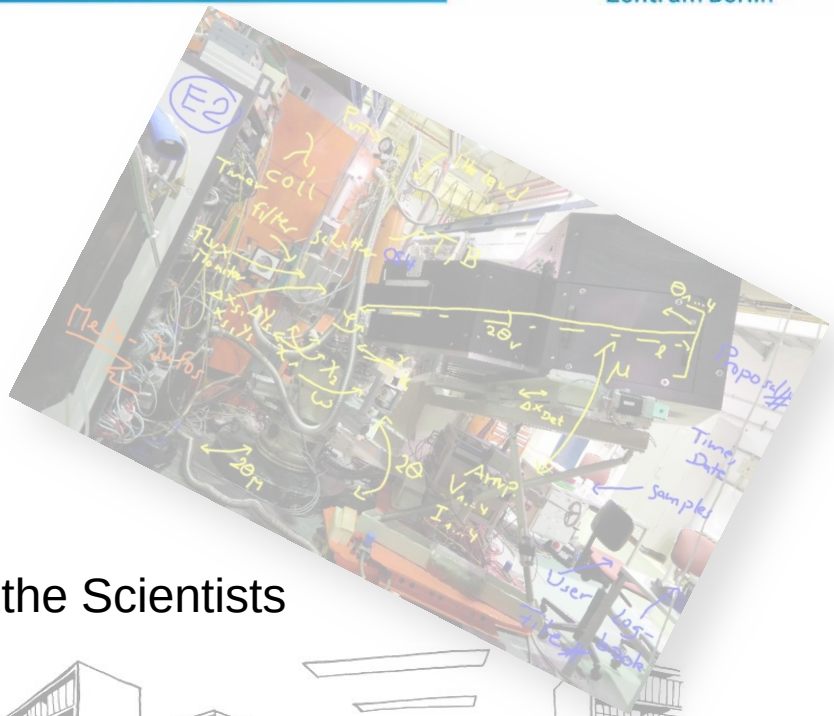
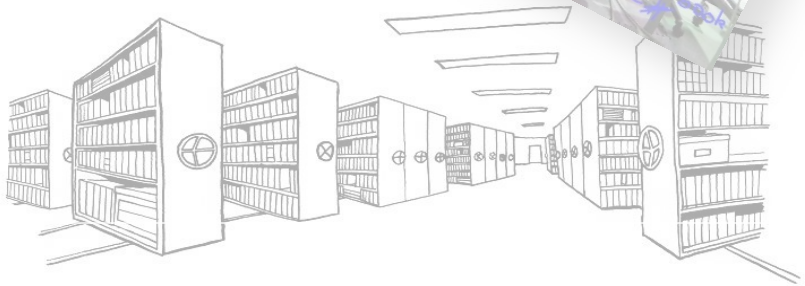
Data curation at HZB

WHAT TO EXPECT

- FAIR Data Principles
  - **What metadata are required**
  - **What metadata standards help us**
- RDM approach at HZB IT
- Implementations
- What IT can offer ←→ What we need from the Scientists





HZB Helmholtz Zentrum Berlin

EXCHANGE AND ARCHIVE

- Requirements described in FAIR Data Principles
  - **Findable, Accessible, Interoperable, Re-usable**

- Prerequisite for data exchange:
  - **standardisation and usage on agreed conventions: formats, vocabulary, units**
  - **extensive metadata**

HZB Helmholtz Zentrum Berlin

SEARCH, SORT, FILTER, QUERY, BROWSE

- Data catalogue vocabulary
  - **Bibliographic Metadata (DataCite)**
  - **Disciplinary vocabulary**
- Persistent identifiers
  - **Linking research data, persons, instruments, software, samples,… together**
- Technologies
  - **Metadata catalogues based on SQL data bases, geometrical databases, JSON, ontologies**

Structure search requests

browse linked data

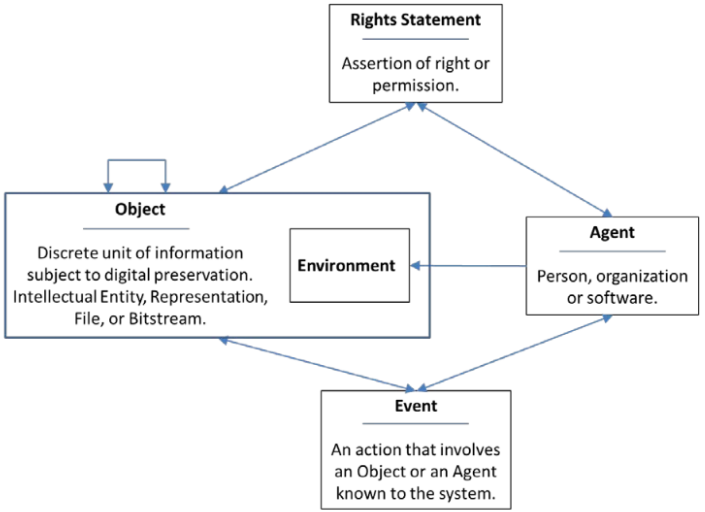
search:
- key/value
- propositional logic
- vector space model
- image search

