DAta for PHoton and Neutron Experiments

Bridget Murphy
CAU - Kiel

C A U
Christian-Albrechts-Universität zu Kiel

DAPHNE4NFDI aims

Christian-Albrechts-Universität zu Kiel

to make the growing volume of valuable measured data FAIR for the DAPHNE4NFDI community, for the whole NFDI and the scientific community.

These key objectives will be achieved within DAPHNE:

Reuse data by:

1. **Definition and collection** of **metadata**
2. **Databases** of data
3. **Curated repository of managed software**
4. **Multidisciplinary data platform**
5. **Education** and **training** in research data management.

Nationale Forschungsdaten Infrastruktur

DAPHNE 4NFDI

Bring together

- Universities
- Research Institutes
- Users Organisations
- Facilities

DAPHNE4NFDI

Data for Photon and Neutron Experiments

Consortium

Brings together 18 partners:

- University user groups
- Large scale facilities
- In addition: KFN + KFS
- > 60 participants (without funding)

Task Area Leaders:

Anton Barty (DESY, Speaker)
Bridget Murphy (CAU, Speaker)
Astrid Schneidewind (FZJ, Deputy speaker)

Sebastian Busch (Hereon)

Frank Schreiber (U Tübingen)
Wiebke Lohstroh (TUM)
Christian Gutt (U Siegen)
Jan-Dierk Grunwaldt (KIT)
Tobias Unruh (FAU)

