Towards structured data in electron spectroscopy

Heiko B. Weber and the FAIRmat team

Friedrich-Alexander-Universität Erlangen-Nürnberg

Outline

- The FAIRmat approach for making experimental data FAIR
- Application definitions and NEXUS
- MPES example

FAIRmat

RDM

Open data

Open access (~2001) to peer reviewed literature, being enforced by funding agencies

FAIR principles[1] (2016); statement at G20 summit in China[2]

Data as resource[3]: A European strategy for data (2020)

Expectations: DFG and other funding agencies expect FAIR data from us researchers

(funding is connected to FAIR data pledges)

FAIRmat

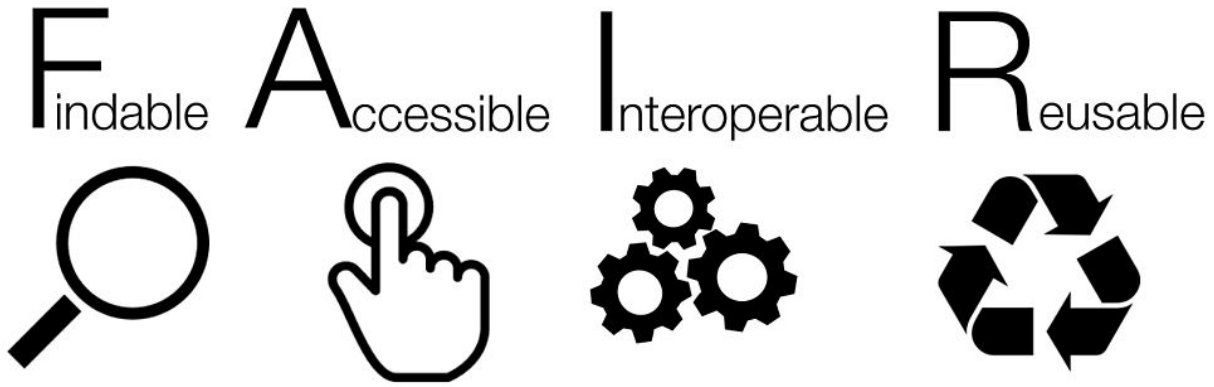
[1] Scientific Data 3, 160018 (2016)

[2] https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967

[3] https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_en.pdf

RDM

Underlying principles for open research data



Findable and **A**rtificial **I**ntelligence **R**eady

FAIRmat

RDM

FAIR data: benefit or burden?

Rich opportunities

- Big data opportunities in science
- Data analysis as novel discipline (besides experiment and theory)
- Accelerated materials discovery
- Enhanced data competence of scientists
- Reduction of misleading interpretations and fraud

Significant investment

- The transition to FAIR data demands for reorganizing scientists‘ workflow entirely

FAIRmat

RDM

Nationale Forschungsdaten-Infrastruktur (NFDI)

(polycentered bottom-up network)

1st round (10/2020)

- DataPLANT: Plant research data
- GHGA: German Human Genome–Phenome Archive
- KonsortSWD: Consortium for the Social, Educational, Behavioural and Economic Sciences
- NFDI4Biodiversity: Biodiversity, Ecology and Environmental Data
- NFDI4Cat: NFDI for sciences related to catalysis
- NFDI4Chem: Chemistry consortium for the NFDI
- NFDI4Culture: Consortium for Research Data on Material and Immaterial Cultural Heritage
- NFDI4Health: NFDI personal health data
- NFDI4Ing: NFDI for Engineering Sciences

2nd round (10/2021)

- BERD@NFDI: NFDI for Business, Economic and Related Data
- **DAPHNE4NFDI: Data from PHoton and Neutron Experiments for NFDI**
- **FAIRmat: FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids**
- MaRDI: Mathematical Research Data Initiative
- NFDI4DataScience: NFDI for Data Science and Artificial Intelligence
- NFDI4Earth: NFDI Consortium Earth System Sciences
- NFDI4Microbiota: NFDI for Microbiota Research
- **NFDI-MatWerk: NFDI for Materials Science and Materials Engineering**
- PUNCH4NFDI: Particles, Universe, NuClei and Hadrons for the NFDI
- Text+: Language and text-based research data infrastructure

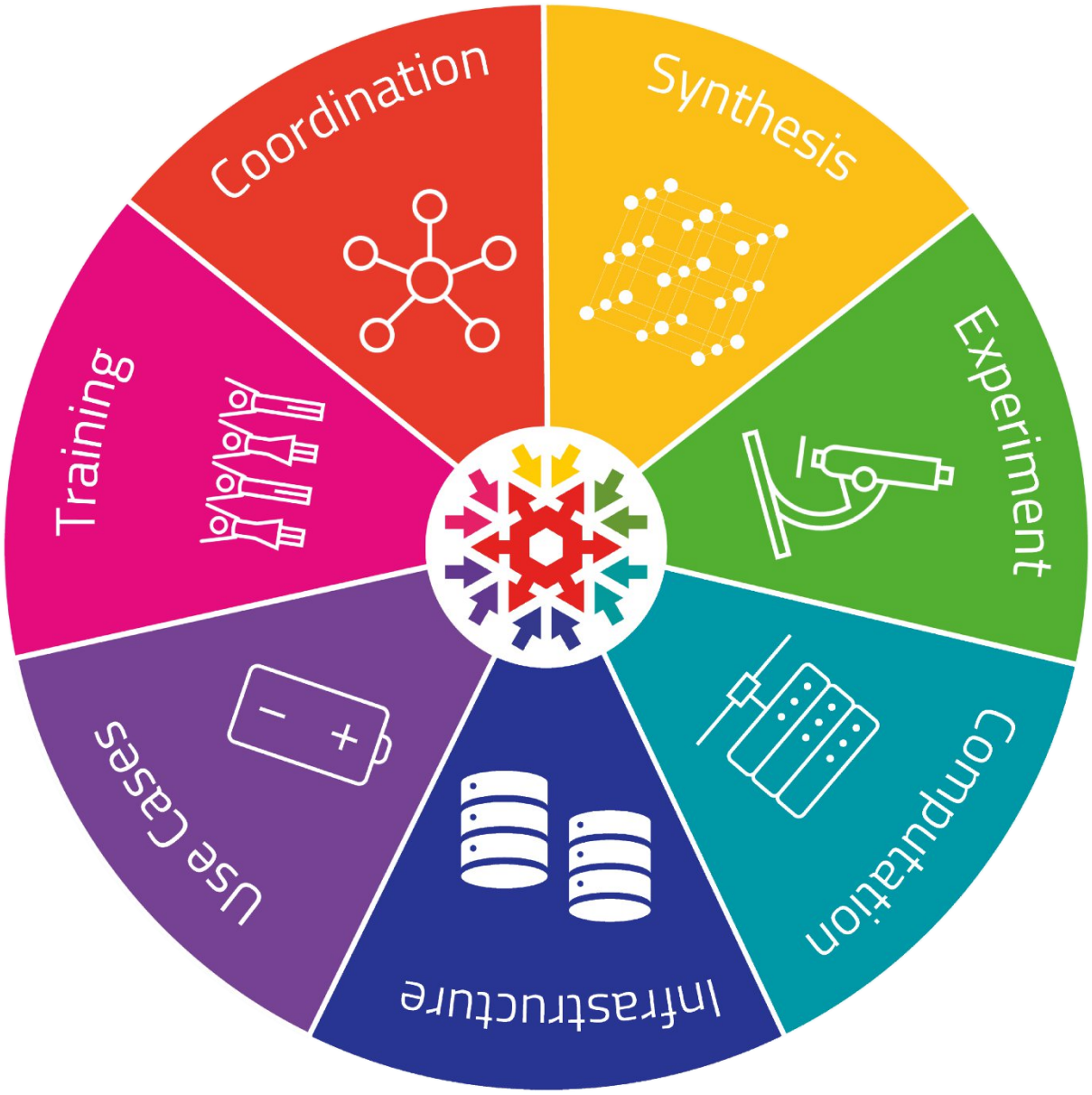
3rd round (3/2023)

- FAIRagro: FAIR Data Infrastructure for Agrosystems
- NFDI4BIOIMAGE: National research data infrastructure for microscopy and bioimage analysis
- NFDI4Energy: National Research Data Infrastructure for Interdisciplinary Energy System Research
- NFDI4Immuno: National Research Data Infrastructure for Immunology
- NFDI4Memory: The Consortium for the Historically Oriented Humanities
- NFDI4Objects: Research Data Infrastructure for the Material Remains of Human History
- NFDIxCS: National Research Data Infrastructure for and with Computer Science

