DAPHNE4NFDI

Consortium Proposal

National Research Data Infrastructure (NFDI)

The DAPHNE4NFDI consortium

30 September 2020

1 General Information

1.1 Name of the consortium

EN: DAPHNE4NFDI: **DA**ta from **PH**oton and **N**eutron **E**xperiments for NFDI

DE: DAPHNE4NFDI: DAten aus PHoton- und Neutronen Experimenten für NFDI

1.2 Summary of the proposal

English summary

The photon and neutron science community encompasses users from a broad range of scientific disciplines facing a common need for high-level, rapid data analysis and the challenge of implementing research data management. The communities of photon and neutron users are represented by the KFS and KFN committees, which have worked together for many years and are coming together here to meet the common challenges imposed by the digital transformation. The DAPHNE4NFDI consortium serves the broad community of researchers employing a wide range of photon and neutron techniques, comprising more than 5500 scientists throughout Germany. In addition, DAPHNE4NFDI reaches deeper into the science communities of its users, e.g. solid state physics, molecular /atomic physics, protein crystallographers, chemistry, catalysis and others (over 50 000). The users produce ca. 28 PB of data per year, which are in turn used by the wider science community from which the users come. Individual experiments can produce millions of files and in some cases over 1 PB data per week, depending on the experimental configuration. Moreover, the community is currently witnessing a fundamental change in both the amount of data recorded and the corresponding data rates triggered by the increase in the brightness of the sources themselves (x-ray free-electron lasers, high-brightness storage rings and new neutron facilities) and by the rapid increase in the size and speed of modern detectors.

DAPHNE4NFDI brings together users representing key scientific application domains with the large-scale research facilities in photon and neutron science in order to advance the state of data management in the community. Uniquely, DAPHNE4NFDI engages directly with the user community to develop user-driven data solutions to advance science experiments. Broadly, we will provide the following tangible infrastructure through DAPHNE4NFDI for the wider photon and neutron community:

1. Improve metadata capture through consistent workflows supported by user-driven online logbooks that are linked to the data collection, thus enabling a richer capture of information about the experiments than is currently possible;
2. Establish a community repository of processed data, new reference databases and analysis code for published results, linked, where possible, to raw data sources, to sustainably improve access to research data and enable data and software re-use; and

3. Develop, curate and deploy user-developed analysis software on facility computing infrastructure so that ordinary users can benefit from and repeat the analysis performed by leading power user groups through common data analysis portals.

DAPHNE4NFDI builds on a tradition of very close interaction between user communities driving the scientific and technical developments of the facilities which is a key element of DAPHNE4NFDI. The consortium consequently comprises universities, user facilities and users at large-scale instruments such as the x-ray sources PETRA III and FLASH at DESY, BESSY II at HZB, the European-XFEL, the ELBE Centre for High-Power Radiation Sources at HZDR, as well as the Heinz Maier Leibnitz Centre with its neutron source and laboratory instruments with related data. Furthermore, it will have strong positive effects for international facilities with German participation, such as the ESRF and ILL (Grenoble, France), as well as the ESS (Lund, Sweden).

German summary

*Photonen- und Neutronenstreuung wird für sehr viele verschiedene Wissenschaftsfelder angewendet. Diese sehr verschiedenen Nutzer stehen vor der gemeinsamen Herausforderung den steigenden Bedarf von schneller Datenanalyse anzugehen, die ein effizientes Forschungsdaten-Management erfordert. Die Komitees Forschung mit Synchrotronstrahlung und Forschung mit Neutronen arbeiten als Repräsentanten dieser Nutzergemeinschaft seit Jahren zusammen und stellen sich hier den Herausforderungen der digitalen Transformation. Das Konsortium DAPHNE4NFDI vertritt das breite Spektrum der Nutzer von Synchrotron- und Neutronen-Techniken mit mehr als 5.500 Wissenschaftler in Deutschland. DAPHNE4NFDI hat außerdem Anknüpfungspunkte mit den Wissenschaftsbereichen Festkörperphysik, Molekül/Atomphysik, Protein Kristallographie, Chemie, Katalyse und anderen (über 50 000). Die Nutzer produzieren mehr als 28 PB Daten pro Jahr. Einzelne Experimente erzeugen dabei Millionen von Dateien, in einigen Fällen bis zu 1 PB pro Woche, je nach der Konfiguration des Experimentes. Die aktuellen technischen Entwicklungen beeinflussen diese Nutzergemeinschaft dabei grundlegend, sowohl hinsichtlich der Bewältigung von rasant steigenden Datenmengen als auch durch die steigenden Datenraten – z.B. in Folge der höheren Brillianz der Quellen (*Freie-Elektronen-Laser*, neue Speicherringe und Neutronenquellen) oder schnellerer sowie größerer Detektoren modernster Bauart.*

DAPHNE4NFDI bringt Universitäten, Nutzer und Betreiber der Großforschungseinrichtungen für Photonen- und Neutronenforschung hier in neuer Weise zusammen: Um den Herausforderungen im Bereich Daten-, Metadatenmanagement und hoher Datenraten zu begegnen, werden

Lösungen für herausragende wissenschaftliche Experimente gemeinsam mit der Nutzergemeinde entwickelt. Dafür entwickelt DAPHNE die folgende Infrastruktur:

Verbesserung der Metadatenerfassung durch Nutzer-initiierte Online-Logbücher, welche mit der Datenerfassung vernetzt sind und damit einen deutlich weitreichenderen und nachhaltigeren Informationsgehalt über die Experimente als derzeit ermöglichen;

1. *Erstellung von Katalogen/Repositories für die von der Gemeinschaft erzeugten Daten und Analysecodes sowie von neuen Referenzdatenbanken für publizierte Ergebnisse, um eine Nachnutzung von Daten und Auswertungs-Software zu ermöglichen. Wo immer möglich, sollen die Einträge in den Katalogen und Datenbanken mit den Quellen der Roh- bzw. reduzierten Daten verknüpft werden; und*
2. *Entwicklung, Kuration und Verwendung von nutzerentwickelten Auswertewerkzeugen über die Infrastrukturen der Großforschungszentren, so dass alle Nutzer die Auswertungen, welche von den „Power“-Nutzergruppen durchgeführt wurden, nachvollziehen und von ihnen profitieren können.*

DAPHNE4NFDI baut auf der traditionell engen Zusammenarbeit innerhalb der Nutzergemeinschaft auf, die die wissenschaftliche und technische Entwicklung an den Zentren vorantreibt. DAPHNE4NFDI entwickelt diese Kooperation als das Schlüsselelement des Konsortiums weiter und vereint Universitäten, Nutzer und Großforschungseinrichtungen – PETRA III und FLASH an DESY, BESSY II am HZB, European-XFEL, ELBE am HZDR als Photonenquellen, das Heinz Maier-Leibnitz-Zentrum mit seiner Neutronenquelle und Laborinstrumente mit vergleichbaren Daten. Durch die starke internationale Vernetzung werden die Ergebnisse von DAPHNE4NFDI positive Effekte auf die internationalen Einrichtungen mit Deutscher Beteiligung haben, so wie ESRF und ILL (Grenoble, Frankreich) sowie ESS (Lund, Schweden).

1.3 Applicant institution

Applicant institution	Location
Deutsches Elektronen-Synchrotron (DESY) Notkestrasse 85 D-22607 Hamburg Germany	Hamburg, Germany

1.4 Spokesperson

Spokesperson	Institution, location
Anton Barty DESY Photon Science	DESY Hamburg, Germany

1.5 Participants

Contribution of participating national institutions:

The participating institutions will be active participants and provide considerable in kind contributions to the task areas as listed above.

Contribution of participating international institutions:

The participating institutions will be active participants and provide considerable in kind contributions to the task areas as listed above.

Contribution of industry collaborators:

Discussion and consultation on the topics of Metadata nomenclature and capture and high speed and large and file formats in TA1 and TA2. Dissemination and transfer of standards and best practice in TA4. The full list of task areas is listed above

List of abbreviations

CXIDB	Coherent X-ray Imaging Databank
DAPHNE	Synonym for DAPHNE4NFDI
DPG	Deutsche Physikalische Gesellschaft
ELN	Electronic Laboratory Logbook(s)
ESS	European Spallation Source ERIC
EuXFEL	European XFEL
ExPaNDS	European Open Science Cloud (EOSC) Photon and Neutron Data Service (ExPaNDS) project
ILL	Institut Laue-Langevin, Grenoble, France
ISPyB	Information System for Protein Crystallography
KFN	Komitee Forschung mit Neutronen
KFS	Komitee Forschung mit Synchrotronstrahlung
KFSI	Komitee Forschung mit nukelaren Sonden und Ionenstrahlen
KPI	Key Performance Indicators
LSF	Large Scale (research) Facility
MX	Macromolecular Crystallography
TA	Task Area
PaNOSC	Photon and Neutron Open Science Cloud
PDB	Protein Data Bank
PI	Principal Investigator
PO	Project Office
PSI	Paul-Scherrer-Institut, Villigen Schweiz
RI	Research Infrastructure, related to LSF
x/n, n/x	synchrotron-x-ray and neutron, respectively, photon and neutron
IUCr	International Union of Crystallography

2 Scope and Objectives

2.1 Research domains addressed by the consortium, specific aim

2.1.1 Research domain

The DAPHNE4NFDI consortium represents the community of researchers who use data measured at large-scale x-ray and neutron facilities in Germany. The community extends way beyond direct facility users to include researchers who wish to make use of the unique data measured at the facilities in their research. The DAPHNE4NFDI community impacts on research from diverse fields ranging from biology and medicine through materials science and engineering to cultural heritage science, chemistry and physics.

DAPHNE4NFDI is a collaborative platform linking researchers and large-scale facilities producing and injecting more than 28 PB of data annually into the NFDI. The x-ray and neutron science communities derive from a broad range of scientific disciplines, yet face a common need for high-level, rapid data analysis and the challenge of implementing research data management along FAIR principles. Measurements at the facilities support research from a broad variety of scientific disciplines, such as physics, biology, chemistry, materials science, engineering, geophysics, archaeology to name a few. Accordingly, DAPHNE4NFDI will have an impact across a wide range of scientific fields through its diverse community of data users. **The data management problem addressed by DAPHNE4NFDI is tractable across this breadth of communities because of the commitment of the large-scale facilities to working with users to advance their data management needs.**

2.1.2 The research community served by Daphne

Research at large-scale photon and neutron facilities (LSF) is performed by researchers who need **access to specialist measurement facilities** in order to further their research goals. These researchers include a sizeable number of PhD students at universities, which connects the LSFs strongly to young researchers and their training. The procedure for gaining access to the limited amount of available measurement time generally involves a process in which the **researcher submits a proposal**, and experimental time is granted based on **competitive peer review**. When a proposal is accepted, the researcher becomes a **facility user** and is scheduled one or more blocks of time to perform their experiment. Due to the limited time available at each instrument, available **measurement time is generally oversubscribed** by a factor of 2-10 (2-10 times more proposals than can be scheduled). Access to measurement time is highly competitive, thus **measured data is extremely valuable** to the wider research community beyond direct users: there is often no other way to obtain the desired information. After their experiment, researchers generally have a period of exclusive access to the data during which time they hopefully produce

a publication detailing their findings to the wider scientific community. Since the facilities are publicly funded, the data should become open to other researchers after some period of time – typically 3 years. **Published and unpublished research findings should be traceable back to the original measurement data** and **other researchers should be able to reanalyse that data** either to repeat the analysis or make new discoveries. In other words, data measured by users should abide by FAIR principles (Finable, Accessible, Interoperable, Reusable).

