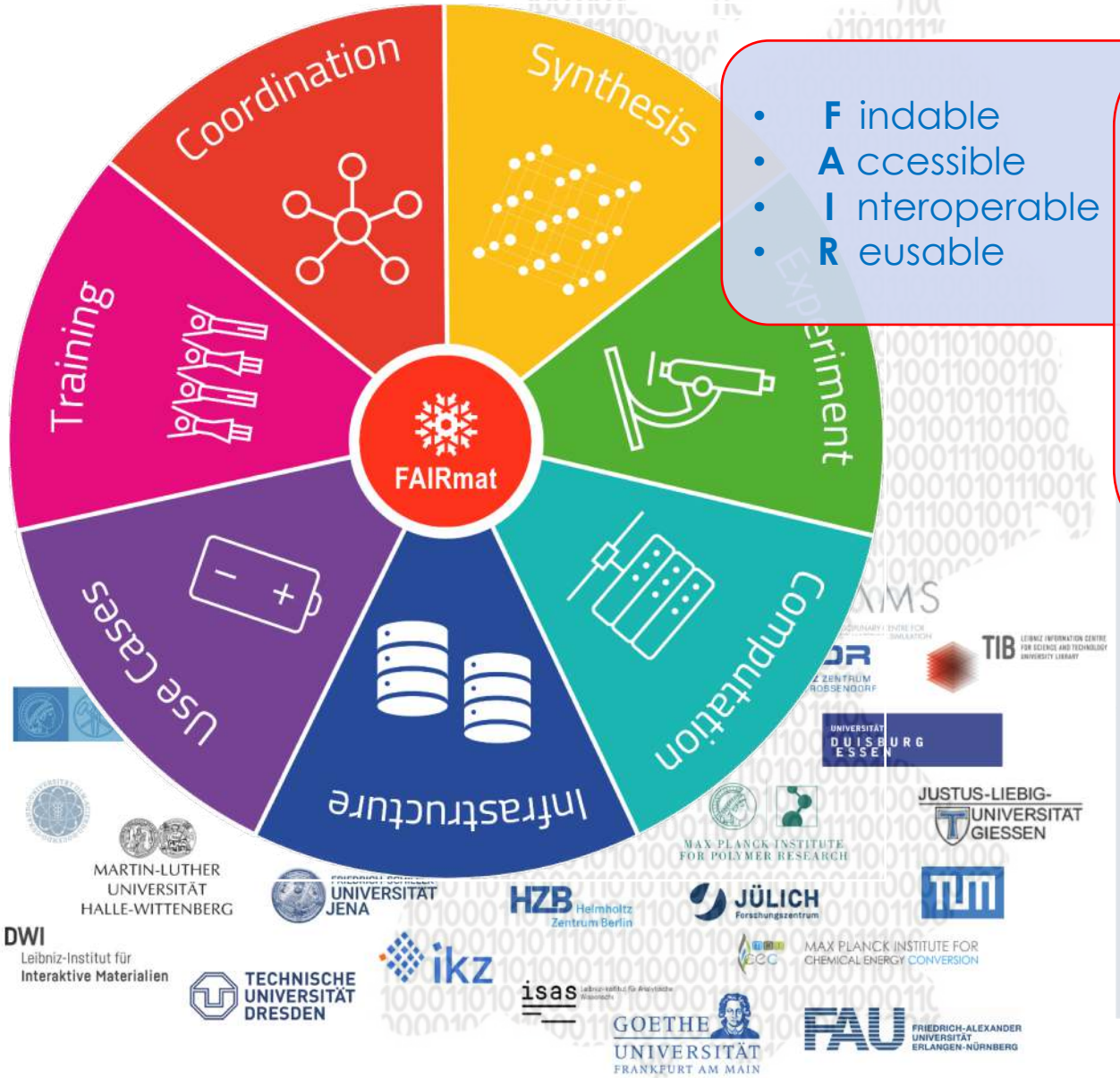


- **F** indable
- **A** ccessible
- **I** nteroperable
- **R** eusable

FAIRmat

FAIR Principles

Findable



- **F** indable
- **A** ccessible
- **I** nteroperable
- **R** eusable

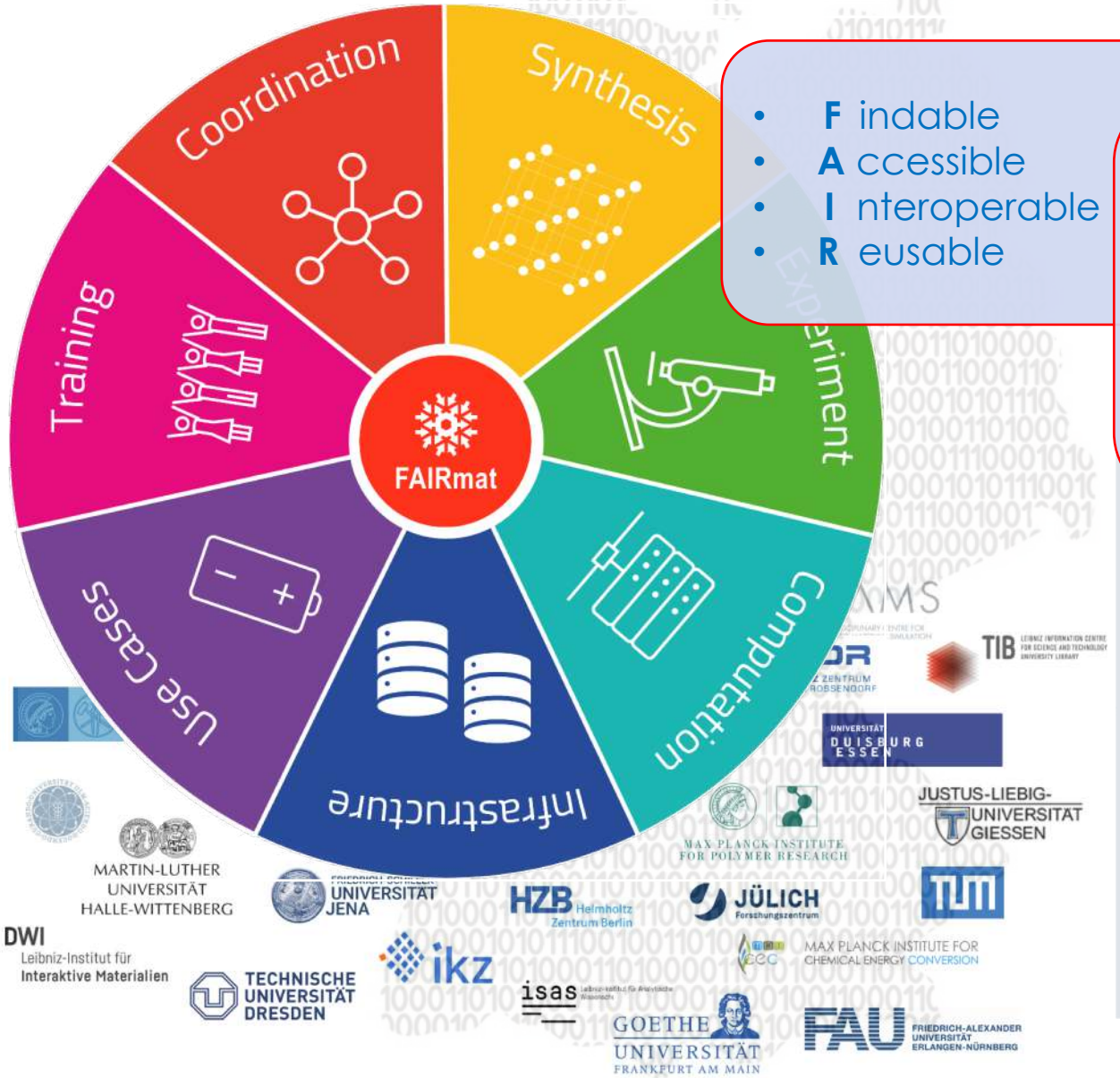
Findable

- Search platform
- who / what / which project / ...
- Computation metadata:
  Input parameters / Output
- Experiment/Synthesis parameters
  ELN integration
- Materials properties
- Encyclopedia

FAIRmat

FAIR Principles

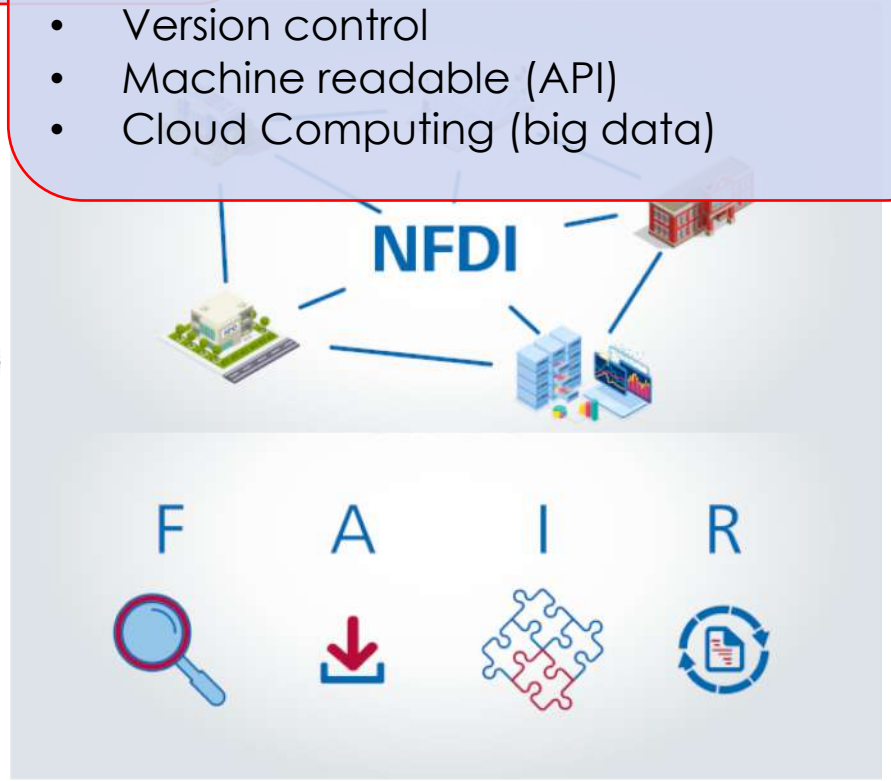
Accessible



- **F** indable
- **A** ccessible
- **I** nteroperable
- **R** eusable

Accessible

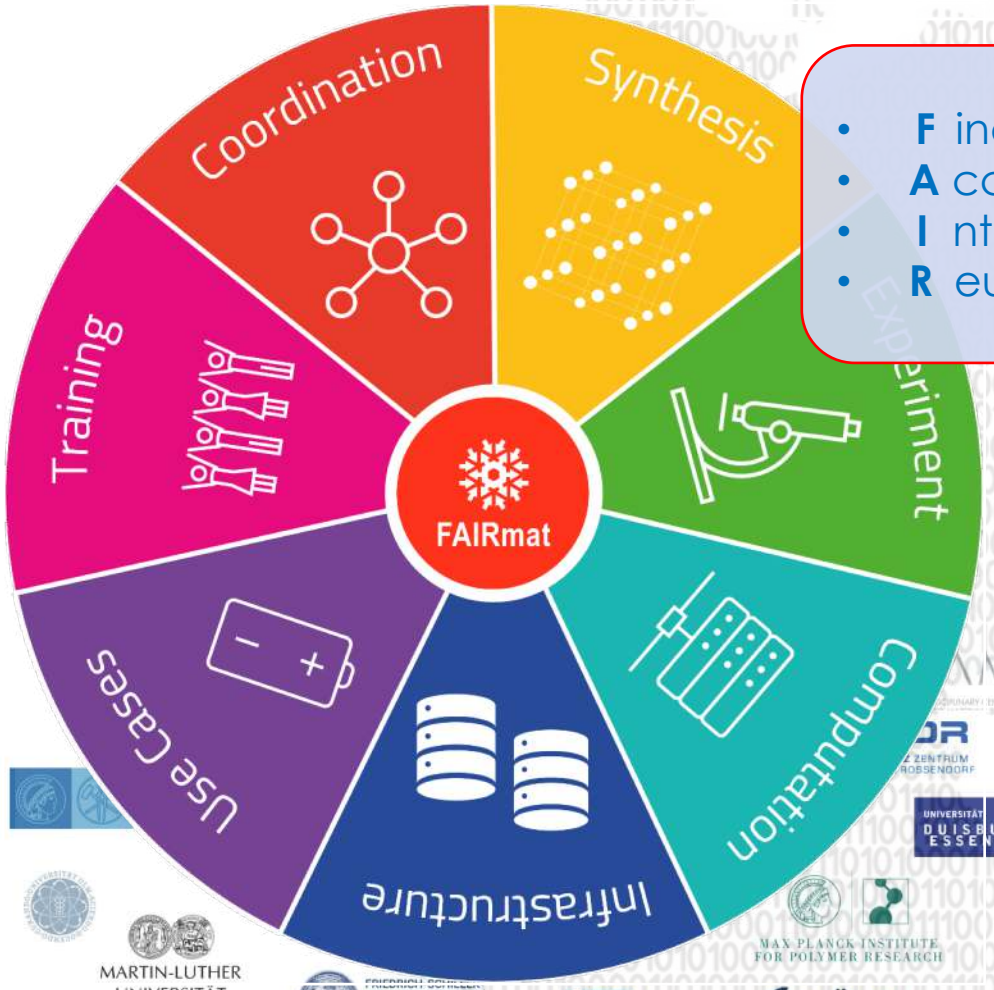
- Access & Download portal (UX)
- Data Exploration & Visualisation
- DOI
- Version control
- Machine readable (API)
- Cloud Computing (big data)