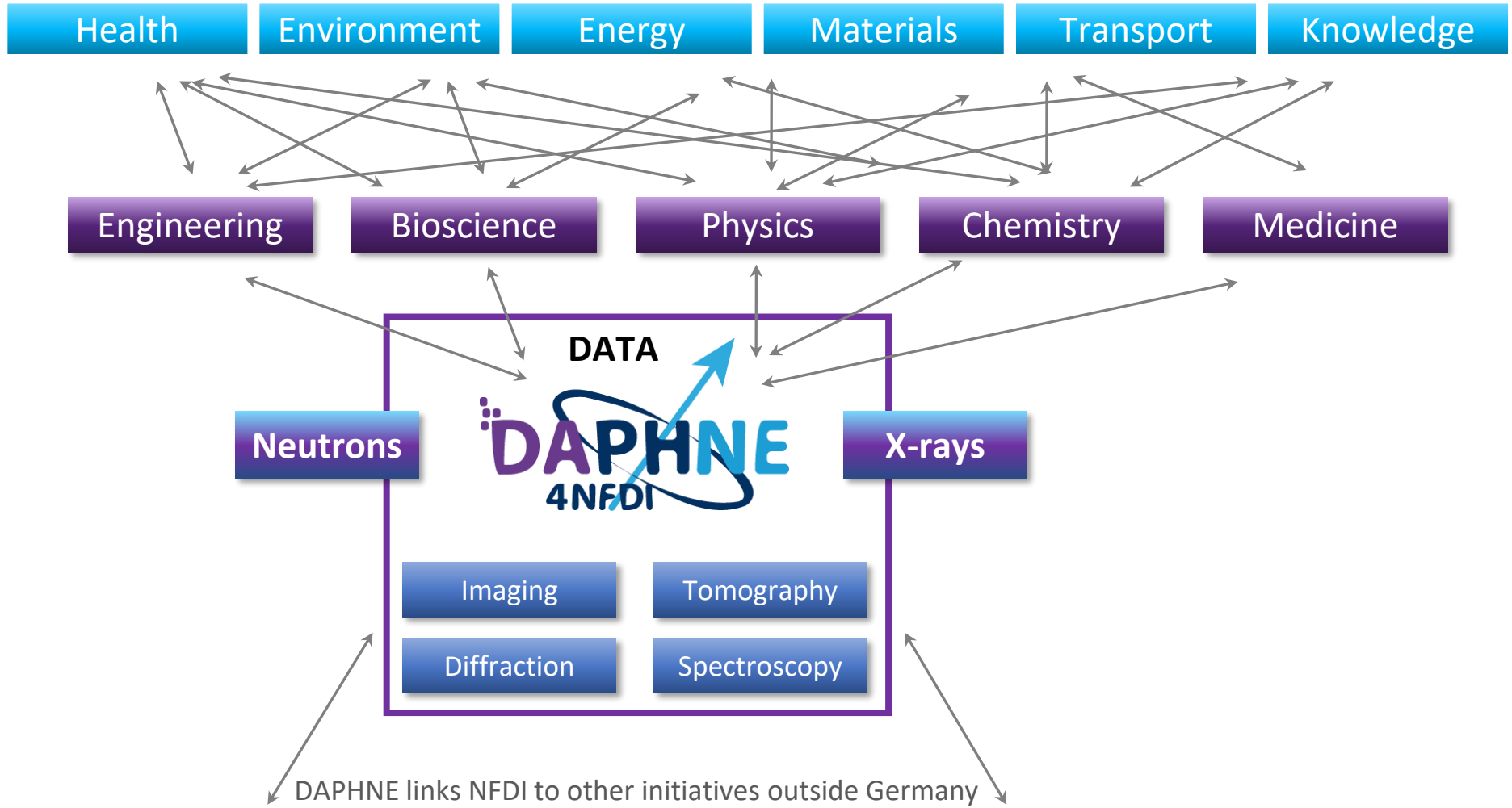
Coordinator: lisa.amelung@desy.de

DAPHNE4NFDI.de

DFG - NFDI

Impacts on global challenges

Impact extends far beyond the physics or materials science community

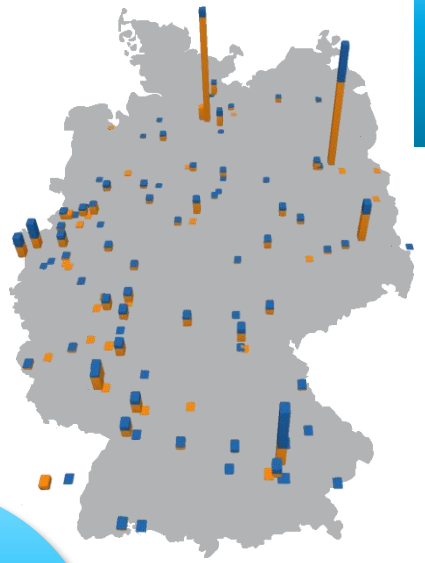


DAPHNE links NFDI to other initiatives outside Germany

Research with photons and neutrons in numbers

per year in Germany

8 sources in Germany
33 sources in Europe
94 sources worldwide



3000
participants at facility user meetings

3000 experiments

50 PB data

3000 publications

250 patents
In Europe

5500 users

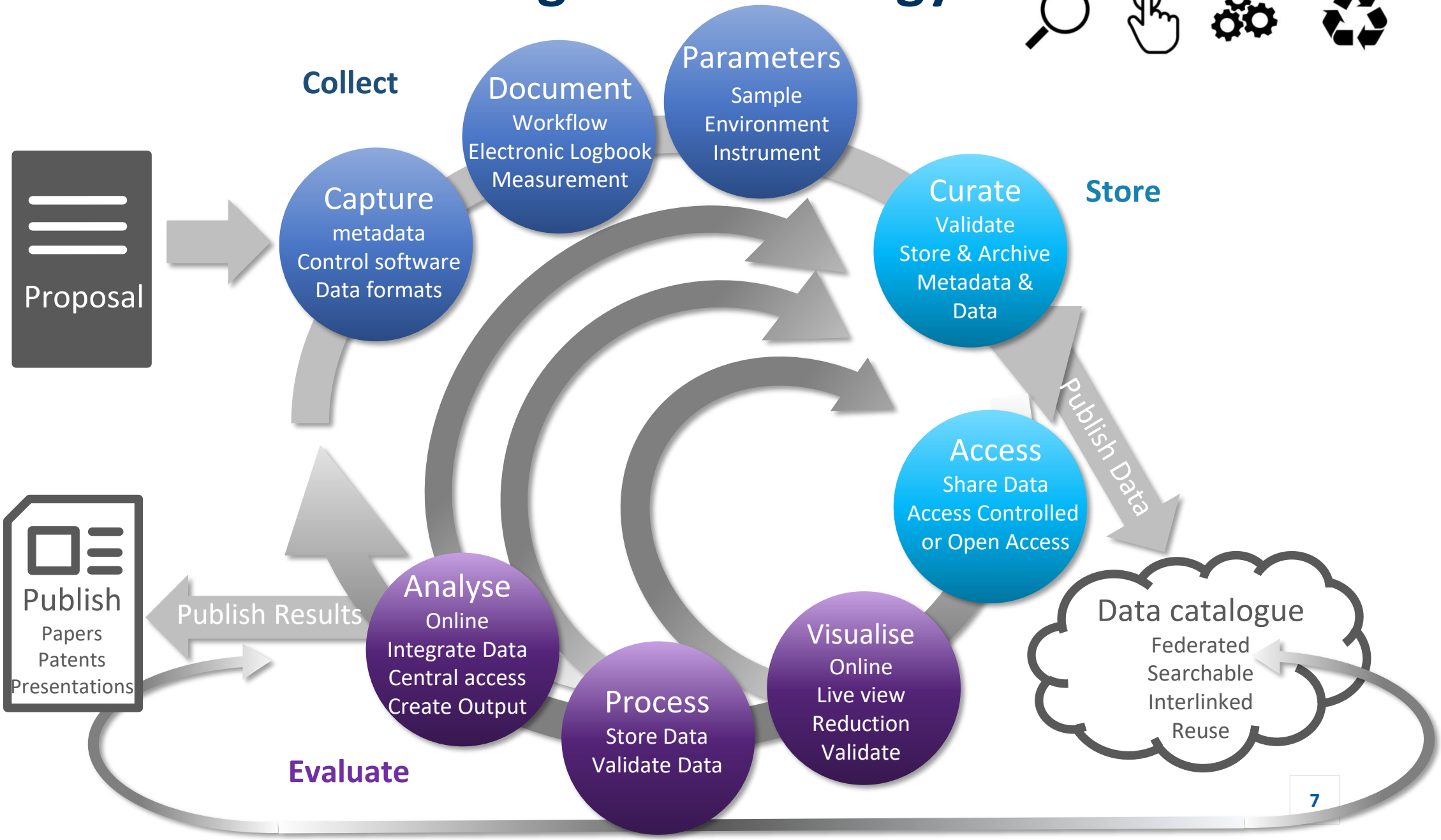
from 50 companies,
100 universities,
115 research institutions

Reaching a community of over 50 000

Christian-Albrechts-Universität zu Kiel

Research Data Management strategy

Findable Accessible Interoperable Reusable



Managing data and metadata collection

Task Area leaders: Wiebke Lohstroh (TUM) and Bridget Murphy (CAU-Kiel),

supported by Philipp Jordt (CAU-Kiel)