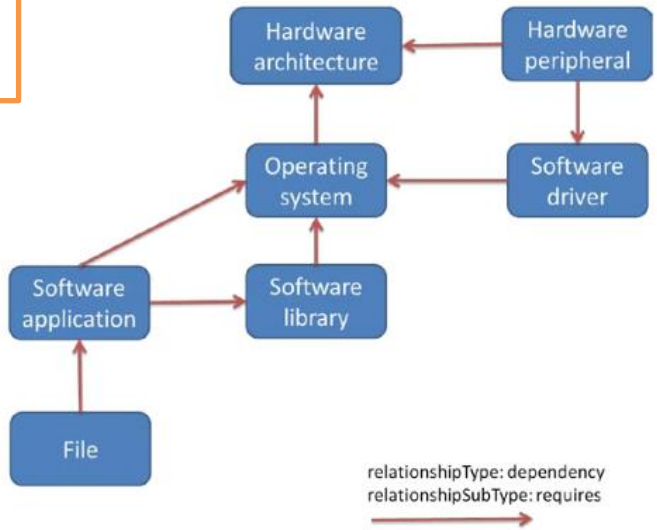
HZB Helmholtz Zentrum Berlin

PROVENANCE, PERMISSIONS/LICENSE, HOW TO OPEN

- Not disciplinary
- PREMIS

Trust the data you find and know how to use it





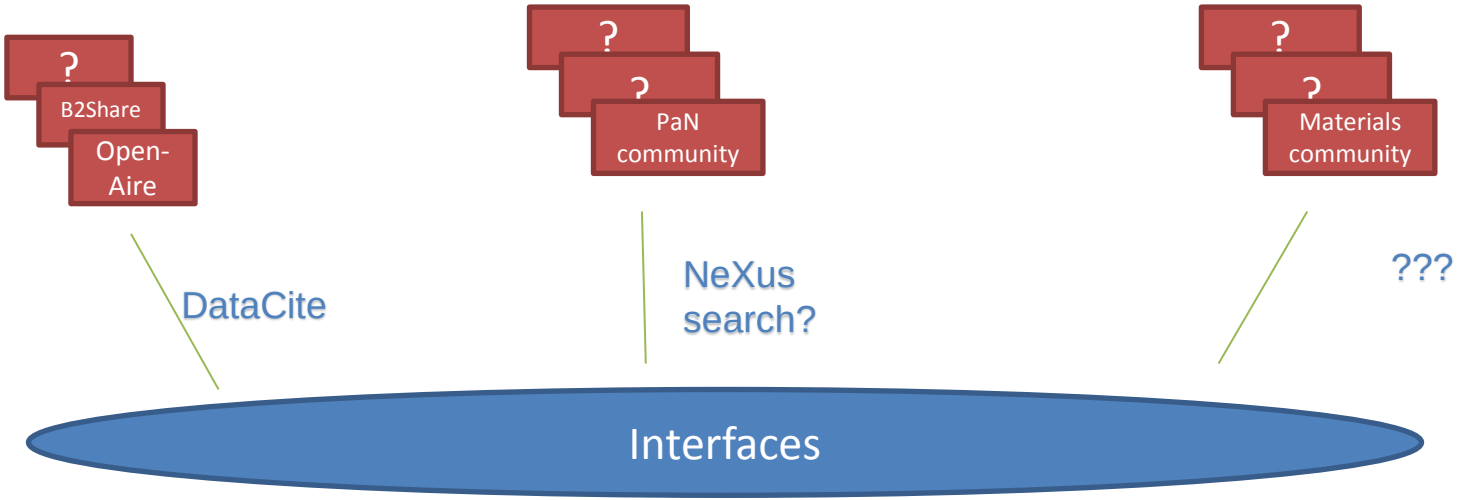
HZB Helmholtz Zentrum Berlin

OPEN, VIEW, UNDERSTAND, (RE-)USE, CHANGE, AUTOMATION

- What is required to use your experiment data for analysis?
- Is this enough information to understand your experiment?
- What else is required to simulate or reproduce the experiment?

- For light sources we can use NeXus
- For materials is it NOMADS?
    - **What else is relevant?**

HOW DO WE GET THERE?