FAIRmat

FAIR Principles

Interoperable



- **F** indable
- **A** ccessible
- **I** nteroperable
- **R** eusable

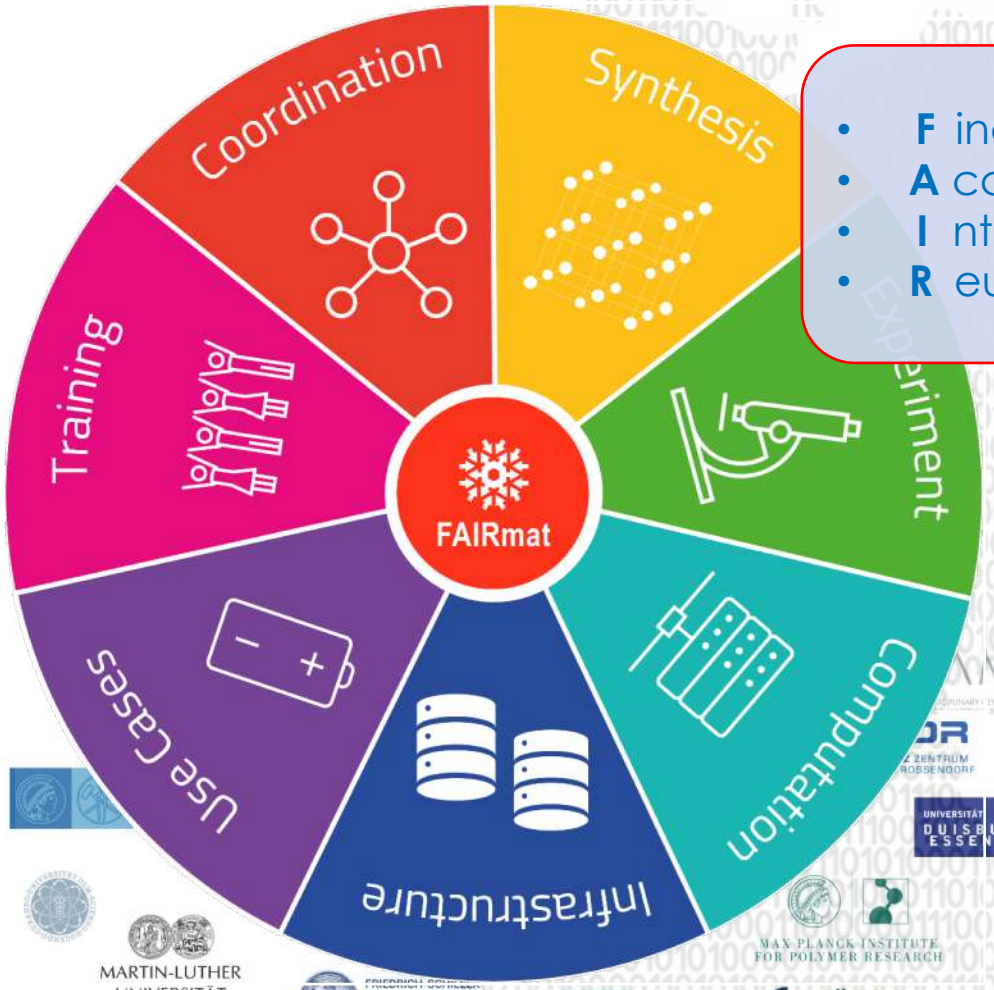
Interoparable

- Glossary with defined vocabulary
- Machine interpreatable
- Ontology relationships
- Connection to external ontologies

FAIRmat

FAIR Principles

Reusable



- **F** indable
- **A** ccessible
- **I** nteroperable
- **R** eusable

Reusable

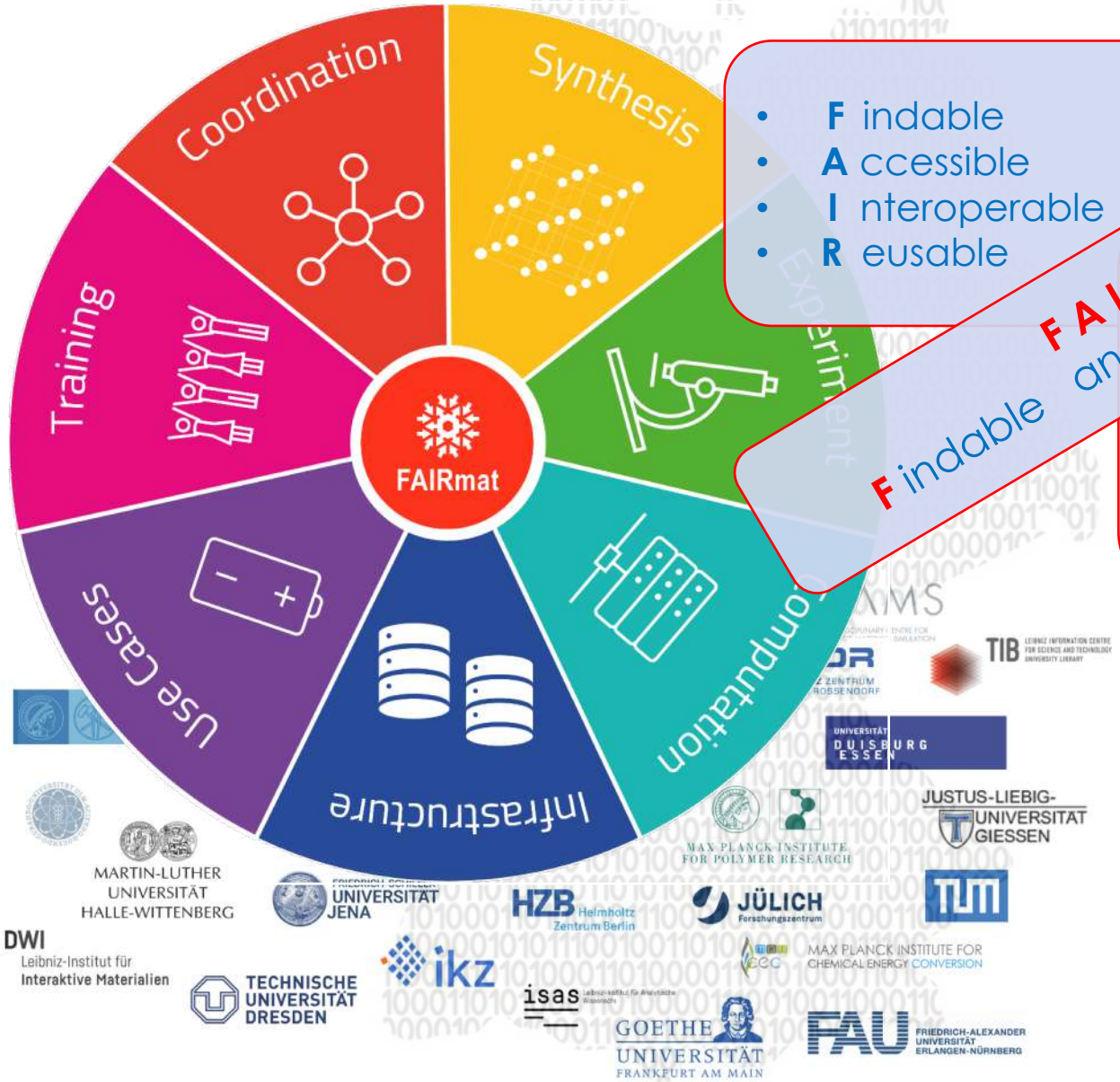
- Data manipulation
- Version control
- Remote Data Analysis Services

- Machine Readable
- Machine Interpretable

FAIRmat

FAIR Principles

AI Ready



- **F** indable
- **A** ccessible
- **I** nteroperable
- **R** eusable

F A I R and **AI R**eady

Findable

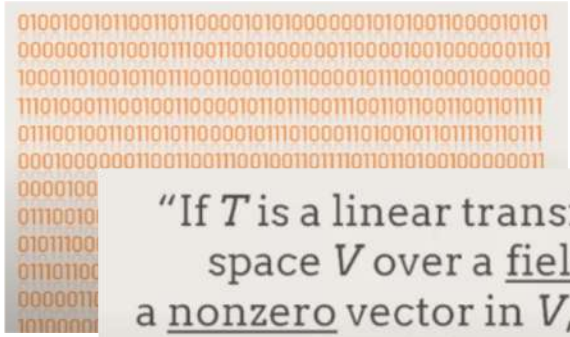
- Data manipulation
- Version control
- Remote Data Analysis Services

- Machine Readable
- Machine Interpretable

FAIRmat

Reproducibility

Data Management and Reproducibility (in Computational Science)



"If T is a linear transformation from a vector space V over a field F into itself and $\mathbf{v}$ is a nonzero vector in V, then $\mathbf{v}$ is an eigenvector of T if $T(\mathbf{v})$ is a scalar multiple of $\mathbf{v}$."

- DATA (information representation)
- Metadata (contextual information)
- Data + Metadata => Input data (for use)
- **Application** (**uses** Data for search, analysis, visualisation, etc.)
- Data + Metadata + Application => Output (new information)
- OUTPUT DATA + NEW METADATA (new information representation)

FAIRmat

Reproducibility

Reproducibility in Experiments and Synthesis

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detectors
  - Data Processing
  - *Notes*

- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

FAIRmat

Reproducibility

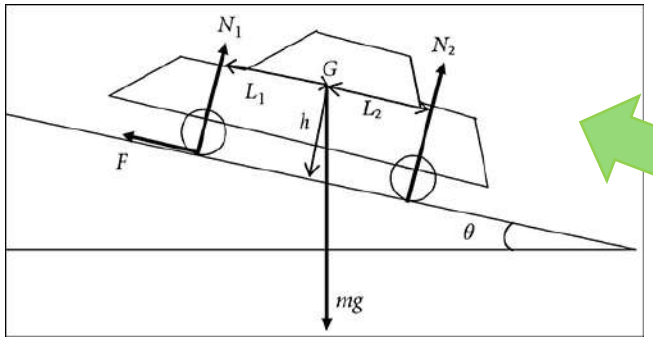
Reproducibility in Experiments and Synthesis

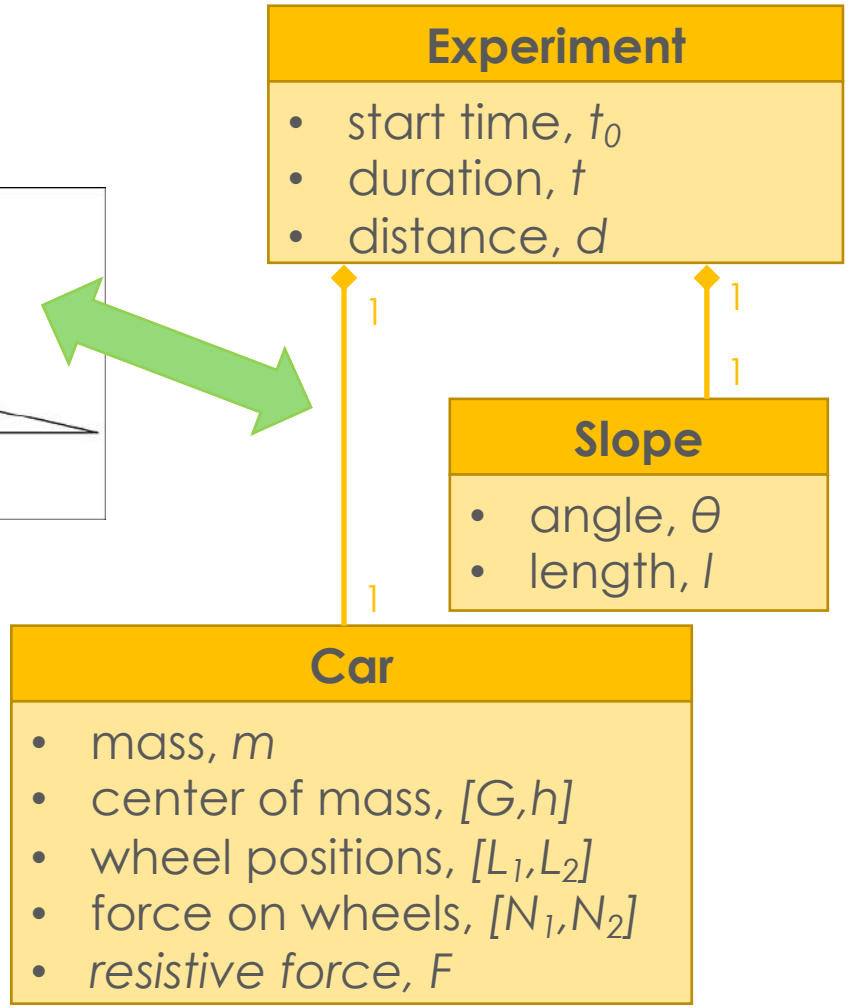
- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detectors
  - Data Processing
  - *Notes*
- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