Instrument and sample data capture

- Capture data and metadata at instrument (sample environment and sample data)
  - Electronic log book
  - Define and capture metadata for catalogue
  - Sample persistent identifier (IGSN)
- Automatic ingestion
- Authentication

High performance data format standards

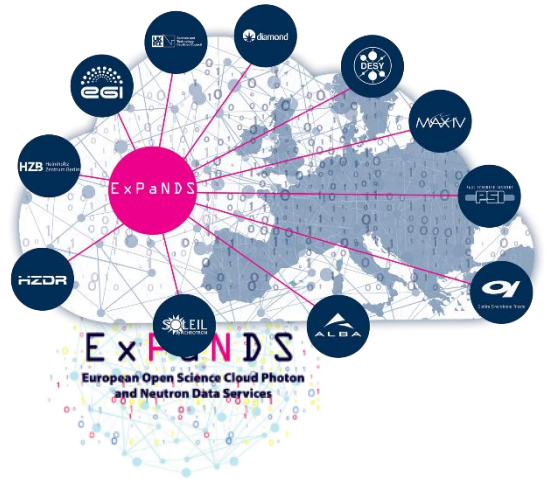
- NeXus data standards and container deployment
  - MLZ, PETRA III, BESSY, HZDR (openPMD), X-spectrum and EuXFEL

Managing data and metadata collection

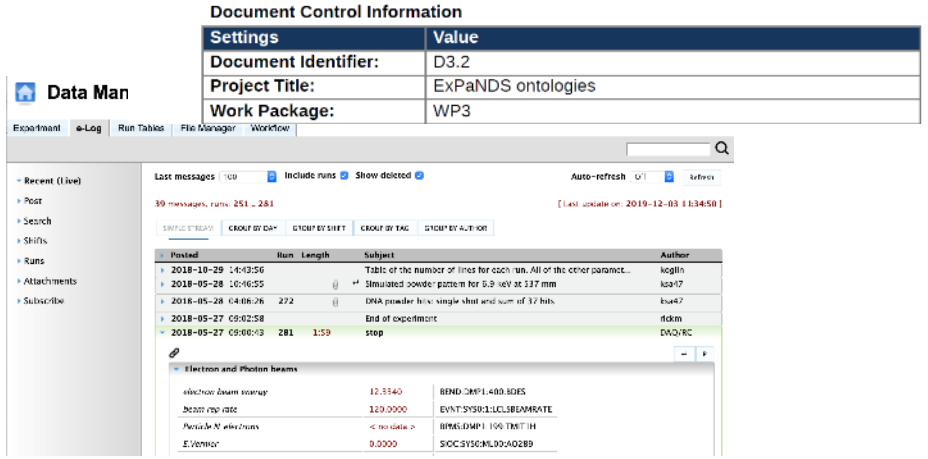
Enabling re-use and repeatability of results, ideally searchable

Meta data schemata and vocabulary

- Specification, develop ExPaNDS ontologies
- Automatic ingestion during experiments
- Implementation (use cases)
- Standards – White paper in progress



ExPaNDS ontologies



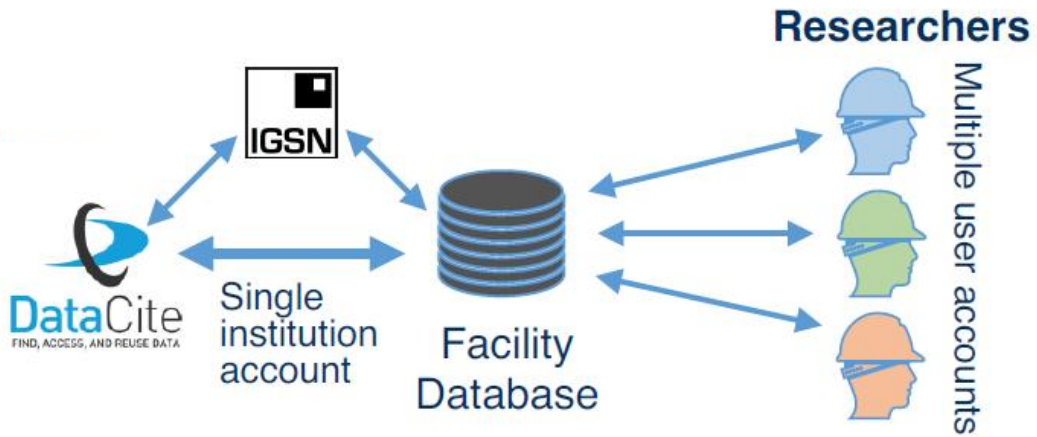
Persistent Sample Identifiers PID

Christian-Albrechts-Universität zu Kiel

Joint working group

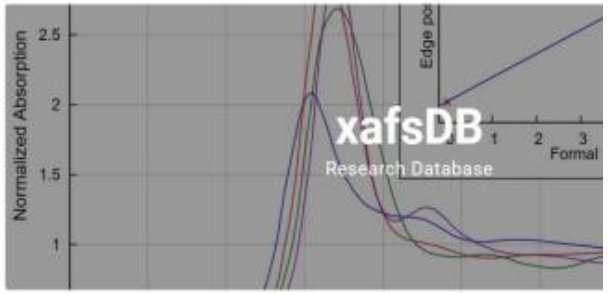
International Generic sample number (IGSN)

- Tests of IGSN at MLZ and CAU.
  - Identifier needs to be unique and persistant
  - Simple to use
  - Which (meta)data information is required?
  - When to catalogue ?