The challenges to making data useable by researchers other than the users who collected the data are:

1. **Collected metadata** about the experiments **are not adequate** for the measured data to be reusable by any researcher in a similar field and beyond;
2. The **absence of curated databases** of raw, intermediate and processed data which can ideally be searched, but at a minimum be traced from published and unpublished results; and
3. The general **absence of curated and managed software** developed by leading research groups for remotely working with large measured data sets so that others can repeat the data analysis pipelines.

Addressing these three challenges forms the heart of DAPHNE. Currently, facility users form the critical bridge between experiments made at the photon and neutron facilities and the wider research community who use the scientific output. DAPHNE therefore focusses on engaging with the community of users, as will become evident through the proposal. Indeed, DAPHNE is driven by the research community of facility users who need a structured way in which to make their data and research results FAIR in the era of exploding data volumes. The goals of DAPHNE are well aligned with the goal of NFDI within which the facility user community plays a central role.

The diverse German communities of large-scale photon and neutron facilities (LSF) users are represented by the Komitee Forschung mit Synchrotronstrahlung (KFS) and Komitee Forschung mit Neutronen (KFN), which have been working together for years and are joining forces here to meet the challenges of the digital transformation. The community performs thousands of individual user experiments at LSF every year, across many disciplines and using a wide range of techniques and a diverse instrumentation. Individual experiments can produce millions of files and in some cases over 1 PB data per week, depending on the experimental configuration. Moreover, the community is currently witnessing a fundamental change in both the amount of data recorded and the corresponding data rates triggered by the increase in the brightness of the sources themselves (x-ray free-electron lasers, high-brightness storage rings and new neutron facilities) and by the rapid increase in the size and speed of modern detectors. Research in the x-ray and neutron communities is therefore experiencing a transformation in the challenges of data processing, storage and management that were previously known only from experiments in

areas such as high-energy physics. This not only concerns the IT infrastructure for storing data, but also the development of new algorithms and software concepts for processing data at the facility both during and after experiments, rather than the previous model in which researchers take their raw data home for later analysis. This revolution in data management calls for a careful consideration of its ethical and legal implications. In order to be successful, this transition requires an investment into efficient research data management for the community. If done right, though, this revolution offers the unique opportunity to transform completely the way the data is used towards a FAIR model.

DAPHNE brings together the large-scale research facilities in photon and neutron science with users, representing typical scientific domains, to advance data management within the community. The consortium is representative of the broader community of users who employ a broad range of x-ray and neutron techniques, which comprises approximately 5500 registered scientists throughout Germany. The community, however, extends way beyond direct facility users to include researchers who make use of results and data measured at the facilities in their research, typically in larger collaborations, in which the information from LSFs is a key ingredient to the overall success also for those who do not actively participate on-site at the LSFs themselves. An analysis of DAPHNE related publication records and author lists shows that the number of scientists using, reusing, analysing, or participating in publications of DAPHNE related data exceeds more than 30.000 scientists.

User facilities traditionally interact very closely with the user communities, and this drives the scientific and technical developments. This interaction and connection between users and facilities is a key element of DAPHNE, since implementing data management requires a joint and coherent approach, in which the facilities act as data custodians and the user communities act as data curators. The consortium therefore comprises the LSF and their users such as the x-ray sources PETRA III and FLASH at DESY, BESSY II at HZB, the European-XFEL, the ELBE Centre for High-Power Radiation Sources at HZDR and the ESRF in Grenoble, as well as neutron sources at the Heinz Maier-Leibnitz-Centre (MLZ), the Institute Laue-Langevin (ILL) in Grenoble and the European Spallation Source (ESS). Included are of course plans for PETRA-IV and BESSY-III upgrades

2.2 Objectives and measuring success

2.2.1 Key objectives of DAPHNE4NFDI

Members of the DAPHNE4NFDI consortium currently provide more than 28 PB of measured data annually from thousands of user experiments. Annual data volumes are growing rapidly necessitating a paradigm shift in data management. **The main objective of DAPHNE4NFDI is to make the growing volume of valuable measured data FAIR for the DAPHNE4NFDI community but importantly also to make it FAIR and available for the whole NFDI - with special emphasis on the scientifically closely related NFDI consortia.**

The main objective of DAPHNE is to improve the transparency, FAIRness, and re-usability of data measured at photon and neutron facilities for the wider research community beyond the individual group which performs the measurement.

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 curated databases of raw, intermediate and processed data which can ideally be searched, but at a minimum be traced from published and unpublished results;
3. Develop a curated ecosystem of managed software developed by leading research groups, and to make this available 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.

These key objectives will be implemented through the community of facility users, which form the key bridge between measurement at the facilities and accessibility of that data by the wider research community.

In order to achieve these objectives DAPHNE will:

1. Link electronic laboratory log books (ELNs) and automatic metadata capture to the data collection, adding developments that addresses the needs of users and instruments in a technically advanced and efficient manner, aiming to collect information in structured and digital form early in the data generation process **(TA1)**.
2. Create repositories of processed (reduced) data and analysis code to go with each publication in order to maximise data reuse and transparency **(TA2)**.
3. Implement structured databases and catalogues for user experiments, samples, experimental data and calculations, thereby providing the opportunity to search for and find information before, during and after the experiments **(TA2)**.

4. Create, curate and foster analysis software that can be deployed on ‘cloud-like’ services so that ordinary users can repeat and benefit from the work of power users, and to make the analysis of ‘big data’ technically simple, reproducible and sustainable, including the accessibility to machine-learning strategies **(TA3)**.
5. Develop a common data policy between users and large-scale facilities which addresses common needs such as data curation, archiving standards and embargo policies. Align data policies and standards also on a European level by corporation with our European partners **(TA4)**.
6. Develop and promote efficient data flows and metadata definitions in agreement with other communities and NFDI consortia which enable and foster the reuse of all photon and neutron data in the NFDI **(TA4)**.
7. Establish and enhance the awareness of FAIR principles and needs for research data management in our community, especially within university curricula **(TA5)**.
8. Manage and oversee the development and curation of software packages **(TA6).**
9. Manage and coordinate the financial and organisational aspects of the consortium **(TA6)**.

2.2.2 Measuring the success of DAPHNE4NFDI

We will measure the impact of DAPHNE not only within our community but also within the NFDI through **key performance indicators (KPI)**. Typical performance indicators count numbers such as data sets downloaded, the number of DOIs published, number of users or number of publications citing our services. The following list comprises KPIs typically used in European projects such as PaNOSC and ExPaNDS with adaption to the needs of DAPHNE. The KPI will be inspected twice a year by the DAPHNE government body yielding regular feedback to the success of DAPHNE. Task 6 will manage the oversight on KPI and assignment of collecting the different KPIs to the relevant task areas.

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	

2.2.3 First year implementation of the project

The first year of the project is mainly foreseen as to gather awareness and, to build momentum in the user community, which is crucial for building the broad base on which the projects of the different task areas will grow. Aside from **recruitment of personnel, this time period will be used to hold kick off meetings as well as a series of workshops** to establish a broad communication and community platform among all DAPHNE task areas and participants. Connections and links to other NFDI consortia to pursue cross-cutting topics will also be initiated. This will ensure efficient and transparent collaboration within the consortium and the wider community. **Building on that, we aim to have preliminary specifications and detailed implementation plans** (including measures for incorporation of user feedback) **for the tasks ahead after the first 12 months.** This is especially prudent for all areas that stretch across different TAs such as (meta)data and vocabulary specification where common standards are to be established. Preliminary technical implementations will allow expert user groups to use them for “use cases” and to give quick feedback on the “product”, allowing for a lean rapid prototyping approach to the software development. This is a common task for TA1, TA2, TA3 and TA4. TA5 will coordinate this activity with other consortia. **A preliminary report on (meta)data standards, existing metadata schemata and domain specific vocabulary will be released at the end of the year** – to be continuously updated throughout the project. Similarly, draft specifications for data catalogues, requirements for automatic (meta)data capture, electronic logbooks (ELNs) and analysis tools will be made available at the end of the first year. These specifications will go hand-in-hand with prototype implementations that will be continuously updated and adjusted to user feedback throughout the project.

3 Consortium

3.1 Composition of the consortium and its embedding in the community of interest

The consortium DAPHNE4NFDI directly represents 5,500 registered facility users in Germany. The users perform 3000 experiments per year, publish 3,000 papers with an average of nine authors and produce ca. 28 PB of data per year. In addition, data measured by the users is in turn used by the wider science community from which the users come. DAPHNE4NFDI reaches deeper into the science communities of its users: solid state physics (DPG-SKM 19,000 members), molecular/atomic physics (DPG-SAMOP 8,500 members), DECHEMA (5,8000 members), GDCh (31.000 members) and others.

In Europe DAPHNE4NFDI has ties to the 30.000 registered users of x/n facilities which perform >6.000 experiments per year, publish >5.000 paper and produce > 40 PB of data.

The photon and neutron (x/n) user communities play a central role in DAPHNE, acting as a key link between data resulting from measurements made at the facilities and the much wider community of researchers wishing to make use of results from those measurements. The x-ray and photon user communities in Germany consist of user groups that are typically research groups, mainly from universities, with an accordingly large fraction of young researches and a substantial training component and multiplier effects, but also from non-university research organisations such as the Helmholtz Association, the Max Planck Society, the Fraunhofer Society and the Leibniz Association. These groups represent a diverse field of scientific disciplines (see below). Naturally, only a fraction of the user groups can be co-applicants in this proposal. It was possible to win over as co-applicants a selection of groups that are representative for the wide variety of measurement techniques and scientific disciplines and have the necessary long-standing expertise. A wider circle of university groups is involved as participants, but crucially through the involvement of KFN and KFS, essential all neutron and synchrotron users in Germany are linked to this proposal. Most of the facilities and/or beamlines are operated by Helmholtz Centres (DESY, HZB, HZDR, FZJ and HZG); in the case of MLZ, about half the centre is operated by TU Munich. On a European level, Germany contributes to the European Synchrotron Radiation Facility (ESRF) and the Institute Laue-Langevin in Grenoble, to the European XFEL in Schenefeld and to the European Spallation Source (ESS) in Lund, where the facility is currently under construction and will go into operation in 2023. On a local level, the smaller synchrotron radiation facilities DELTA (TU Dortmund) and KARA (KIT) also offer experiments to external researchers.

The photon and neutron facilities at DESY, MLZ, HZB, HZDR, and the European facilities European XFEL, ILL and ESRF (and future ESS) have a peer review proposal system in place for distributing beamtime and related access to the facilities on the basis of scientific merit. Most facilities offer reimbursement for travel and accommodation for 1-3 users per experiment.

Communication between users and facilities is well established, via

- The elected user organisations KFS and KFN with approximately 5500 active members and many more users (see below for details)
- Workshops and annual user meetings organised by the facilities or in collaboration with associations like DPG, DECHEMA etc. They typically comprise keynote talks, dedicated workshops, a highly successful poster session and public evening lectures and is attended by more than 1000 participants (DESY / XFEL annual meeting) every year.
- Users serving on the scientific and governance advisory boards of the facilities.
- Chairs and members of the proposal and beamline review panels, who are experienced users and submit advice on strategic needs, new trends and problems to the facilities.

In addition to the LSFs mentioned above, German user communities also perform experiments at international sources such as SLS, SINQ and SwissFEL in Villigen, Switzerland; ISIS and DLS in Didcot, UK; Soleil in Saint-Aubin, France; Elettra and FERMI in Trieste, Italy; MAX IV in Lund, Sweden; ALBA in Barcelona, Spain; JPARC, Spring8 and SACLA in Japan; SSRL in China; and in the US at the LCLS at SLAC (Stanford), NCNR at NIST (Gaithersburg), APS (Argonne), ALS (Berkeley), HFBR and NSLSII (Brookhaven), HFIR and SNS (Oak Ridge).

