DAPHNE 4NFDI

DAta for PHoton and Neutron Experiments

DPG Tagung 2021

Sept. 28, 2021

Presenting

Anton Barty

DESY photon science

Christian Gutt

KFS

UNIVERSITÄT SIEGEN

Astrid Schneidewind

KFN

JÜLICH Forschungszentrum

Bridget Murphy

KFS

Christian-Albrechts-Universität zu Kiel

Wiebke Lohstroh

KFN

Technische Universität München

DFG - NFDI

DAPHNE 4NFDI

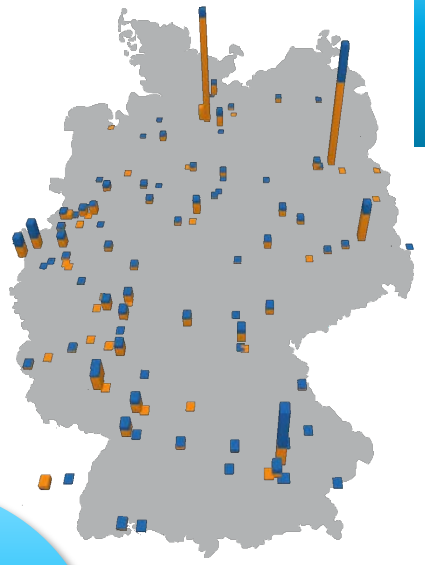
Bring together

- KFS-Committee for Synchrotron Research
- KFN-Committee for Neutron Research
- Large-scale photon and neutron research facilities
- Universities
- Research institutions
- Wider community

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

28 PB data

3000 publications

5500 users

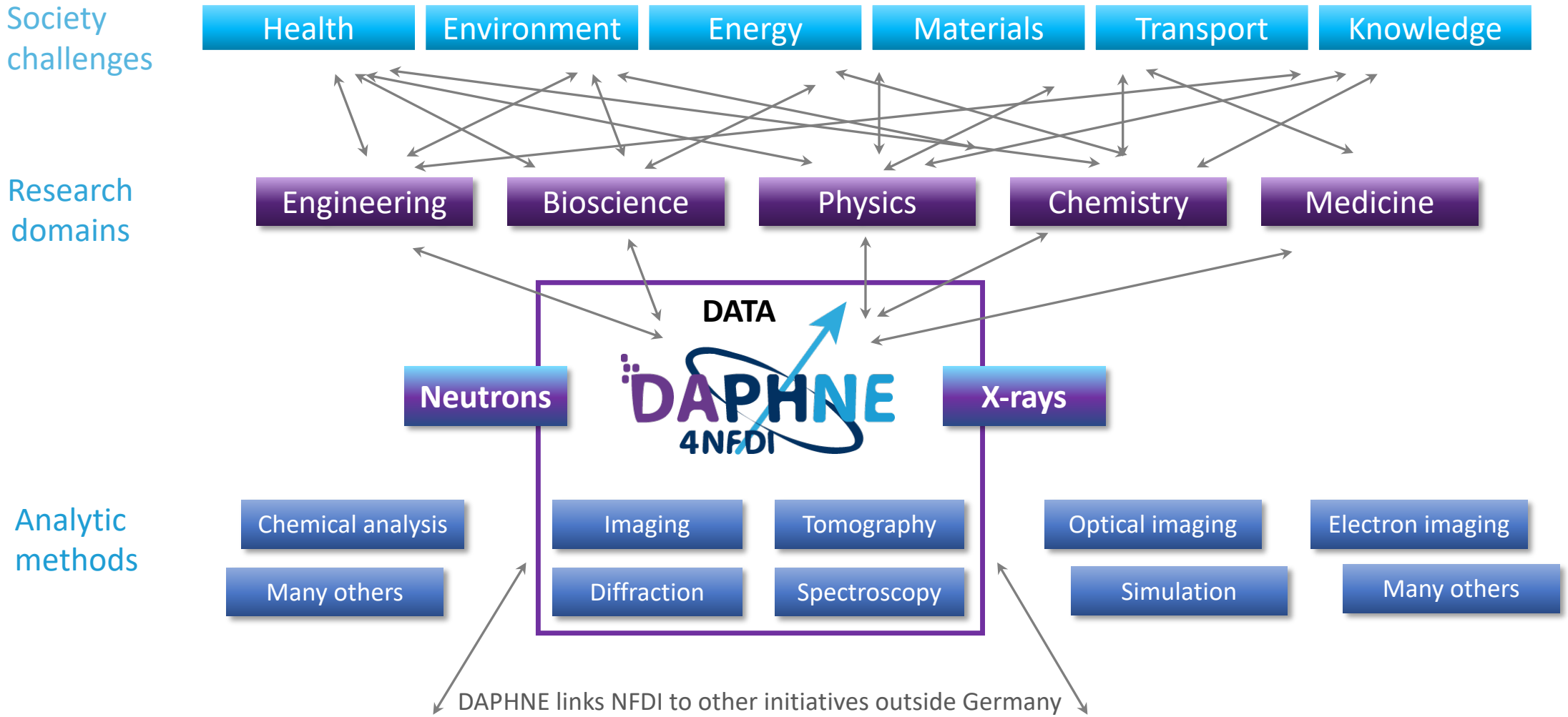
from 50 companies,
100 universities,
115 research institutions