FAIRmat

Experiment

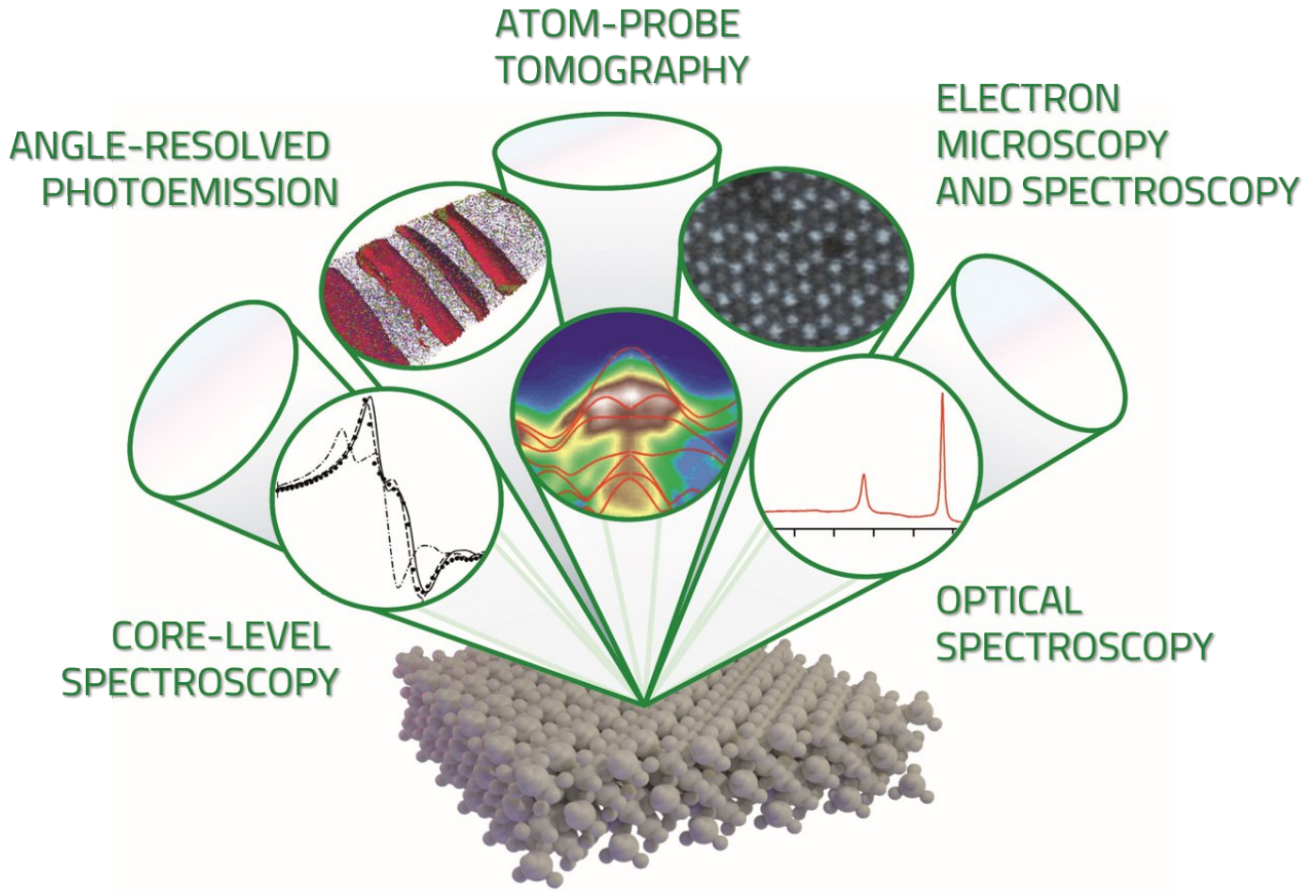
FAIRmat



FAIRmat

Experiment

Area B and its pilot experiments



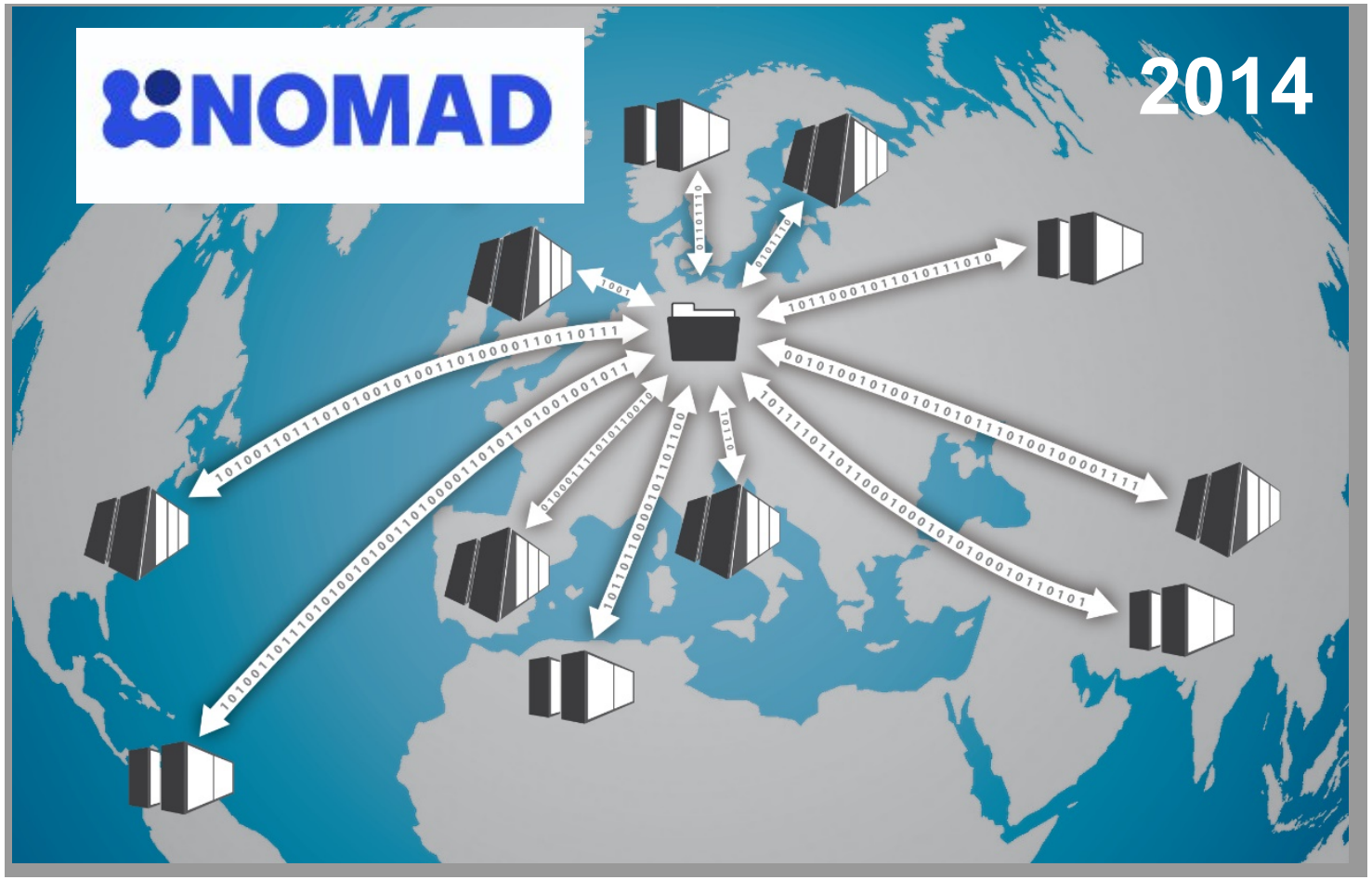
FAIRmat

Forerunners

FAIRmat

Nomad repository

World's largest collection of computational materials science data



Make data repurposable

Recycle the waste!

The Novel Materials Discovery Laboratory

Forerunners

Computational Solid-state science

40 community software packages supported

Thus 40 parsers serve as *converters*

Normalized data in the NOMAD *Archive*

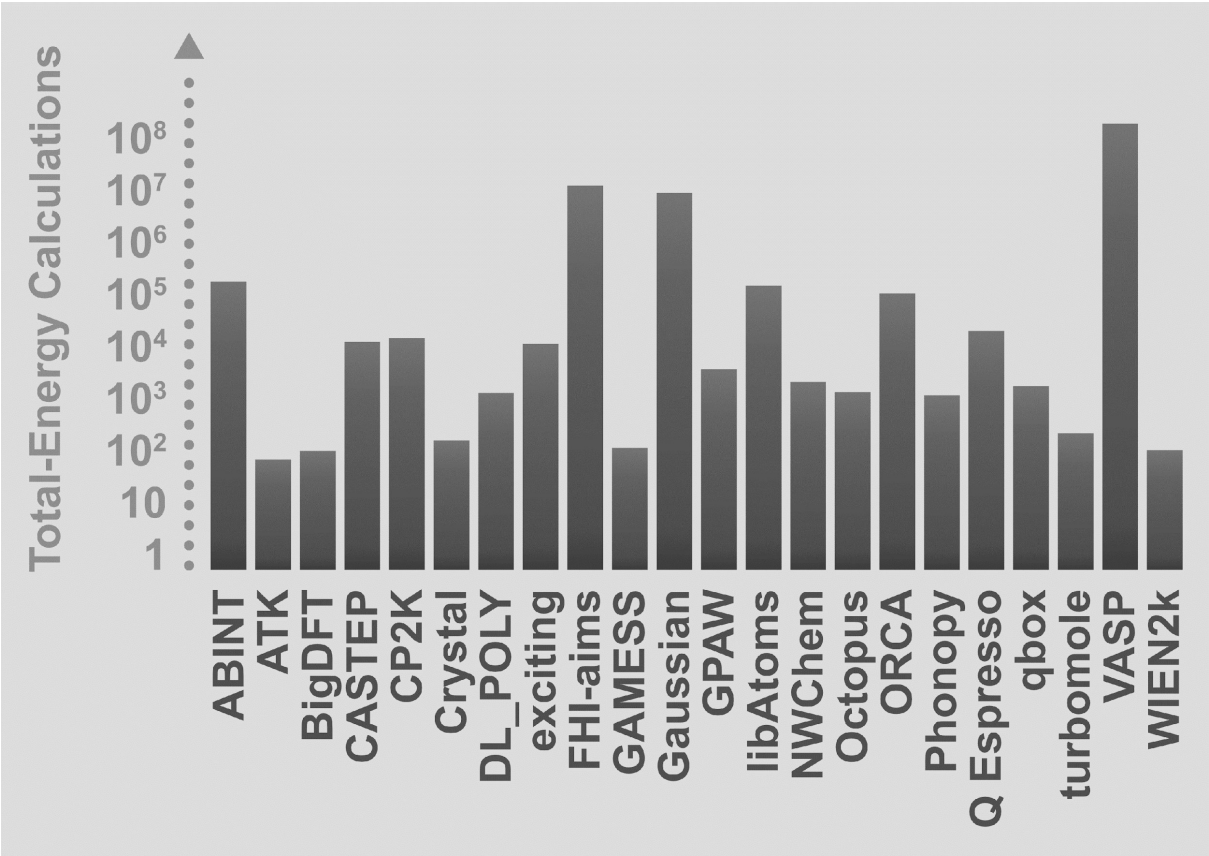
Same units, formats
>100 mio. calculations