In practice, users submit proposals for experiments to the facilities for review approximately 6-9 months before beamtime may start, and beamtime is granted for successful proposals a few months prior to the actual experiment. User groups prepare samples and, in many cases, auxiliary instrumentation and travel to the facility with a team of 1-5 researchers. Depending on the type of experiment, the sample environment and complementary equipment is provided either by the users or by the facility. Typically, beamtimes last between 1 and 14 days and are conducted in collaboration with the beamline personnel on a 24/7 basis. After finishing the experiment, in-depth analysis of the experimental data is usually performed by Master's/PhD students and postdocs. In many cases, this analysis can take a long time, ranging from several months to 1–2 years, depending on the novelty of the experiment, the complexity of the data and further complementary simulations or experiments that may need to be conducted. Some data are revisited at later time in the light of new insights.

The consortium DAPHNE is a joint undertaking from facility users conducting experiments at photon and neutron sources, their representative bodies KFS/KFN and the operators of the beamlines and instruments at the Photon and Neutron facilities.

3.1.1 Breadth of research fields represented

In both photon and neutron research, the fields of study are diverse (Fig. 3.1). In addition, on average 50 % of the users in the KFS survey indicated that they worked in interdisciplinary fields. In neutron research, magnetism (which can be fundamental or applied materials research) is an important field due to the unique properties of neutrons as a probe for magnetic properties. In most scientific fields the two techniques are complementary and significant parts of the user community use both x-rays and neutrons in their research. The systems investigated are very diverse, for example, x-ray users apply their methods in the field of solid state (34 %), surfaces/interfaces (18%), biology (14 %), soft matter (10 %), atoms/molecules/gases (8%), liquids (7 %), optics/lasers (5) and warm dense matter (1 %). Both for the x-ray and the neutron-related communities the studies extend far into the applied science field, even being used by small and large companies in Germany and world-wide. The range of topics studied at the large-scale facilities and the range of techniques used are consequently very diverse.

Further information on all the techniques has recently been published in the KFS user survey, the KFS brochure 2020 or the MLZ annual reports. This diversity of scientific areas and methods is reflected by the present requirements for data management and analysis, and its joint development is one of the key challenges for the community, which is to be addressed by DAPHNE.

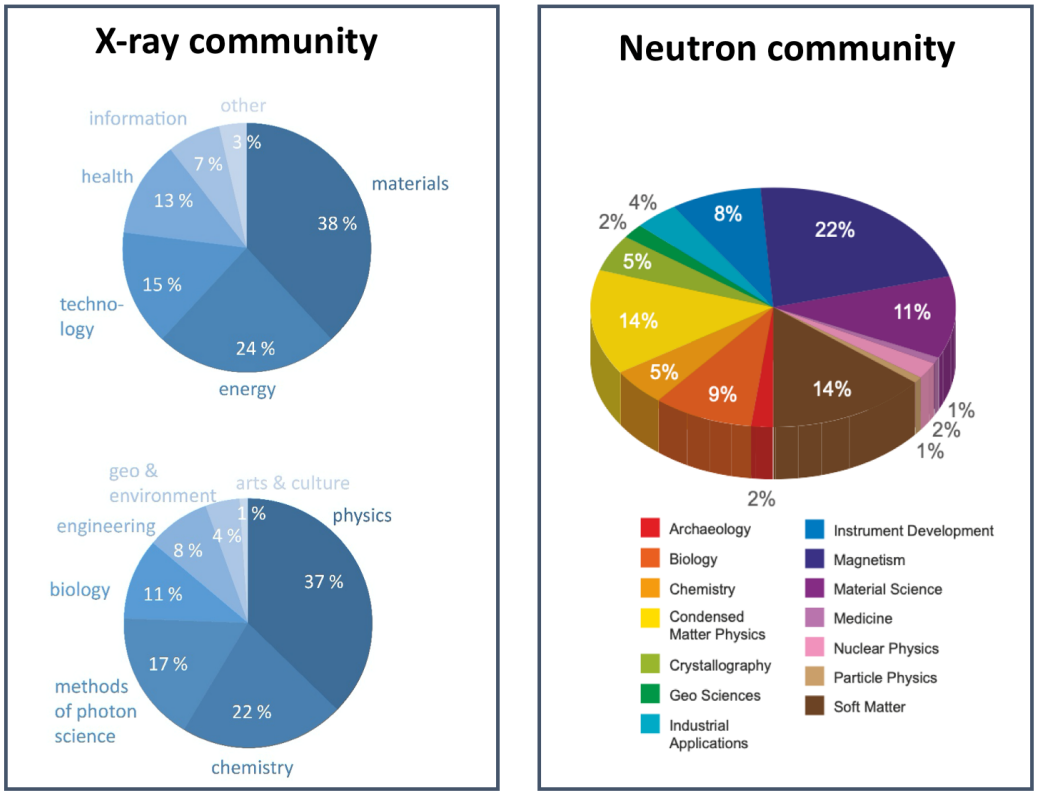

Fig. 3.1: Scientific fields of neutron and x-ray research user groups. Source: KFS survey (2018/2019, 894 users participating), MLZ report for 2014-2017 (Dec. 2018), KFS brochure (2020).

3.1.2 Needs in data management

The KFS recently asked the users the question: “What are your data management requirements? Please evaluate the present situation at the facilities with respect to your needs.”

It turns out that, depending on the specific area, between 30 to 50 % of the participants are not satisfied with the current situation (Fig. 3.2). The issues mentioned include data storage, data transfer, archiving, data reduction, data analysis, analysis software and theoretical support. The evaluation of data and the software used for this appears to be particularly problematic. Transferring the measured data from the facility to the home institution is also an important issue, although this is probably mainly a matter of hardware. Data transfer and storage are particularly important issues for chemists and biologists. The NFDI initiative itself was not part of this survey as the survey was conducted in early 2018.

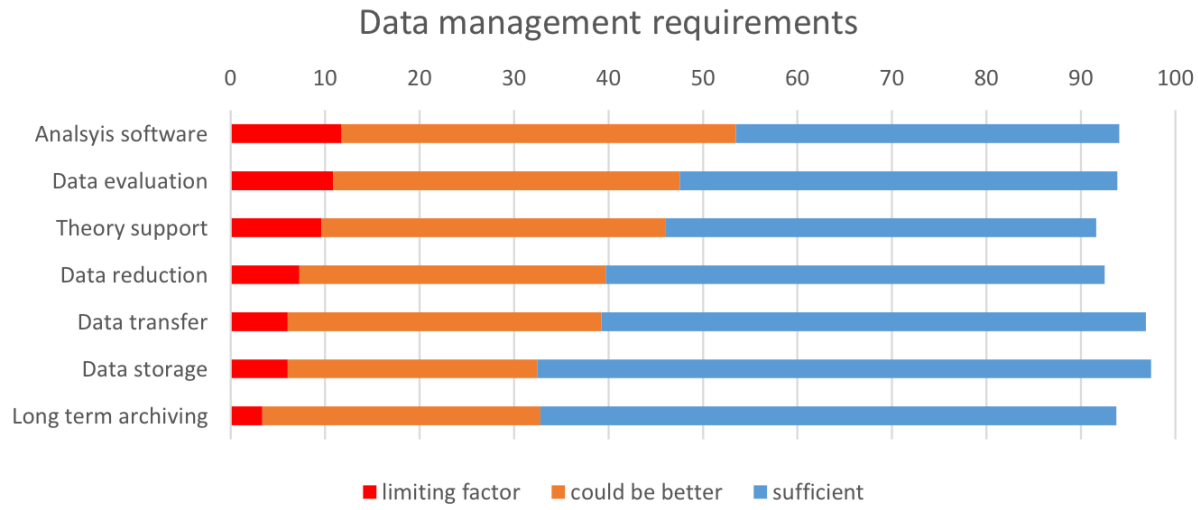

Fig. 3.2: Data management requirements expressed in the KFS user survey.

3.1.3 The KFS and KFN as elected bodies of the German user community

The Synchrotron Radiation Research Committee (KFS) and the Committee for Research with Neutrons (KFN) are elected bodies representing the interests of about 5500 registered users of synchrotron radiation sources (including FELs) and neutron research in Germany vis-à-vis politics (e.g. the BMBF, DFG) and the research centres. We estimate, that at least 45.000 scientists make use of their results every year. Both committees have existed for 30 years now and are elected every three years. The committee members elect a chair, vice chair and board members responsible for different community-related topics, plus up to four co-opted members representing fields of interest not covered by the elected members. The directors of the research facilities and representatives of the BMBF are permanent guests of the KFN and KFS.

The committees hold regular meetings every year, 3 - 6 depending on demand. A list of past meetings and the minutes are published online. They communicate with the user community via email and their webportal,[1] and they present the current status every year at user conferences. In 2006 (Hamburg), 2010 (Berlin), 2014 (Bonn) and 2018 (Garching), the KFS and KFN organised the national SNI conferences (3 days) together with the KFSI, presenting the state of science within the three communities, including dedicated micro-symposia and public lectures. National conferences are held in the field of neutron scattering (Deutsche Neutronenstreutagung, DN) every 4 years, alternating with the SNI.

3.1.4 KFS / KFN user participation in the area of digital transformation

The KFS and KFN conducted two user workshops in 2018 and 2019 and a symposium at the 2018 SNI conference, in order to obtain input from the user community concerning the digital transformation. The KFS and KFN have also participated in numerous BMBF workshops and, together with the other committees organised within the scope of the BMBF topic Science in Universe and Matter (hereafter ErUM, from the German *Erforschung von Universum und Materie*[2]), written the recommendations for ErUM Data (see below for details). Finally, KFS and KFN are naturally driving this DAPHNE proposal and collaborations with other NFDI consortia.

More specifically, the following activities have been conducted:

- Workshop #1: Digital transformation, Feb/March 2018 at DESY, Hamburg
- Symposium at the 2018 SNI conference in Munich
- Workshop #2: Digital transformation, June 2019 at DESY, Hamburg
- Participation in the BMBF ErUM Data process 2018-2019
- Participation in the first NFDI conference 2019 with a neutron and an x-ray oriented proposal, merger after the conference
- Participation in the DAPHNE proposal 2019.

The focus of Workshop #1 was to learn about the needs of and challenges faced by both the user community and the facilities. Accordingly, users were invited as representatives of their specific fields in KFS/KFN to present their needs, challenges and opportunities. In addition, representatives from national and international facilities also presented their perspectives. The outcome of this workshop was a first KFS/KFN white paper[3] for the digital transformation, which served as the basis for the subsequent discussions with the BMBF and within DAPHNE.

To facilitate further discussion with and receive input from the entire user community, a symposium was organised in the framework of the SNI conference (> 500 participants) in Munich, dedicated to the topic of “Digital Matters”. This symposium was well received by the community and provided additional valuable input for the KFS/KFN.

In 2019, both the BMBF ErUM Data initiative and the DFG NFDI programme started to be rolled out with considerable participation on the part of both the KFS and the KFN. The new developments were discussed in a second user workshop in June 2019, leading to the shape of the DAPHNE proposal presented here. Information about and calls for participation in DAPHNE have also been conveyed into the community leading to considerable feedback and participation by users in DAPHNE. As a result, DAPHNE is both aligned to and integrally linked with the activities of the KFS and KFN on behalf of the user community in the field of digitalisation and data management.