Reaching a
community of
over 50 000

250 patents
In Europe

Our research community impacts on global challenges

Impact extends far beyond the physics or materials science community



DAPHNE is integrated into the international research community

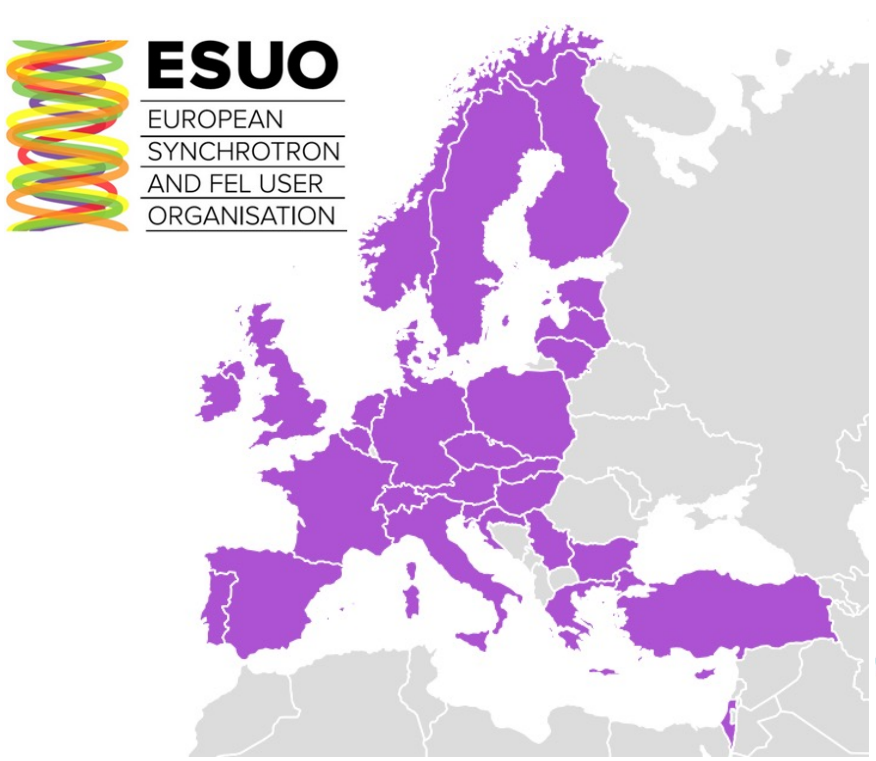
The Neutron and X-ray community is international

Making X-ray and neutron data FAIR in Europe

DAPHNE embedded in a network of > 30.000 synchrotron and neutron users

Organizational structures exists: European user organizations and facility organization

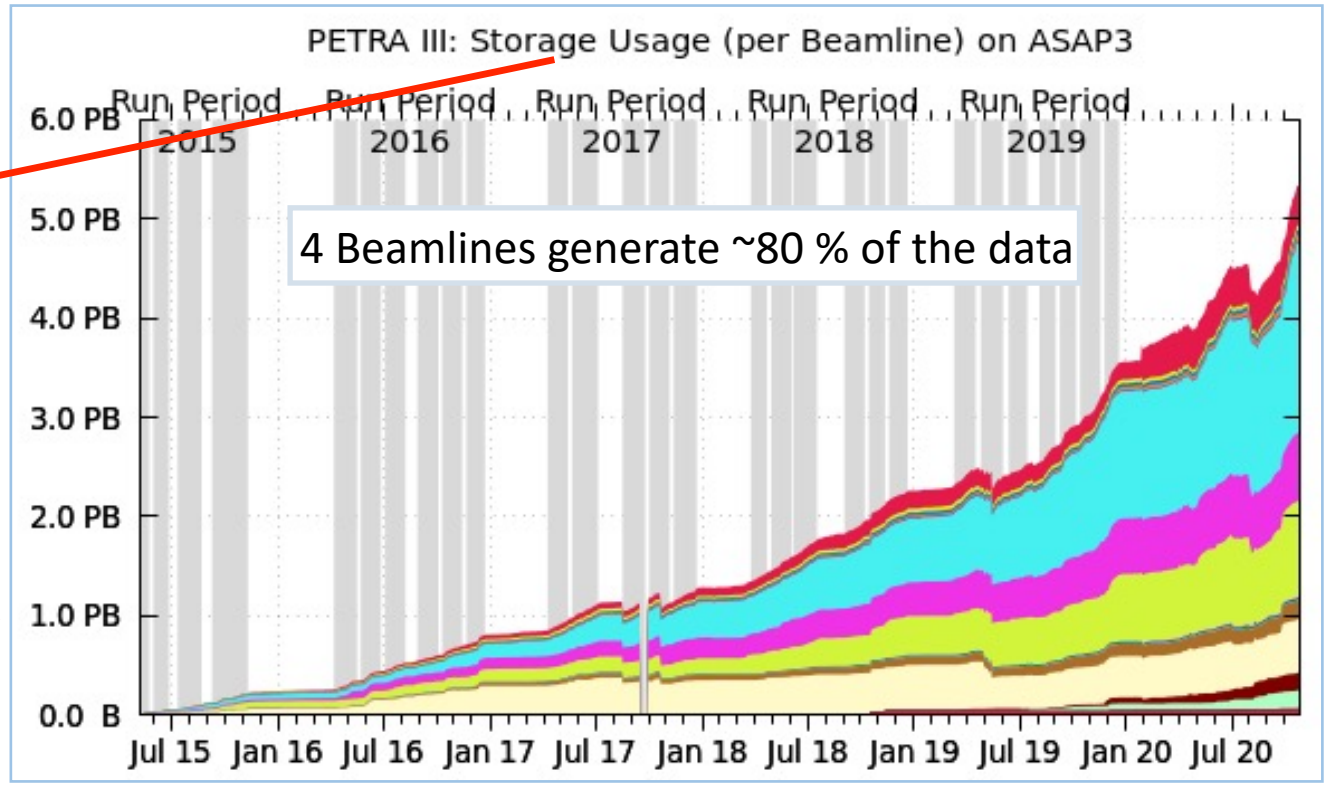
DAPHNE connects to European open science cloud X-ray and neutron data projects PaNosc and ExPands