billions of CPU core hours
at HPC centers, worldwide

Metadata items

several 1000

to uniquely describe all
the computed data

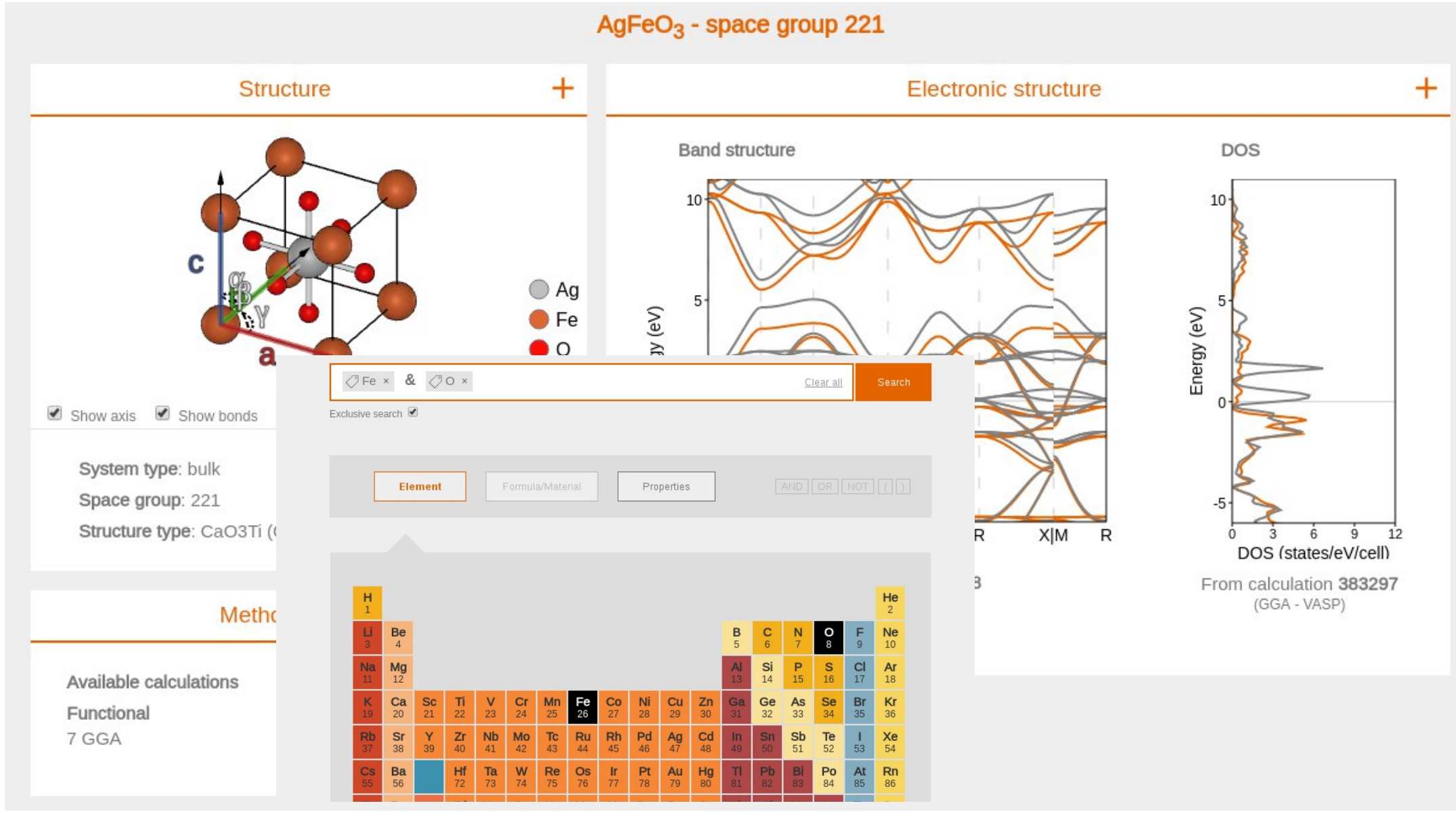


FAIRmat

Forerunners

FAIRmat

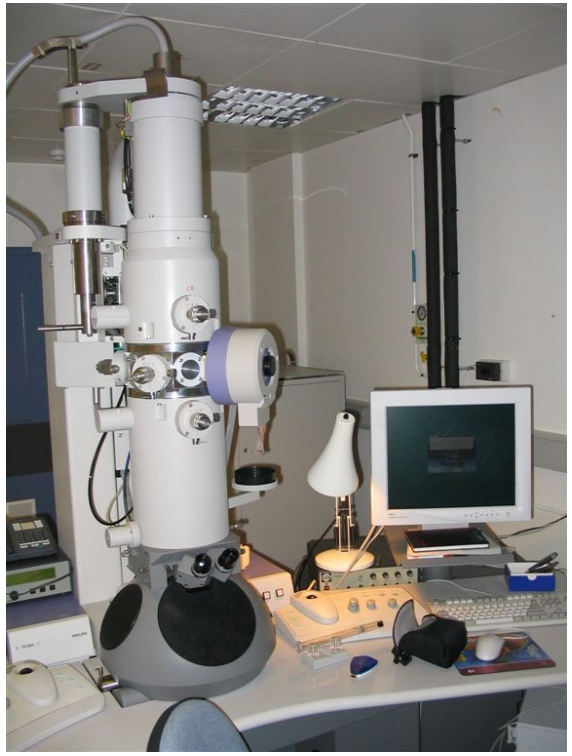
Added value – the NOMAD Encyclopedia

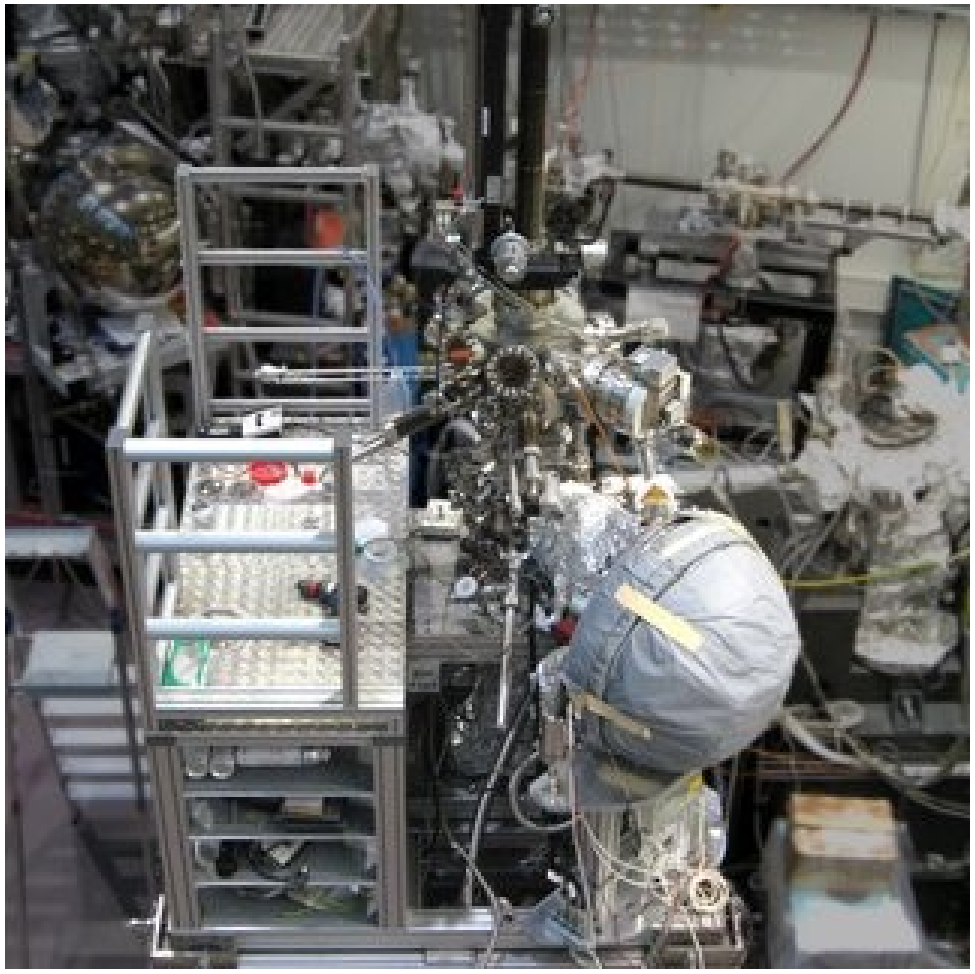


Experiment

Experiments

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





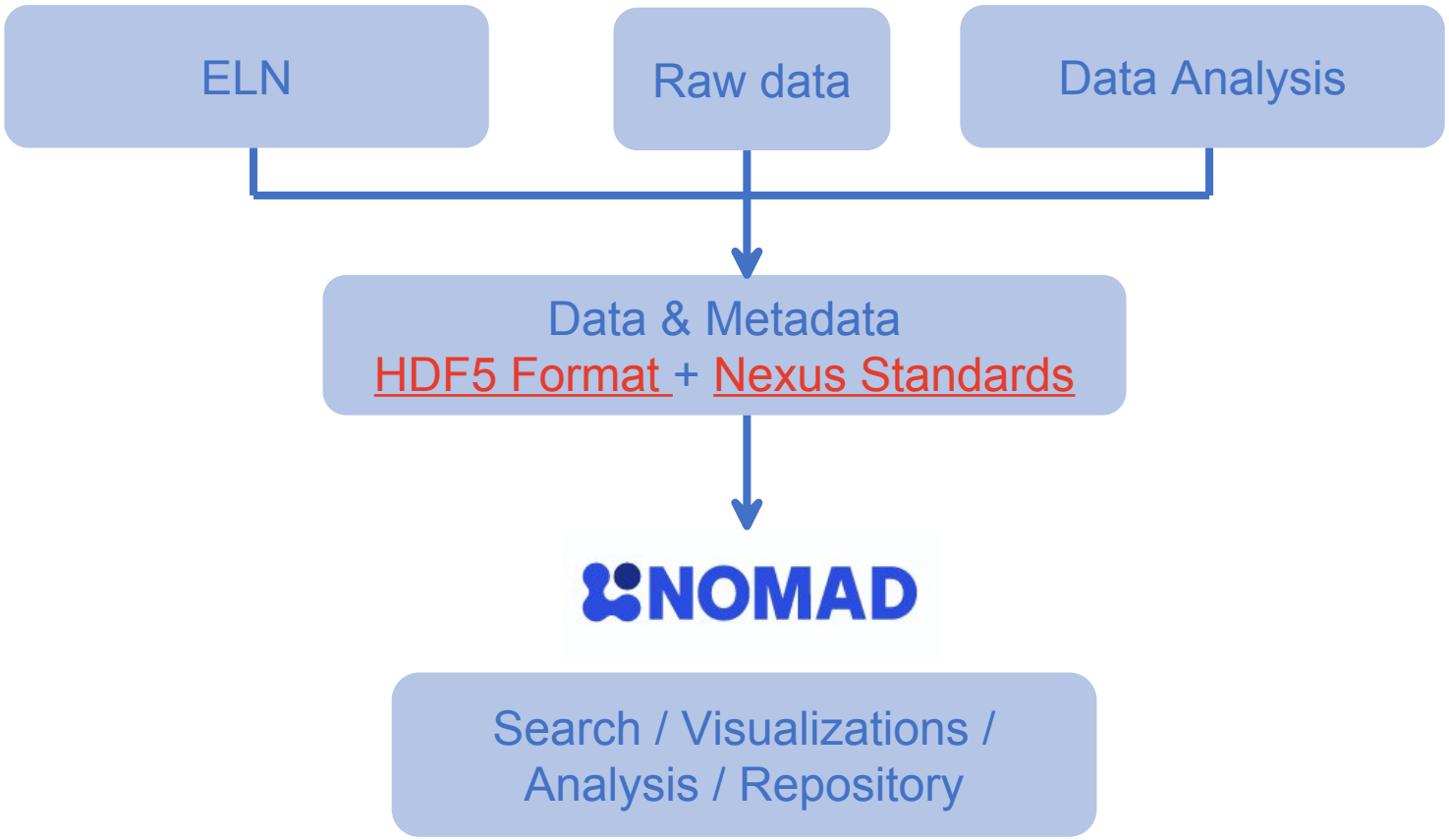
In comparison to large-scale experiments in physics,

solid-state physics is extremely heterogeneous.

FAIRmat

FAIRmat: experimental research data

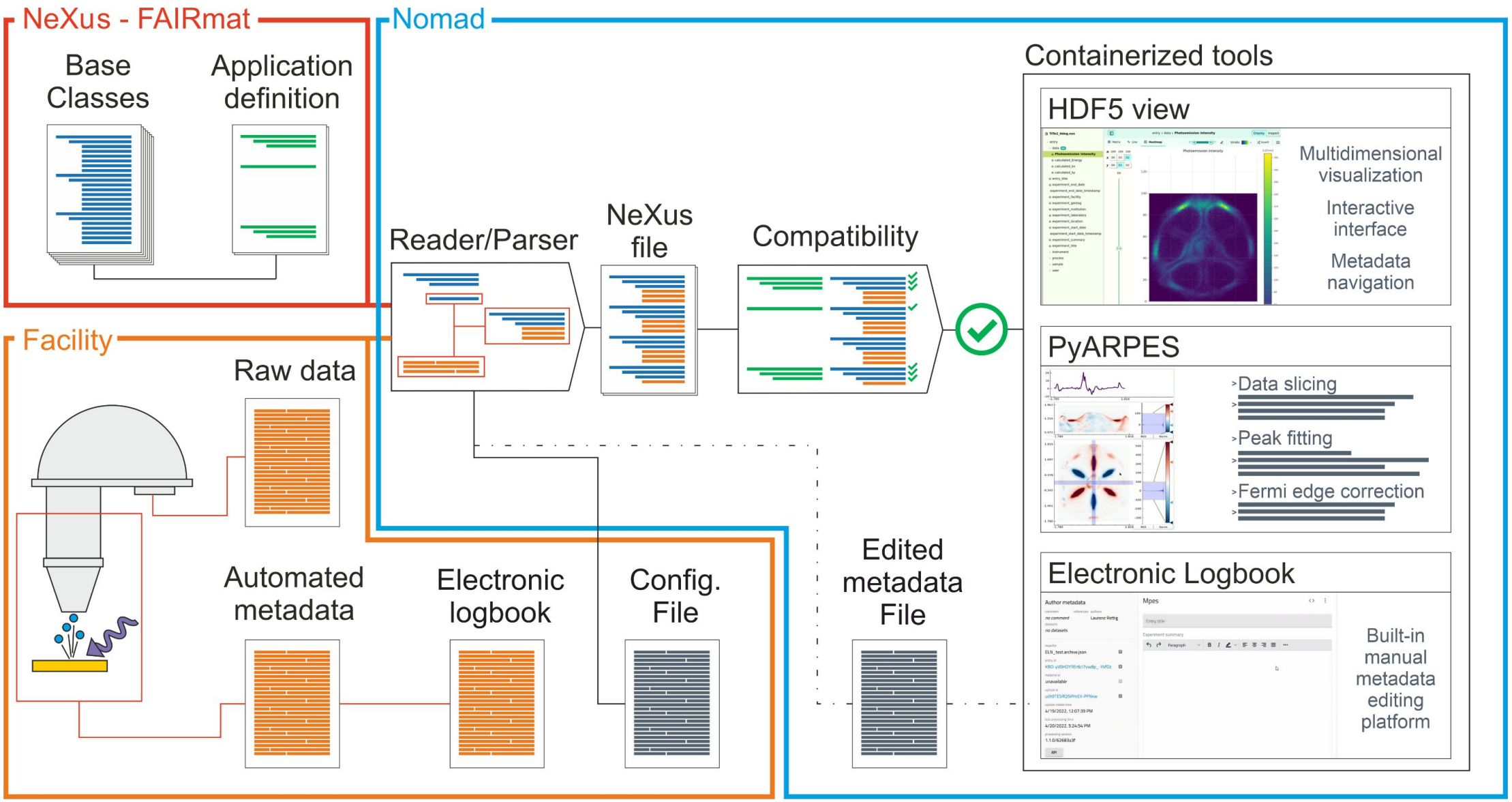
Experiment



FAIRmat

Photoelectron spectroscopy

What to do?

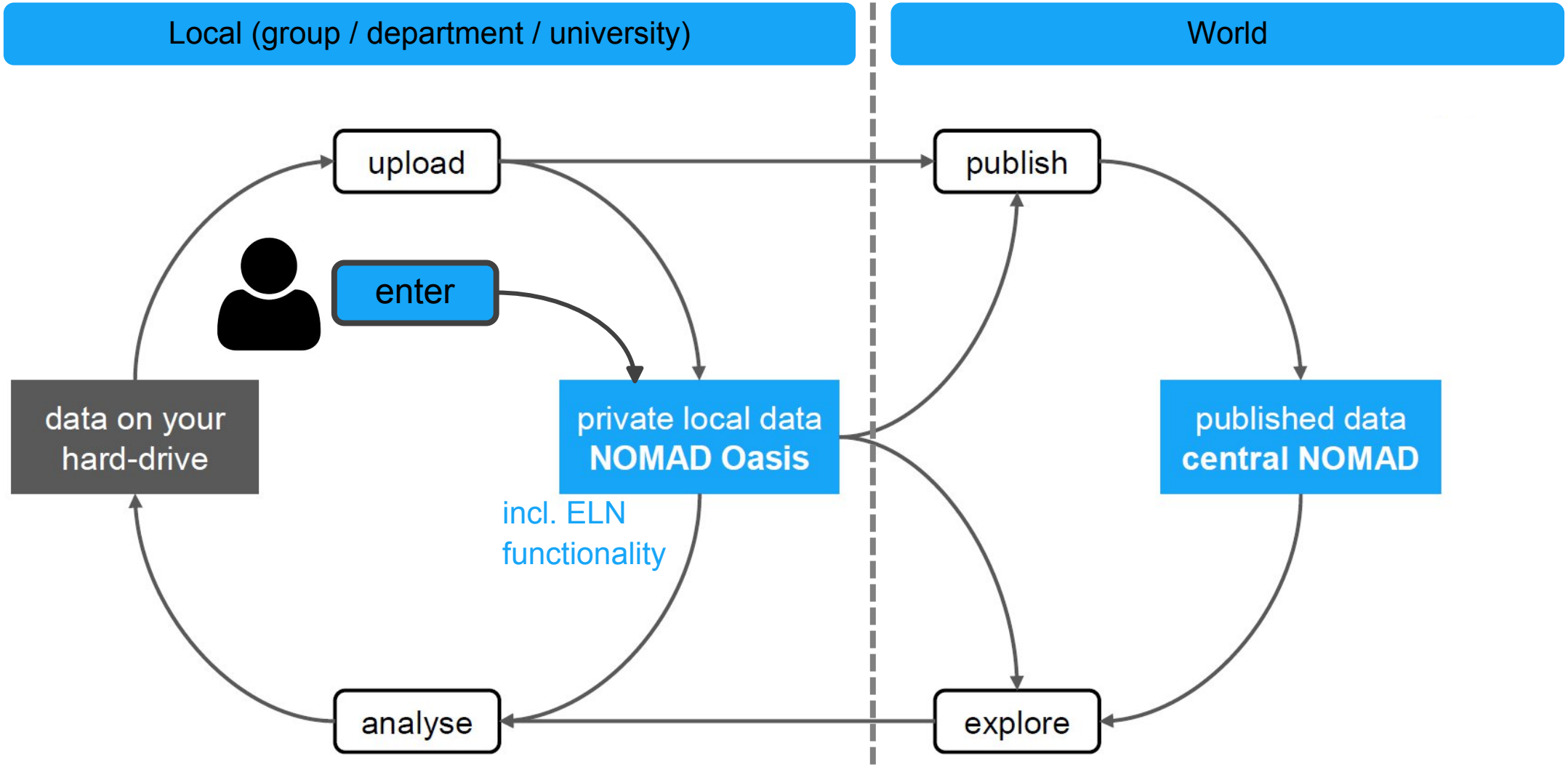


FAIRmat

Local/global

NOMAD OASIS

NOMAD



FAIRmat

FAIRmat's fields of action for experimental physics

Experiment

- Configurable Lab Control Software
- Electronic Laboratory Notebooks (ELNs)
- FAIR-ready data management
- Community standards
- Involving technology partners
- Workflow in the NOMAD environment
- Broad data expertise