3.1.5 Participation of KFS/KFN in DAPHNE4NFDI

The members of the KFS and KFN are elected representatives of their communities and most of them view their role as community organisers and as providing a service for the community. As such, many co-speakers of DAPHNE are also active in the present KFS and KFN and we think that this ensures a good overview of and balance between the photon and neutron communities. After an initial period, in which we feel that the KFN/KFS needs to be active in shaping DAPHNE, we envisage the future role of the KFS/KFN as transforming into that of an advisory body, both for DAPHNE and the BMBF ErUM Data. In addition, within KFS/KFN we have strongly discussed about use cases. The following table gives an overview on use cases that will be addressed within DAPHNE.

Use case area	Allocation and short explanation
Biological Matter **x-ray imaging**	Data and metadata capture Re-use of data and software
Dynamics **Coherent Scattering XPCS**	Data and metadata capture
Amorphous materials and catalysis **x-ray absorption spectroscopy (XANES/EXAFS)**	Data catalogues, metadata specification, re-use of data, analysis software, quality assurance
Chemical systems **x-ray emission spectra, RIXS etc.**	Data catalogues, RDB, builds up on XAS database, combination of theoretical and experimental data
Dynamics in correlated electron systems **Inelastic x/n scattering**	Metadata specification, ELN
Soft matter and liquid interfaces **x-ray reflectivity**	ELN, metadata specification, software catalogue and web-oriented access to repeatable, overall accessible and reusable analysis software
Magnetic structures **Ultrafast / Magnetic x-ray scattering**	Software catalogue and web-oriented access to repeatable, overall accessible and reusable analysis software
Structure Refinement Accessible and reusable neutron powder refinement	Software catalogue and web-oriented access to repeatable, overall accessible and reusable analysis software
Engineering Materials, Catalysis, Battery Materials **Tomography with neutrons and photons**	Metadata specs, catalogues overall accessible and reusable analysis software
Proteins and food science, soft matter **Diffraction (small and wide angle) & Spectroscopy**	Metadata specification, analysis software, data catalogues
Electrochemistry and –catalysis, and particle accelaration **High energy x-ray diffraction**	Metadata, ELN, software catalogue

The co-speakers listed in the use case table above will work together with community scientists and the facility staff across facilities to push development in the science area outlined. For illustration, **three cases** are described in more detail. The other use cases will be treated in a similar manner.

X-ray imaging in biological matter: Data management has over the past years become one of the main bottlenecks of synchrotron x-ray imaging experiments. The use of micrometric and sub-micrometric spatial resolutions combined with multi-technique approaches (e.g. x-ray phase contrast, x-ray fluorescence and x-ray diffraction imaging) together with the need of large or full sample coverage, leads to critical still unsolved challenges. Data production during experiments

can reach several TB per day and appropriate data transfer/storage/archiving solutions cannot be addressed individually by user groups. Ad-hoc image processing tools are required to handle datasets of hundreds of GB and perform quantitative 3D data analysis. Raw and processed data from multi-scale (zooming-in studies) and multi- technique experiments need: (i) the use of metafile formats including all relevant information on the measured sample, sample preparation and the experimental settings; ii) reliable identifiers for data recovery and re-use; and iii) a reliable cross-filing system linking all the data collected on a given sample.

X-ray absorption spectroscopy: X-ray absorption spectroscopy is important to analyse solid materials, in particular amorphous materials. XAS (in terms of XANES and EXAFS) builds on comparison with experimental and theoretical spectra. Power users usually have their own individual reference databases, but certified shared reference databases hardly exist. Present databases for XAS suffer from problems such as not providing detailed information about the sample itself, unknown and inconsistent data formats, difficulties in adding to the database, non-standard organisation of the database, and a review process to ensure and assess the quality of submitted data. Note, that by a simple comparison to reference data, the user may easily trace whether or not the data quality achieved in a running experiment is comparable to accepted standard criteria. Hence, this is an ideal use case as start for further databases in spectroscopy (e.g. x-ray emission data, resonant inelastic x-ray scattering, and further novel photon-in/out techniques). Power users from university will collaborate with several dedicated beamlines.

Dynamics in correlated electron systems: Inelastic scattering is used to study lattice dynamics and magnetic excitations for materials science applications and condensed matter research. Historically a domain of neutrons, nowadays, new possibilities at modern synchrotrons provide complementary information for exciting science – e.g. experiments under high pressure at tiny single crystals and in strongly neutron absorbing materials. Within all objectives of TA1, the project will define relevant metadata vocabularies for INS/IXS, suggest extensions/adaptions to existing ELN implementations, to be tested during real experiments on collaborating beamlines. In the second half of the project, the focus will be on installing the new tools at selected instruments and gather feedback from the user community for the continuous development.

3.1.6 State of the art of user community for topics relevant to the NFDI

In view of the broad range of disciplines and techniques present in the research community, we also face quite a diverse situation in terms of data management. Facility users can be categorised roughly into three groups in terms of sophistication of data management. Group 1 collects relatively small amounts of data and analyses these at home. These users often still resort to handwritten logbooks, and metadata is not captured systematically or only generated later at home in order to identify potentially interesting data. Once processed, the data is usually not stored in archives or repositories. It is virtually impossible for other groups to process the data at a later time. Group 2 are users who collect huge amounts of data in single experiments with a

high degree of variability making metadata capture difficult. The data cannot be transferred home, nor can it be analysed/processed without the help of the facility's IT infrastructure. Processing of data is only possible for the group that performed the experiment. Group 3 are users who perform standardised measurements rather than conducting experiments (with an element of instrumentation). Macromolecular crystallography (MX) users are a prime example of this group. Here, the standardisation has led to a high degree of international organisation in the form of repositories, such as the famous Protein Data Bank.[4] The Protein Data Bank and the handling of data by this community are one example for an international success story, providing guidance and an idea of best practices that serve as examples for DAPHNE**.** The microscopy/ imaging community is another example of a user group that performs standardised measurements (although not exclusively). Here the scheme for handling data has not yet reached the high level of standardisation seen in MX.

3.1.7 Large-scale facilities in Germany linked to DAPHNE4NFDI

Per year (2019)	BESSY II	PETRA III	FLASH	EuXFEL	ELBE	MLZ	Total
Beamlines (BL) Measuring stations (MS)	27 BL 37 MS	22 BL 42 MS	2 BL 7 MS	3 BL 6 MS	1 BL 7 MS	26 BL 26 MS (+7 BL)	81 BL 125 MS
Experiments/ year	800	1400	38	25	70	620	approx. 2950
Individual Users	1500	3000	350	499	100	600	approx. 6050
User visits	3000	5000	500			>1000	>9500
Publications	500	500	30	94	40	345	approx. 1500
Data generated	1-2 PB	5 PB	1 PB	20 PB	0.5 PB	6TB	28.5 PB
Expected 2025	8 PB	20 PB	4 PB	100 PB	2 PB	0.5 PB	134.5 PB

PETRA III: The third-generation synchrotron radiation source PETRA III delivers bright beams, mainly in the high-energy x-ray range, enabling users to exploit the high brightness and coherence for *in-situ* and *operando* x-ray experiments. With its unique experimental capabilities, it serves users from a wide spectrum of research fields.

As of 2019, PETRA III is one of the brightest storage-ring-based x-ray source for high-energy photons worldwide. It operates at a particle energy of 6 GeV and a beam current of 100 mA in top-up mode, both with a continuous bunch-filling and in a timing mode of operation. PETRA III serves experiments with photon energies between 150 eV and 200 keV. A maximum spectral brightness of 10^{21} photons/(s·mm^2·mrad^2·1% BW) is achieved thanks to the low horizontal and vertical emittances. This makes PETRA III an exceptional synchrotron radiation light source with unique applications in physics, chemistry, and biology, materials, biomedical, and nanoscience, as well as cultural heritage. A total of 22 insertion device beamlines are in operation with more than 40 different experimental stations or instruments. In conjunction with PETRA III beamtimes, users can apply for access to the DESY NanoLab, which offers full support for certain nano-characterisation and -preparation methods and dedicated sample environments for *in-situ* and *operando* experiments.

European XFEL: The European X-Ray Free-Electron Laser Facility GmbH is a limited liability company under German law. At present, 12 countries are participating in the project: Denmark, France, Germany, Hungary, Italy, Poland, Russia, Slovakia, Spain, Sweden, Switzerland and the United Kingdom. The company is in charge of the operation and construction of the European XFEL, a 3.4 km long x-ray free-electron laser facility extending from Hamburg to the neighbouring town of Schenefeld in the German federal state of Schleswig-Holstein. Civil construction started in early 2009, and the user operation in September 2017. With its repetition rate of 27 000 pulses per second and a peak brilliance a billion times higher than that of the best synchrotron x-ray radiation sources, the European XFEL will allow the investigation of still open scientific problems in a variety of disciplines (physics, structural biology, chemistry, planetary science, study of matter under extreme conditions, and many others).

HZB: BESSY II is Germany's synchrotron radiation facility optimised for experiments in the spectral range of soft to tender x-rays and is operated by the Helmholtz-Zentrum Berlin for Materials and Energy (HZB). The operation and development of the BESSY II photon source is explicitly aligned to the needs of its national and international, multidisciplinary user community. The BESSY II facility is a leader in high-resolution spectroscopy, flexible pulse patterns for advanced time-resolved experiments, and in innovative instrumentation for *in-situ* and *operando* measurements in the soft x-ray range. These are powerful tools for advanced analytics and therefore play a key role for research and development in materials and energy science. The unique Energy Materials In-Situ Laboratory (EMIL@BESSY II) is of special relevance in this context, allowing for *in-situ*, *in-system* and *operando* investigations of materials and processes. Looking towards the long-term future and extrapolating the scientific portfolio served by BESSY II, HZB is currently sharpening the properties required of a next generation source in the soft and tender x-ray regime ("BESSY III"). Smart optimisation and specialisation of this facility and its

associated infrastructure will yield experimental capabilities that are unique in the world. In December 2019, HZB discontinued the operation of the neutron source BER II.

ELBE FREE-ELECTRON LASER Facility at HZDR: The heart of the ELBE Centre for High-Power Radiation Sources is the 40 MeV superconducting electron accelerator ELBE (Electron Linear accelerator with high Brilliance and low Emittance), generating a variety of secondary radiation, such as MeV bremsstrahlung, neutrons and positrons, and feeding two free-electron lasers (called FELBE). These are the only FELs in Europe open to users that provide a continuous pulse train (cw mode) at MHz repetition rate (13 MHz). In addition, two beamlines provide high-power, coherent terahertz radiation between 0.1 and 3 THz at a 100 kHz repetition rate (TELBE). Roughly 40% of the total beamtime is given to the IR and THz beams. The average power of FELBE is in the range of tens of Watts, corresponding to pulse energies of a few µJ and peak powers of the order of 1 MW. The pulse duration ranges from sub-ps to several tens of ps.

The combination of FELBE and the Dresden High Magnetic Field Laboratory (HLD) enables unique experiments to be conducted with an FEL in a pulsed magnetic field of up to 70 T. The low-frequency THz facility TELBE provides coherent, high-field THz radiation with either full 100% bandwidth from a diffraction radiator or ~20% bandwidth from an 8-period undulator. Most of the research done is in solid-state physics, focusing on semiconductors or correlated-electron materials.