ENSA · LENS League of Advanced European Neutron Sources · pandata europe · panosc photon and neutron open science cloud · LEAPS League of European Accelerator-based Photon Sources · ExPaNDS · ESUO European Synchrotron and FEL User Organisation · European Open Science Cloud

5

Data rates



New detectors

Eiger2 at P11
(installed 2020)

500 fps, 12 MPix
12MB/image
21 TB/hr (6 GB/s)

2PB free space
'full in a few days'

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

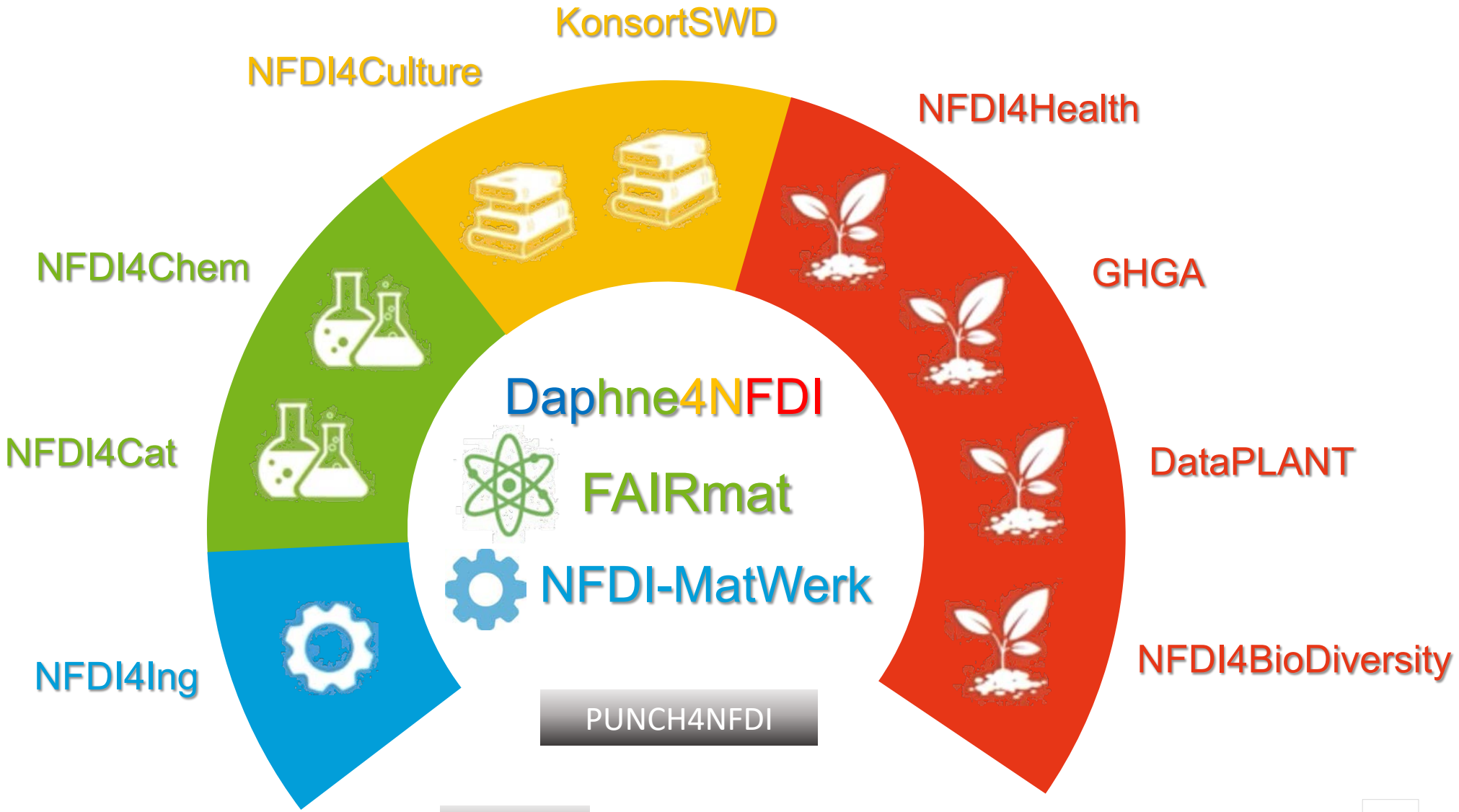
The main objective of DAPHNE4NFDI

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

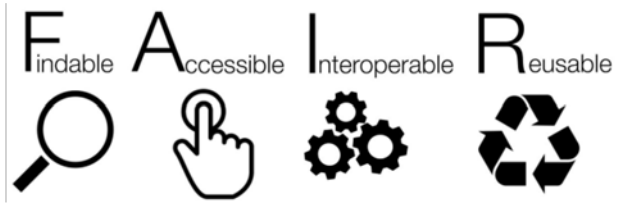
These key objectives will be achieved within DAPHNE:

1. Improve the **collection** of **metadata** about the measurement so that the **measured data** is **reusable** by the wider research community;
2. Implement **searchable curated databases** of raw, intermediate and processed data, **traceable** from **published and unpublished results**;
3. Develop a **curated repository of managed software** developed by leading research groups, accessible to any researcher so that others can repeat the data analysis pipelines and **re-use** the code in their own research;
4. Develop a **multidisciplinary data platform** for NFDI cross-consortia actions;
5. Provide **education** and **training** in research data management.