Metadata and data repositories and catalogues

A place to find published data - and in some cases the ability to reprocess data



Repository and catalogue roll-out and development

- Presentation later

(Meta)data standardisation and sample identification

- Focus on which information should be available
  - Overlap with TA1 (joint working group)
- PaNOSC search API
- (Meta)data standardisation and sample identification
  - NeXus/HDF5
  - openPMD (HZDR)
- Insertion of additional (meta)data into repository/catalogue

SciCat - DAPHNE4NFDI standard

Christian-Albrechts-Universität zu Kiel

Cross TA development with a focus working group

May 2023: new backend

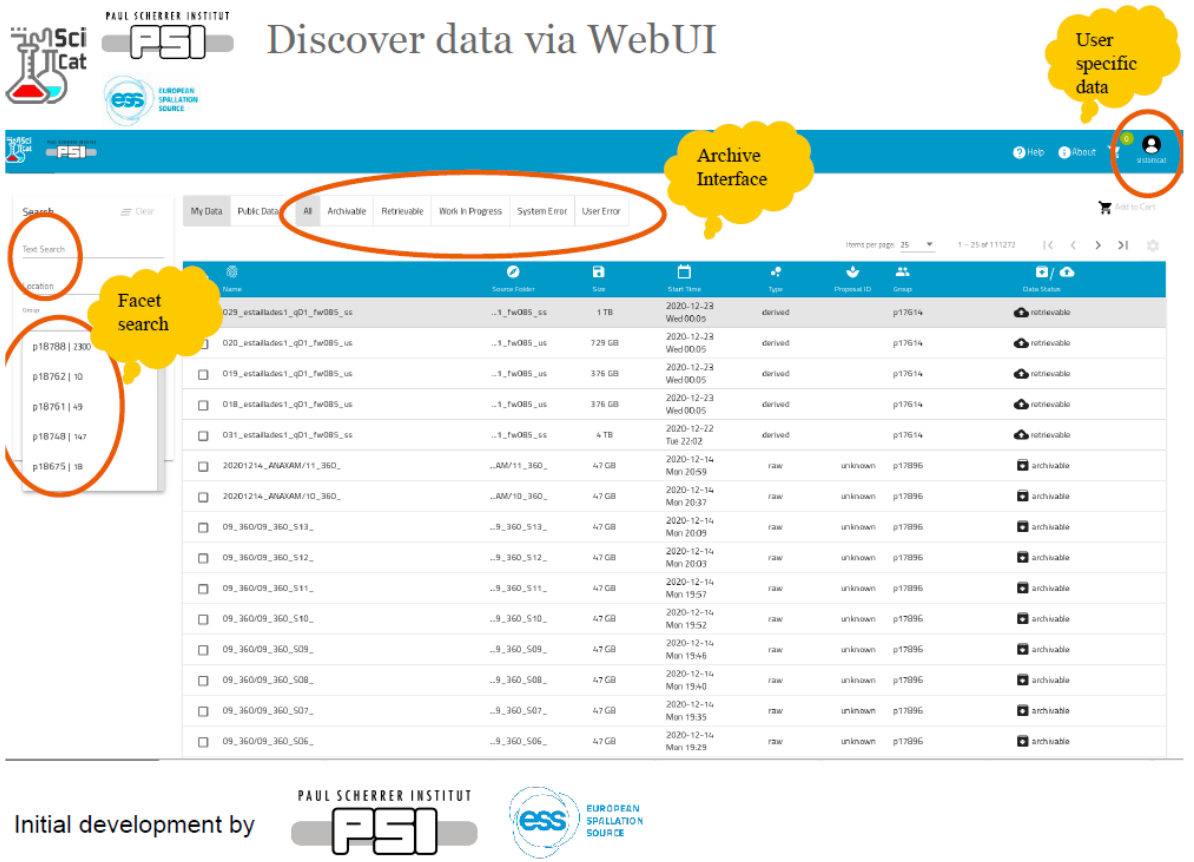
Facilities

- MLZ, DESY, and HZDR installed SciCat
- BESSY(ICAT), EuXFEL (MyMDC)

- Automatic ingestion
  - Test user version running at P08 DESY
- Authentication
  - User rights difficult at DESY
  - keycloak working at MLZ

Universities

- Local installations led by Tübingen
- Currently @ KIT, TUB, BUW, CAU, FAU
- Central IT installations slower



Active in SciCat developers group

Metadata and data repositories and catalogues

Task Area leaders: Sebastian Busch (hereon) and Tobias Unruh (FAU),

supported by Jonas Graetz (FAU)

Activities and Achievements

- catalogues,
- (meta-)data formats
- sample databases

- SciCat installation training (DAPHNE4NFDI standard)
- X-ray absorption spectroscopy reference database
- Tests of IGSN sample identifiers at MLZ and CAU.