Heinz Maier-Leibnitz-Centre (MLZ): MLZ is the leading centre for cutting-edge research with neutrons and positrons in Germany and one of the leading centres in the world. Operating as a user facility, the MLZ offers a unique suite of high-performance instruments using neutrons and positrons as probes. The MLZ cooperation includes the Technical University of Munich (TUM), Forschungszentrum Jülich (FZJ) and the Helmholtz-Zentrum Geesthacht (HZG) and it is funded by the German Federal Ministry of Education and Research, the Bavarian State Ministry of Education, Science and the Arts, and by the partners of the cooperation. Based on its long-standing tradition in science with neutrons, operating neutron sources and developing innovative neutron instrumentation, TUM owns the high flux neutron source “Forschungsneutronenquelle Heinz Maier-Leibnitz, FRM II”, the scientific exploitation of which happens under the umbrella of the Heinz Maier Leibnitz Centre, (MLZ). MLZ currently has 26 instruments in user operation another seven are under construction (16.5 operated by TUM, 12.5 operated by FZJ, 2 operated by HZG, 2 operated by MPG). The MLZ offers high flux experiments using neutrons with energies ranging from neV to MeV and a high brilliance positron source.

3.2 The consortium within the NFDI

The DAPHNE4NFDI consortium generates each year more than 28 PB of curated data from thousands of user experiments in a broad range of scientific fields. These high quality and well-defined data sets will be made available for the whole NFDI on levels of both raw and processed data. With this, DAPHNE4NFDI provides data and information for the NFDI ranging from protein structures, crystals, magnetic and electronic properties, dynamic properties, biophysics, health, life science, material science, engineering and cultural heritage, etc. This diversity in the field in turn enables connection and cross-cutting activities with many other NFDI consortia.

During discussions with other potential NFDI consortia, the following three categories of NFDI consortia have been identified as being relevant to DAPHNE:

- **Category 1**: Consortia which are connected to DAPHNE via their scientific scope and a potential overlap between their user communities. These are the consortia FAIRmat, NFDI4CAT, NFDI4Chem, NFDI4Ing, NFDI-MatWerk and possibly NFDI4PHYS (no proposal this round). First steps have been taken here and outlines for cross-correlation have been drafted at a meeting in Berlin on 6 May 2019. In addition, the connection with Math4NFDI is promising in terms of schemes for analysis, such as artificial intelligence (AI) or machine learning (ML) methods.
- **Category 2**: Consortia which are connected to DAPHNE via their use of large-scale facilities, the large data sets generated and their collaboration within the framework of the BMBF ErUM Data. This is currently the consortium PUNCH4NFDI, comprising the communities for astrophysics, and hadron and high energy physics. The co-speakers of these consortia include many of the persons working on the strategic recommendations for the digital transformation in the research field ErUM, who have already been collaborating in formulating a portfolio of measures over the past year. We want to explicitly use this synergy for the NFDI and develop a coherent approach that includes actions on the BMBF funding side (still to be defined).
- **Category 3**: These are consortia in the broader context of the NFDI. Here we see a generic overlap in terms of data policies, metadata schemes and the generic outline of user services. One example of this is the Berlin declaration on the scope of the NFDI.

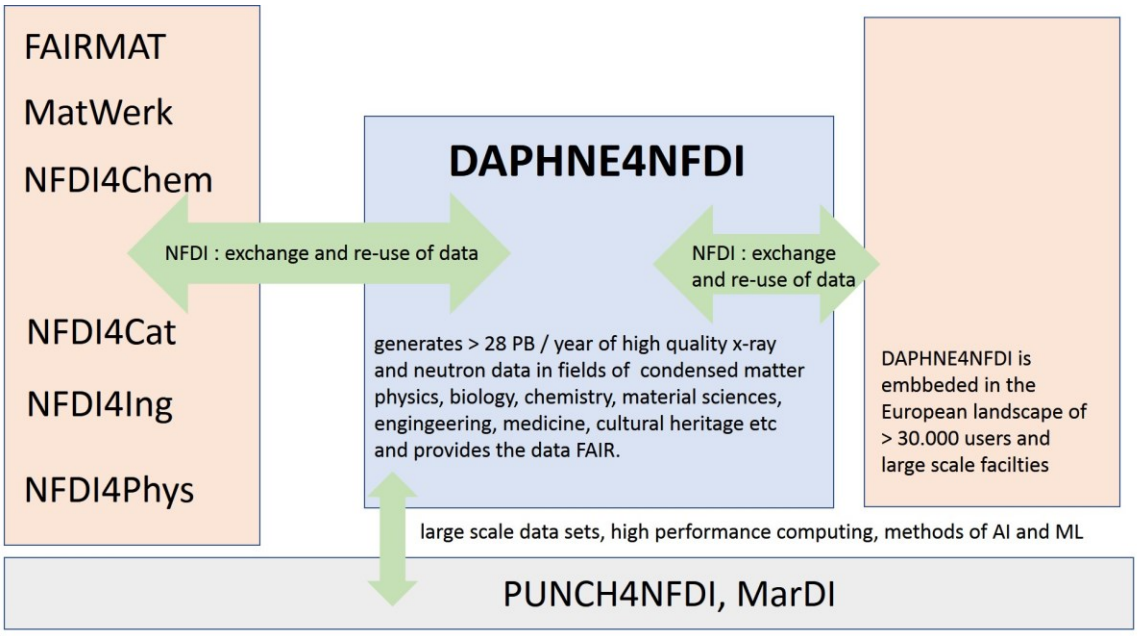

Fig. 3.3: DAPHNE4NFDI in the context of the NFDI: interactions and exchange of data, methods and workflows.

3.2.1 Relationship to other consortia

Relationship to FAIRmat

DAPHNE focuses on its multidisciplinary user community at photon and neutron sources and bridges various areas form physics, chemistry, biology to engineering. FAIRmat aims for making data FAIR in the material sciences and focuses on laboratory data also with strong links into theory and simulations. The approaches and user community of both consortia complement each other in the area of condensed matter physics, where data from both consortia including theory and simulations are of mutual scientific interest. We will establish, harmonise and foster exchange with the FAIRmat consortium.

Relationship to NFDI4CAT

Scientists at NFDI4CAT use synchrotron and neutron facilities for catalysis research due to the unique possibilities for in situ/operando spectroscopy and structure determination on a hierarchy of length scales. Scientists in DAPHNE investigating in situ/operando processes are interested in catalysis data from NFDI4CAT. Use cases will be developed between the two consortia with strong interrelation.

Relationship to NFDI-MatWerk

NFDI-MatWerk aims to set up a digital platform for the materials community investigating the microstructure of materials and its relation to mechanical and functional properties. Its community constitutes an important user sub-community of DAPHNE. Consequently, MSE-related metadata and catalogue specifications will be reconciled. While DAPHNE focuses on managing provided

raw/reduced data, NFDI-MatWerk will develop ontologies for materials definition along the complex details of a sample life path, which could be inherited in the DAPHNE (MSE related) infrastructures and effect the developments for other sub-communities.

Relationship to NFDI4CHEM

NFDI4Chem is highly complementary to DAPHNE as it mainly deals with the chemistry and structure on the molecular level. DAPHNE will benefit from the ELN as NFDI4Chem will develop it further to inorganic samples. DAPHNE data repositories are highly interesting for NFDI4Chem - and vice versa - and will be interlinked via the sample identifier, e.g. the NMR-database for molecular structures.

3.2.2 Integration into the wider NFDI landscape

Metadata and data management:

Outreach and dissemination into the broader scientific community and industry by DAPHNE: The partners of DAPHNE are broadly distributed in terms of their fields within the natural sciences, but still have well-defined data providers determined by the large-scale facilities involved. Thus, we see the potential for DAPHNE to efficiently develop coherent standards for (meta)data, data management and best practice examples that are acceptable/interesting to the broader user community, and therefore disseminating these in the broader context of NFDI. DAPHNE also has considerable experience in handling and managing large data sets. DAPHNE reaches into and is of interest to smaller university groups involved in a wide range of activities within the natural sciences. DAPHNE also represents industry users.

Helping to implement data policies/data structures by structuring user communities: A large part of the NFDI is about communication. Communication is needed between data consumers and data providers so as to establish data policies such as open data, for example, or to define (meta)data standards. DAPHNE is well prepared to tackle this problem due to the structured user organisation and the coherence of the data flow coming from a few large-scale data providers only. Here, DAPHNE can contribute to the overall NFDI by best practice examples and successful workflows in defining data policies and organising a larger community.

Implementation of FAIR principles on a European level

The German user organisations KFS/KFN are active in DAPHNE and the photon and neutron facilities, but are also embedded in a larger European context with well-organised structures – see section 3.4. We therefore envisage that DAPHNE can contribute substantially to the international/European aspect of the entire NFDI.

Education/training

The education and training of future scientists and the aspect of generating awareness for data management matters are of considerable importance for DAPHNE. Introducing these issues into university curricula in the natural sciences is a great challenge. Here, DAPHNE will set up university courses with the help of organisations such as DPG, in cooperation with the universities and data science schools and other NFDI consortia. Our broad user community combined with the science-driven needs of the DAPHNE user groups will help to generate enough (science-driven) interest in advancing university curricula. The context of ErUM Data funding is also important here. Within the ErUM Data framework, we aim to introduce a tenure track university program for data scientists.

3.3 International networking

DAPHNE is linked to and has connection with the following European and international organisations of photon and neutron sources and their respective user organisations.

ESUO - The European Synchrotron User Organisation is the elected body of the European photon user community. ESUO comprises 30,000 users of synchrotron and FEL radiation from 30 European countries (including Israel and Turkey). In Germany, the KFS sends representatives to ESUO. The general mission of ESUO is to coordinate the activities of synchrotron and FEL radiation users in Europe and to provide support to the users in order to get access to synchrotron and FEL beamlines in Europe. ESUO coordinates activities on a European level, with a special emphasis in recent years on ensuring financial support for European users when travelling to European synchrotron and FEL sources.

ESUO will also play an important role as a user organisation in European data projects such as PaNOSC, ExPaNDS and others. Hence, DAPHNE's collaboration with ESUO leads to an ideal connection to the European user community and will be able to act on a truly European level.

ENSA – the European Neutron Scattering Association. With approximately 4500 registered researchers, representing about 7000 researches, the European neutron user community is the largest neutron scattering community in the world. In total, 20 national associations built up ENSA by national delegates (for Germany sent by the KFN). As the community's association – comprising of academic and industrial users who exploit neutrons beams in various ways for different applications – it contributes to define the community's scientific needs with regard to

instrumentation, availability and infrastructure, as identified in the different countries. ENSA organises the ECNS conference every four years and awards three prizes biannually. The current initiatives focus on the coordination of activities related to the chances in the scientific environment, the neutron landscape and the financial support of users in Europe.

LEAPS – the League of European Accelerator-based Photon Sources – is a strategic consortium initiated by the directors of the synchrotron radiation and free electron laser (FEL) user facilities in Europe. LEAPS comprises all major European synchrotron and free-electron laser facilities, including all the facilities that are partners of DAPHNE. Its primary goal is to actively and constructively ensure and promote the quality and impact of the fundamental, applied and industrial research carried out at their respective facility to the greater benefit of European science and society. Part of the mission of LEAPS is to engage with stakeholders and organisations such as the European Commission and national funding agencies in all matters relevant to the development and long-term sustainability of synchrotron and FEL based research, with the objective of informing and shaping future policies and funding opportunities. LEAPS will engage with the current and potential user communities to discuss their respective needs and anticipate and meet future challenges and promote greater coherence in developing data policy, handling, storage, analysis and access, and in promoting open science.

LEAPS is associated with SESAME (Synchrotron-light for Experimental Science and Applications in the Middle East), the first synchrotron facility in the Middle East, which is located in Jordan. LEAPS hosts the web page www.leaps-initiative.eu and is in communication with user organisations such as the KFS and ESUO.

LENS - The League of advanced European Neutron Sources (LENS) is a not-for-profit consortium formed to promote cooperation between European-level neutron infrastructure providers offering transnational user programs to external researchers. The network of international and national neutron sources in Europe is a world leader and serves a scientific community of more than 7,000 researchers, with over 32,000 instrument days per year. Nine of these neutron sources have formed a strategic consortium with the aim of strengthening European neutron science by enhancing collaboration among the facilities. LENS places emphasis on the relationship between user communities and funding organisations, continuous improvement of source facilities, optimising resources between and aligning policies among partners – all to ensure excellence for the communities they serve.