Within NFDI


KonsortSWD
NFDI4Culture
NFDI4Health
NFDI4Chem
GHGA
Daphne4NFDI
NFDI4Cat
FAIRmat
DataPLANT
NFDI-MatWerk
NFDI4Ing
NFDI4BioDiversity
PUNCH4NFDI
MaRDI

RDM strategy



Task area 1: Managing metadata collection

Enabling re-use and repeatability of results, ideally searchable

Before experiment

Proposal
- *Proposal number (ID)*
- Science motivation
- Experiment concept
- Technique(s)
- Sample(s)
- Sample environment
- Instrumentation
- Science team

+

Facility
- *Which facility?*
- Beam parameters
- Instrumentation
- Detectors

Digital Sample ID
- 'DOI for samples'
- Cross link to other consortia

During experiment

Facility logs
- *Experiment ID (directory)*
- Beam parameters
- Motor positions
- Instrument configuration
- Sample environment
- Detector calibration

+

User record of experiment
- Actual samples used
- Actual sample environment
- Instrumentation configuration
- Changes to original plan
- What happened when
- Run log (data lookup table)
- Paper logbooks, google sheets, confluence, and more

After experiment

can be 750 TB

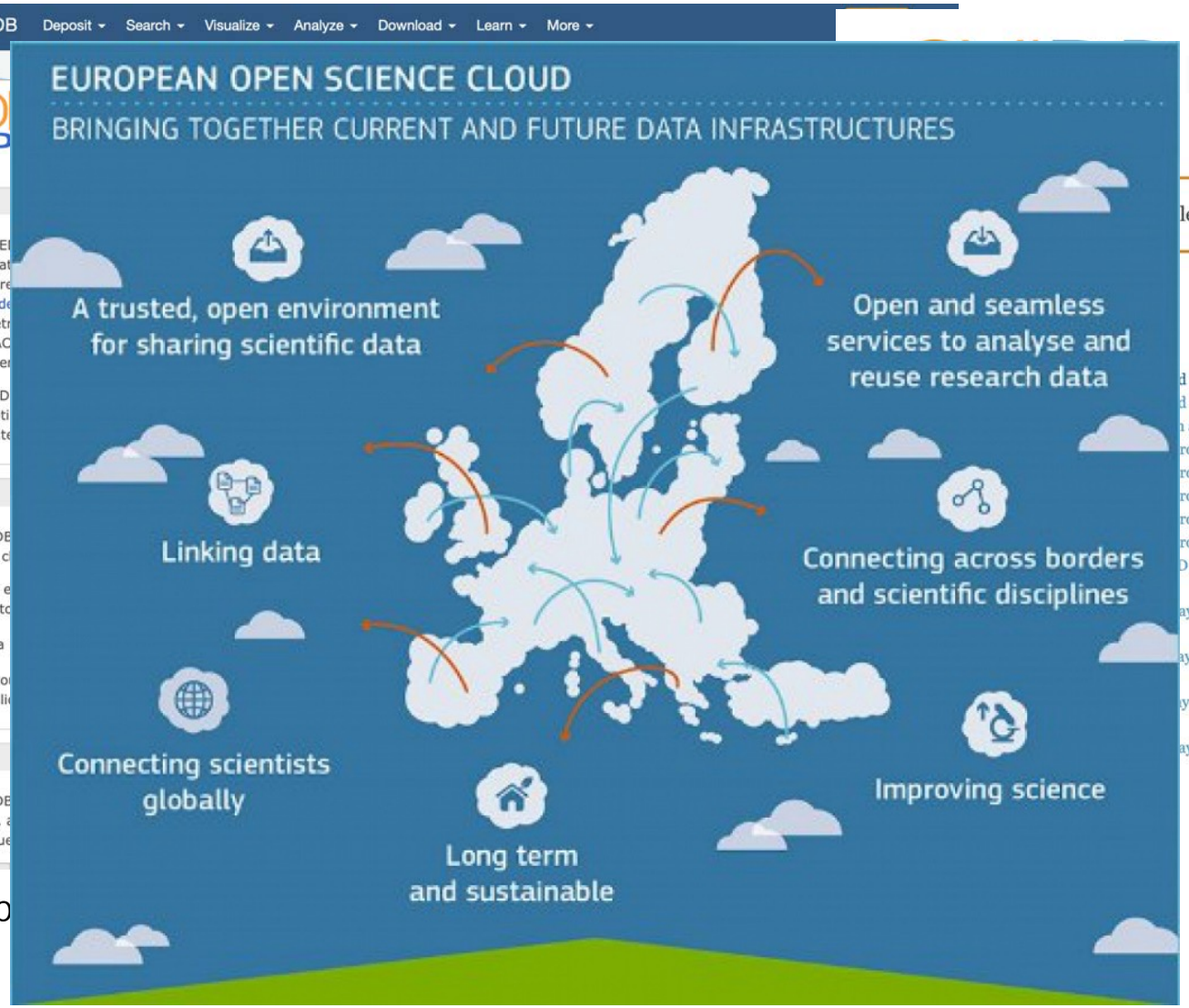
Analysis
- Data analysis steps
- Not all data is useful (runs)
- Intermediate data, code, scripts

Publication
- Findable and searchable
- Should describe what was done
- *Citation, DOI reference*
- Sometimes data is deposited (PDB, CXIDB)
- May use a subset of data, or data from many experiments

Re-use
- Check and verify results
- Improve the analysis
- Re-use code for new work
- Build on past data