NOMAD CAMELS

NOMAD OASIS

NOMAD

FAIRmat

Application Definitions and the NeXus Standard

Sandor Brockhauser and the FAIRmat team

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

Data Modelling

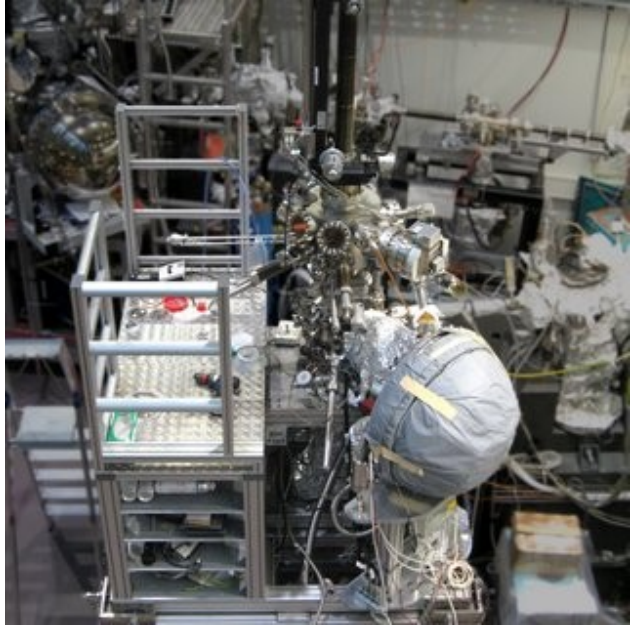
What to Model

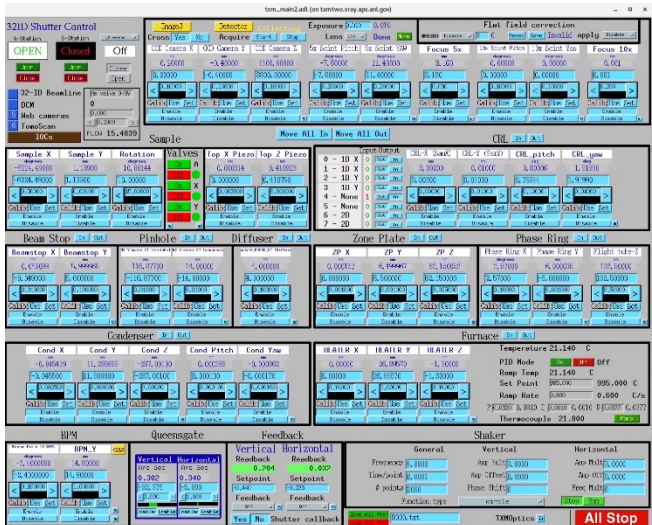
- **Experiment**
  - *Instrumentation*
  - *Sample* Preparation
  - Sample Environment
  - Monitors and Detectors
  - Data Processing
  - *Notes*

- **Observation**

in

- **Well-controlled environment**





FAIRmat

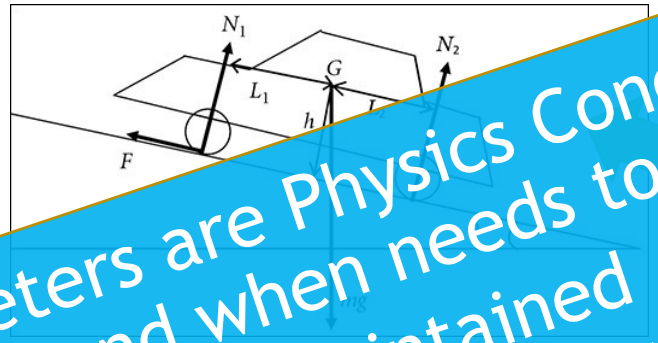
Data Modelling

What to Model

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

Experiment Data Models

Physics Model



Experiment Parameters are Physics Concepts to describe what and when needs to be
- Set or Maintained
- Measured or Calculated

Experiment
- start time, t_0
- duration, t
- distance, d

Slope
- angle, θ
- length, l

Car
- mass, m
- center of mass, $[G,h]$
- wheel positions, $[L_1,L_2]$
- force on wheels, $[N_1,N_2]$
- *resistive force*, F

Experiment Parameters

FAIRmat

Data Modelling

What to Model

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

Conceptual Design Model

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

Samples:
- Composition
- Geometry
- History
- …

Sample Environment.:
- Temperature
- Pressure
- Magnetic field
- …

Instrument:
- ARPES experiment
- XRD
- XRF measurement
- I-V measurement
- …

Data Processing:
- Data correction
- Calibration
- Data reduction
- Scientific analysis
- …