PaNOSC project - The Photon and Neutron Open Science Cloud (PaNOSC) is a European project (financed by the INFRAEOSC-04 call) for making FAIR data a reality in 6 European research infrastructures (RIs), which develops and provides services for scientific data and connects these to the European Open Science Cloud (EOSC). The following research infrastructures contribute

to PaNOSC: European XFEL, ESRF, ILL, ESS, ERI, the European Grid Infrastructure and CERIC. DAPHNE is directly connected to PaNOSC via the common partner EuXFEL.

PaNOSC has the following objectives

1. Participate in the construction of the EOSC by linking with the e-infrastructures and other ESFRI clusters.
2. Make scientific data produced at Europe's major photon and neutron sources fully compatible with the FAIR principles.
3. Generalise the adoption of open data policies, standard metadata and data stewardship from 15 photon and neutron LSFs and physics institutes across Europe
4. Provide innovative data services to the users of these facilities locally and the scientific community at large, via the European Open Science Cloud (EOSC).
5. Increase the impact of LSFs by ensuring that data from user experiments can be used beyond the initial scope.
6. Share the outcomes with the national LSFs, who are observers in the proposal, and the community at large, to promote the adoption of FAIR data principles, data stewardship and the EOSC.

ExPaNDS is a federation of 10 European photon and neutron research infrastructures and the e-infrastructure EGI. The goal of the project is to set up a platform for data analysis, as a service for users from research institutes, universities, industry, etc., thus enabling EOSC services and providing coherent FAIR data services to the scientific users of photon and neutron sources. Within the framework of ExPaNDS, partners will maintain and develop a catalogue of data and analysis software for PaN data, and will cooperate with the EOSC governance bodies to further improve the EOSC. Sharing and exchanging the benefits of ExPaNDS with other clusters connected to the EOSC is also an activity of key importance.

DAPHNE has close ties with ESUO/ENSA and LEAPS/LENS and is therefore ideally suited to actively participate in European initiatives such as PaNOSC and ExPaNDS. Solving FAIR data and other challenges, such as metadata creation and curation, on a European level is a realistic goal for the photon and neutron community.

3.4 Organisational structure and viability

The DAPHNE consortium is governed by an executive committee (EC), a steering committee (SC), the Consortium Assembly (CA), and advised by an International Advisory Board (IAB) and a Technical Advisory Board (TAB).

The executive committee (EC) is formed by the task area leaders (co-speakers) who are responsible for execution of the project. The EC is ultimately responsible for resource allocation within the project on the advice of the SC and is thus responsible for meeting deliverables within budget and on schedule. The applicant institution is in charge of day to day project management and additionally plays the role of chief executive within the EC and SC. The EC is also responsible for allocation of the in-kind contributions of the partners to the work packages of DAPHNE.

The steering committee (SC) is formed by the spokespersons representing each of the funded institutions. The KFN and KFS each send one representative. The SC advises the EC on all matters of the consortium, including providing advice regarding the distribution of funds to participants and efforts allocated to the tasks.

The consortium assembly (CA) is composed of all participants independent of whether or not they receive funding. The CA will set up science clusters, within which participants define specifications and organise workflow tests for all tasks, discuss and suggest the topics that are relevant to their area of expertise within the framework of DAPHNE, and invite external experts when necessary. They set up and follow use cases, dealing with all the relevant aspects of TA 1-5. The topics to be covered include, for example, metadata schemata, data formats and bets practice workflows. The science clusters act as advisory panels for TA 1-5. The number of science clusters and the topics covered are to be decided and confirmed at regular intervals in the CA.

The technical advisory board (TAB) is formed by the IT experts of the facilities and advises on state-of-the-art solutions as well as the overarching strategy concerning the facilities involved and the European partners. The TAB advises on technical solutions of software management, roll-out, version control etc. and IT infrastructures hosted at the facilities. The TAB also advises on allocation and distribution of the facility in-kind contributions.

For the **International Advisory Board (IAB)** external experts and members from the KFS, KFN, ESUO, ENSA, LEAPS and LENS will be invited, balancing the different scientific and IT fields. It advises the Steering Committee on all issues relating to the international cooperation and ongoing strategic and scientific developments.

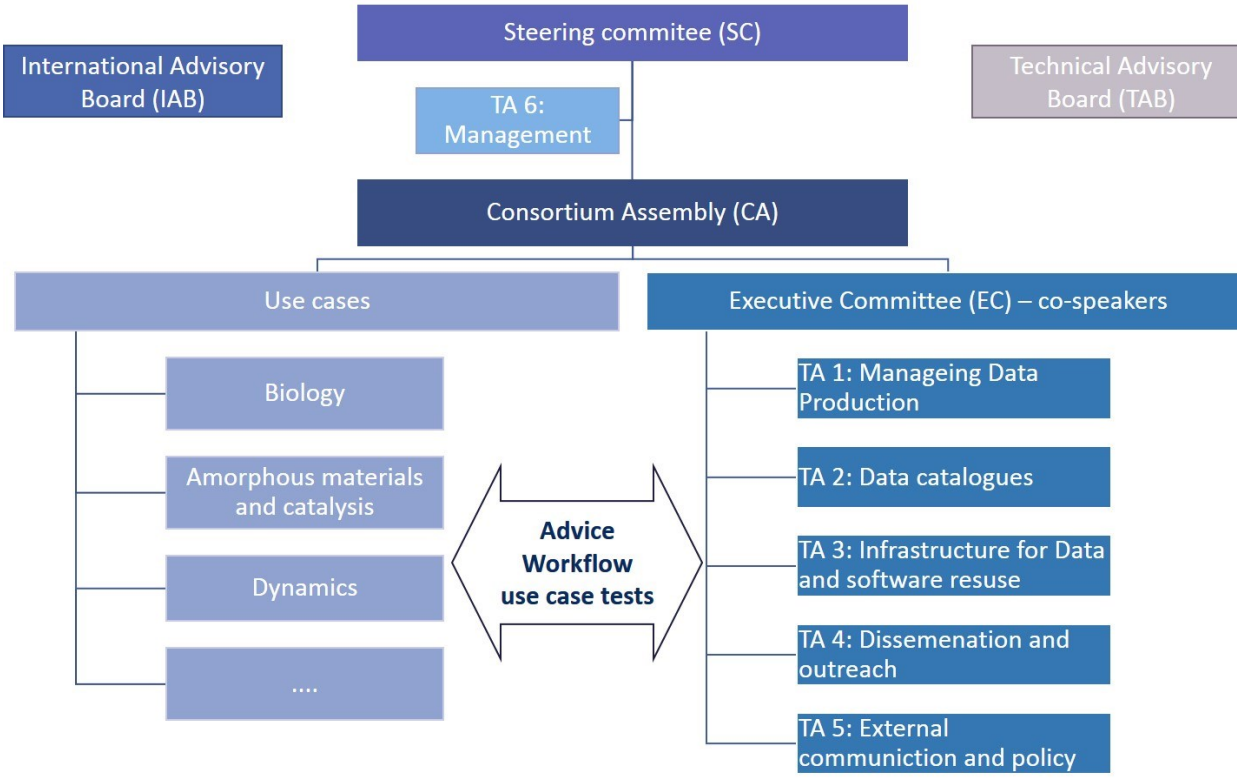

Fig. 3.4: Suggested governance structure of the DAPHNE consortium

The running of the consortium, management of tasks amongst the participants and allocation of resources is managed in TA6: Management (i.e. reports, finances, follow-up of deliverables, tracking of participant progress towards deliverables, and distribution of resources). Responsibility for meeting deliverables within budget and schedule lies with the body responsible for resource allocation within the project, which in the case of DAPHNE is the executive committee (EC).

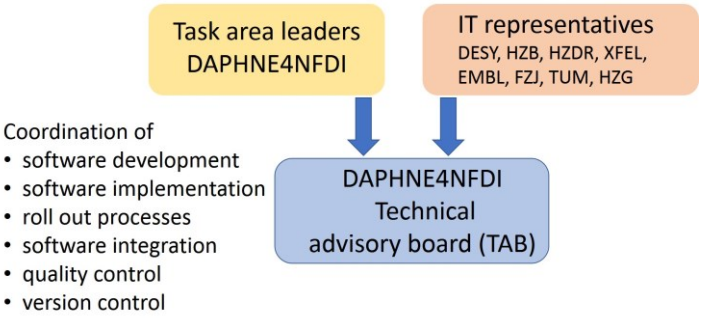

Fig. 3.5: The DAPHNE4NFDI technical advisory board (TAB) comprises of representatives (IT specialists) from the facilities and of the consortium to ensure sustainable software development and data curation.

Organisational structure: cooperation users - facilities in the task areas (TAs)

A close collaboration between users and facilities is key for the success of DAPHNE. We enforce this interaction process by involving both parties into the task areas. This will ensure that (a) the

user needs and demands of the specific science/technical questions are met and (b) that these demands are translated into technical solutions that can be incorporated at the experimental stations in a proper way. This will be achieved with the work model displayed in Fig. 3.6 involving all the key players: beamline scientists, users and IT specialists. *Here the in-kind contributions and the IT centres from the facility side play an important role as they enable us to implement sustainable software solutions into DAPHNE.*

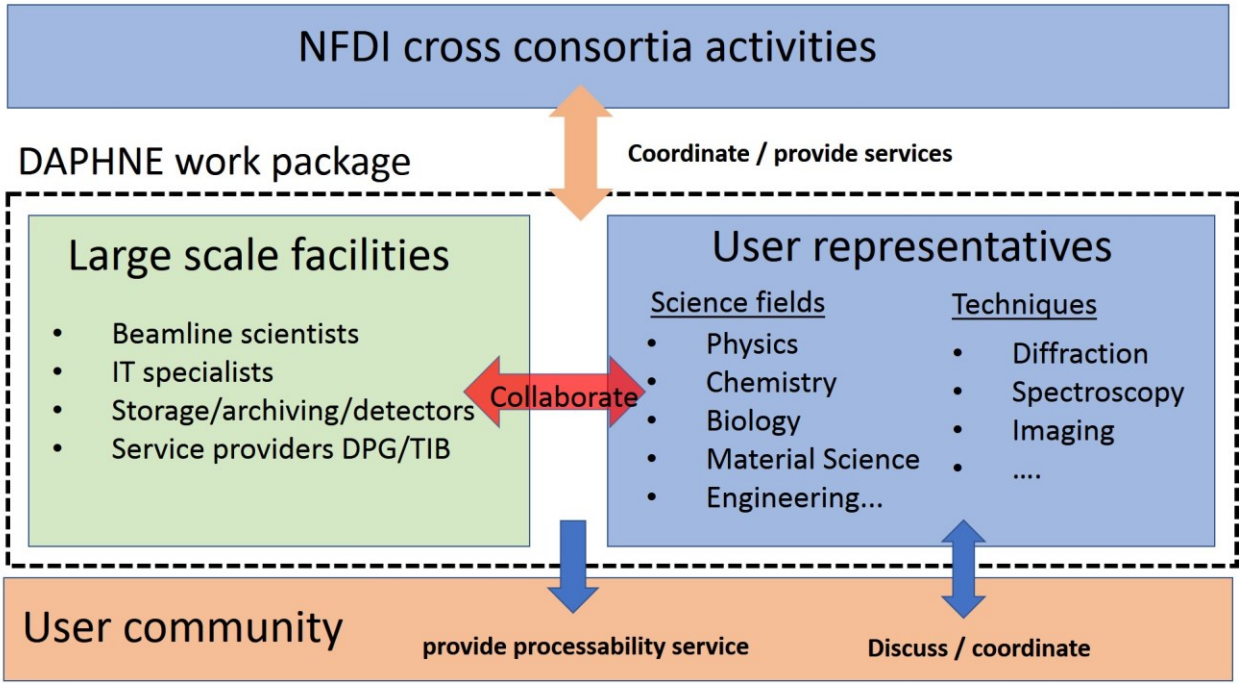

Fig. 3.6: Interaction within DAPHNE4NFDI between facilities and users via the DAPHNE4NFDI representatives.