FAIRmat

Experiment Data Models

Physics Model





Experiment Parameters

Reproducibility in Experiments and Synthesis

Reproducibility

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detectors
  - Data Processing
  - *Notes*
- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

Conceptual Design Model

- *Implementation independent Physics Model*
- *Does not tell how exactly it is performed*

Samples:
- Composition
- Geometry
- History
- …

Measurements:
- ARPES experiment
- XRD
- XRF measurement
- Hyperspectral PL imaging
- THz spectroscopy
- I-V measurement
- …

Sample Environment.:
- Temperature
- Pressure
- …

Processes:
- Spin coating
- PVD deposition
- Solution preparation
- PLD deposition
- Hot plate annealing
- …

FAIRmat

Reproducibility

Reproducibility in Experiments and Synthesis

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detectors
  - Data Processing
  - *Notes*

- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

Technical Design Model

- *Implementation specific*
- *Not only Physics Model,*
  But also Technical Details

E.g.
what was the pressure in a chamber

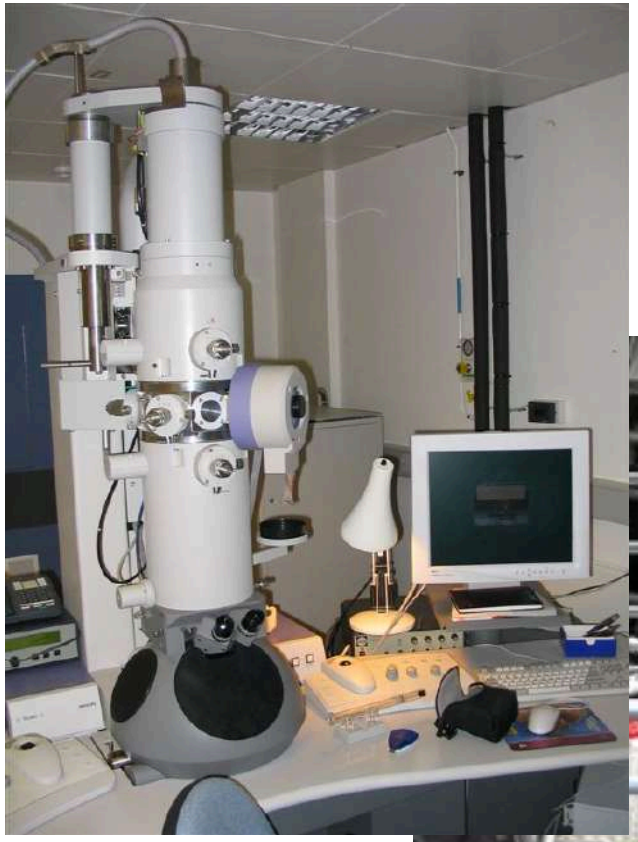
how it has been produced and maintained

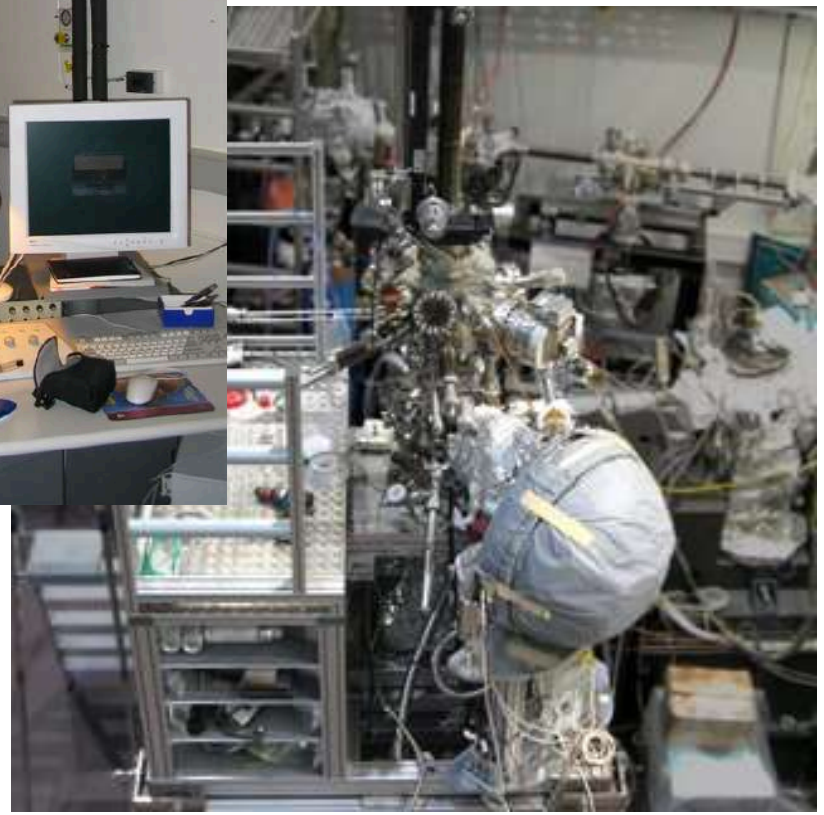
FAIRmat

Reproducibility

Reproducibility in Experiments and Synthesis

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detectors
  - Data Processing
  - *Notes*
- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*



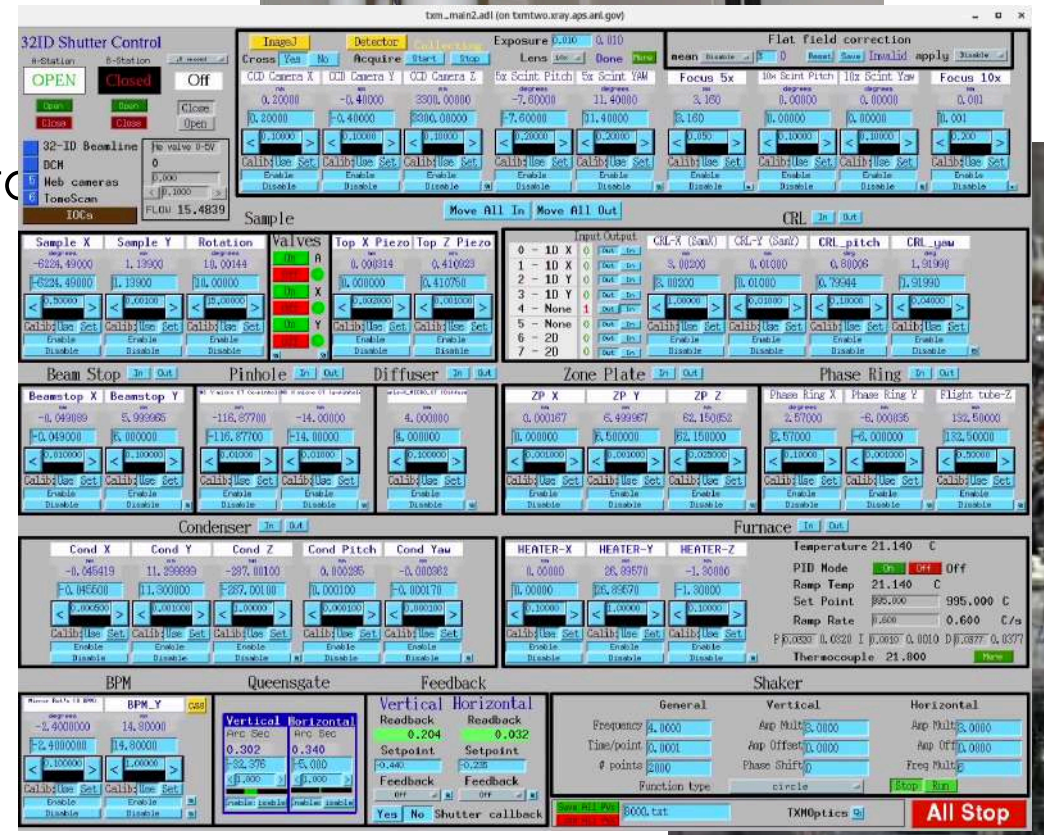


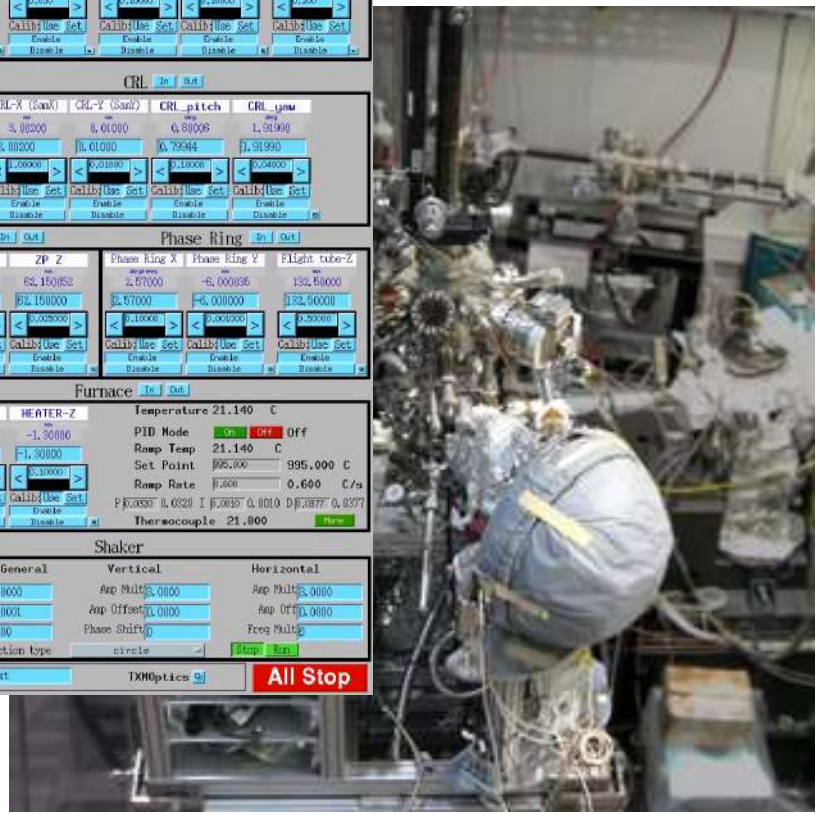
FAIRmat

Reproducibility in Experiments and Synthesis

Reproducibility

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detecto
  - Data Processing
  - *Notes*





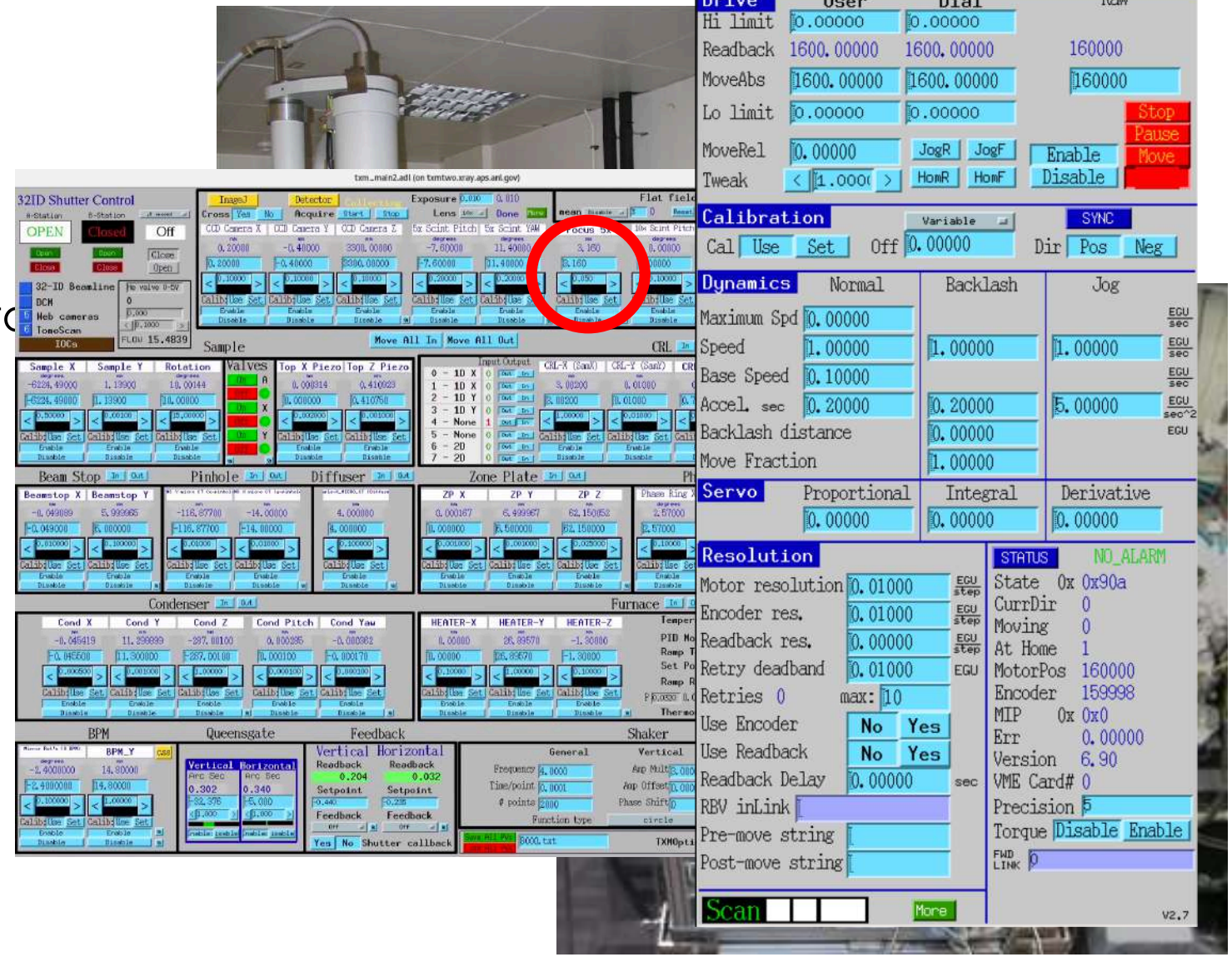
- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