Metadata and data repositories and catalogues

A place to find published data - and in some cases the ability to reprocess data

Christian-Albrechts-Universität zu Kiel

Insertion of additional (meta)data into repository/catalogue at a later stage

- X-ray absorption spectroscopy (XANES/EXAFS) database first
  - XFS next phase
  - KIT, Uni. Wuppertal and TU Berlin, ESRF
- Round robin sample
- SciCat as backend

Search data

- Using GID data set as test case (HZDR)

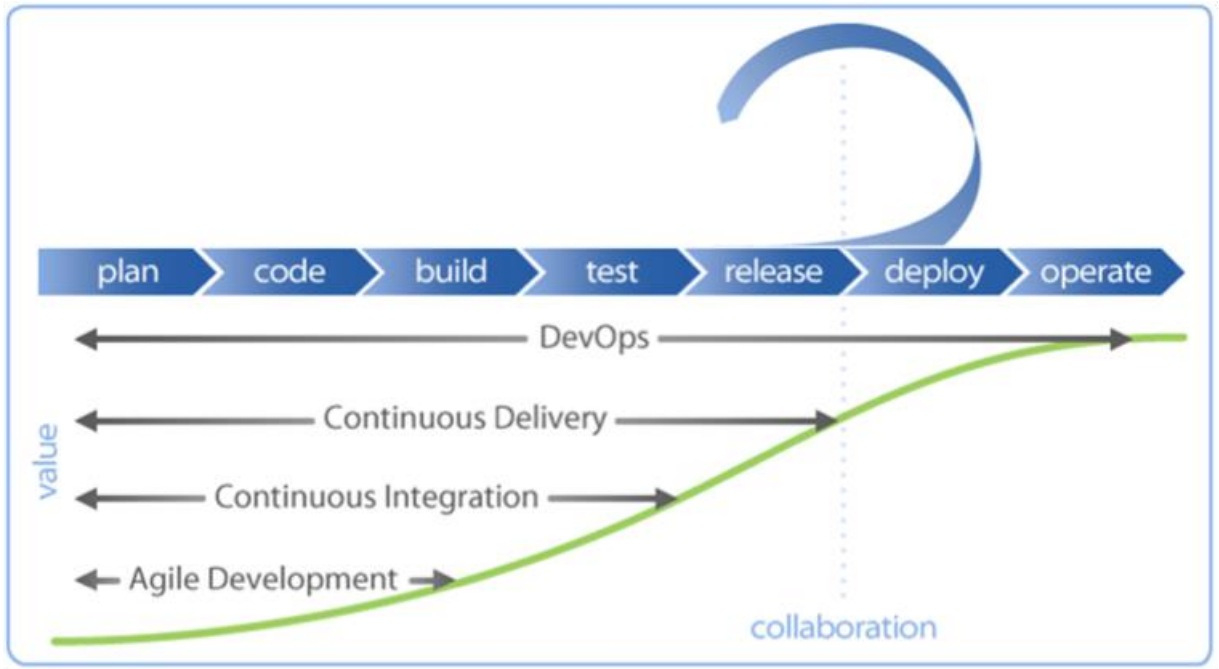
Open infrastructure for data and software re-use

Task area leaders: Frank Schreiber (Uni. Tübingen) and Anton Barty (DESY),

supported by Linus Pithan (Uni. Tübingen)

Activities and Achievements

- Infrastructure and training for professional research software engineering practices
- Providing interfaces for machine learning software
- Developing community software for use cases



Open infrastructure for data and software re-use

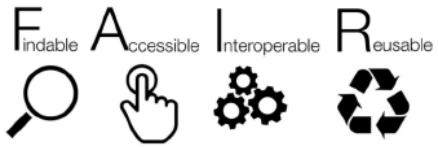
Christian-Albrechts-Universität zu Kiel

Community data analysis software and data mining strategies including machine learning

Supporting infrastructure and related developments

GitLab

- GitLab
- Continuous Integration & Continuous Deployment (CI/CD)platform
- Support for software development. (e.g. CrystFEL use case)
- Data analysis platforms based on the VISA developments of ExPaNDS and PaNOSC
- Interfaces for machine learning using SciCat to access data
- Aim to define common language and file formats