findability

ICAT

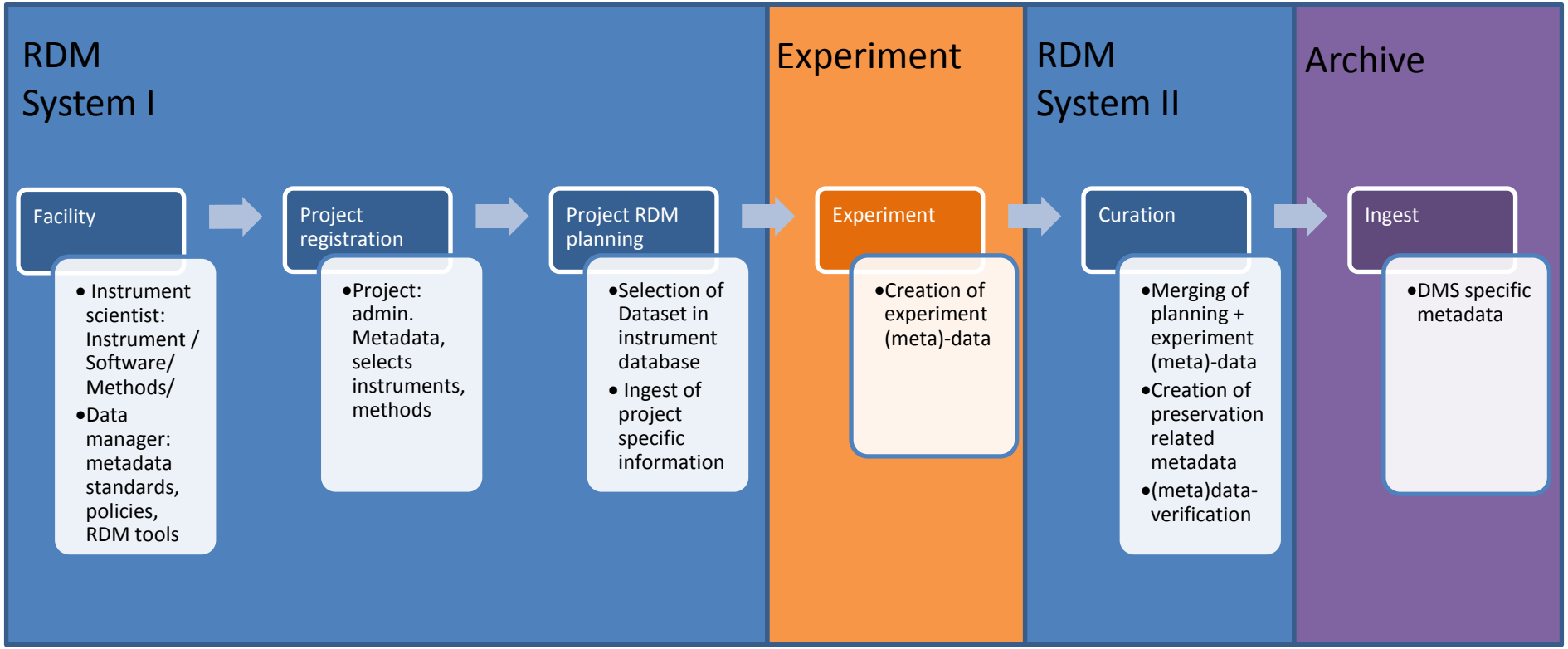
interoperability
re-useability

HZB Helmholtz Zentrum Berlin

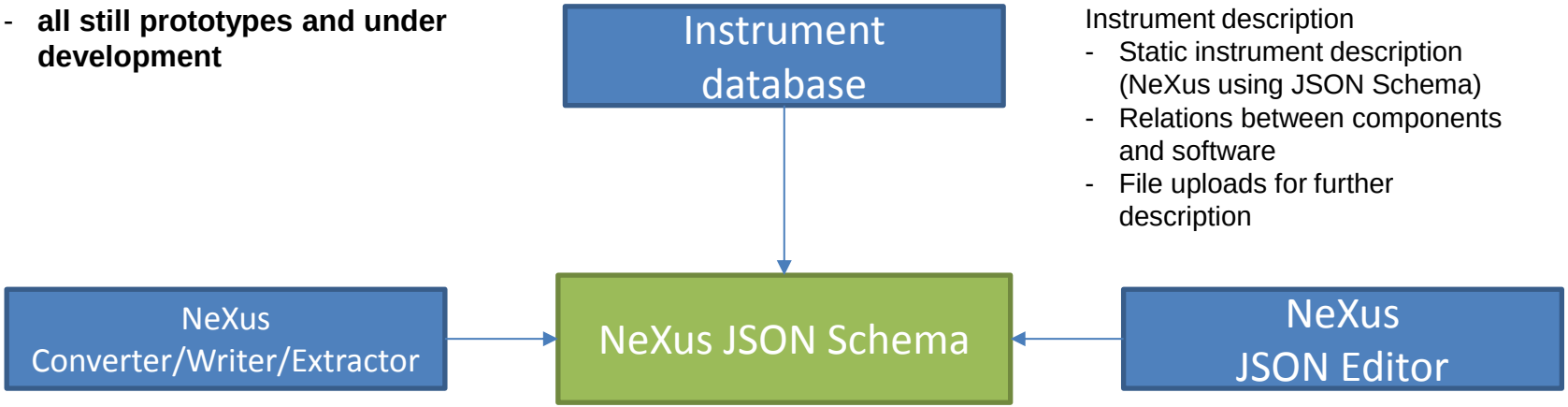
RE-USE INFORMATION, USE STANDARDS AND AUTOMATE COLLECTION

- Re-use information about data: collect static, not so static and metadata about instruments and software
- Collect and share information about standards
- Automate collection as far as possible

WORKFLOW


RDM System I
Experiment
RDM System II
Archive
Facility
• Instrument scientist: Instrument / Software/ Methods/
•Data manager: metadata standards, policies, RDM tools
Project registration
•Project: admin. Metadata, selects instruments, methods
Project RDM planning
•Selection of Dataset in instrument database
• Ingest of project specific information
Experiment
•Creation of experiment (meta)-data
Curation
•Merging of planning + experiment (meta)-data
•Creation of preservation related metadata
•(meta)data-verification
Ingest
•DMS specific metadata