3.5 Operating model

3.5.1 Concept of DAPHNE4NFDI

- The goal of DAPHNE4NFDI is to make x-ray and neutron data FAIR for the users AND for the NFDI. **The tools generated will be open source, publicly available and free of charge.**
- DAPHNE4NFDI has access to the scientific infrastructure from six large-scale research facilities generating > 28 PB of data annually.
- The research facilities who operate the x-ray and neutron sources will provide their IT infrastructure for realising DAPHNE4NFDI ("own contribution" of the facilities, consortia members will offer these services free of charge). Positions assigned to the facilities within DAPHNE4NFDI will be used to set up and operate the corresponding parts of the IT infrastructure, or to execute technical aspects closely connected to the facility such as interfaces to instrumentation.
- The IT centres of the facilities will contribute considerable in-kind contributions in terms of IT infrastructure and 38 FTEs/year for software development, release and coordination management, data curation and archiving to the project.
- The in-kind contributions from the universities amount to 9.5 FTE/year plus access to university computing centres.

The DFG funding is needed to realise this unique consortium between users and facilities. Only the funding ensures that the process is science driven and bottom-up with the users as important player. The challenges of realizing online logbooks (ELN), common (meta)data schemes and data catalogues which are accepted from all users and used on a daily basis in our science are enormous and needs the NFDI funding.

In general, DAPHNE will be conducted along the following principles:

- Prospective users of the photon and neutron facilities in DAPHNE need to accept data policies in agreement with the FAIR principles for obtaining measurement time at the facilities.
- The users accept and adhere to good scientific data standards such as providing meaningful and high-quality meta-data.
- The tools of DAPHNE are user friendly and developed from and within the science fields from the user community.
- During an embargo period of typically 2-5 years (see TA5, policies) the users who proposed and performed the research have exclusive access to the data. After this period the data will be made publicly available under licenses such as CC-By-4.0.[5]
- DAPHNE establishes automatic work flows from proposal → experiment → raw and processed data → data repositories/catalogues, collecting and saving the relevant

(meta)data along the entire chain. DAPHNE will provide persistent identifiers (DOIs) for data sets, including bibliographic data of the researchers that collected the data originally.

- Databases, catalogues and repositories will be available to the whole NFDI upon registration at DAPHNE.
- DAPHNE will develop tools and services with other NFDI consortia to allow the cross-use of data between the consortia.

4 Research Data Management Strategy

4.1 State of the art and needs analysis

Currently, the large-scale photon and neutron facilities which host the instrumentation and where the data is generated have their own data management strategies. Here, the effort is focused on ensuring that measured data can be saved at an adequate data rate and preserved on disk for some period of time. The focus at the facilities is very much on collecting and saving data from instrumentation. Recording of metadata related to sample information, sample environment, or intention of the experimenter is largely left up to the individual user groups and is often in paper form. In many cases, the experiment is recorded in hand written physical logbooks that of course have no standard format and are decoupled from the data taking. There are a few exceptions – for example, in macromolecular/protein crystallography (MX), the established workflows result in largely standard measurements and data acquisition which is amenable to metadata capture through databases. In some cases, MX is offered as a mail-in service. These kinds of measurements can serve as a role model for the rest of the community. Currently, at the majority of instruments users are responsible for defining the collection of their own metadata, while the facility records the data for the user to take home. In most cases, users have only a short period of time in which to copy their data home before it is deleted at the facilities. User groups then become responsible for the long-term preservation and archiving of the data, which is far from standardised. Only a few instruments or facilities promise to centrally archive all data with curation limited to a list of directories corresponding to proposal IDs. Currently it is rare that facilities assign data DOIs for tracking data provenance.

Therefore, there is great need to improve and ideally standardise data curation across the photon and neutron facilities. This is particularly acute in the areas of:

- Metadata collection attached to the measured ‘raw’ data and collected at the time of measurement in a standard and searchable manner;
- Preservation of raw data alongside intermediate results and with links to publications so that published results can be traced back to raw data and analysis in a standardised manner;
- A curated ecosystem of analysis software available to the entire research community, without which it is impossible to re-use or even make sense of the data.

Research data production

Research data in the x/n communities is taken at the large-scale facilities. In a typical workflow, a specific material (sample) is grown or produced and subsequently characterised both at the home laboratory as well as at the large-scale facility, employing one or more beamline. The

(meta)data created at this home laboratory stage mostly existing as hand-written logbooks. The interesting samples are investigated in detail at the facilities by applying specific techniques such as crystalline structure, dynamics, electronic and magnetic properties etc. Users are, by and large, composed of small user groups operating on a university level or research institute, with the data analysis performed by PhD and postdoctoral students. **These individual groups currently process and store data locally.** They perform their experiments at the facility and then take their data home for analysis. **This is a challenge in itself as using** two-dimensional detectors may involve transferring TBs of data to the home institutes. With these enormous data volumes and the increasing complexity of data analysis methods (and underlying software development), the concepts of data stewardship, data policies, online and offline data reduction and analysis, reproducibility as well as a metadata catalogue are essential for the successful operation and support of excellent science.

Although the research data at the facilities is digitally stored, many intermediate experimental steps and data processing scripts and the sample descriptions are often not properly digitized or in a standard format. The current workflow comprises of the use of hand-written logbooks by both users and facility staff, in addition to the digital capture and storage of (meta)data and analysis programs, and of processed data. Further, the use of data catalogues and repositories is not well established in the community, mostly because science driven solutions are non-existent. This results in manual/digital hybrid situations as the current status quo.

One example of advanced FAIR data management in our context is the field of MX. We aim to learn from this successful example and to adopt similar procedures and workflows for other areas within the different TAs in DAPHNE. In MX, diffraction data is typically collected on two-dimensional area detectors by illuminating a crystal with photons or neutron, while the crystal is rotated around a single axis. With a current-day setup, several thousand data frames of **raw data** with a total volume of GB are collected in few minutes at synchrotrons.[6] For well-diffracting crystals, subsequent **automatic processing** reduces the diffraction intensities measured on the images acquired by the detector into structure factor estimates (**reduced data**) in minutes, reducing the data volume to typically to MB. **Models** for the structure of the molecules present in the crystal are then be build, refined and validated against the **reduced data**.[7] Eventually, the **final model** and the reduced data are deposited into the **Protein Data Bank (PDB[4])**, where models and their fit to the data are validated and the results of the validation become attached to the database entry in the form of a **validation report**. For publication of structural results obtained for MX, for most journals, **deposition of the model and the raw data is obligatory**; inspection of the validation reports is part of the peer-review process.

Example: Protein Data Bank

The deposition of models and reduced data became possible when the Protein Data Bank was established in 1971.[8] While a policy for data deposition was published by the International Union of Crystallography in 1989, [9] deposition of data became a requirement for publication in major journals only in the late 90's of the last century.[10,11,12,13] Deposition of reduced experimental data became obligatory in 2008.[14] By now, for every structural model referenced in a publication, the PDB-id is stated in the publication and allows to link into the relevant PDB-entry.

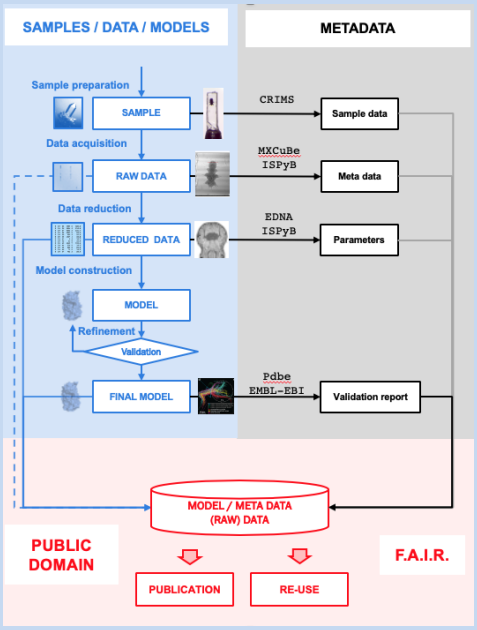

Fig. 4.1: Workflow and metadata management in the field of macromolecular crystallography.

The IUCr Diffraction Data Deposition Working group (DDDWG) in their final report[15] has formulated recommendations with respect to the archival of raw diffraction data including that such data should be permanently accessible following FAIR principles via a Digital Object Identifier (DOI). Meanwhile, several repositories are accepting raw diffraction data. These include general purpose facilities such as ZENODO[16] field-specific resources such as BioStudies[17], or facilities specific to the domain of Structural Biology, such as SBGRID,[18] proteindiffraction.org,[19] or CXIDB.[20] Synchrotron and neutron facilities have started to provide services for raw data storage and access (ESRF, PSI, DLS,ILL) embedded into European initiatives such as PANOSC and ExPaNDS.

Other examples are the Coherent X-ray Imaging Databank (CXIDB)[20] which is, however, comparatively small and far less structured. In x-ray absorption spectroscopy, XANES and EXAFS databases mostly exist with power users and they exchange data but certified reference databases hardly exist.[21] Present day databases for x-ray spectroscopy (EXAFS), such as the Farrel Lytle database[22] and the XAFS spectra library at the University of Chicago[23] suffer from problems such as not providing detailed information about the sample itself, unknown and inconsistent data formats, difficulties in adding to the database, non-standard organisation of the database, and no review process to ensure and assess the quality of submitted data. Many other communities are not even on that level because experimental parameters and metadata, experimental setups, sample descriptions and analysis schemes are much more complex and heterogeneous. Reuse of original data files is usually rare and only few researchers have defined workflows and curation standards with final storage procedures often in the hands of the individual research groups.

Large-scale research facilities:

The large-scale facilities in Europe work on and offer solutions for capturing the descriptive part of complex metadata. Examples are SciCAT, a development, driven by PSI, ESS and MAX IV. EuXFEL uses MyMdC, HZB ICAT (a product of the PaNDATA project), and HZDR INVENIO. The ExPaNDS and PaNOSC projects both work to improve the interfaces between all these different implementations, aiming to create a federated data catalogue. However, these initiatives are mainly centred around instrument and beamline parameters, and need to be complemented by meaningful metadata about samples and actual experiment conduction – information that is located with the user and often science field specific. Moreover, common standards for metadata definitions that are meaningful beyond a confined scientific area are largely missing. Here, DAPHNE will make a difference in bringing the users and facilities together and working together.