FAIRmat

Reproducibility

Reproducibility in Experiments and Synthesis

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detecto
  - Data Processing
  - *Notes*
- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

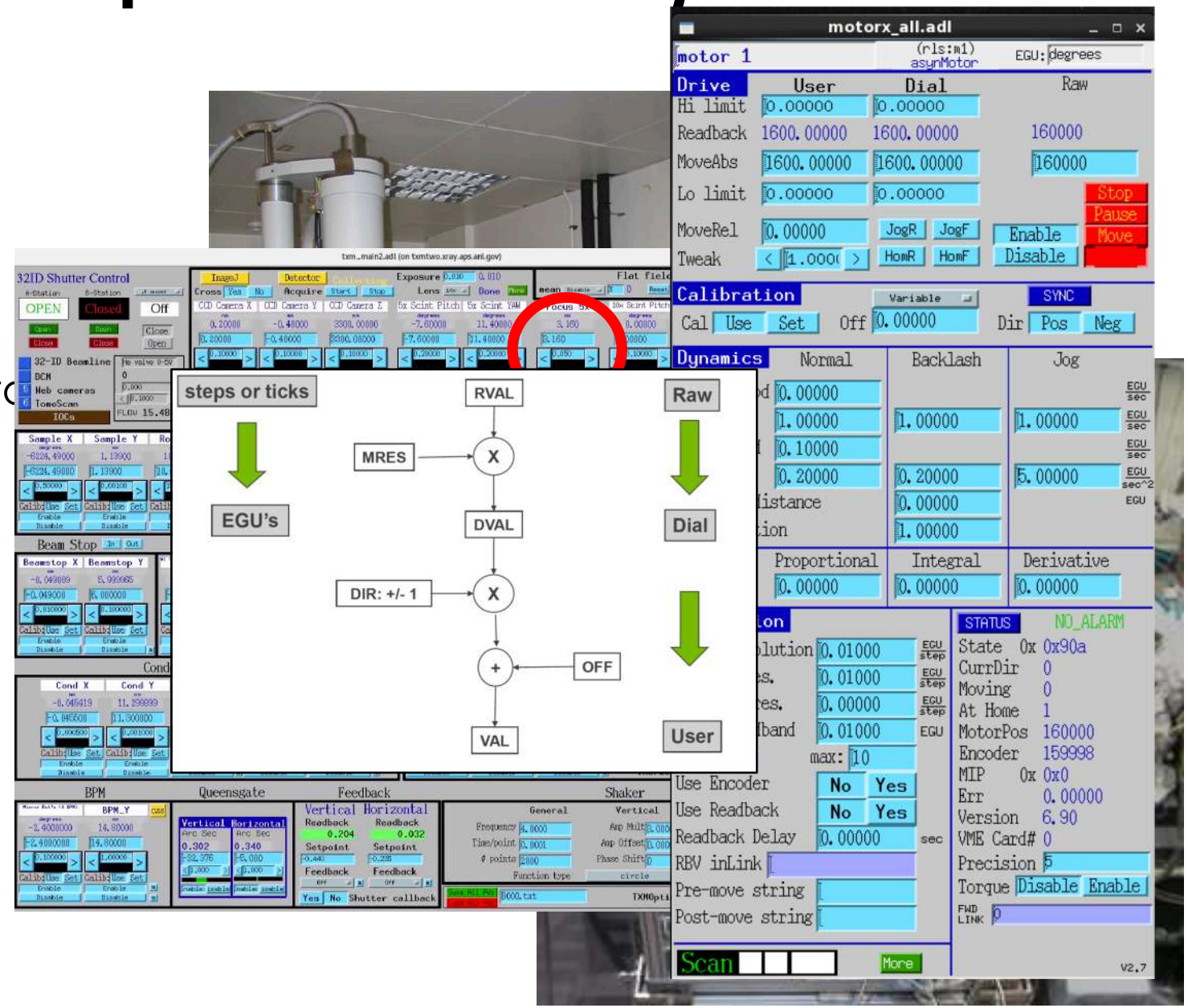


FAIRmat

Reproducibility in Experiments and Synthesis

Reproducibility

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detecto
  - Data Processing
  - *Notes*
- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

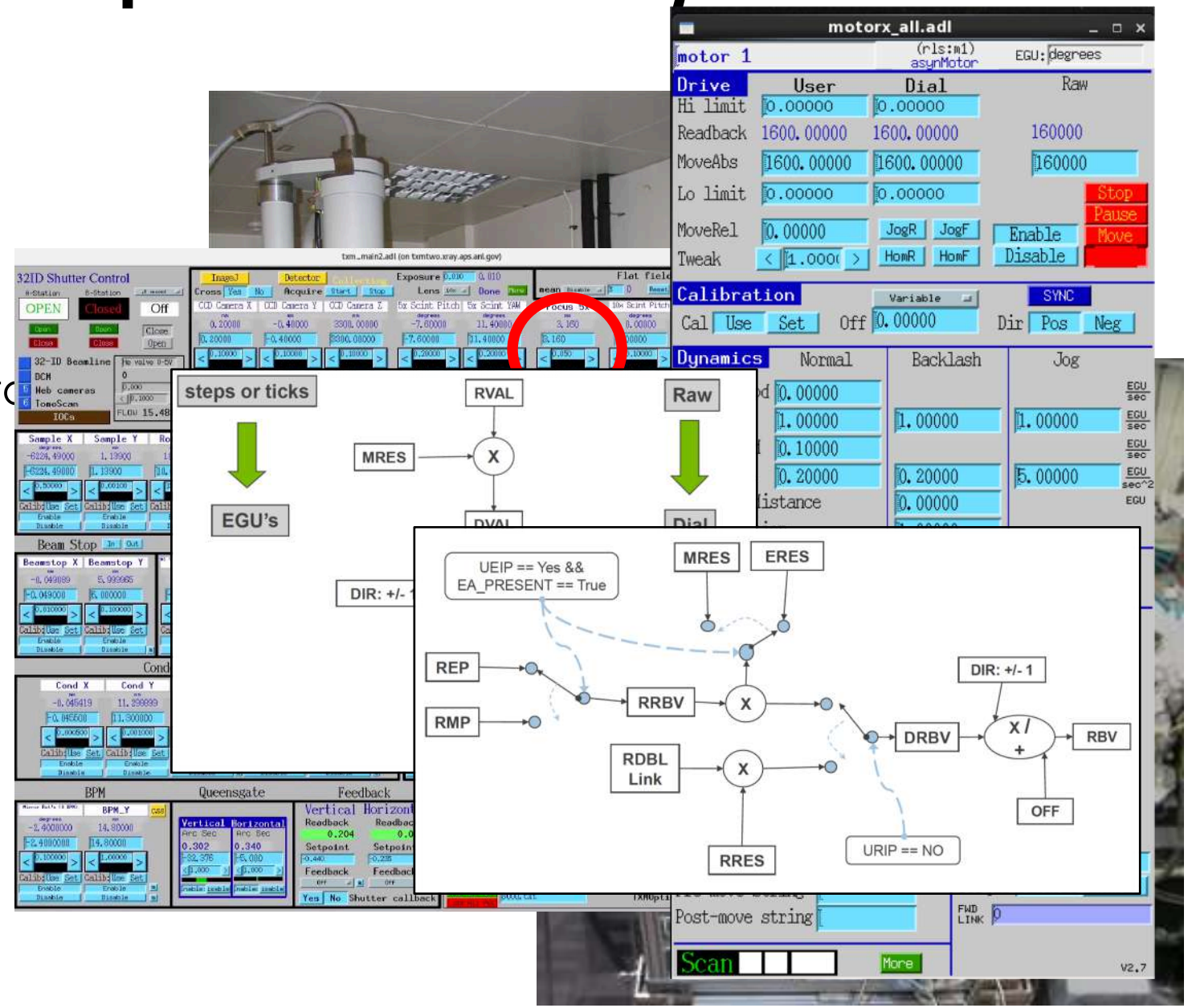


FAIRmat

Reproducibility

Reproducibility in Experiments and Synthesis

- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detecto
  - Data Processing
  - *Notes*
- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*



FAIRmat

Reproducibility in Experiments and Synthesis

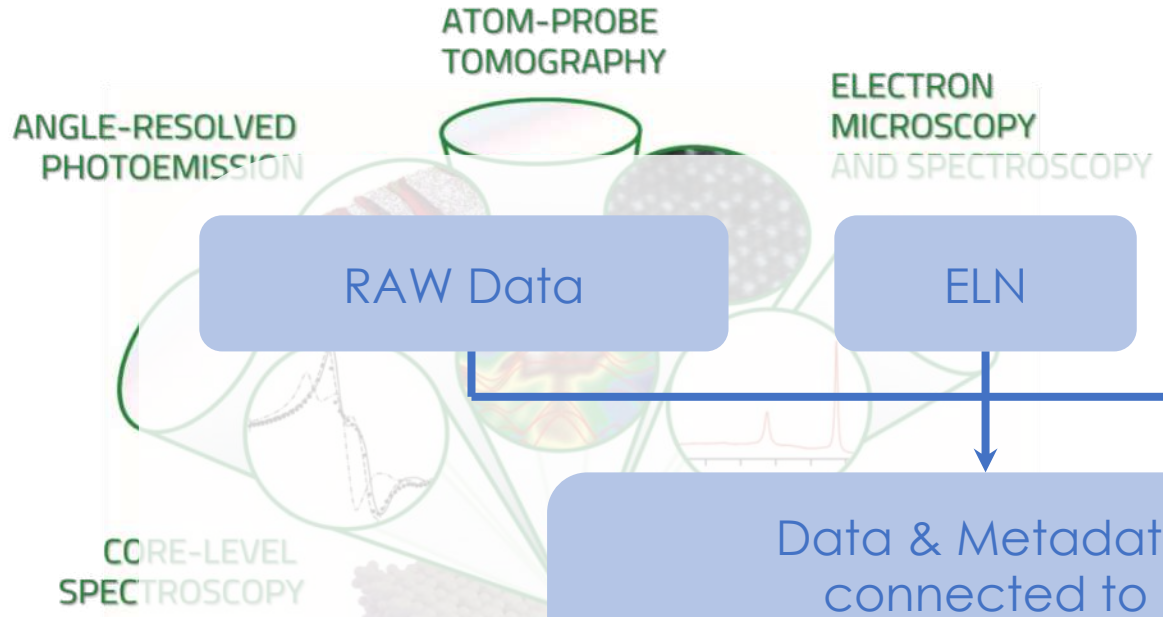
Reproducibility

- **Synthesis**
  - *Sample* History
  - *Processes*
  - *Characterisation*
  - *Notes*