- **all still prototypes and under development**



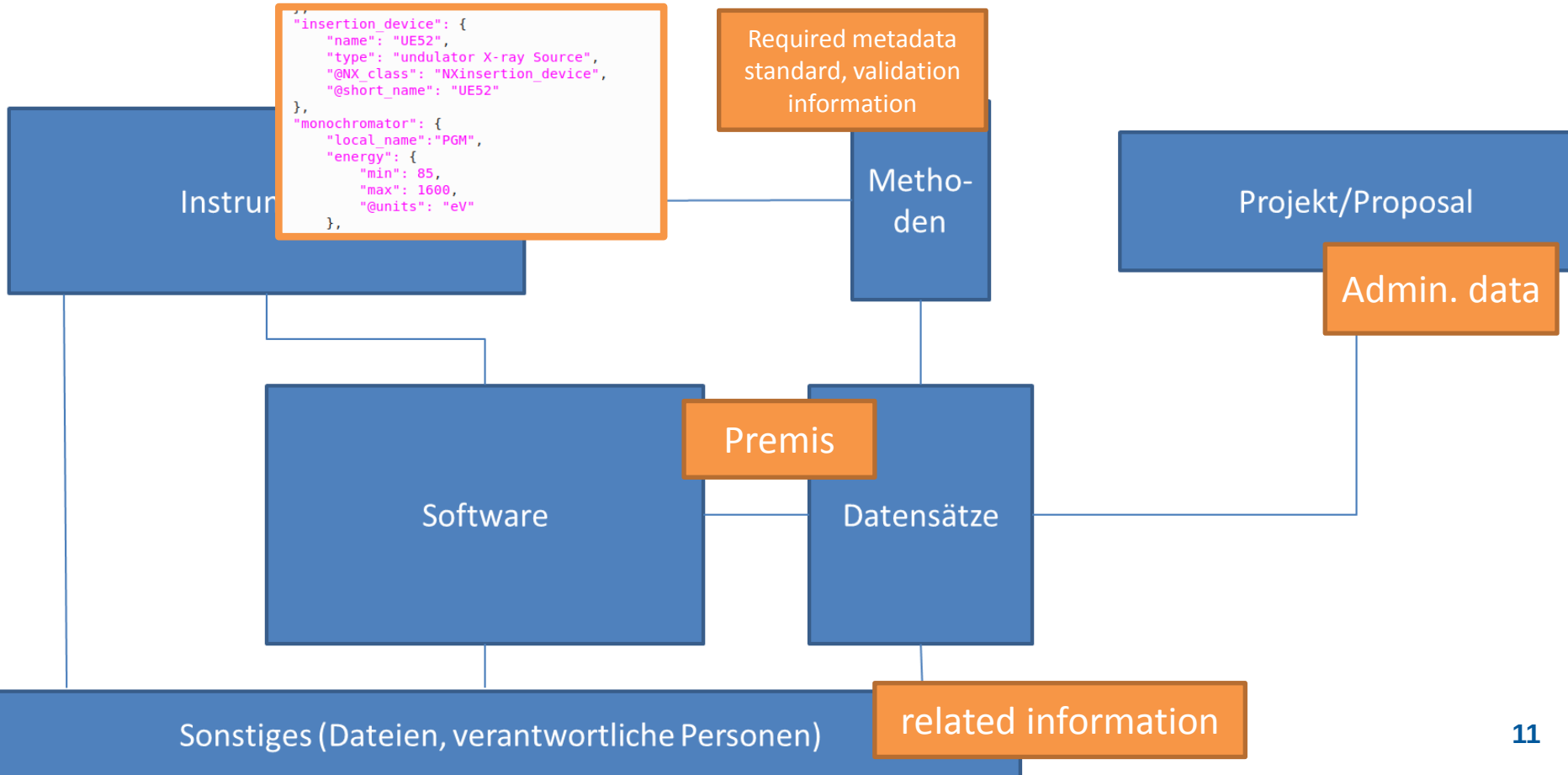
Instrument description
- Static instrument description (NeXus using JSON Schema)
- Relations between components and software
- File uploads for further description

- Creates NeXus files from JSON using JSON Schema
- Reads data from ASCII (key/value, table, xml – with XPath)
- Writes to NeXus path
- Creates XML ingest files for ICAT