Task area 2: Community data repositories

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



https://www.eosc-portal.eu

DataCite
FIND, ACCESS, AND REUSE DATA
https://datacite.org

DataDOI tracks usage of open data

Task area 3: Infrastructure for data and software re-use

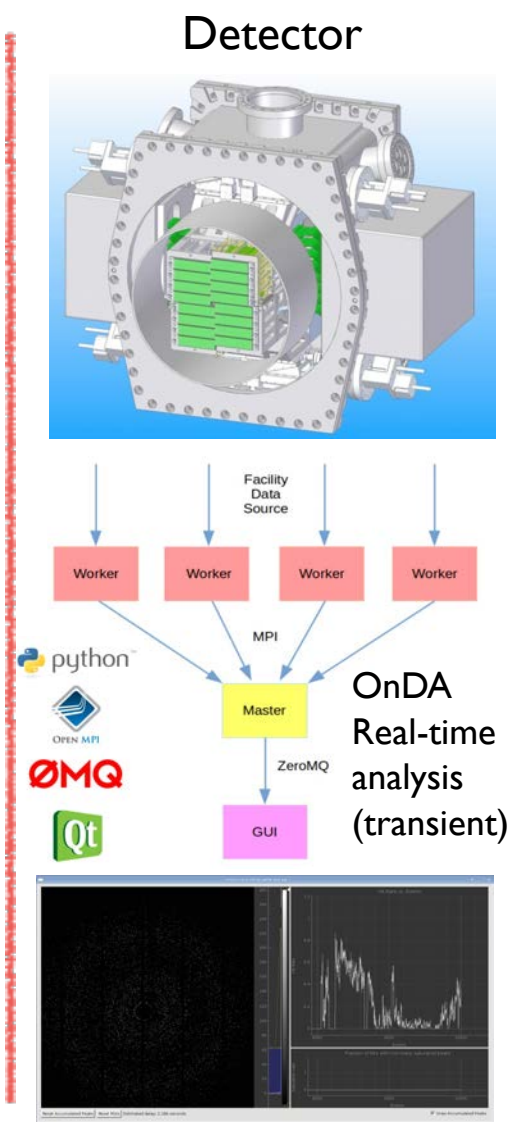
Analysis software infrastructure is currently fragmented and hard to re-use

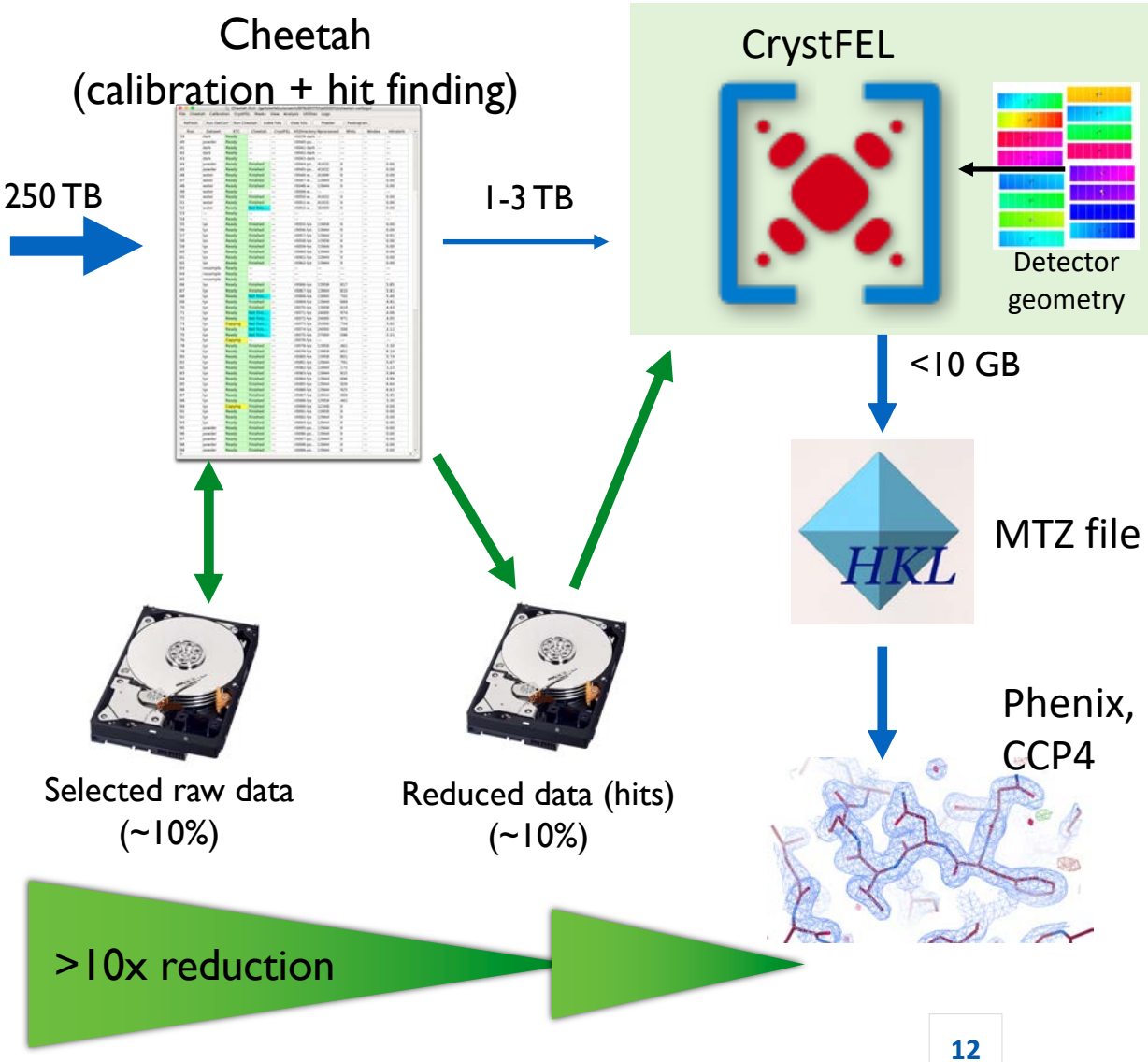
Too much analysis based on bespoke code and scripts