FAIRmat

Data Modelling

Data-modelling for Reproducibility

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

Technical Design Model

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

E.g.
what was the pressure in a chamber

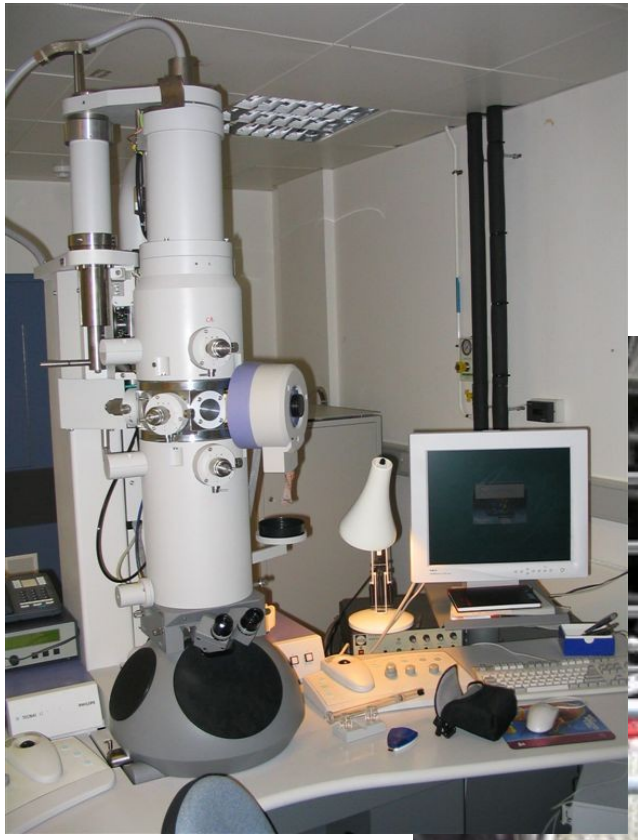
how it has been produced and maintained

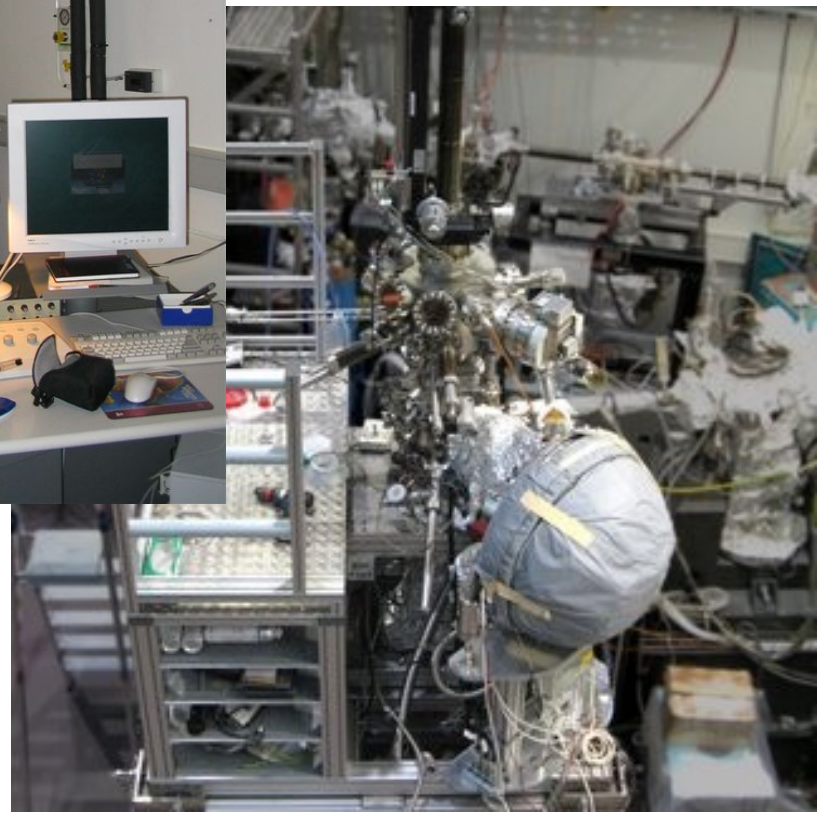
FAIRmat

Data Modelling

Data-modelling for Reproducibility

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



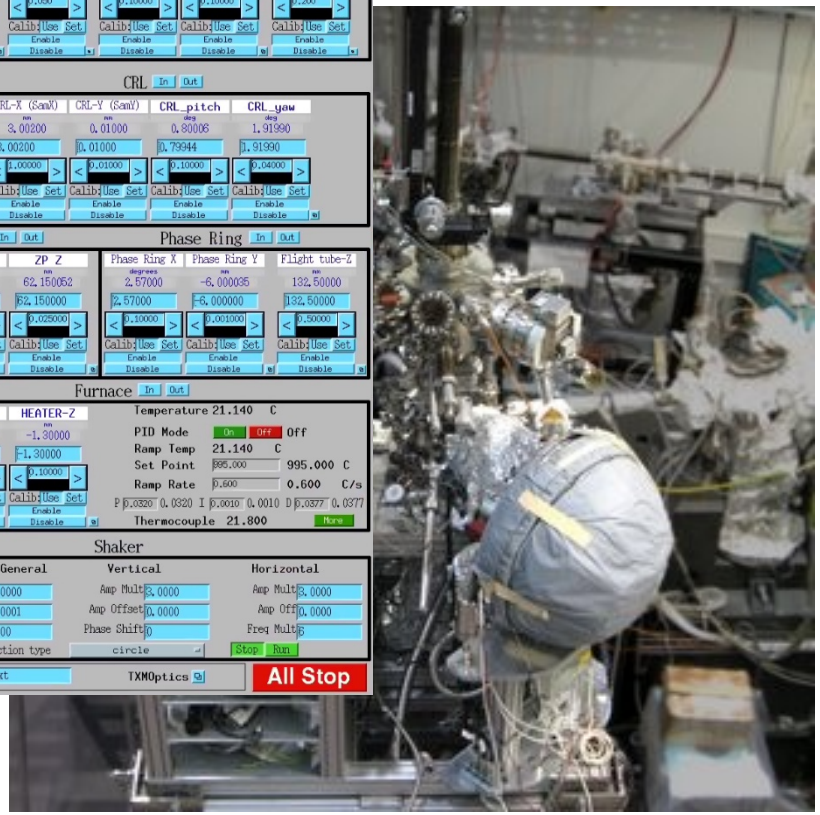


FAIRmat

Data Modelling

Data-modelling for Reproducibility

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

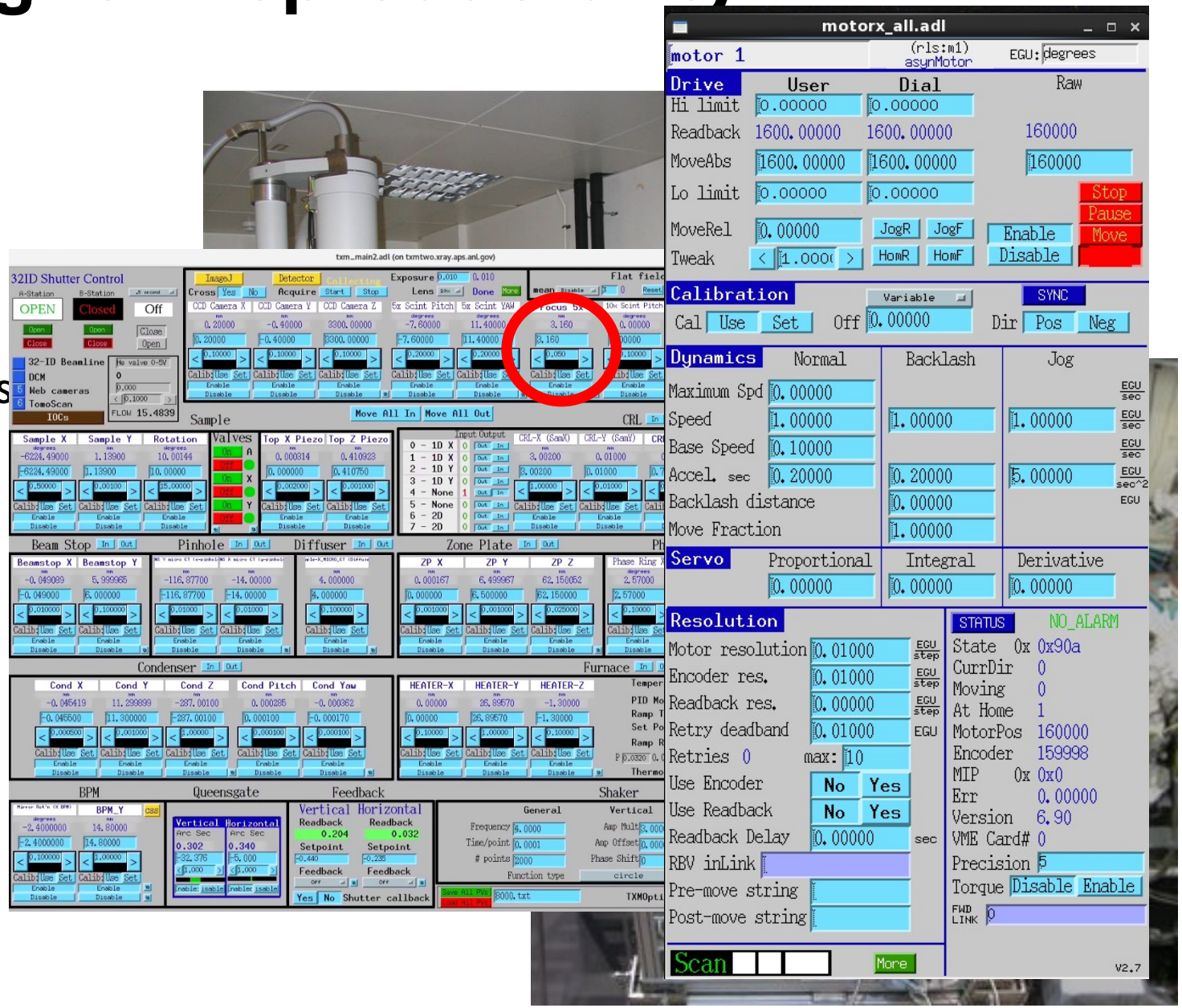


FAIRmat

Data Modelling

Data-modelling for Reproducibility

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

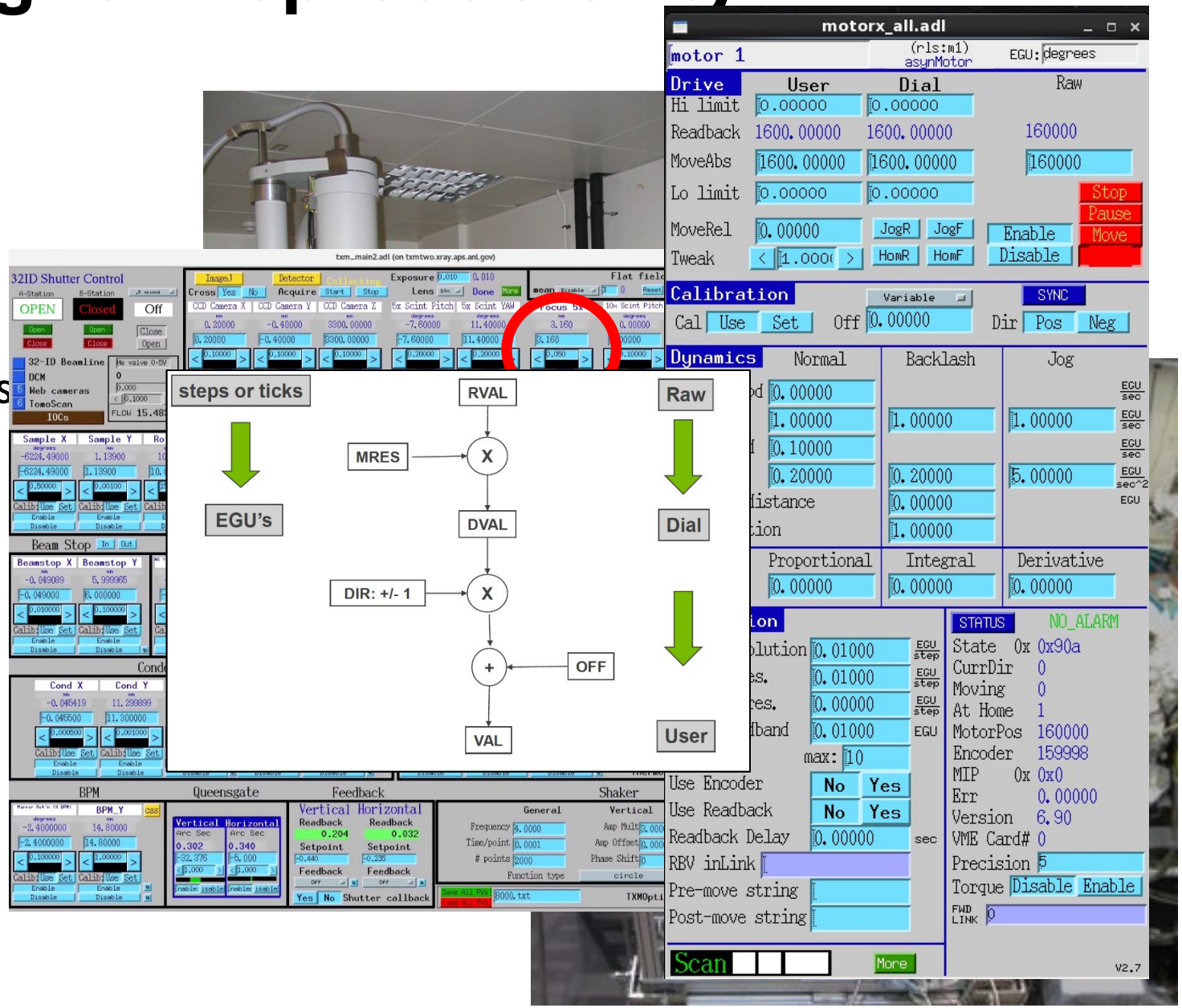


FAIRmat

Data Modelling

Data-modelling for Reproducibility

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

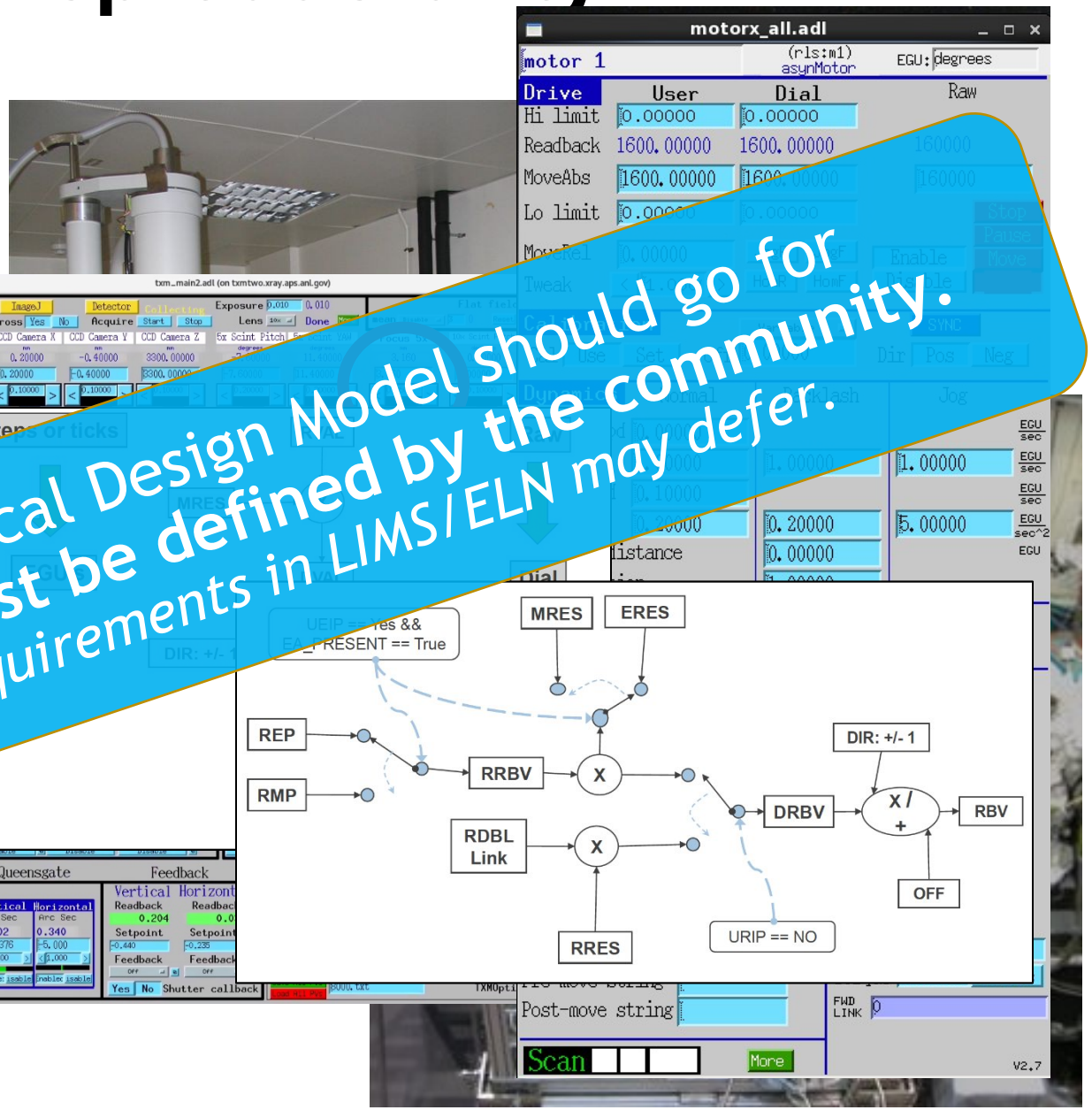


Data Modelling

Data-modelling for Reproducibility

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

How deep a Technical Design Model should go for reproducibility, **it must be defined by the community.** *Note that local requirements in LIMS/ELN may defer.*



FAIRmat

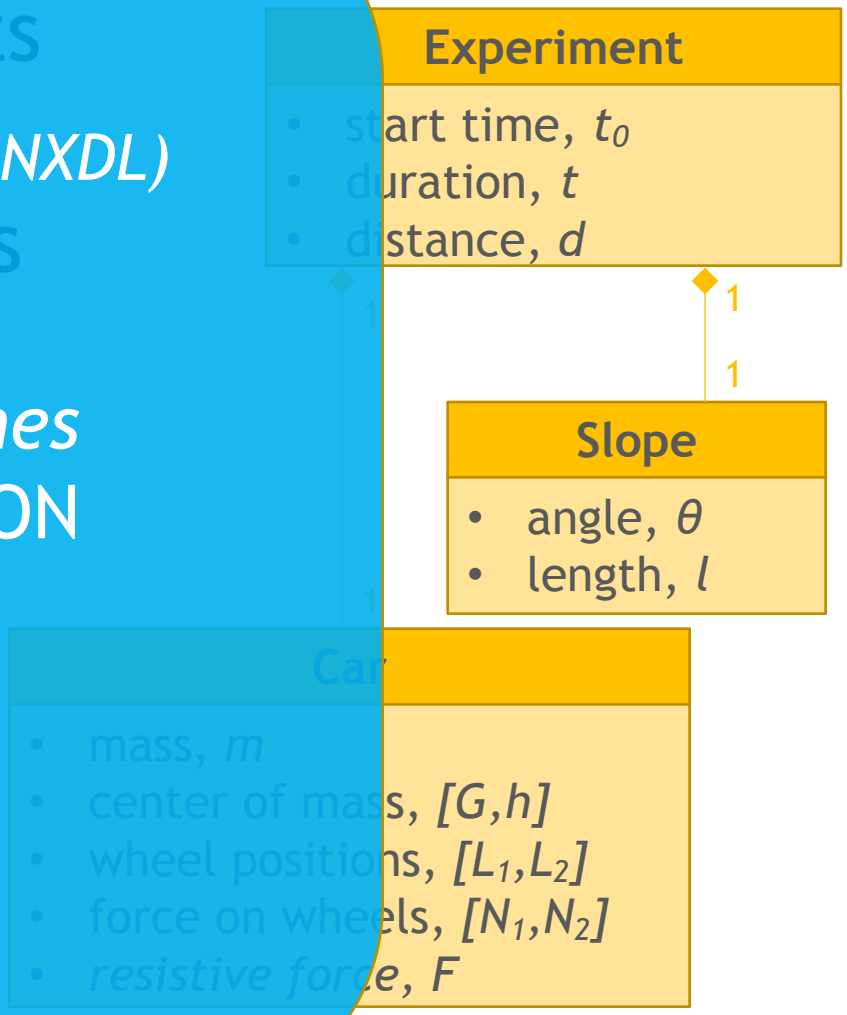
Data Modelling

Data-modelling for a Community

- Contr
- Clear
  conce
- Cardi
- Stand
- Comm
- Agree
- Commo

NeXus *https://www.nexusformat.org/*

- Extendable Data Modelling *(NXDL)*
- Automated documentation:
  HTML vocabulary - Humans
  OWL ontology OWL - Machines
- Community STANDARDIZATION process *(NIAC)*
- Tools
  - Verification and Validation
    NXDL *(schema)* ⟷ NXS *(data)*
  - Visualisation