Open infrastructure for data and software re-use

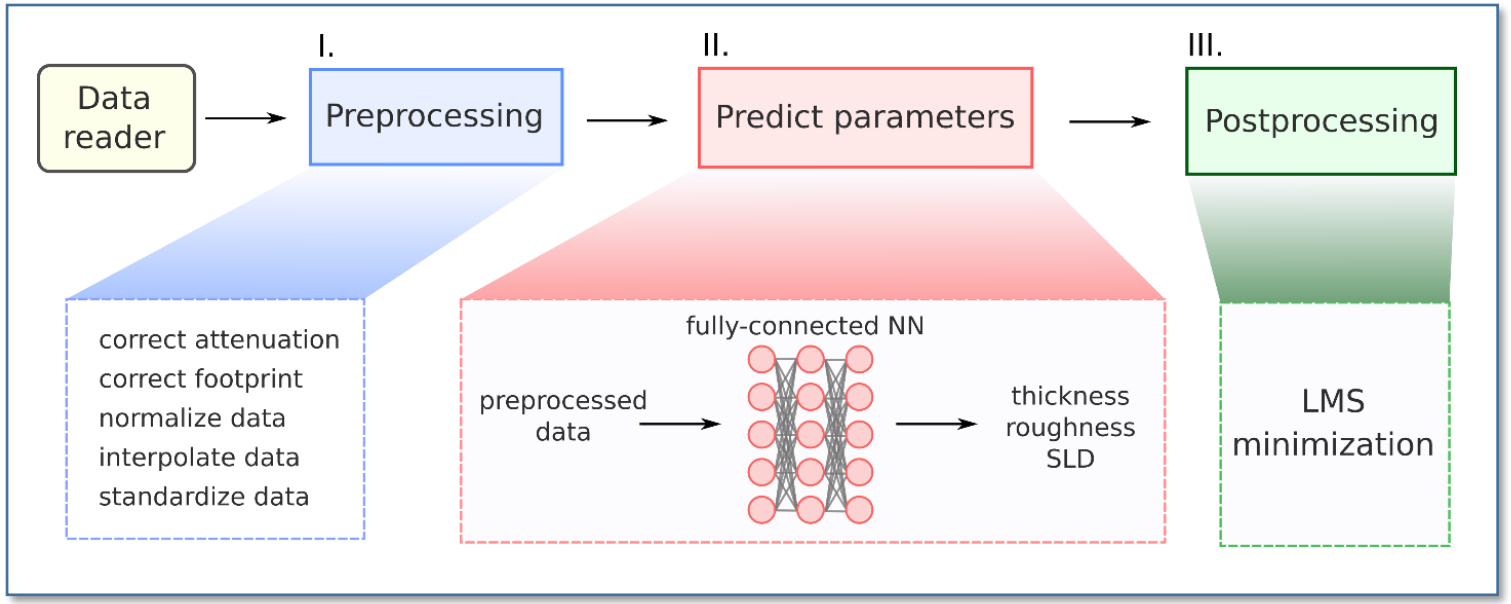
Christian-Albrechts-Universität zu Kiel

Community data analysis software and data mining strategies including machine learning

Machine learning based software interfaces

- Deep learning 2D data analysis pipeline GIWAXS (Uni. Tübingen)
- Reflectivity organic thin films (Uni Tübingen), liquid samples (CAU Kiel)
- Amortized bayesian inference of of GISAXS (HZDR, XFEL, Uni Siegen)
- Powder diffraction (At RWTH Aachen)

The *mlreflect* pipeline

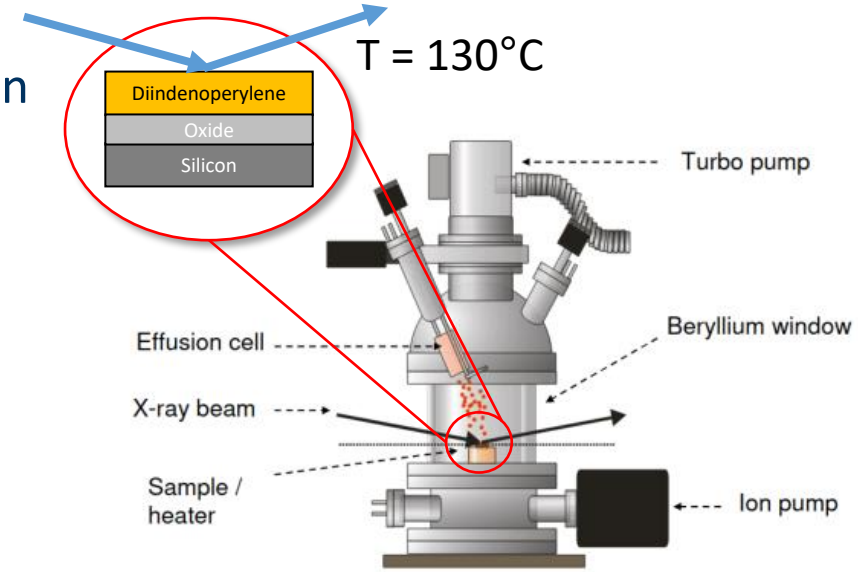


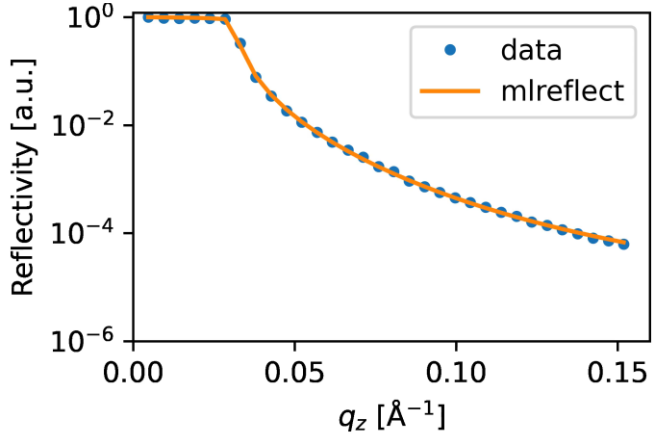
Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