FAIRmat

Data Models for Experiments

Techniques



RAW Data

ELN

Data Analysis

Data & Metadata
connected to
Standardised Data Model

NOMAD
Search / Visualisation / Analysis

Ellipsometry

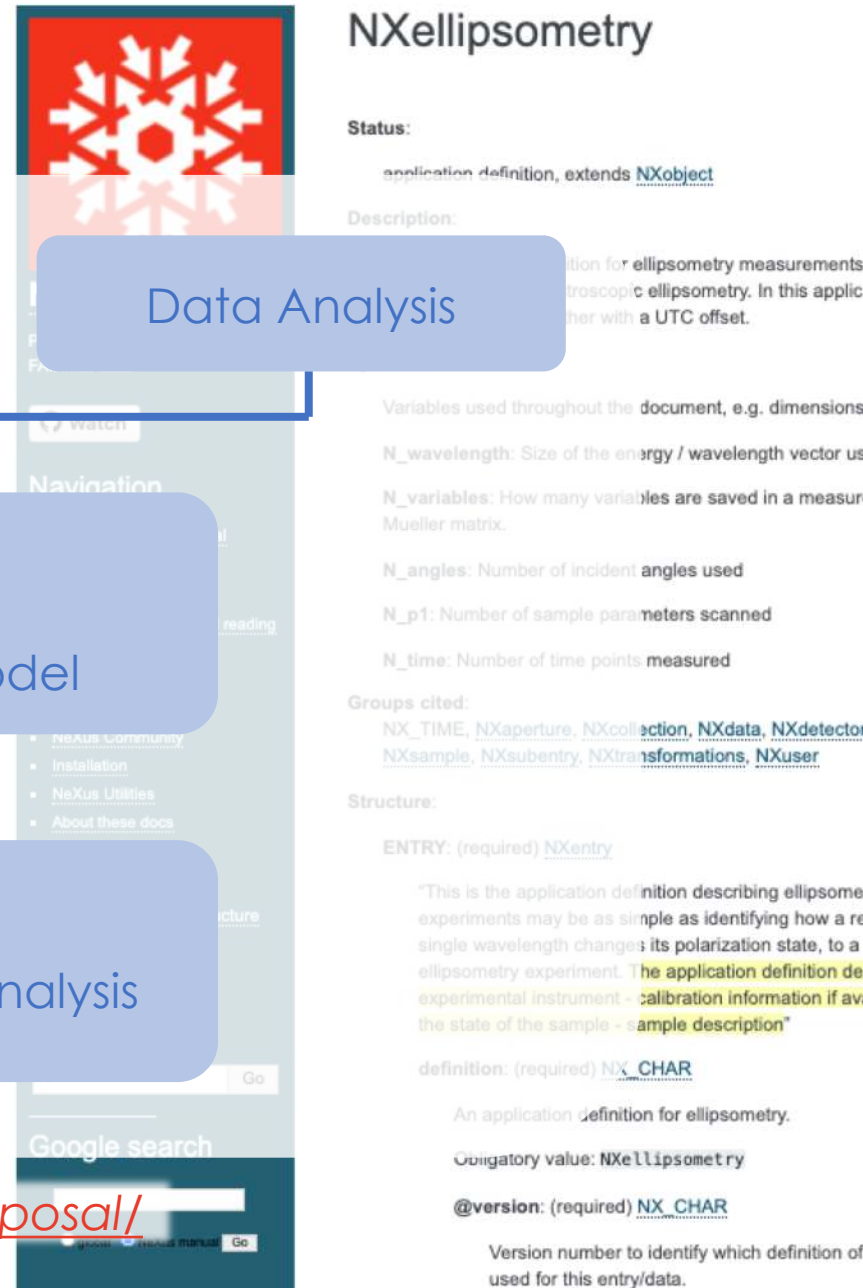
Electron Microscopy

Atom Probe Microscopy

Photoemission Spectroscopy

Temperature dependent I-V measurements

https://fairmat-experimental.github.io/nexus-fairmat-proposal/



FAIRmat

Techniques

Characterisation Methods in FAIRmat

Area B:

- Optical Spectroscopy, EM, AREPS, XPS, Atomprobe Microscopy

Area A:

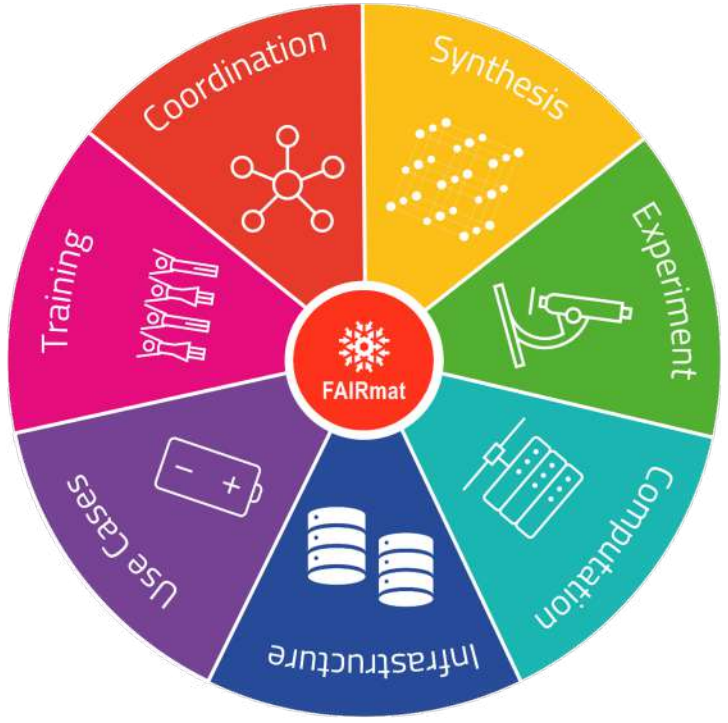
- Magnetic Resonance, Magnetometry, Scattering and Imaging, Electrical and Mechanical methods
- Synthesis processes and Sample description

Area C connection:

- Theory models and Simulation results

Area D Support

- Integration at Control System level
- Infrastructure for Data Management

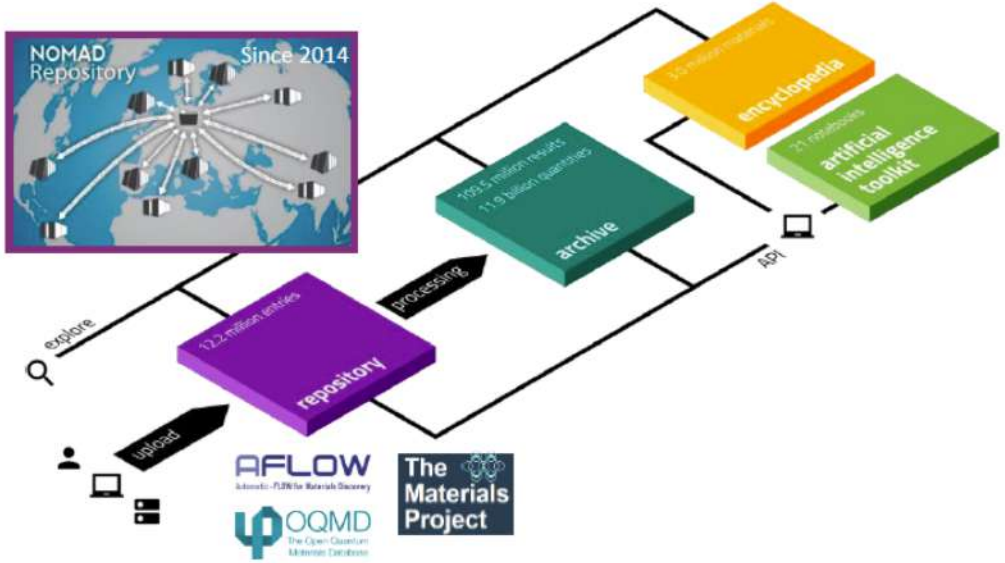


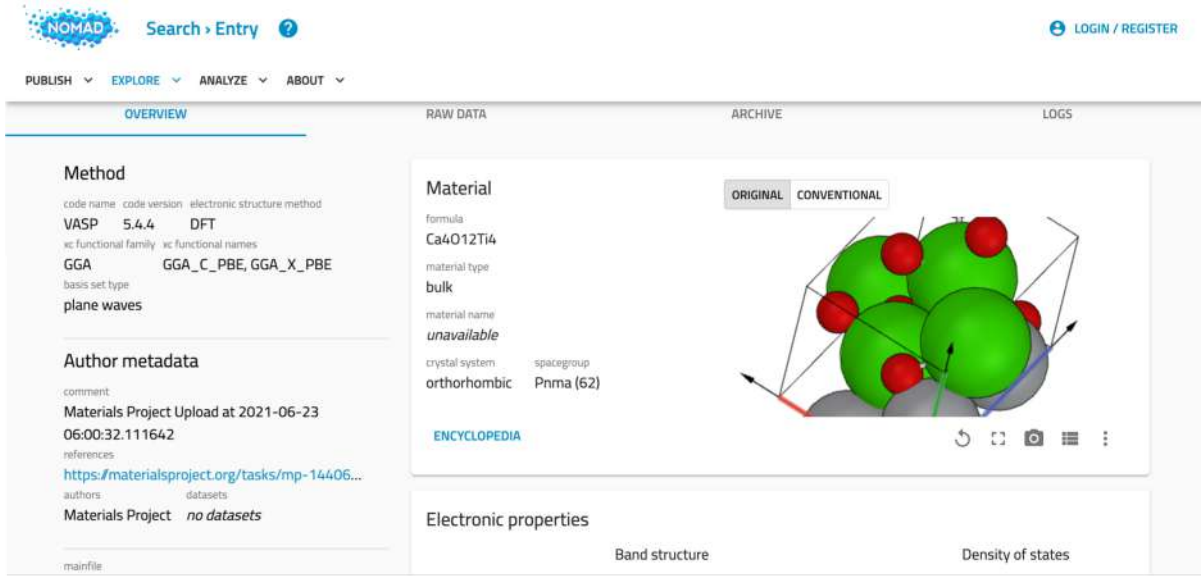
FAIRmat

Data Management

FAIR Theory and Computations

Biggest database in Materials Science *(**https://nomad-lab.eu**)*
more than 100 million high-quality calculations