FAIRmat

Standardisation

Data-modelling in NeXus

- Base Classes
  - NXinstrument, NXsample, NXprocess, NXuser NXnote
  - NXdetector, NXsensor, NXmonitor, NXbeam, NXmirror, NXgeometry,…

57

+106

- Application Definitions
  - NXarpes, NXxas, NXem, NXapm, NXmpes NXellipsometry, NXsensor_scan
    …

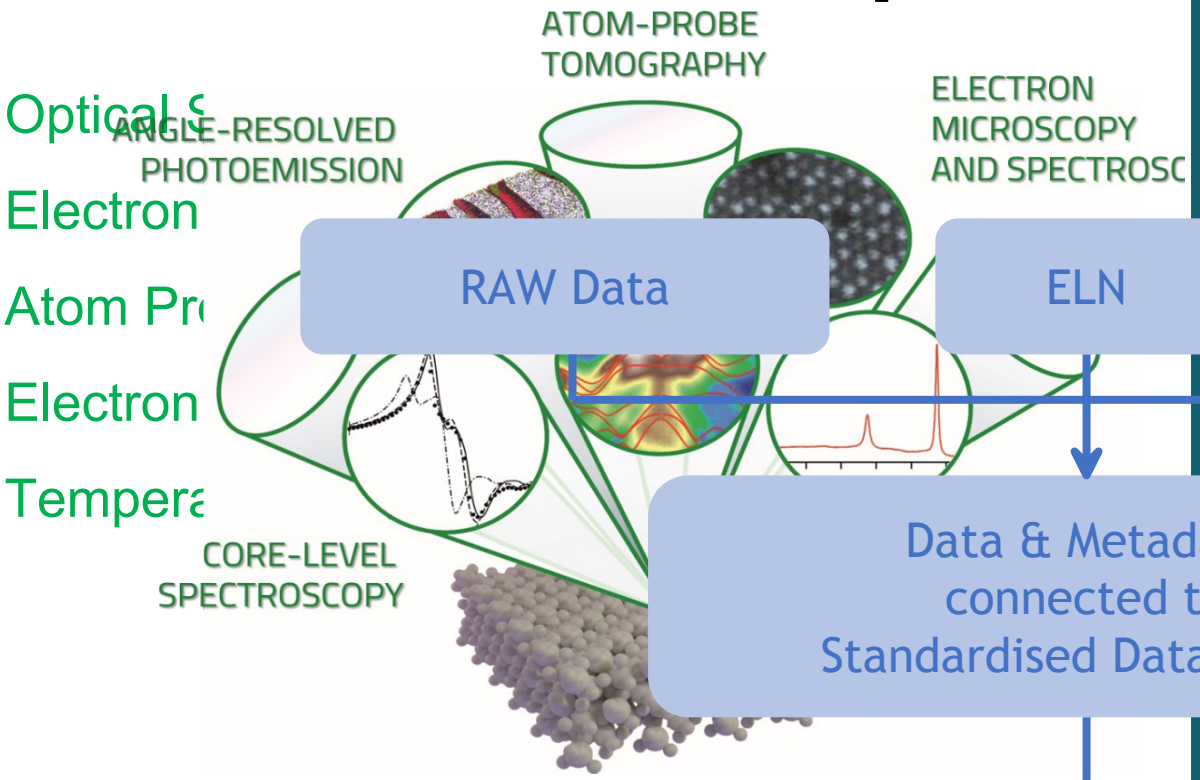
34

+38

FAIRmat

Techniques

Data Models for Experime



RAW Data

ELN

Data Analysis

Data & Metadata
connected to
Standardised Data Model

NOMAD
Search / Visualisation / Analysis

NeXus-FAIRmat

Proposal of NeXus expansion for FAIRmat data.

Watch

vigation

mat-NeXus Prop

s Documentation

Xus: User Manual

- Examples of writing and reading NeXus data files
- NeXus: Reference Documentation
- NeXus Community

Sample preparation

Nomad Remote Tools Hub (NORTH)

NXmpes

Status:

application definition, extends NXobject

Description:

This is the most general application definition for multidimensional photoelectron spectroscopy.

Symbols:

No symbol table

Groups cited:

NXaperture, NXbeam, NXcalibration, NXcollectioncolumn, NXdata, NXdetector, NXinstrument, NXmanipulator, NXnote,

title: (required) NX_CHAR

start_time: (required) NX_DATE_TIME

Datetime of the start of the measurement.

definition: (required) NX_CHAR

Obligatory value: `NXmpes`

@version: (required) NX_CHAR

USER: (required) NXuser

Contact information of at least the user of the instrument or the investigator who performed this experiment. Adding multiple users if relevant is recommended.

name: (required) NX_CHAR

Name of the user.

affiliation: (recommended) NX_CHAR

Name of the affiliation of the user at the point in time when the experiment was performed.

address: (recommended) NX_CHAR

Full address (street, street number, ZIP, city, country) of the user's affiliation.

email: (required) NX_CHAR

Email address of the user.

orcid: (recommended) NX_CHAR

Author ID defined by https://orcid.org/.

INSTRUMENT: (required) NXinstrument

energy_resolution: (required) NX_FLOAT {units=NX_ENERGY}

SOURCE: (required) NXsource

The source used to generate the primary photons. Properties refer strictly to parameters of the source, not of the output beam. For example, the energy of

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

FAIRmat

Roadmap

Data-modelling for a Community

- Initial Proposal with Examples
- Community Feedbacks
- Common Proposal with Technology Partners
- Standardisation

FAIRmat

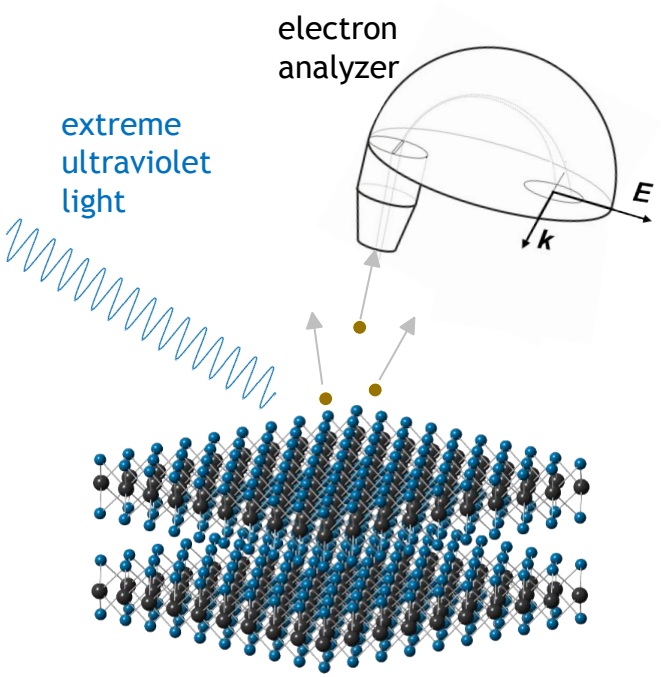
NXmpes: structure and example in NOMAD

Laurenz Rettig and the FAIRmat team

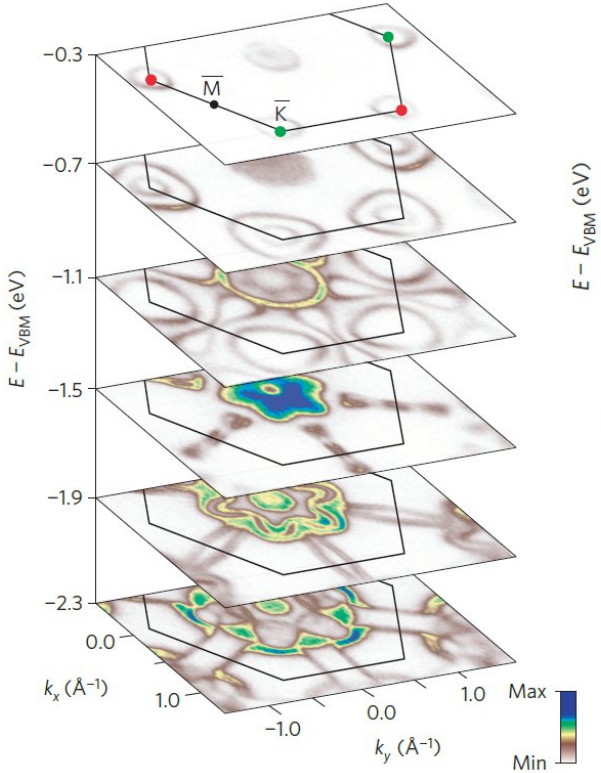
Fritz Haber Institute of the Max Planck Society, Berlin

Multidimensional photoemission spectroscopy

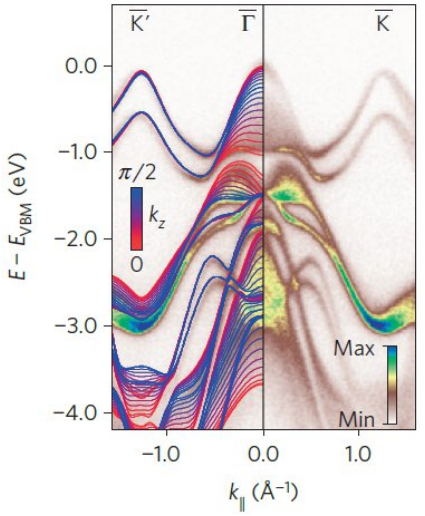
experiment



data: band mapping

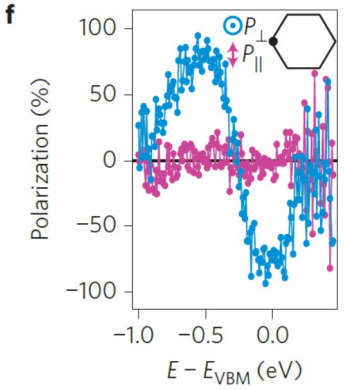


energy band dispersion



Also possible:

Spin polarization



Example: valence band structure of WSe_2
Riley *et al*, Nature Phys. 10, 835 (2014)

- momentum ***k*** *(k_x, k_y)*
- energy *E*
- spin ***S***

} "multidimensional PES" (more than 3D)

Schönhense et al., New J. Phys. 20, 033004 (2018)

Dataformat

FAIRmat

Dataformat

Extending the parameter space

$$I \rightarrow I(\mu_{mat}, E, k_x, k_y, S, t, \omega_p, \sigma, x, y, T, \ldots)$$

- Dependence on materials and preparation procedures (μ_{mat})
- Spin-resolving detectors: S
- Photon energy and polarization ω_p
- Electronic structure of **non-equilibrium states** (time-resolved ARPES): time t, pump/probe photon energy ω_p, polarization, …

à Dependence on sample parameters:
strain σ, sample position x,y, temperature T

- …

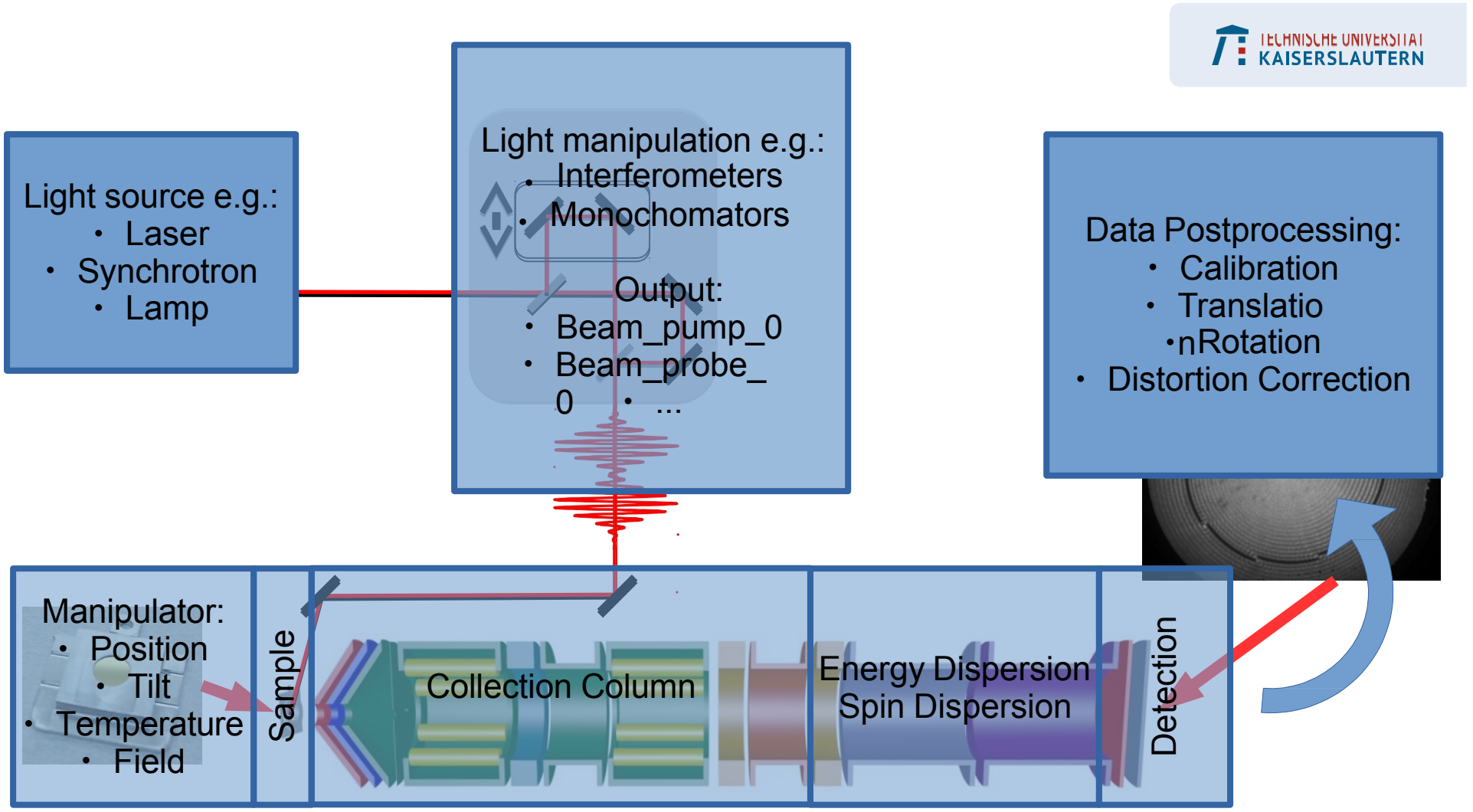
à huge parameter space
à Mostly only subspace both experiment-tally accessible and interesting