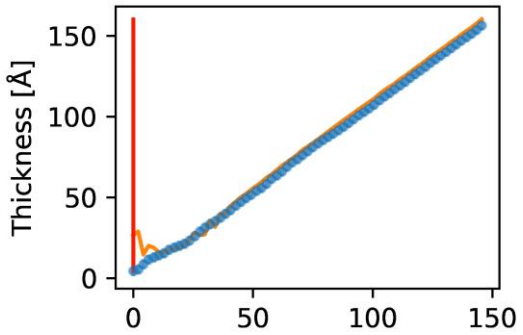
In situ applications of mlreflect

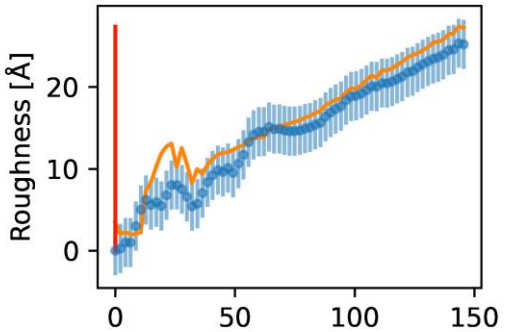
In situ XRR during film deposition

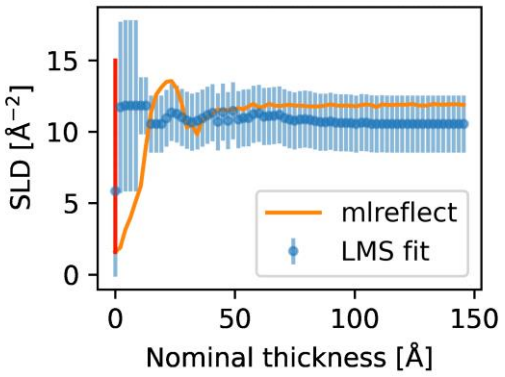
- Real-time parameter prediction useful for in situ experiments
- After training, no human input is necessary
- Results are obtained within <1s per curve
- Ideal for monitoring and feedback loops











Hinderhofer *et al. Europhys. Lett.*, 2010, **91**, 56002
Kowarik *et al. Phys. Rev. Lett.*, 2006, **96**, 125504
Bommel *et al. Nat. Comm.*, 2014, **5**, 5388
Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

Pithan *et al. submitted* 2023, **https://arxiv.org/abs/2306.11899**

Open infrastructure for data and software re-use

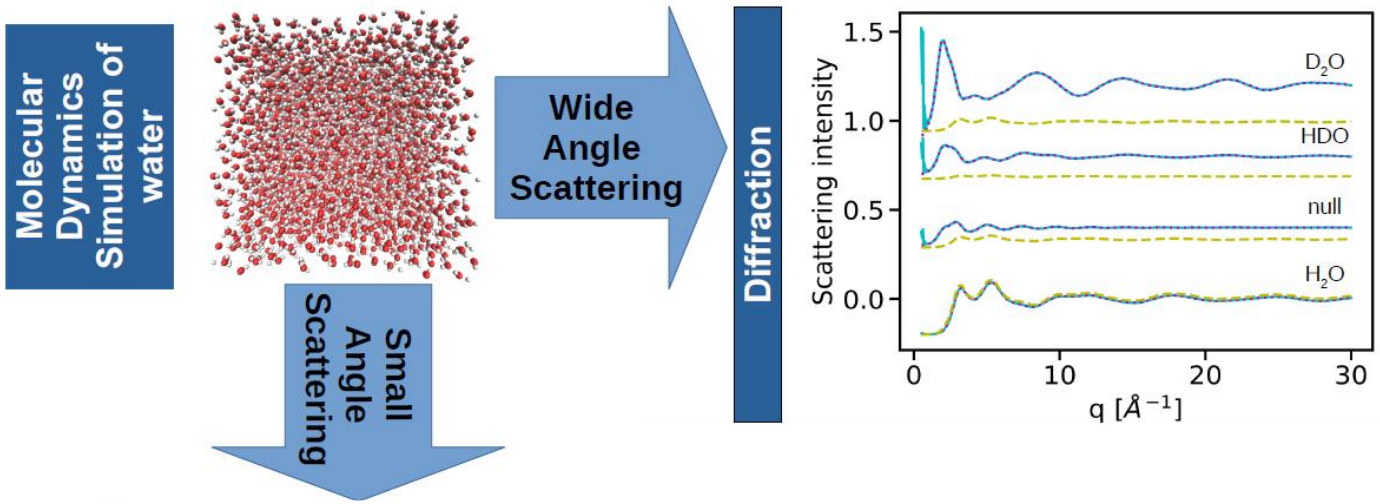
Christian-Albrechts-Universität zu Kiel

Community data analysis software and data mining strategies including machine learning

Development of scientific software

- Grazing incidence small angle scattering (AIXTAL@RWTH Aachen)
- X/n diffraction and quasielastic neutron scattering (Sassena@MLZ)
- X/n reflectivity (XRR @CAU-Kiel, Made2Reflect@MLZ)
- CrystFEL (DESY)