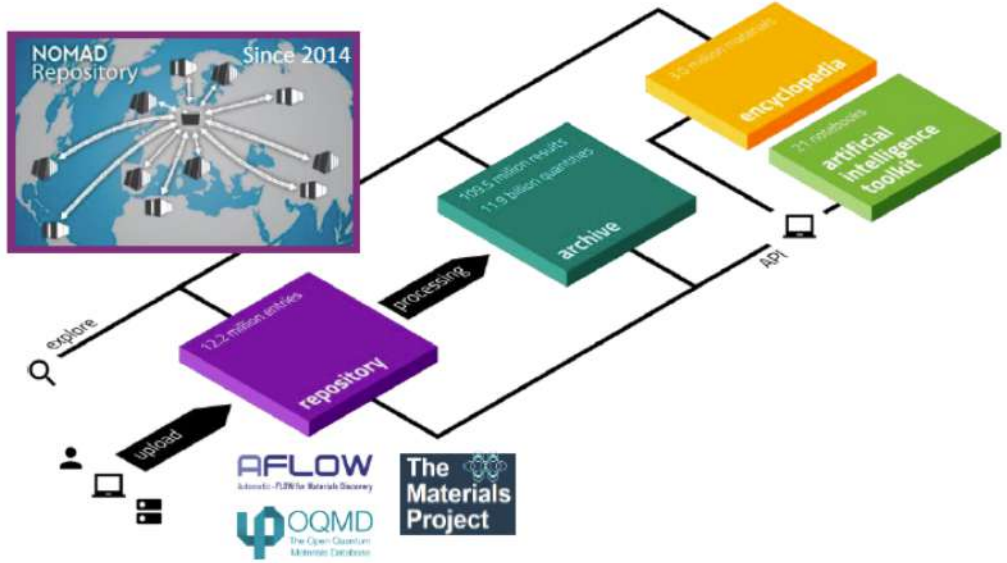


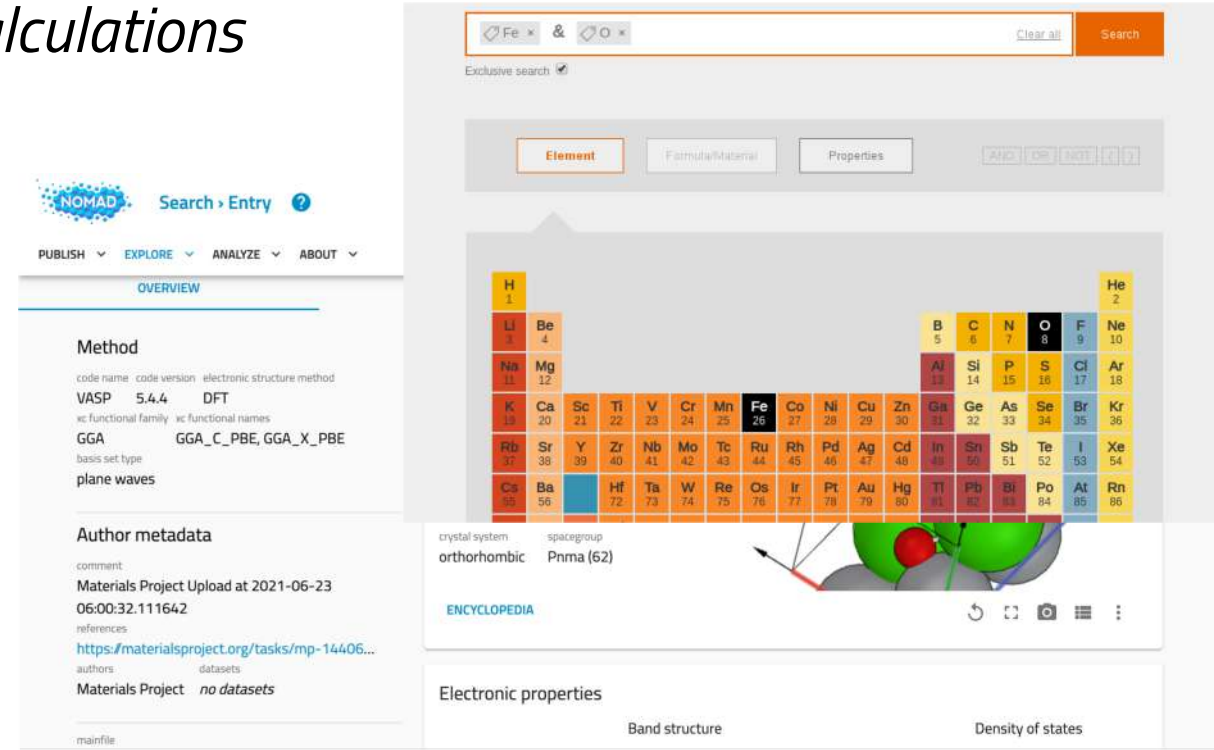
FAIRmat

Data Management

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations



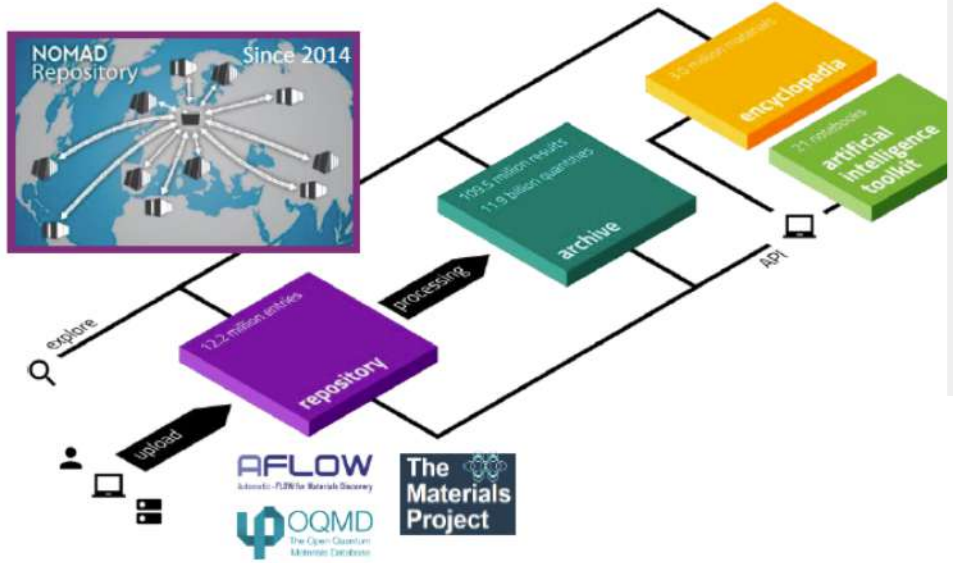


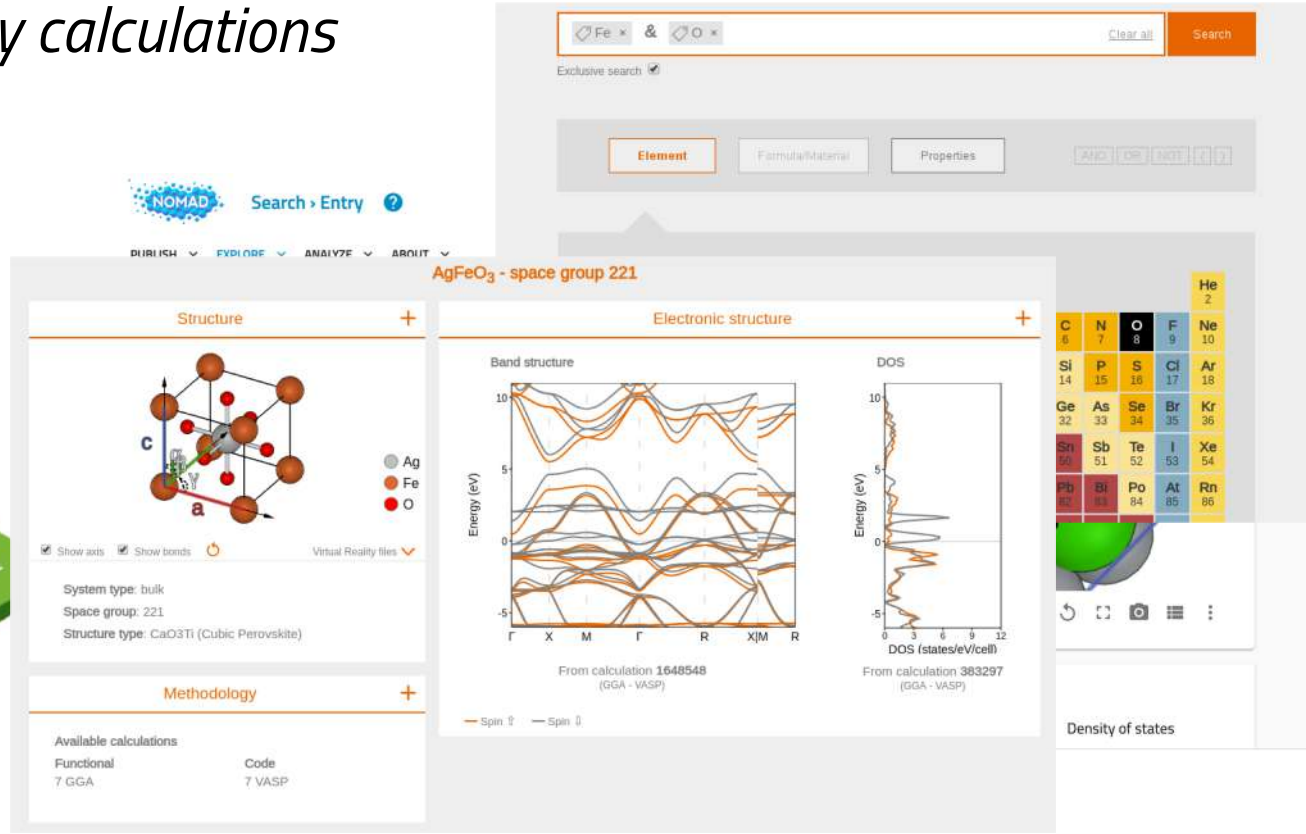
FAIRmat

Data Management

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations



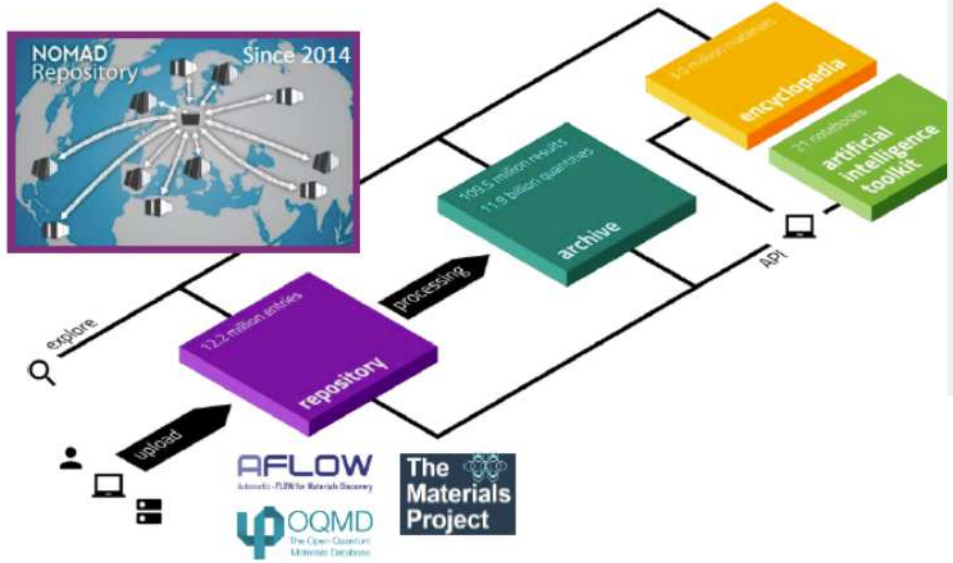


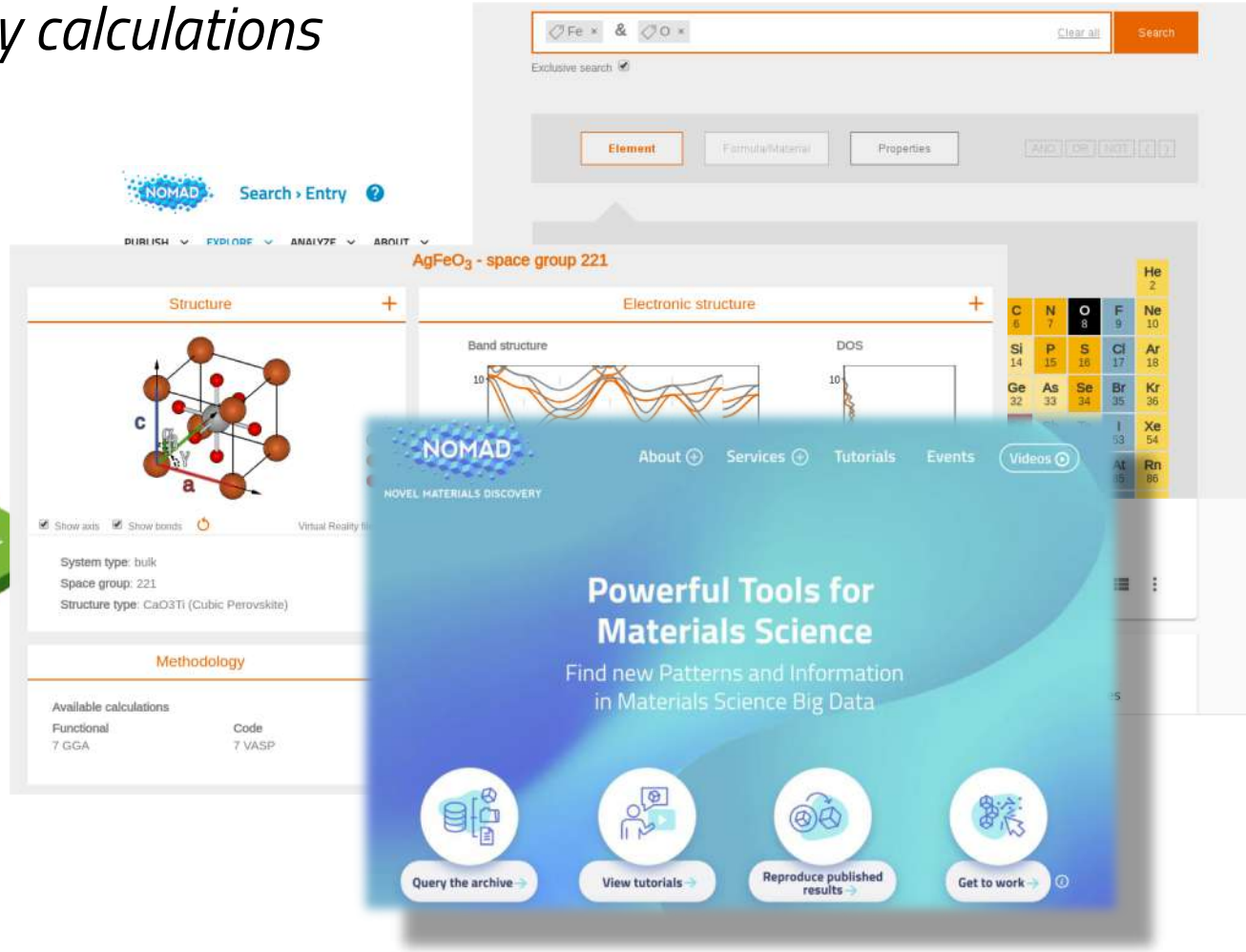
FAIRmat

Data Management

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations





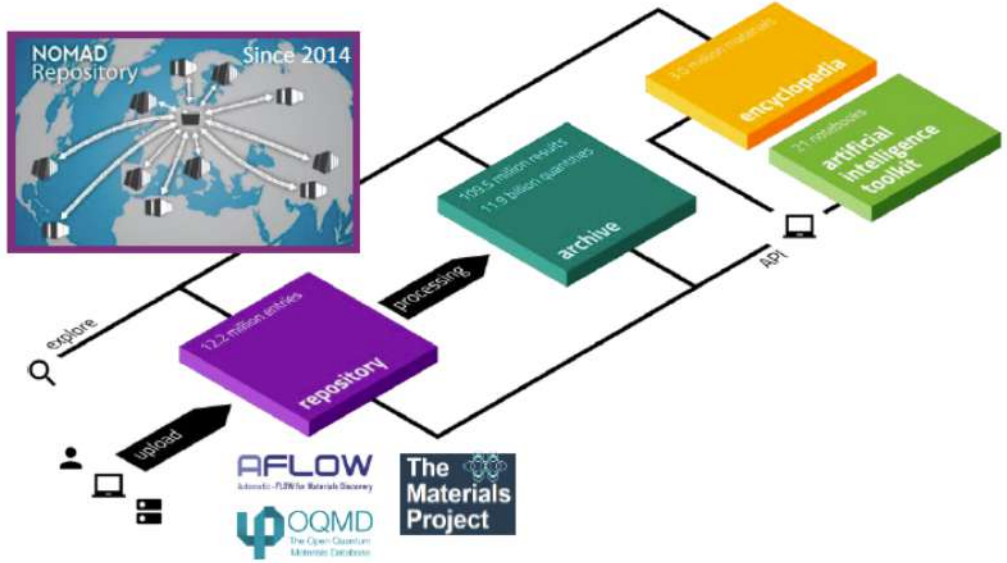
FAIRmat

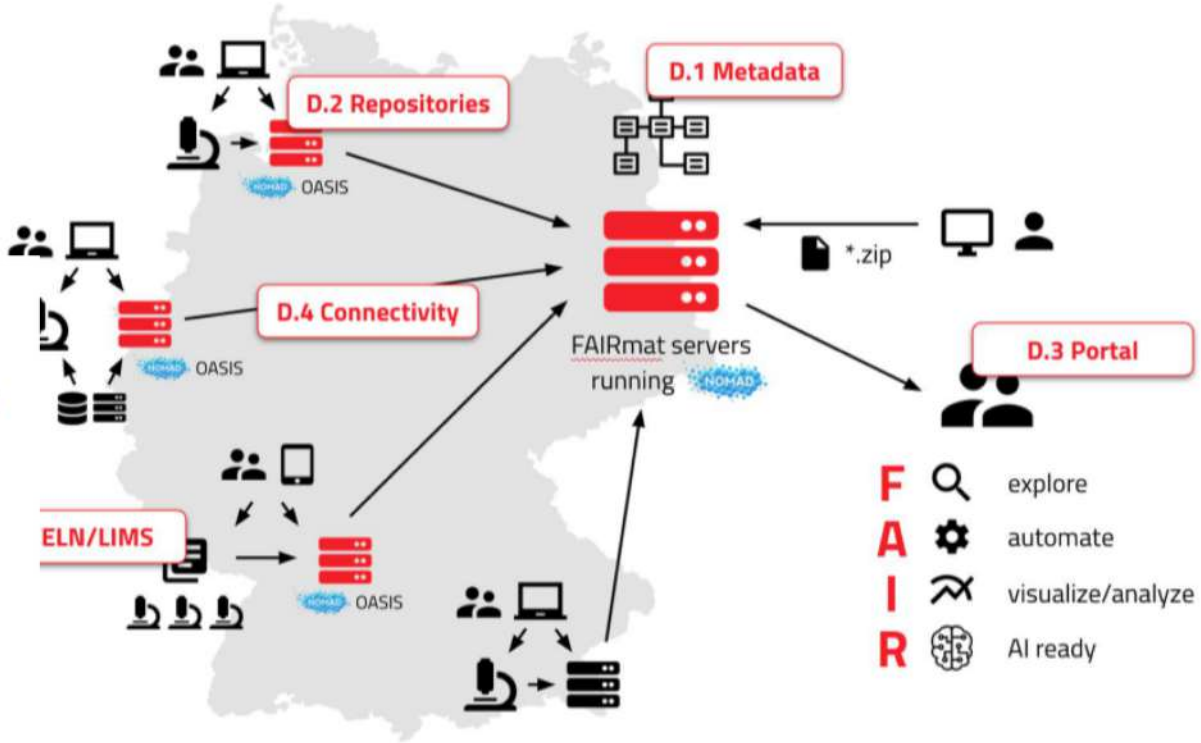
Data Management

FAIRmat

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations



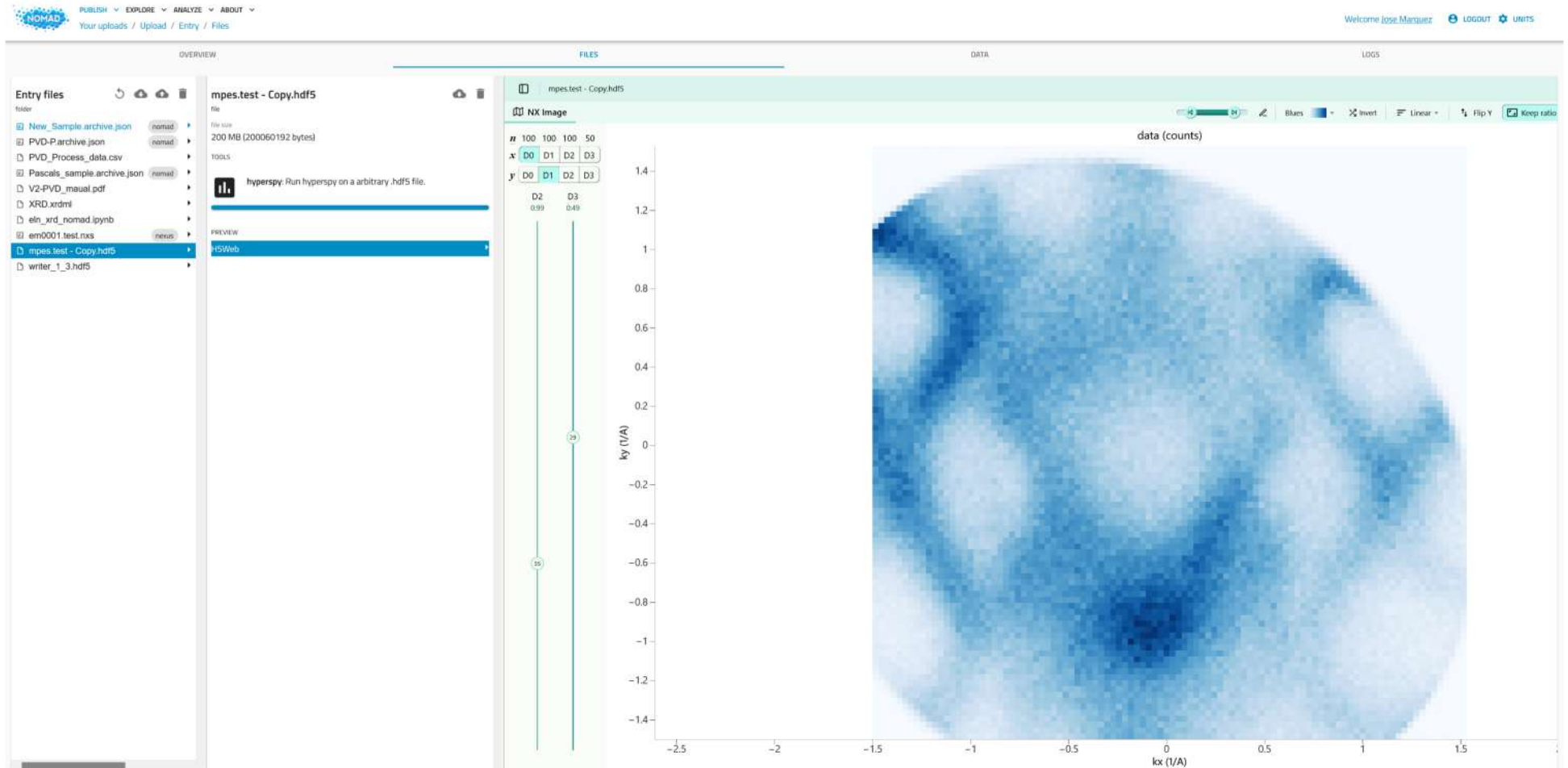


Data Management

FAIRmat

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations

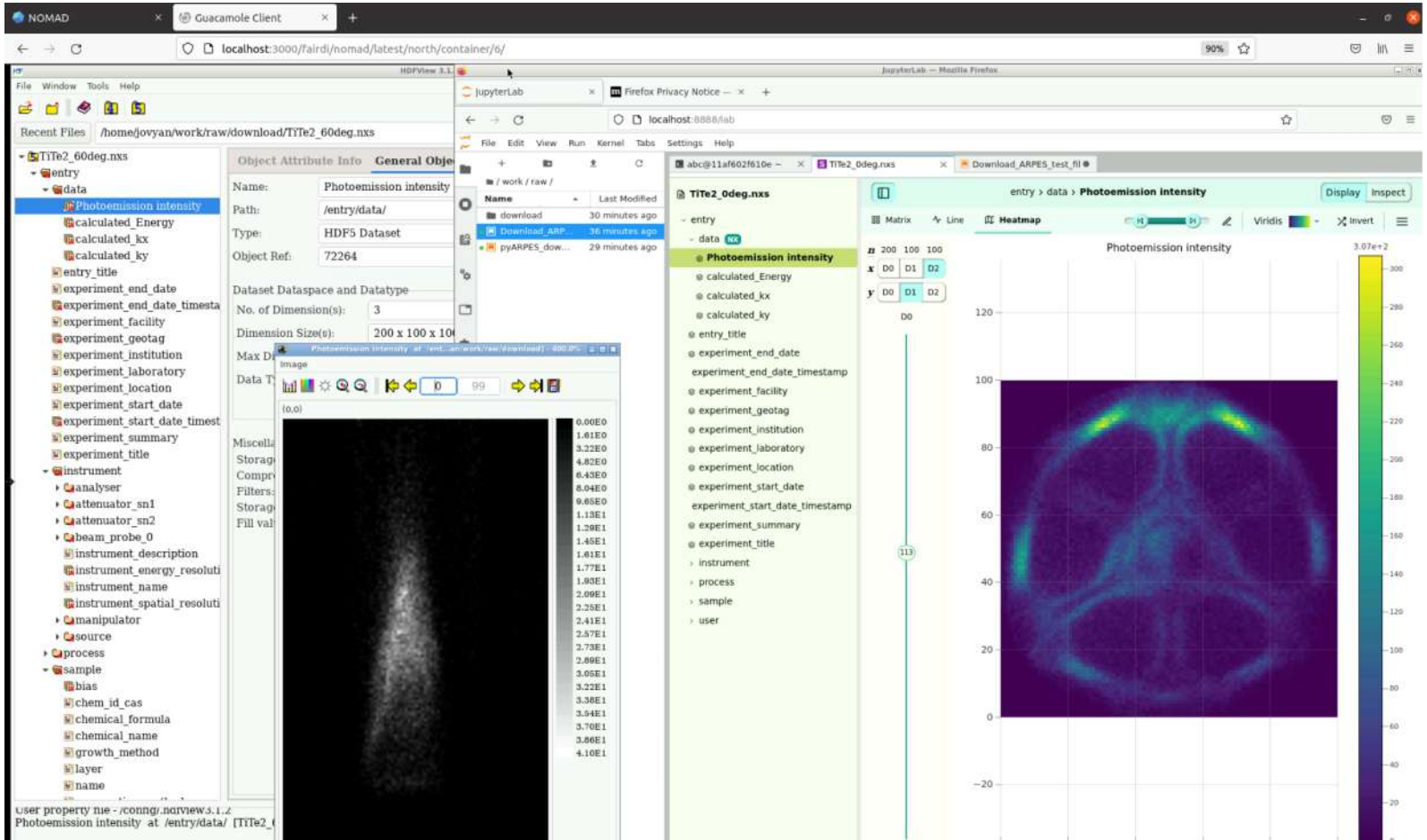