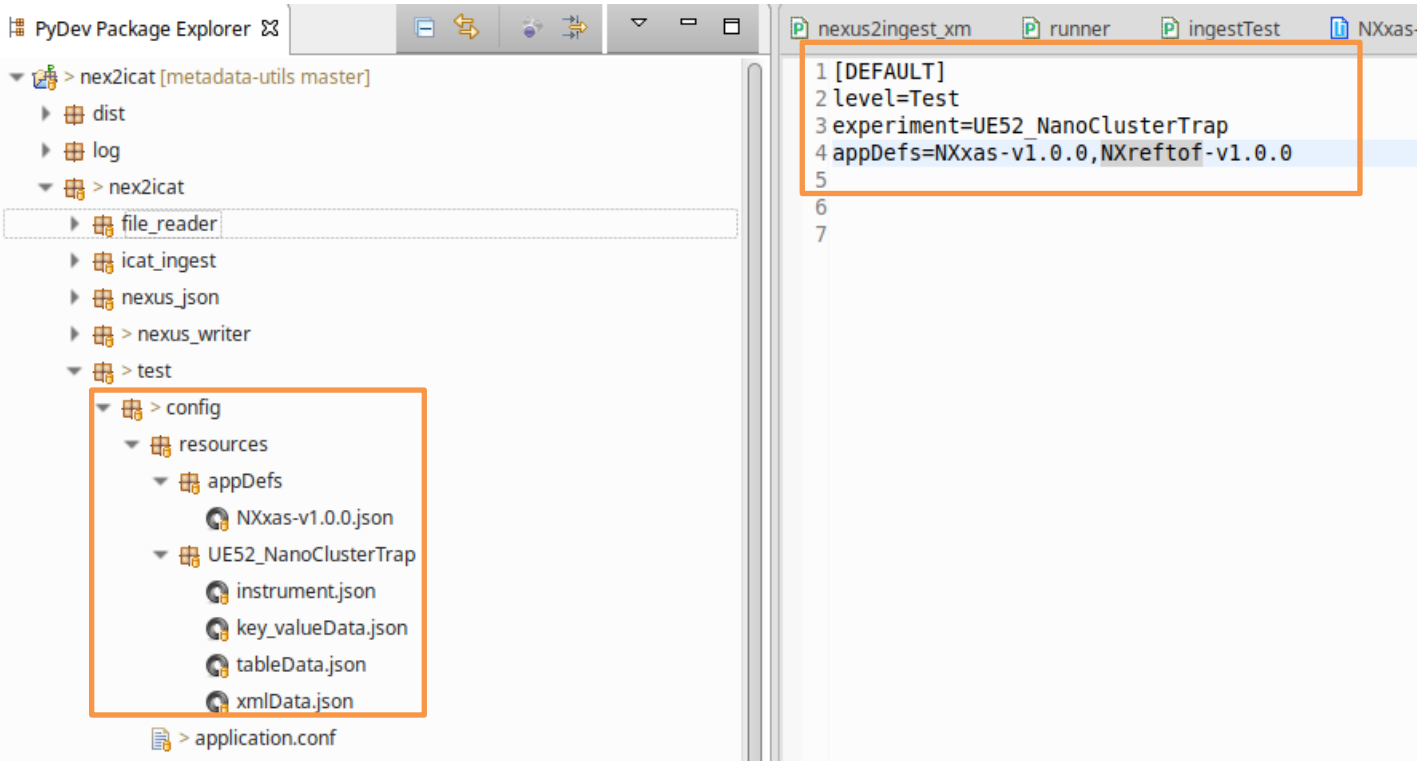
- Used in all applications as common format
- Instrument descriptions and application definitions
- Modular composition

- Reads JSON Schema application definitions or instruments
- Inserts initial data
- Edits data