Sassena

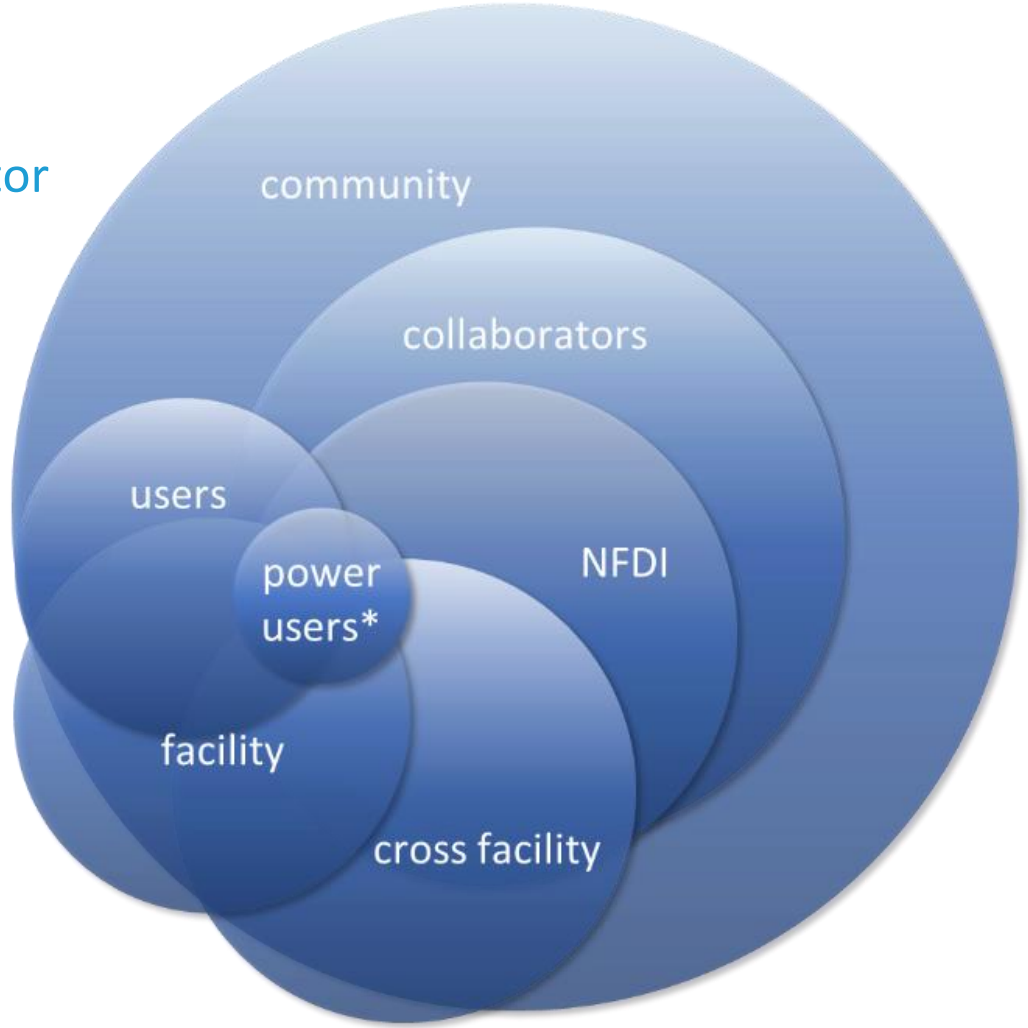


Dissemination and outreach

Task area leaders: Astrid Schneidewind (FZJ) and Jan-Dierk Grunwaldt (KIT), supported by Paolo Dolcet(KIT)

The NFDI consortium as a role model and educator

- Workshops and community building - (meta)data definition and ontologies
- Discuss and Support use cases
- Data managment as part of the curriculum
- Status meetings and TA meetings
- Webpage/ highlight reports
- University open lecture series



Use Cases

DAPHNE as a role model

Biomaterials
x-ray imaging
LMU – Uni Göttingen

Energy and battery materials, catalysis
Tomography
TUM – MLZ – BAM – hereon – HZB - KIT

Amorphous materials for catalysis
x-ray absorption spectroscopy
KIT – TUB – Uni Wuppertal

Soft matter and liquid interfaces
x-ray reflectivity
Uni Kiel – Uni Tübingen

Proteins & Food science
Diffraction (small and wide angle)
Spectroscopy
FAU– Uni Tübingen – EMBL - Uni Kiel

Reusable powder refinement
Neutron TOF diffraction
FZJ – MLZ – ESS - RWTH

Dynamics
Correlation spectroscopy - XPCS
Uni Siegen - EuXFEL

Correlated electron systems
Spectroscopy
KIT – FZJ - MLZ

Magnetic structures
Ultrafast / Magnetic x-ray scattering
DESY – Uni Siegen

Chemical systems
x-ray emission spectra, RIXS etc.
KIT – ESRF - DESY

Electrochemistry & Catalysis
High energy x-ray diffraction
HZDR – Uni Kiel - DESY

External communication and policy

Christian Gutt (U Siegen) and Astrid Schneidewind (FZJ)

- DAPHNE is embedded in worldwide network > 30.000 synchrotron and neutron users
- Cross-consortia activities – interfaces and interlinkages to other NFDI consortia
- Organizational structures exists: European user organizations and facility organization
- Connects to European open science cloud X-ray and neutron data projects

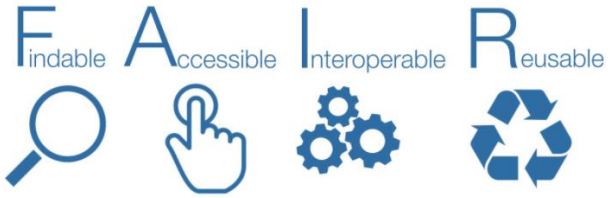
Nationale Forschungsdateninfrastruktur (NFDI) e.V.

NFDI Vision

Research data are available in a FAIR* manner

For everyone**

For always***.



- Findable, accessible, interoperable, reusable.
- **depending on access class
- *** Alice; “How long is for ever”
    - White rabbit “sometimes for one second”

York Sure-Vetter Director NFDI associaton

The main objective of DAPHNE4NFDI

is to make the growing volume of valuable measured data FAIR for the DAPHNE4NFDI community, for the whole NFDI and the scientific community.

These key objectives will be achieved within DAPHNE:

1. **Collection** of **data** and **metadata** so that the **measured data** is **reusable**
2. **Searchable curated databases** of raw, intermediate and processed data
3. Develop a **curated repository of managed software >> re-use** the data
4. **Education** and **training** in research data management
5. Develop **multidisciplinary data platforms** for NFDI and international collaborations

https://www.daphne4nfdi.de


Thank you!
DAPHNE
4NFDI