Data Management

FAIRmat

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations

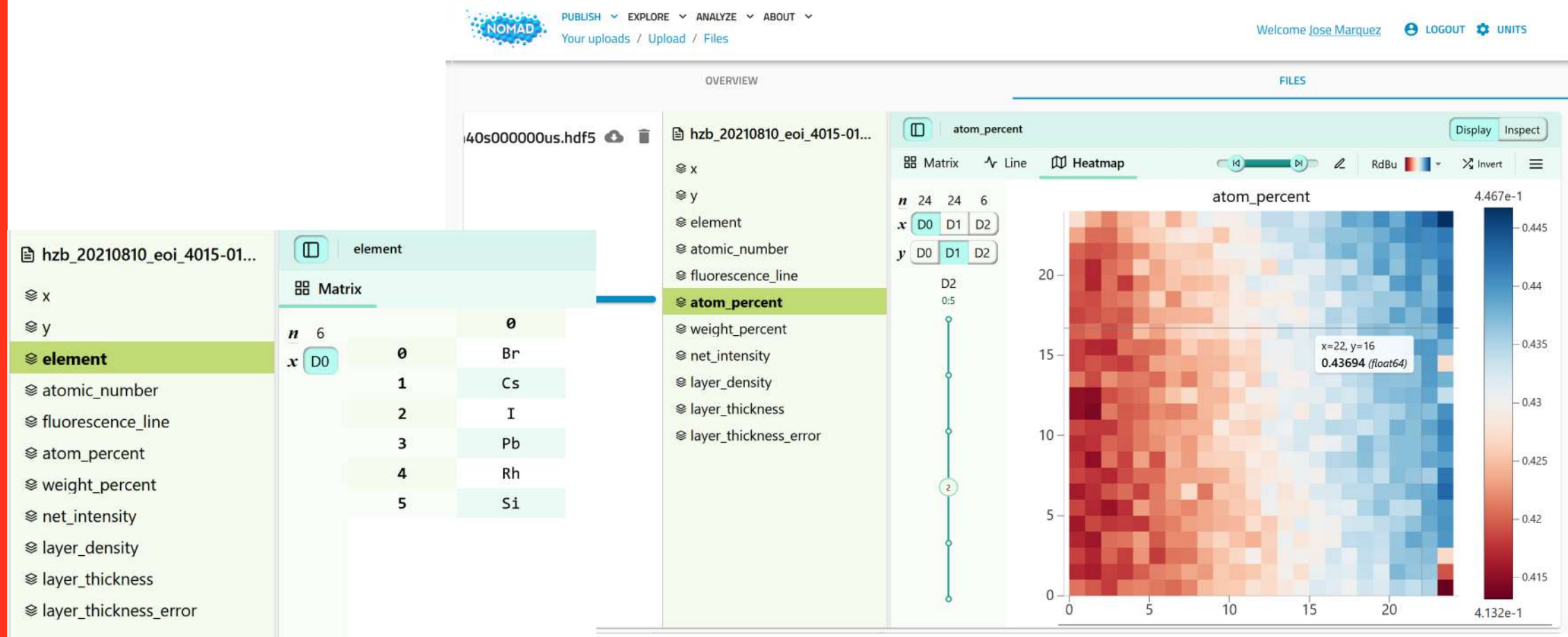


Data Management

FAIRmat

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations

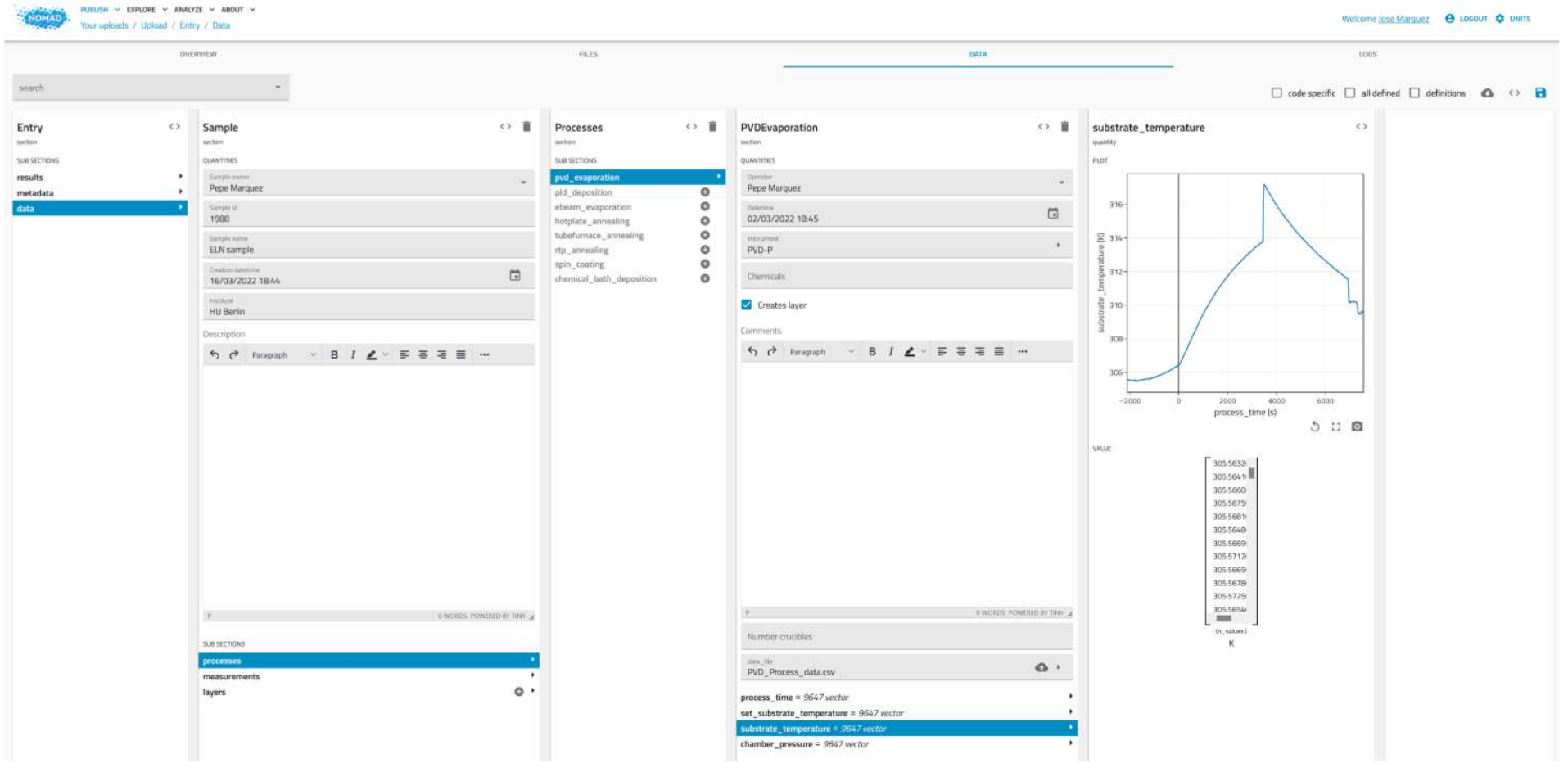


Data Management

FAIRmat

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations

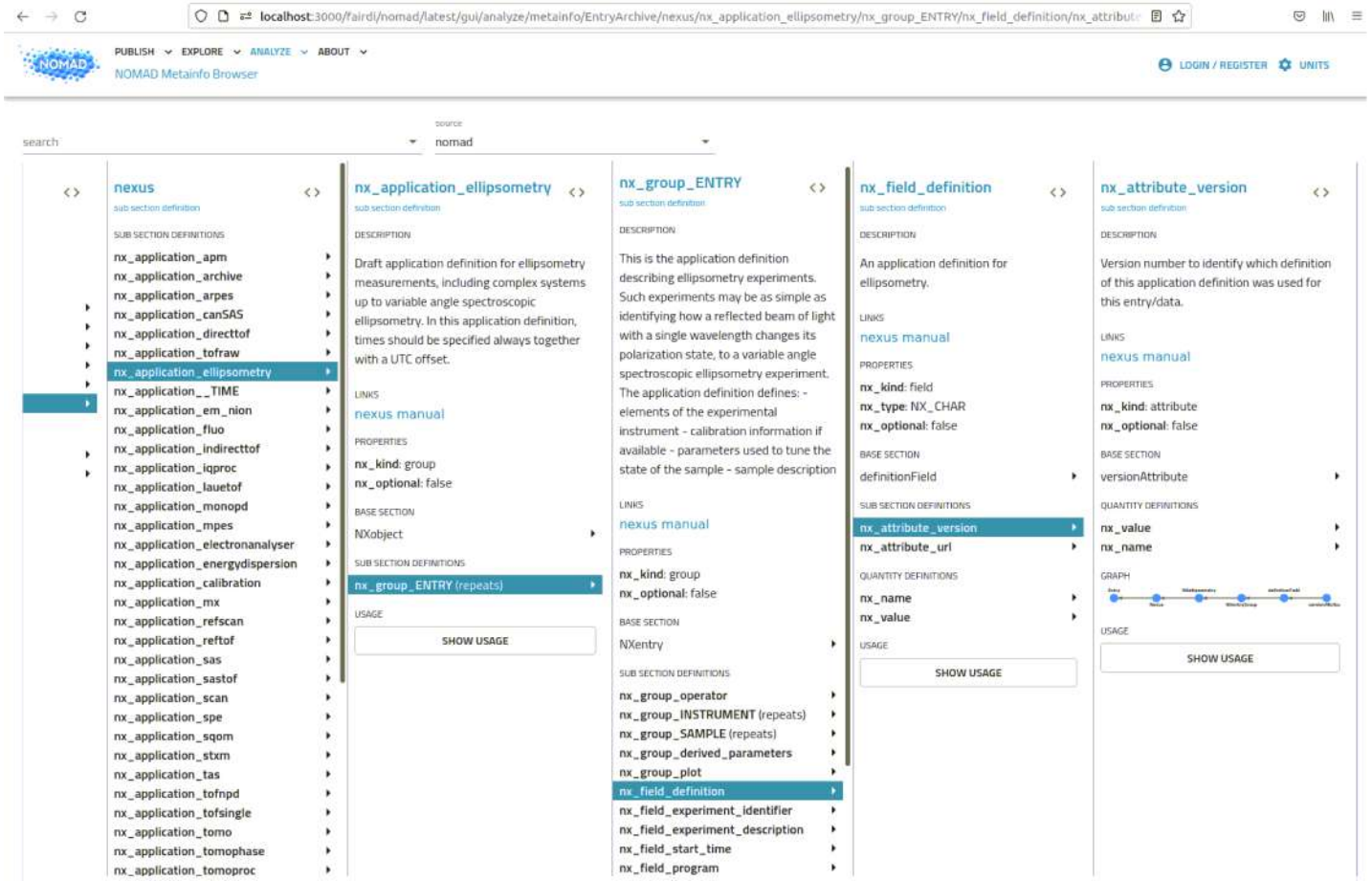


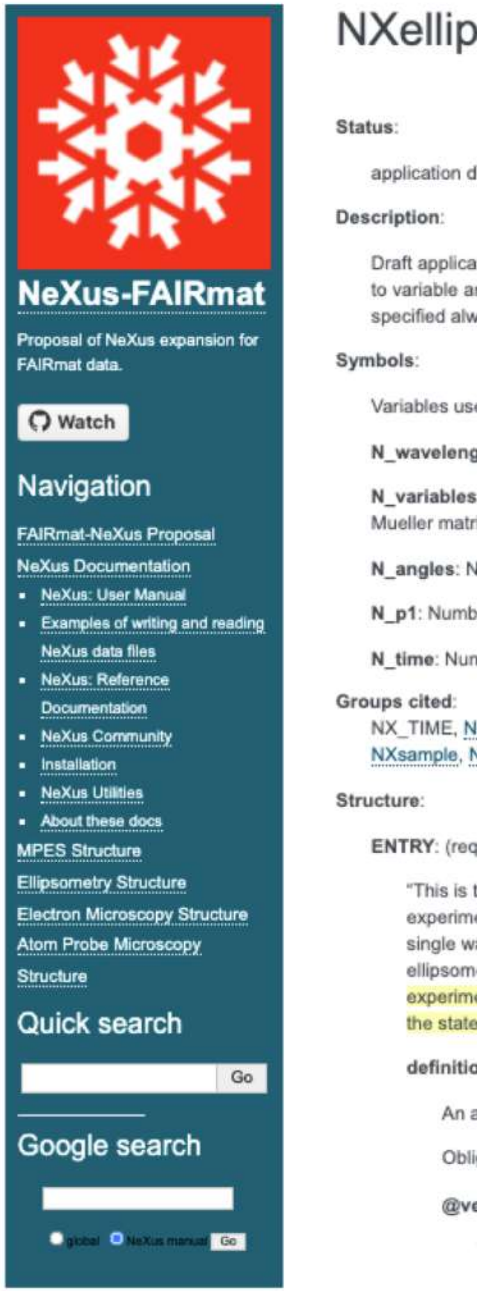
Data Management

FAIRmat

FAIR Theory and Computations

Biggest database in Materials Science ***(https://nomad-lab.eu)***
more than 100 million high-quality calculations





RAMAN

Application Definition

- NXuser
- NXinstrument
  - NXexcitation
  - NXdetection
- NXsample
  - Wavelength_detection[detection]
  - Measured_data[time,parameter,excitation,detection]
  - Timepoints[time]
  - Angle_of_incidence[angles]
  - Sample_stage
  - ...
- plot

FAIRmat


```
doc: "Draft application definition for general optic
symbols:
  doc: Variables used throughout the document, e
  N_wavelength_detection: Size of the energy / v
  N_wavelength_excitation: "Size of the array of
                  excitation in the case of photolumi
  N_parameters: Number of sample parameters s
  N_time: Number of time points measured

(NXopt_experiment):
  (NXentry):
    doc: |
      This is an application definition describing gen

      A general optical experiment consists of a ligh
      beam path, a sample + stage + environment, a
      are reflection or transmission measurements,
      spectroscopy etc.

      This application definition defines:

        * information on excitation and detection
        * description of the beam path
        * sample description
        * N-dimensional data field

    definition:
```