INSTRUMENT DATABASE

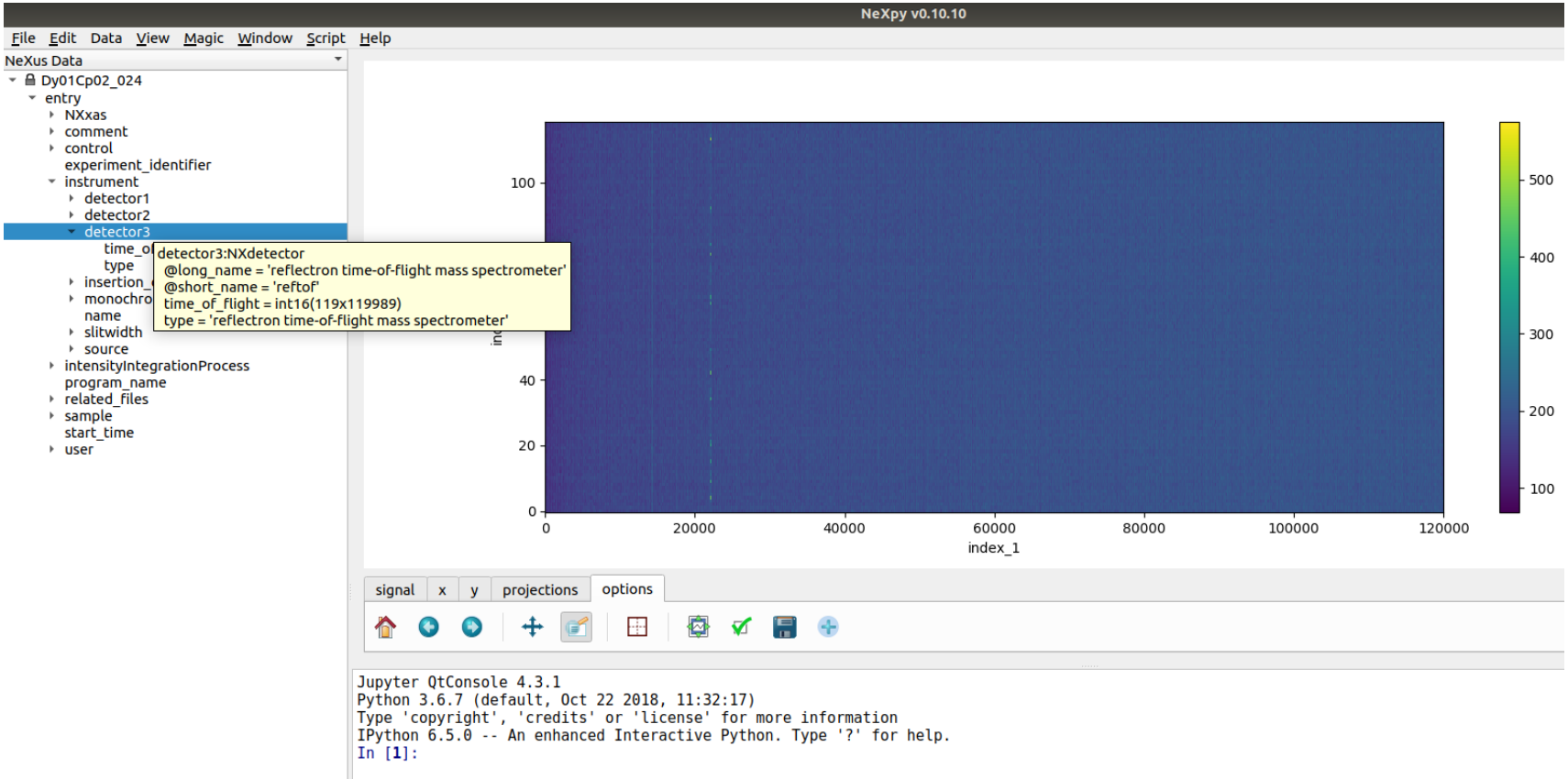


NEXUS WRITER I



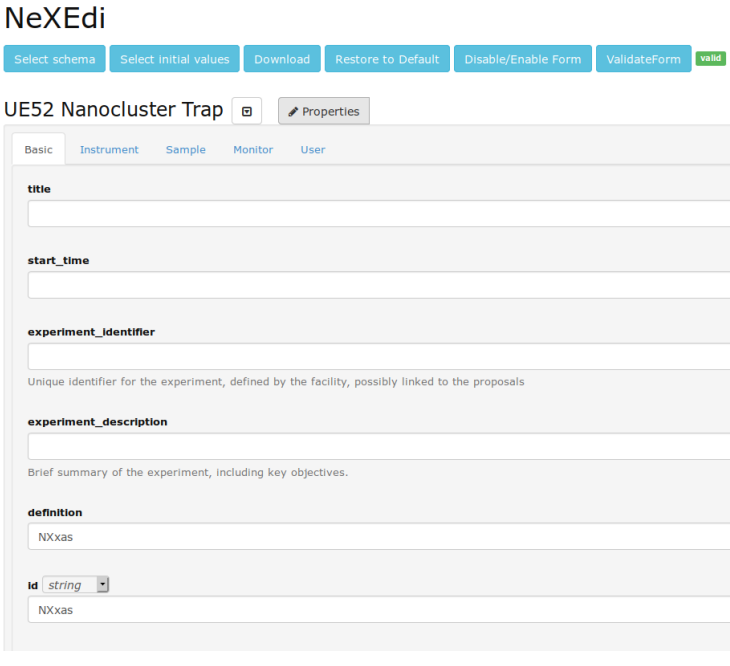
NEXUS WRITER II

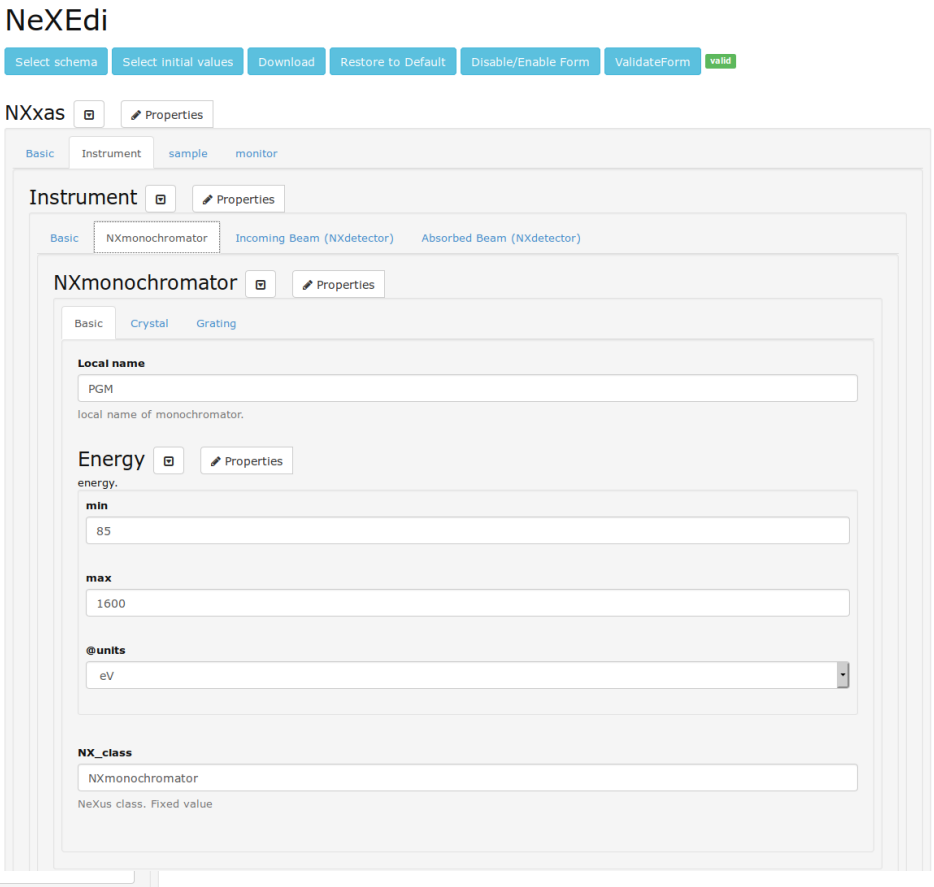
HZB Helmholtz Zentrum Berlin



HZB Helmholtz Zentrum Berlin