Goal to develop flexible, community-driven data + metadata format for such MPES data

FAIRmat

Dataformat

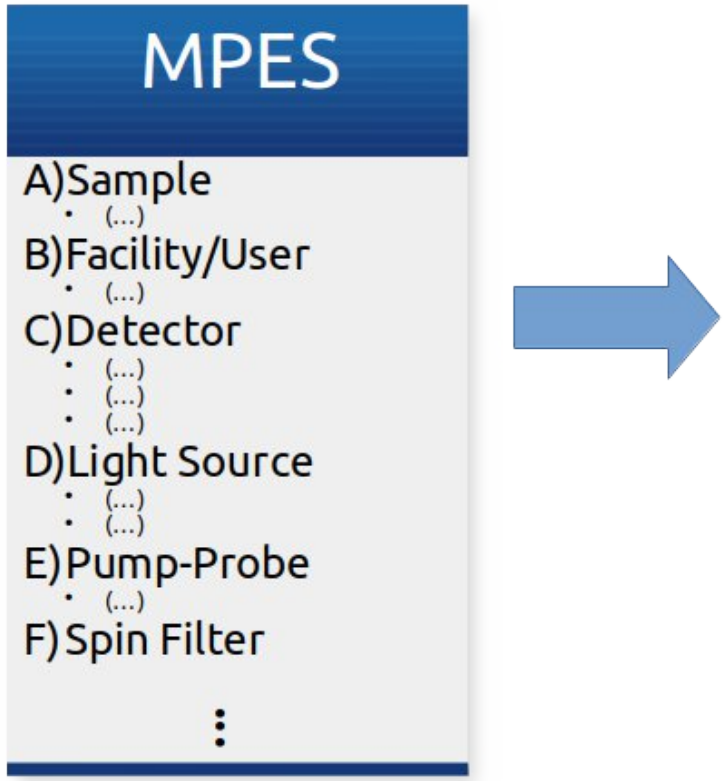
Ingredients of a photoemission experiment



FAIRmat

Hierarchy of application definitions

Universal MPES
metadata dictionary
(in NeXus: base classes)



Dataformat

FAIRmat

Dataformat

Simplify analysis

NeXus

Look-up Dictionary

Application definition

Facility

Raw data

Raw metadata

Parser

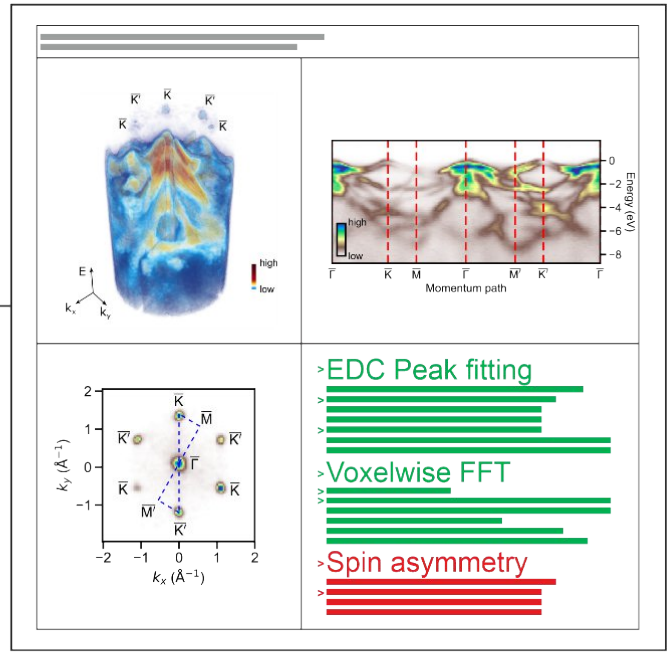
NeXus file

Compatibility

Hierar-chical structure, F.A.I.R. compliant

Modularity: a common platform deploys application-dependent tools

Automated access



Figures from: Xian, R. P. et al. arXiv:1909.07714 (2019).

FAIRmat

Dataformat

Application definition: NXmpes

Application Definitions

We created two new application definitions:

NXmpes:
A general appdef with minimalistic metadata requirements, apt to describe all photemission experiments.

Base Classes

We developed entirely new base classes:

NXelectronanalyser:
A base class to describe electron kinetic energy analizers. Contains the collective characteristics of the instrument such as energy resolution, and includes the following subclasses:

NXcollectioncolumn:
Base class to describe the set of electronic lenses in the electron collection column (standard, PEEM, momentum-microscope, etc.).

NXenergydispersion:
Base class to describe the energy dispersion sytem (hemispherical, time-of-flight, etc.).

NXspindispersion:
Base class to describe the set of electronic lenses in the electron collection column.

NXmanipulator:
A base class to describe the complex manipulators used in photoemission experiments, often with > 4 degrees of freedom, cryogenic cooling and other advanced features.

FAIRmat

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

Application definitions and base classes

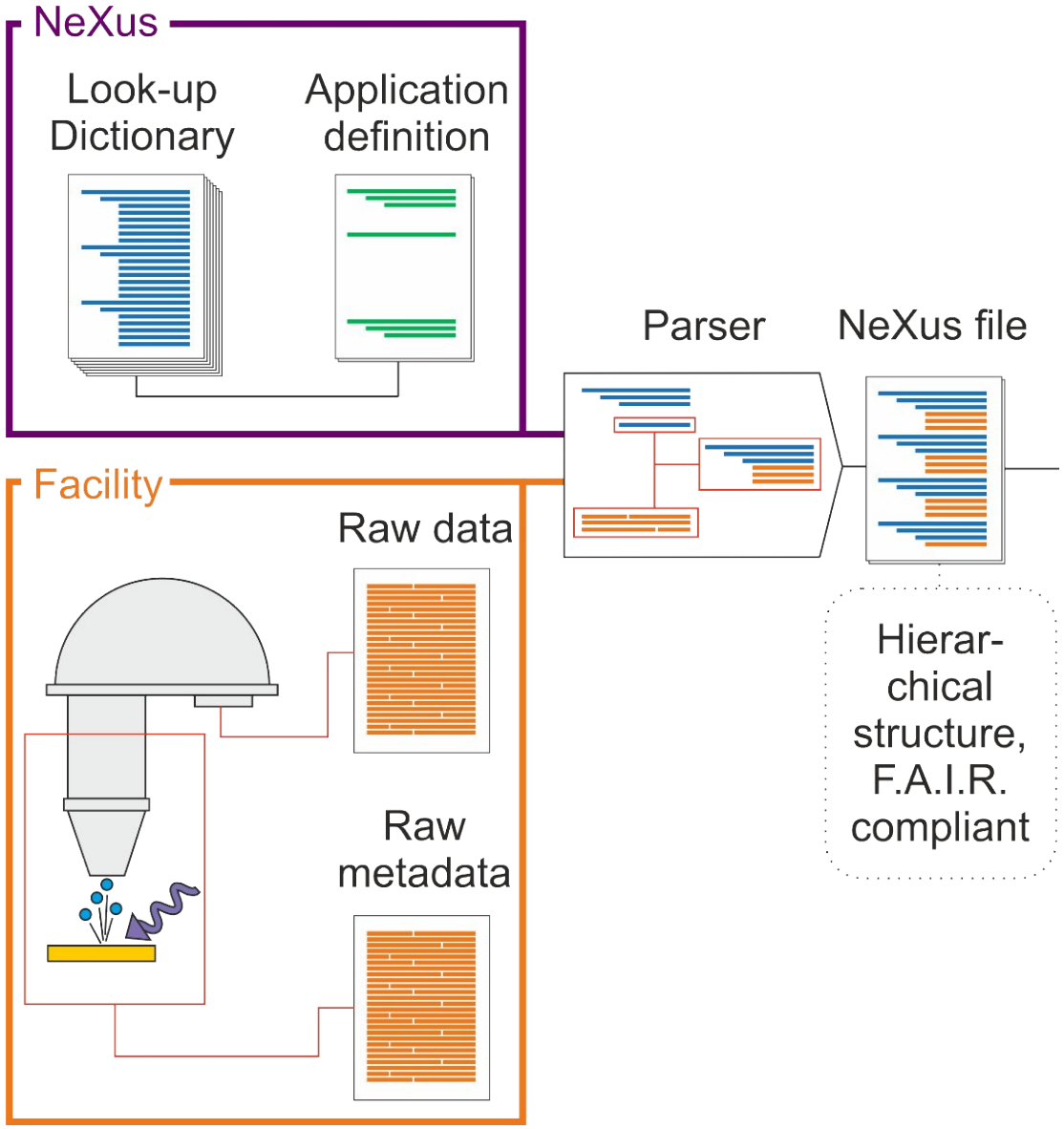
Dataformat

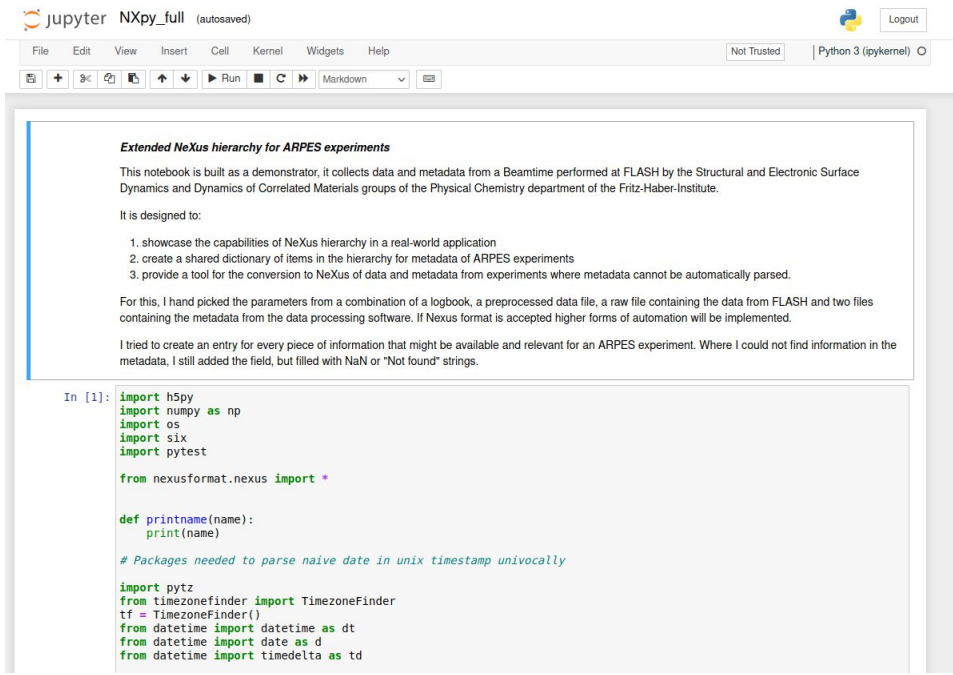
FAIRmat

Detailed discussion of base classes and application definitions tomorrow morning.

Dataformat

Practical examples





Converter/parser

- reads data + metadata from experiment
- Assigns them to the correct NXmpes paths
- Creates a verified NeXus file according to a provided application definition

FAIRmat

Reader + parser structure

Data.h5

- Volumetric data + axes
- Tree of metadata

config_file.json

```
"/ENTRY[entry]/definition": "NXmpes",
"/ENTRY[entry]/definition/@version": "None",
"/ENTRY[entry]/title": "@attrs:metadata/entry_title",
"/ENTRY[entry]/start_time": "@attrs:metadata/timing/acquisition_start",
"/ENTRY[entry]/experiment_institution": "Fritz Haber Institute - Max Planck Society",
"/ENTRY[entry]/experiment_facility": "Time Resolved ARPES",
"/ENTRY[entry]/experiment_laboratory": "Clean Room 4",
"/ENTRY[entry]/entry_identifier": "@attrs:metadata/entry_identifier",
"/ENTRY[entry]/end_time": "@attrs:metadata/timing/acquisition_stop",
"/ENTRY[entry]/duration": "@attrs:metadata/timing/acquisition_duration",
"/ENTRY[entry]/duration/@units": "s",
"/ENTRY[entry]/collection_time": "@attrs:metadata/timing/collection_time",
"/ENTRY[entry]/collection_time/@units": "s",

"/ENTRY[entry]/USER[user]/name": "@attrs:metadata/user0/name",
"/ENTRY[entry]/USER[user]/role": "@attrs:metadata/user0/role",
"/ENTRY[entry]/USER[user]/affiliation": "@attrs:metadata/user0/affiliation",
"/ENTRY[entry]/USER[user]/address": "@attrs:metadata/user0/address",
"/ENTRY[entry]/USER[user]/email": "@attrs:metadata/user0/email",
```

Python-based reader interprets data file, based on the information of the config file.
Additional information from other meta data sources.
A custom reader for each file type/technique, but a common infrastructure.