Engage power users to accelerate science outcomes by enabling data and software re-use

Make user software remotely available for re-use by all users on facility infrastructure

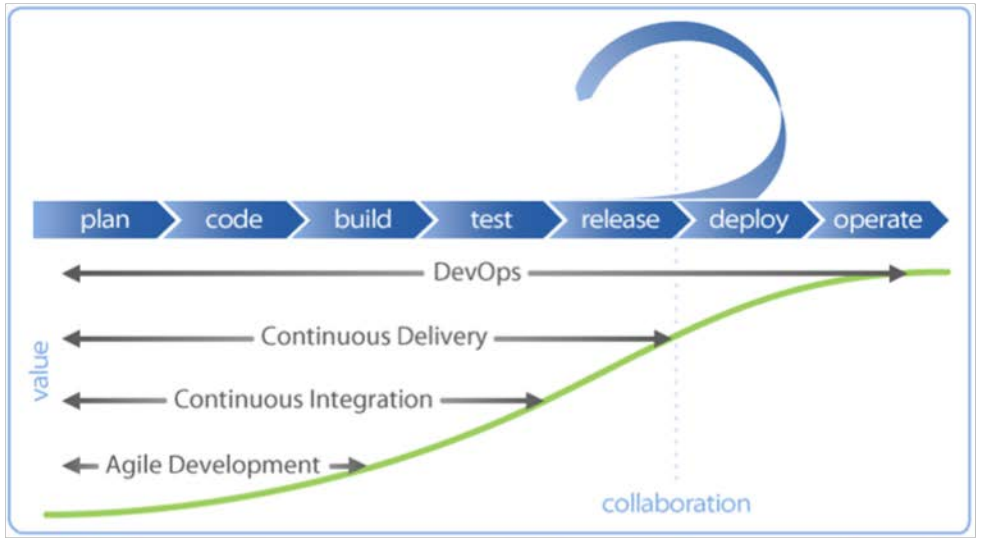
Lead: Anton Barty, Frank Schreiber





Task areas 1-3: Sustainable software development

Strike a balance between agile user-focussed design with sustainability and integration with existing infrastructure



For example: *DevOps* model of software collaboration

Teach and practice sustainable software development

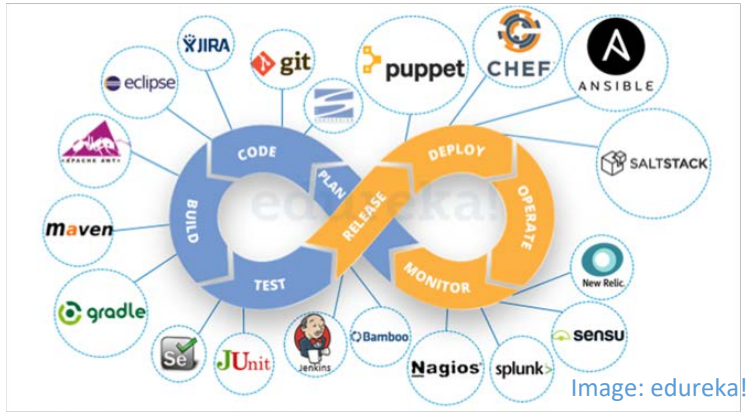
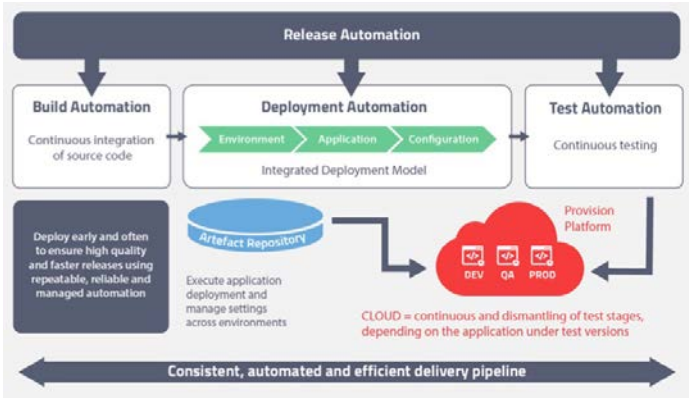
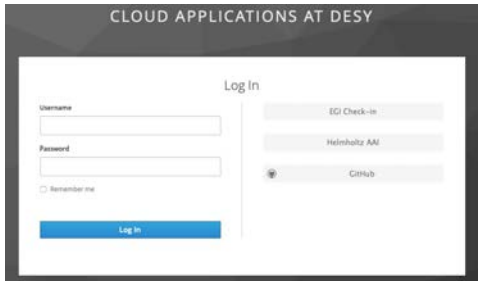


Image: edureka!