- **Loads fixed values from instrument database (to be implemented)**
- **Can be edited before measurement**





HZB Helmholtz Zentrum Berlin

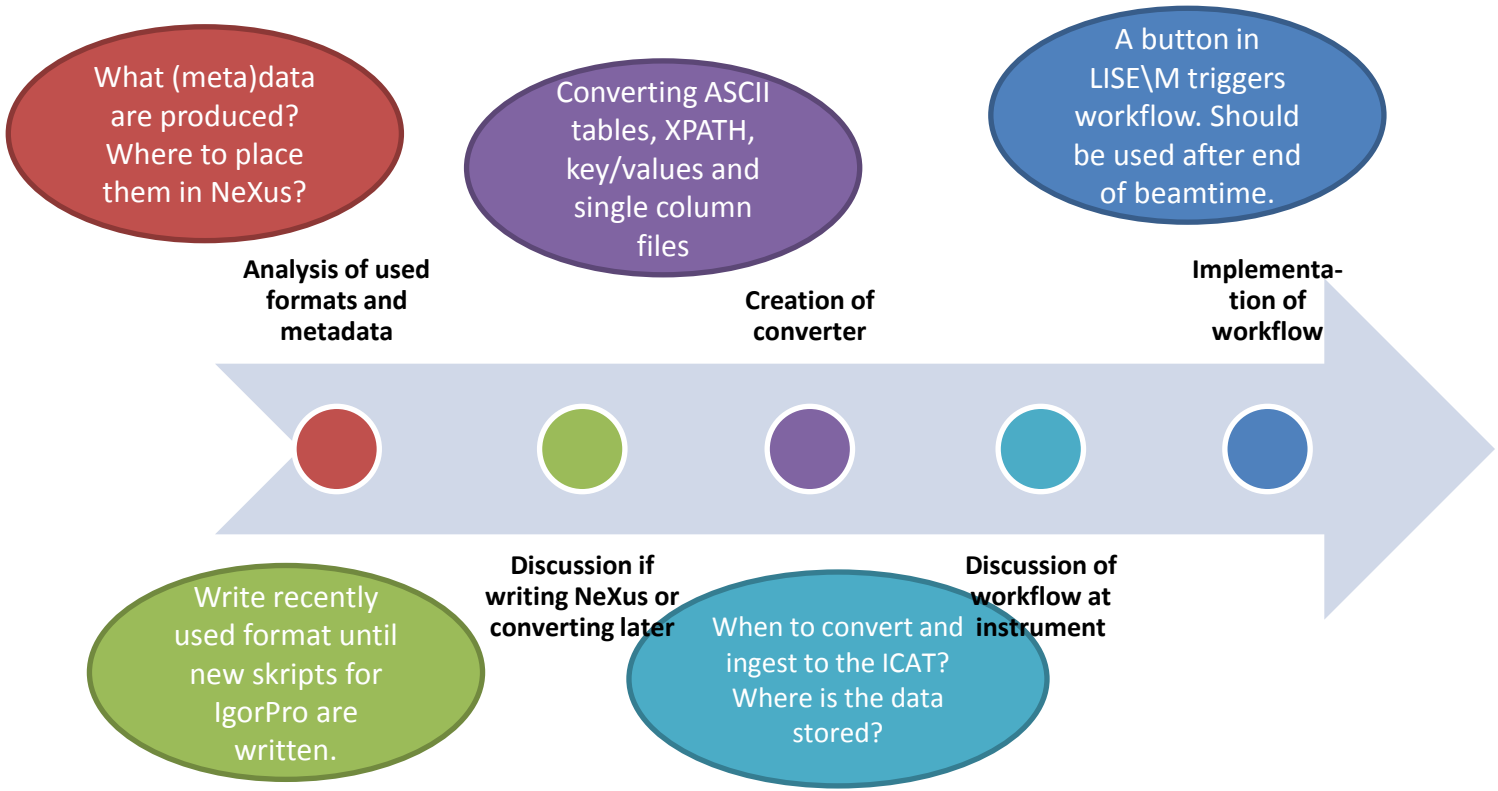
- RDMatDB project with HZDR
  - **NeXus paths as parameter names for ICAT**
- Key words
- Namespaces
  - **nxs/**