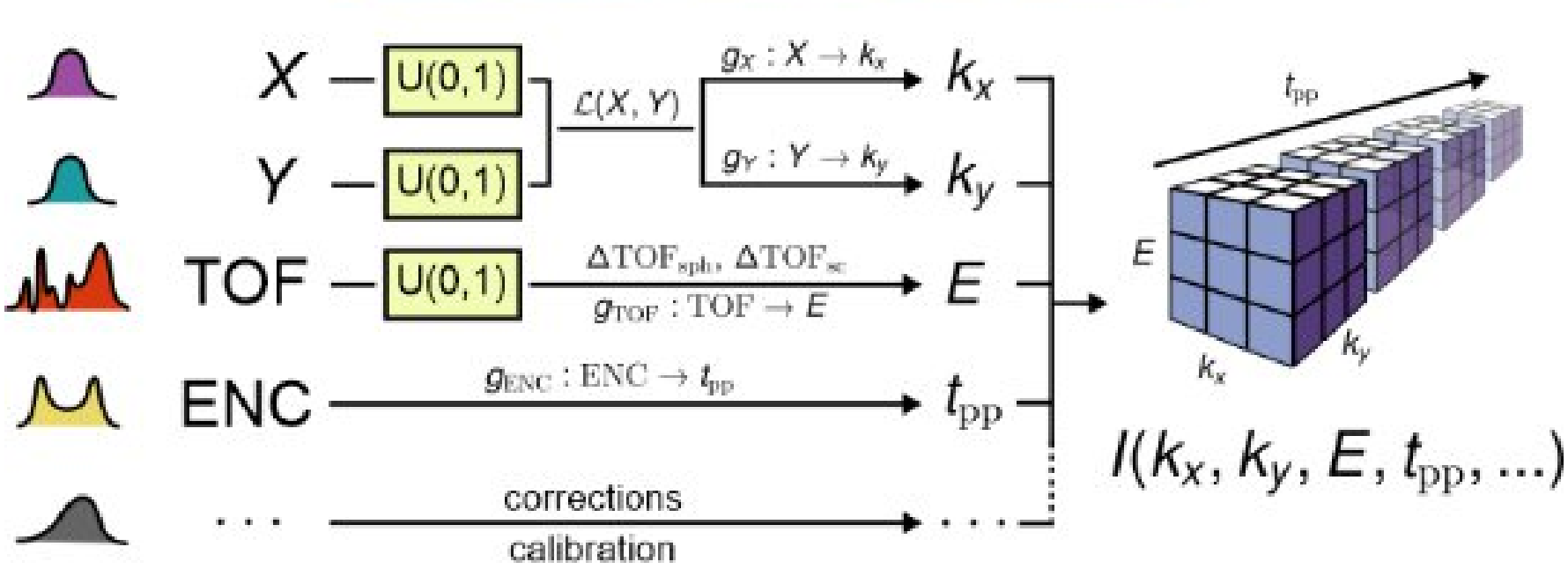
Dataformat

FAIRmat

Dataformat

Single event data processing

Time-resolved Photoemission data, recorded with SPECS METIS 1000 momentum microscope.
Single-event data calibrated and converted into multidimensional volumetric data.



Xian, et al., Scientific Data 7, 442 (2020)
https://mpes.science/

FAIRmat

Practical examples

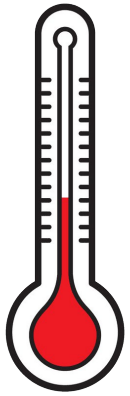
Dataformat

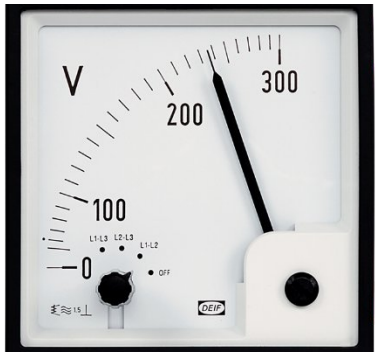
LIVE Example on NOMAD

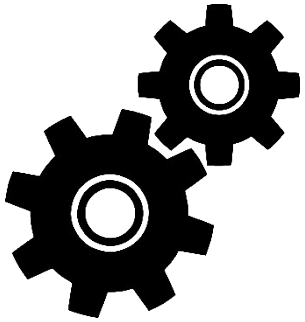
FAIRmat

Sources of metadata

Machine-generated meta data:






Automated recording and assigning to respective fields

"Soft" meta data need structured user input
e.g.:

- User who does the experiment
- Name and composition of the sample
- …

Requires **structured electronic lab notebook** or **user interfaces**

Structured and configurable ELN capabilities available in NOMAD

Dataformat

FAIRmat

MPES reader infrastructure

Lukas Pielsticker and the FAIRmat team

lukas.pielsticker@cec.mpg.de

Data Converters

Storage of photoemission data

SURFACE CONCEPT

ThermoFisher SCIENTIFIC

FOCUS

scientaomicron

SPECS™

…

Problem:

- Each technology partner has its own file format
- Within a technology partner the file format can be different

→ How to go from vendor-specific format to structured NXmpes data?

FAIRmat

Data Converters

NXmpes readers - objectives

- Data- and vendor agnostic Nxmpes files
→ **Parsers, normalizers, and converters**

- Data curation and quality assessment
→ **Validation** against NeXus-FAIRmat application definitions

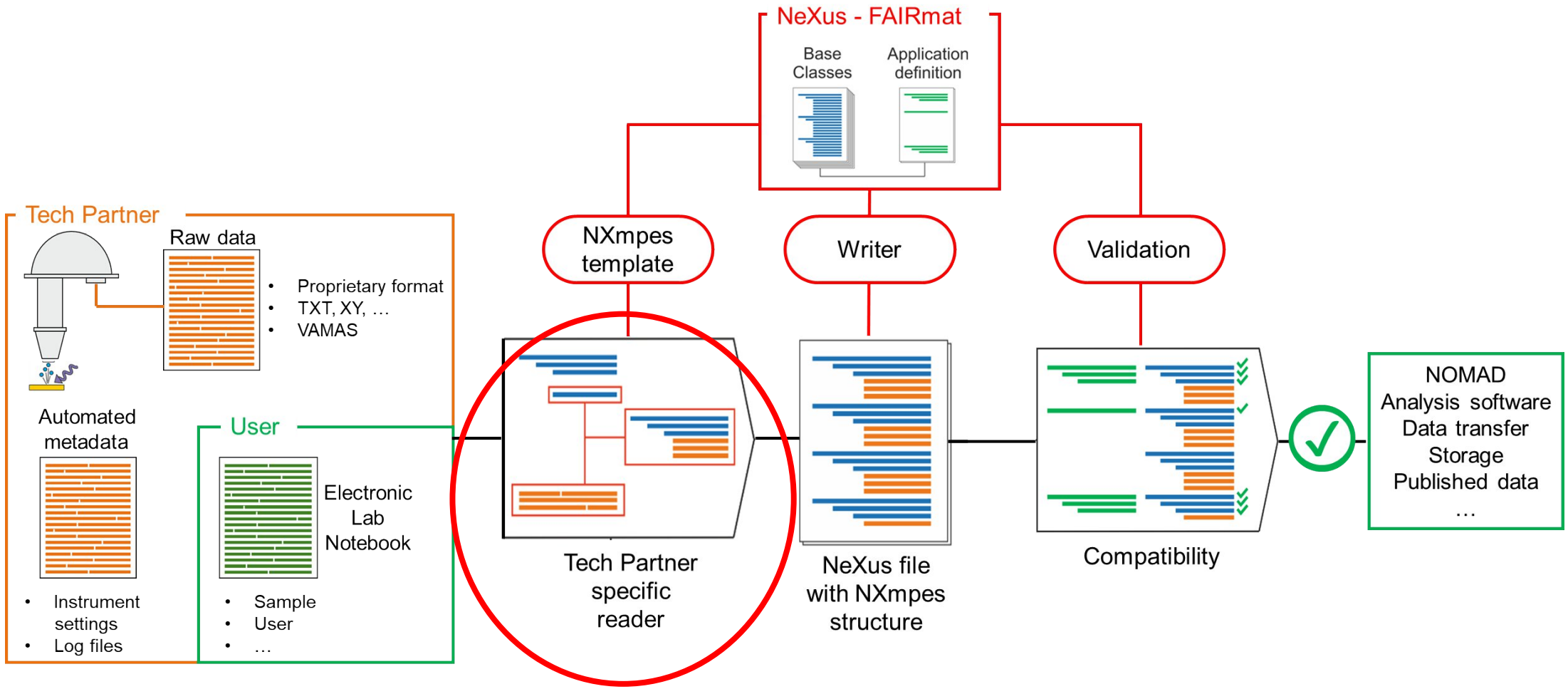
- Capture all **metadata** to produce valid NXmpes files
→ Instrument settings, ELN data, temperature/pressure logs, …

- Interoperability and repeatability
→ Calibration, transmission functions, etc.
→ Storing **raw, processed, and analysed** data

FAIRmat

Data Converters

From vendor-specific data to NXmpes



FAIRmat

Data Converters

Link to technology partners and the community

Choice for technology partners

1. Full integration with native support to read/write NXmpes
2. Integrated option to directly export to NXmpes
3. Provide external parser to convert to NXmpes

What needs to be discussed

- Which (meta-)data would we like to store in NXmpes?
  → Discussion of MPES application definition (tomorrow, 9 am)
- Which are the limits of open data?
- Metadata sources
- Do the technology partners need support?

What we can provide

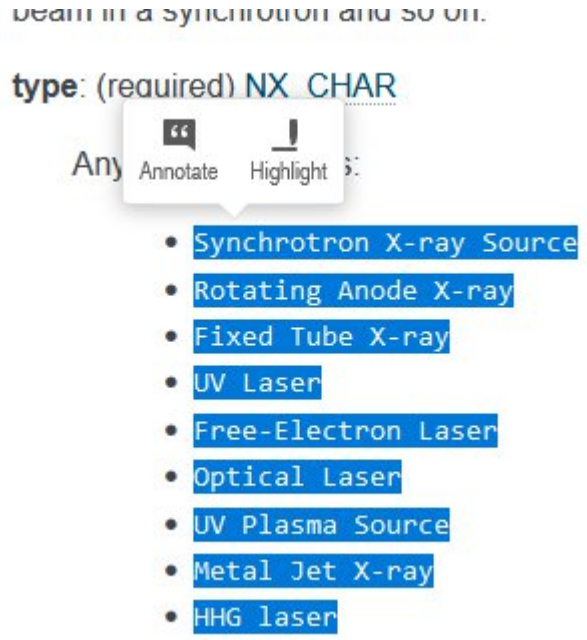
- Detailed description of application definitions
- Reader examples
- Validation infrastructure

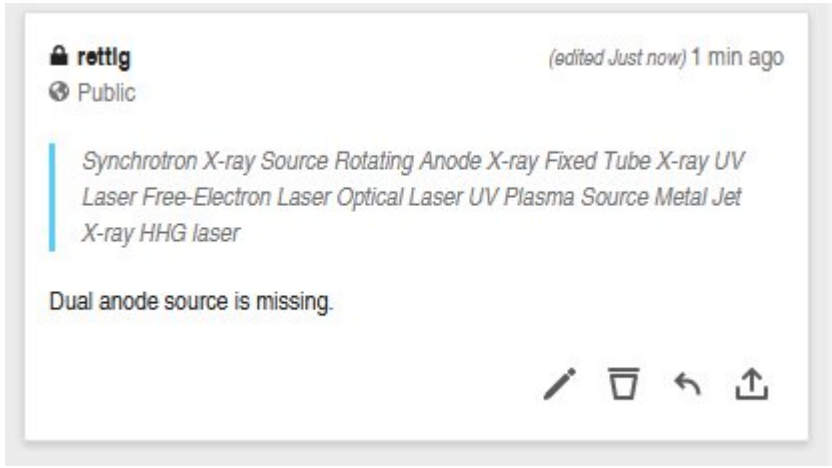
NeXus-FAIRmat application definitions:
https://fairmat-nfdi.github.io/nexus-fairmat-proposal/

Example readers and validators:
https://github.com/FAIRmat-NFDI/pynxtools

FAIRmat

How to contribute further





We are very happy to integrate your input into the Nxmpes and base classes definitions!

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

Dataformat

Discussion points and questions:

- Do our proposed definitions fit your use case? What is missing, needs to be changed?
- Requirement/optionality of parameters
- Which parameters can be provided, which not?
- How are “soft” metadata acquired e.g. in vendor software?
- What are legal constraints, IP restrictions?
- Interplay of user-specific data and instrumental metadata
- …

FAIRmat