AAI infrastructure eg: UmbrelaID + Keycloak



umbrella

KEYCLOAK

OpenID

ORCID

Track guests using ORCID-ID?
('login with facebook' for researchers)

Task Area 4: Dissemination and outreach

The NFDI consortium as a role model and educator

Inside DAPHNE

- Explain, discuss and disseminate the advances of FAIR principles, common **data formats and** cross-community standards for **metadata capture and metadata management,** changed requests related to long-term **data preservation** and the capture of **data provenance**
- Organize workshops and schools to increase **awareness**, to **combine interests of scientists and facilities**, transmit technical developments and opportunities
- Provide a platform for exchange about **electronic logbooks**, metadata standards and agreements, workflows, reference database and catalogue specifications (workshops … web-portal)
- Organize and support **pilot / demonstrator projects**

DAPHNE within NFDI

- Support Task2: cross-community standards for **metadata capture and metadata management,** promoting common **data formats,** cross-community **user portals**
- Connect specialists, use experience

Outreach to society and industry

- Inform about the **increased efficiency due to NFDI**, especially using LSF, and **exciting results**
- Connect industry and society to highlight developments and encourage to re-use

Lead: Jan-Dierk Grunwaldt, Astrid Schneidewind

Task area 5: External communication and policy

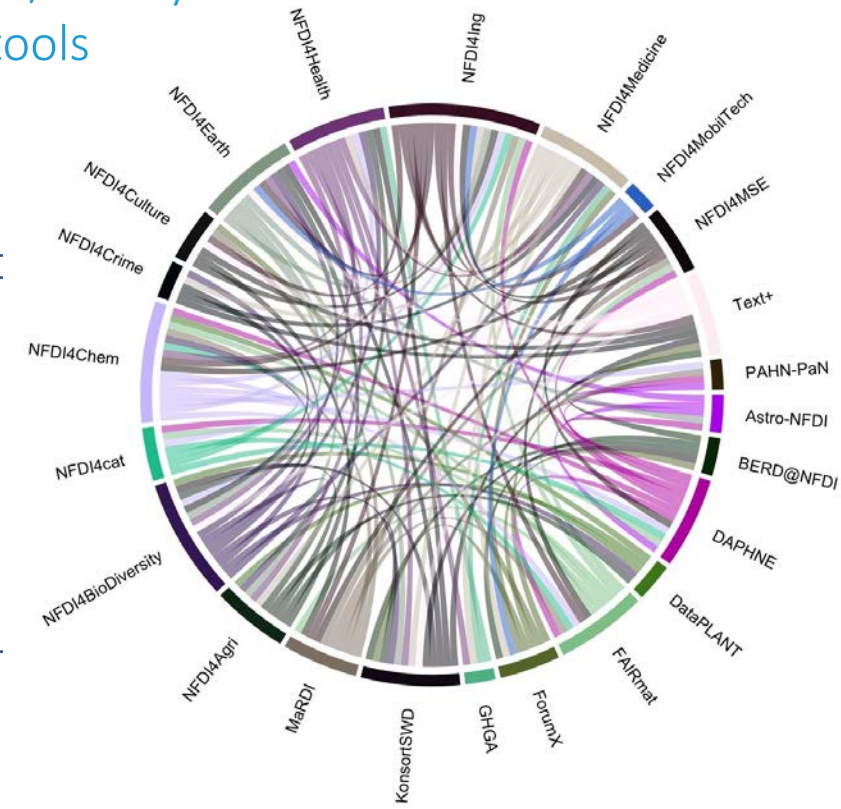
Linking DAPHNE to other NFDI consortia and other research areas

DAPHNE

- Atomic structures
- Material structures
- Electronic, magnetic structures
- Dynamic information spanning 103 – 10-15 s
- Large 100s TB data sets for exploring AI/ML algorithms
- Processed and raw data

FAIRMAT, NFDI4CAT, NFDI4CHEM, NFDI4ING, NFDI4MSE, MaRDI

- Complementary data sets
- Simulation/theory data
- Analysis tools



Approach:

Define science driven pilot projects between the consortia for establishing communication on a NFDI level.

Best practice examples, identify challenges, standards of meta data, converter algorithms, ontologies, etc.

Lead: Christian Gutt, Astrid Schneidewind

Task area 6: Project management

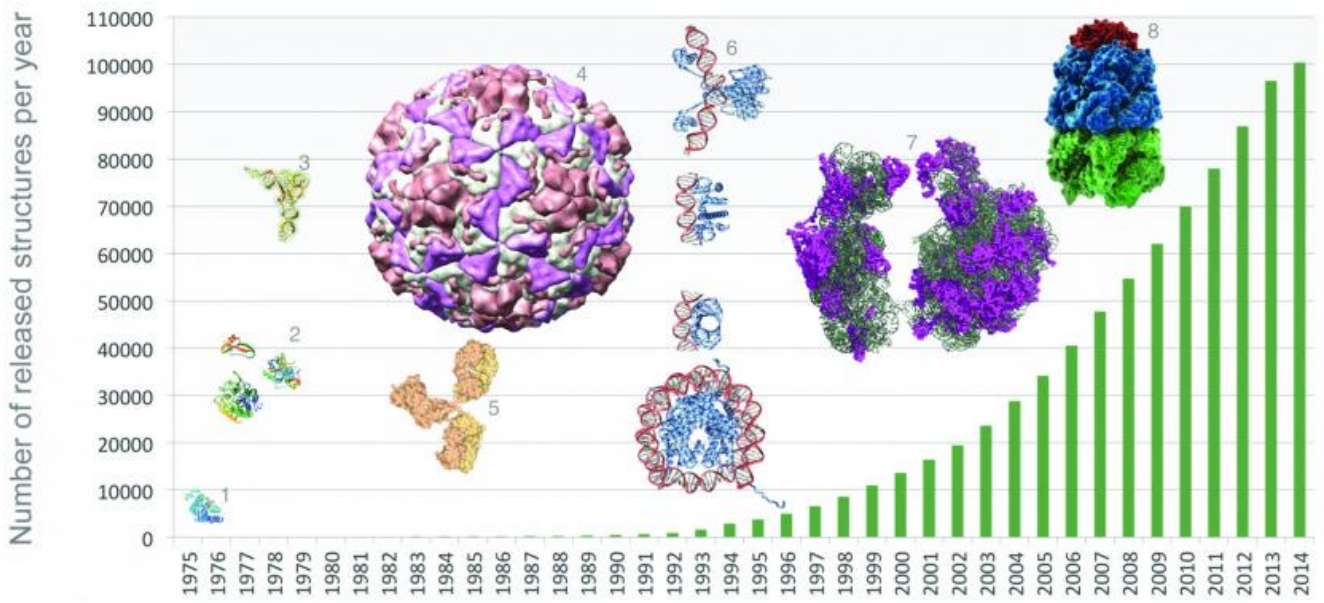
Measuring success

Key performance indicators (KPI)