Are there other?
May be facilitate
ingest to NOMAD

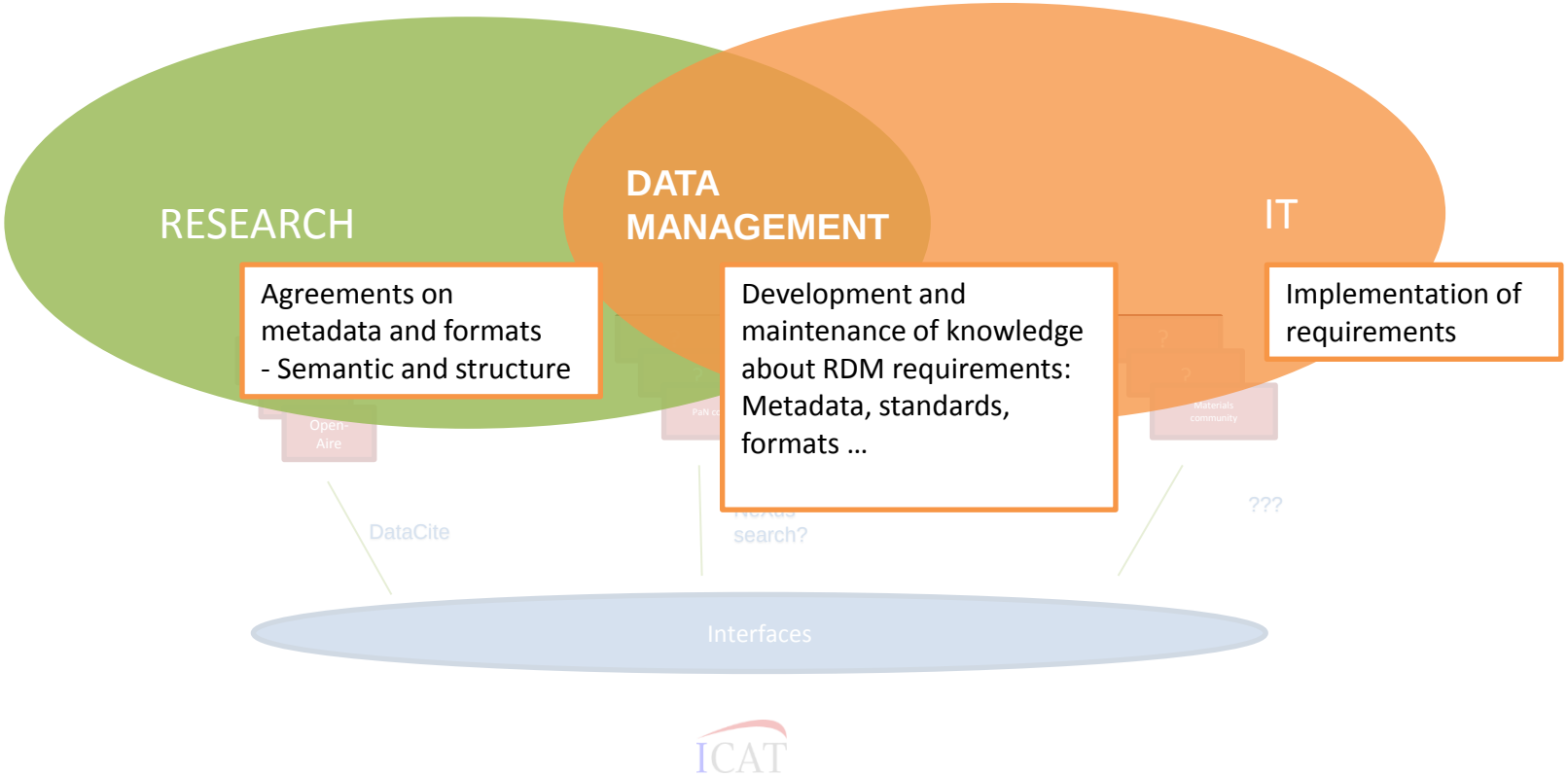
```
−<datasetParameter>
    <stringValue>incoming_beam</stringValue>
    <dataset ref="Dataset_1"/>
    <type name="nxs/entry/instrument/detector1/@long_name"/>
  </datasetParameter>
−<datasetParameter>
    <stringValue>xas1</stringValue>
    <dataset ref="Dataset_1"/>
    <type name="nxs/entry/instrument/detector1/@short_name"/>
  </datasetParameter>
−<datasetParameter>
    <numericValue>3.385958474576271e-05</numericValue>
    <rangeBottom>2.7892e-05</rangeBottom>
    <rangeTop>4.1027e-05</rangeTop>
    <dataset ref="Dataset_1"/>
    <type name="nxs/entry/instrument/detector1/data" units="A",
  </datasetParameter>
−<datasetParameter>
    <stringValue>photo diode</stringValue>
    <dataset ref="Dataset_1"/>
    <type name="nxs/entry/instrument/detector1/type"/>
  </datasetParameter>
−<datasetParameter>
    <stringValue>absorbed_beam</stringValue>
    <dataset ref="Dataset_1"/>
    <type name="nxs/entry/instrument/detector2/@long_name"/>
  </datasetParameter>
```

HZB Helmholtz Zentrum Berlin

WORKFLOW FOR CREATING CONNECTION



HOW DO WE GET THERE?



THANK YOU !