Key performance indicator (KPI)	Detailed measure	How to measure
DOIs published	count the number of DOIs published within DAPHNE4NFDI	this number is monitored within the large-scale facilities and transferred to DAPHNE4NFDI
DOI citations	count the number of citations with DAPHNE4NFDI DOIs	publishers and web resources such as web of science etc.
Number of beamlines using DAPHNE4NFDI specifications and software	count the number of beamlines	number is monitored by the large-scale facilities
Downloads of DAPHNE4NFDI software packages	number of downloads	measured by the IT providers involved in DAPHNE4NFDI
Downloads of DAPHNE4NFDI data sets	number of downloads	measured by the IT providers involved in DAPHNE4NFDI
Visits to DAPHNE4NFDI webpage	number of visits	measured by the IT providers hosting the web-page
Acceptance in the user community	number of users logged into DAPHNE4NFDI pages and regular user surveys	measured by the IT providers hosting the web-page
Best practice at the facilities	Number of beamlines recording metadata	number is monitored by the large-scale facilities
Database	Use and reuse metrics	measured by host IT providers
DAPHNE4NFDI use within NFDI	count the number of data sets used within other NFDI consortia	needs cross counting capabilities between the NFDI consortia
European impact	count how many European users access DAPHNE4NFDI software	measured by the IT providers involved in DAPHNE4NFDI
DAPHNE4NFDI publications	DAPHNE4NFDI related publications, services, talks at conferences, posters etc.	input from the DAPHNE4NFDI participants
Implementation in university curricula	Count the number of curricula addressing FAIR data principles related to DAPHNE4NFDI	input from the university groups in DAPHNE4NFDI
Awareness	Count the number of DAPHNE4NFDI related events and attendance	

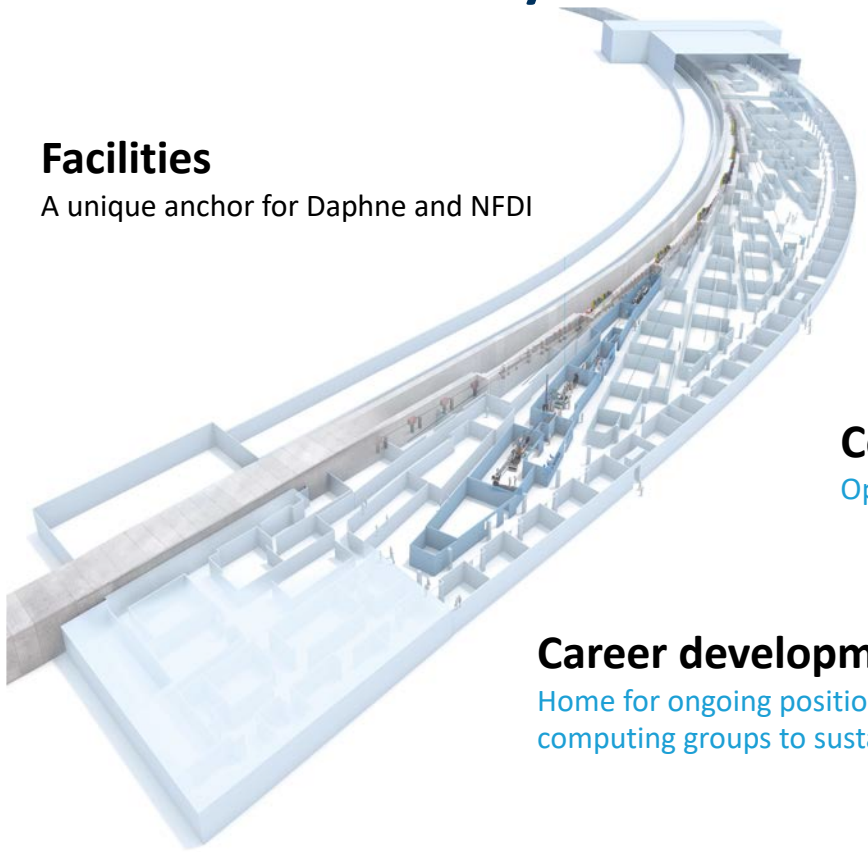
(listed in the proposal - p16-17)

Quantitative performance indicators: counting acceptance by deposition rates and amounts



Example: Protein Data Bank
PDB at EMBL-EBI is accessed **1.7 million times per day**.
Total daily access to EMBL-EBI data bases are > 60 million a day

Sustainability within NFDI beyond initial DFG funding



Facilities

A unique anchor for Daphne and NFDI

Integration

Daphne data infrastructure will be closely integrated with facilities, ensuring continued impact long after NFDI is over.

Continued operation of services

Operation of services and maintenance will be provided by the facilities after funding is over.

Career development

Home for ongoing positions, including long term, within facility computing groups to sustain Daphne after initial funding.

Education

Students and future generations: a culture of data curation and scientific software development will be embedded in University courses

Lead by example

Power user groups act as role models leading the community in entrenching best practices.

Science domains

A large and diverse research community