Furthermore, some facilities provide on-site data processing and limited data curation. For example, the EuXFEL is one of the more modern facilities, having started operation in 2017. About 19 PB of raw experimental data have been collected so far (June 2020). Data analysis and processing capabilities are provided on-site to users through the Maxwell computing cluster, which has a theoretical peak performance of approximately 600 TFlops, and which can be accessed using a cross-platform remote desktop service or through a JupyterHub, for notebook-based analyses. Due to the size of the data sets, the (initial) data analysis and reduction is generally carried out on the Maxwell cluster, which is used by staff and users, often from remote destinations. In addition, raw or reduced data sets can be exported using FTP or Globus-based data transfer services. European XFEL works closely with DESY, in particular in the fields of data management and data analysis that are relevant to this call. EuXFEL currently has storage capacity for about 30 PB of experimental data, and by necessity plans to expand storage capabilities to 100 PB by 2023, accompanied by an upgrade of computing resources. Data production is expected to grow faster, and it will be essential to have data reduction mechanisms in place in the future. Apart from the resources provided on site, EuXFEL is currently investigating how to extend its computing model to include remote resources provided by external partners or as part of European Open Science Cloud. The EuXFEL is a prime example of where taking raw data home is impractical, yet the effort on-site remains focussed on getting the data to disk, with the level of data curation (metadata capture, availability of on-site analysis software) leaving much room for development.

Other facilities are also putting measures in place to improve the research data management. Examples include the quality management system at user beam time projects at BESSY II, the unified (meta)data collection and instrument control system NICOS at MLZ which already captures most auxiliary (meta)data during experiments. Sample description and unique identification is also pursued via databases such as the well-established Information System for Protein Crystallography Beamlines (ISPyB),[24] the sample and measurement database LAMMB at

HZB, or SampleDB at FZJ, and the Helmholtz Metadata Collaboration (HMC) which is part of the Helmholtz Data Federation.

In summary, first steps into the direction of FAIR data have been taken in the photon and neutron community. The large-scale facilities in their role as data generators and data curators are very active in this area and are making progress. Pioneering university groups have started their own data catalogues (KIT), introduced ELNs (U Kiel) and many have developed analysis software. However, the initiatives are still scattered and the vast majority of the x/n community is far away from an overall FAIR concept of their data. User and facility overarching solutions are missing in Germany. Long-term data repositories, efficient data reduction and analysis procedures connected to rich and annotated sample information has to be developed. The transformation to FAIR data requires not only technical developments but also a cultural change and good dissemination within the community including education of the next generation of researchers are central to the success of DAPHNE.

4.2 User interaction and integration in DAPHNE4NFDI

The needs of the community have been monitored by surveys and workshops (see chapter 3).[25]

Communication

The communication between users and DAPHNE is established, via (see chapter 3)

- The elected user organisations KFS and KFN
- Workshops and annual user meetings organised by the facilities. For example, the annual DESY/XFEL user meeting comprises keynote talks, dedicated workshops, a very successful poster session and public evening lectures, and is attended by more than 1000 participants. Similar events take place at all large-scale facilities.
- Users serving on the scientific and governance advisory boards of the facilities.
- Chairs and members of the proposal and beamline review panels, who are experienced users and submit advice on strategic needs, new trends and problems to the facilities.

Feedback mechanism with the community

In the framework of DAPHNE we implement the following feedback mechanisms

1. TA4 and TA5 ensure a continuous communication towards and feedback from the scientific community who want to use the data by means of workshops, surveys etc.

2. The users who measured the data are also involved in DAPHNE directly via their elected user representatives KFS and KFN.

3. The user groups actively participating in DAPHNE represent typical power-user groups which are deeply rooted in their community and have long standing connections to the facilities and the beamline scientists. Those DAPHNE groups are internationally active and perform experiments not only across Europe but also in the US and in Asia. Thus, the groups within DAPHNE will also give highly qualified feedback to the leaders of the task areas.

4. Use case demonstrations and prototype implementations will be made available at selected beamlines to collect feedback from the wider community.

The results of the feedback are to be presented and discussed in the consortium assembly annually. A report (TA4 and TA5) is to be presented to the steering committee, to the technical advisory board and to the international advisory body. The steering committee will develop appropriate actions by implementing changes and adjustments to the overall workflow or to technical implementations. The actions are presented to the advisory bodies who will advise the steering committee. Following assessment changes will be implemented by the management task TA6.

4.3 Metadata standards

The creation, storage and management of metadata differ between scientific fields within the neutron and photon communities. Aside from the sheer amount of data, its complexity is also increasing. At present, the community is experiencing a transition in data curation from the users to the facilities in terms of responsibilities. This also affects data policy and data ownership with a transition from users to the public. Hence, metadata and open data are a very important issue for researchers and the providers of information and services.

Metadata at photon and neutron facilities can be split into more generic parts that are largely independent of research field, and parts that depend strongly on a specific field/discipline. X/n facilities are generally well equipped to deal with the former and information on instrument settings and bibliographic metadata are typically collected in digital form and part of the data aggregation created and stored for each experiment. Metadata standards, common vocabulary and common data formats exist to various degrees, mainly along similar instrument classes, i.e. beamlines that employ similar methods. Some of the facilities already ingest these data into catalogues and assign DOIs for tracking provenance. However, adoption is patchy and there are at times multiple standards. Due to the wide range of different communities that are using photon and neutron facilities, systematic capture of the sample description (such as sample persistent identifier, PID) and the actual conduction of the experiment is less advanced compared to the automatically captured instrument data. There are some exceptions, for example ISPyB, which was established at DLS and is used at ESRF for MX beamlines. A big potential is seen in the field of x-ray absorption spectroscopy, where many groups have their own databases but there is presently no

national or international platform for sharing that data. Such an initiative, also initiated by IUCr, can be considered a seed for further data catalogues. Different approaches/standards (e.g. keyword catalogue for metadata, etc.) need to be discussed for sub-community dependent metadata, in particular. One of the challenges for DAPHNE will be to provide the necessary infrastructure and tools to connect the relevant sample metadata with the instrument metadata in a systematic manner. Metadata schemata and ontologies need to be discussed and developed in cross-community activities. In this context, data formats need to be compared as well as ways of coping with the data from next generation x-ray detectors, which inevitably requires the evolution of HDF5 data formats. The same applies for the validation of standard scalable solutions for high data rates.

The minimum metadata standards required for photon and neutron data should enable later re-use of the measured or deposited data. This includes most importantly a clear description of the method used, the sample studied and auxiliary parameters specific to the experiment, all information necessary for data processing (or the already performed processing steps), raw data provenance, and information on ownership. DAPHNE will address these topics mainly in TA1 and TA2 – working towards well specified, community wide accepted and registered (meta)data vocabularies and schemata, the introduction of sample PID (in coordination with other NFDI consortia), assignment DOIs for experimental data, as well as improvement of automatic capture of metadata and the adaptation of (machine readable) ELNs.

- **Metadata standard for photon and neutron experiments requires**: sample persistent identifier (PID) and description, electronic logbooks (ELNs), user deployed experimental setup parameters, beamline parameters, photon and neutron data set identifier (DOI), and processing routine/software.
- **DAPHNE4NFDI coordinates, integrates** and also represents the existing user community via its consortium structure and is ideally suited to **integrate the sub-specific metadata standards**.
- DAPHNE4NFDI aligns the metadata standardisation on an international scale by closely collaborating and integrating the international partners.

4.4 Implementation of the FAIR principles and data quality assurance

Making photon and neutron data FAIR to all fields is a major challenge. DAPHNE serves the neutron and x-ray user community, and interfaces to related consortia in the natural sciences. However, making FAIR data accessible to other consortia across the various disciplines is a considerable challenge. The interoperability, storage, archiving, definition and use of complex

metadata structures beyond a single discipline need to be addressed by the NFDI as a whole. DAPHNE aims to set the stage for the photon and neutron community to address the possible technical solutions as well as to promote the acceptance and the cultural change associated with the adoption of FAIR data. The MX community is advanced in the application of FAIR data principles. One of the objectives of DAPHNE is to transfer this best practise example to other areas of the x/n communities.

All task areas of DAPHNE are therefore aligned along one or more aspects of the FAIR principles. TA1 is dedicated to the automated capture of metadata standards, thereby ensuring to introduce as much information and meaningful annotation as needed in digital form early in the data generation process. Developing common standards and well-aligned architecture of data formats will improve the interoperability of data sets and analysis tools.

TA2 centres on the definition of (meta)data vocabularies and their registration along with the establishment of searchable and interlinked data catalogues and repositories which will advance the Findable, Accessible and Re-use aspect of photon and neutron data within DAPHNE but also for the wider community. Making software and analysis tools accessible will be a further focus of TA3.

TA4 and TA5 are committed to the intra- and inter-consortia communication, education and outreach to the community. Within this context, DAPHNE will actively pursue policy development, establish best-practices, and promote the FAIR data principles within the communities, with the goal of harmonising data practices across the facilities. We will seek to introduce data management into the university curricula so as to educate the new generation of users.

Additionally, DAPHNE will provide the measures listed in section 2.2 which are currently not available to the user community.

The partners involved in DAPHNE are broadly distributed in terms of their fields in the natural sciences or engineering but still have well-defined data providers, determined by the large-scale facilities involved. Thus, we see the potential for DAPHNE to efficiently develop coherent standards for metadata, data management and best practice examples, which are accepted within a variety of communities, and to disseminate these within the broader context of the NFDI.

Communication and outreach form a large part of the goals of NFDI. Communication between data users and data providers is essential for establishing data policies such as open data, for example, or for defining metadata standards. DAPHNE is well suited to tackle this problem in view of its structured user organisation and the coherence of the data flow, coming from a few large-

scale data providers only. Here, DAPHNE can contribute to the overall NFDI by best practice examples and successful workflows in defining data policies and organising a larger community.

DAPHNE4NFDI will implement the FAIR principles in the following way

- Findability: DAPHNE4NFDI will implement measures to capture, generate and store rich and readable (meta)data for the photon and neutron community but also for the scientific adjacent NFDI consortia. This comprises (meta)data and vocabulary specification, online logbook integration, automatic metadata capture, assignment of PIDs/DOIs and standard setting for high performance data formats (TA1). Metadata will be domain specific and DAPHNE4NFDI aims to provide a platform to promote common workflows and common standards for metadata collection at beamlines (TA1). (Meta)data formats and contents will be developed within the German community and harmonised with European projects and efforts (TA5) and disseminated into the community (TA4) and internationally.
- Accessibility: Data provided by the DAPHNE4NFDI consortium is open-access-accessible on our IT infrastructure via standardised communication protocols. This comprises also access to metadata, analysis software and data catalogues/repositories (TA1,TA2,TA3). Data will become automatically open access after an embargo period. Documents, documentation and software generated by DAPHNE4NFDI will be openly accessible e.g. on the DAPHNE4NFDI webpage or GitHub.
- Interoperability: Metadata and data will be available in standardised and broadly accepted open data formats (TA1). Identifiers allow for references to other sets of metadata and data also beyond DAPHNE4NFDI (TA2). Analysis software will be made available by DAPHNE4NFDI also in standardised forms and in script/languages broadly accepted (such as Python/Jupyter notebooks) and open access (TA3).
- Reusability: DAPHNE4NFDI will apply, develop and strengthen our domain-specific data standards by communication with the users (TA4). The consortia will implement these standards into metadata, data, data catalogues and repositories. DAPHNE4NFDI will ensure reusability also on a European level by harmonizing policies and formats with the European partners.

4.4.1 Data selection and data quality management

As outlined in 4.3.1 the recording of metadata by the users and facility staff is crucial for effective data storage and for the implementation of FAIR principles. The facilities provide metadata information about the instruments and facility status. The user provides metadata information about the sample: sample identifiers, description of sample and sample production and handling routines, chemical composition, pre-characterisation information in addition to experimental parameters from user supplied equipment that is not monitored by the facilities (temperature,

pressure, voltage, optical properties, electric properties, catalytic performance, sensing activity, battery charging, pressure). The users need to provide also processed data sets and information about the analysis workflow when depositing data into catalogues and repositories.

The facilities foster and encourage the collaboration with the users through access and data policies that require the contribution metadata information for their experiments.

All large-scale facilities grant access to their beamlines through a proposal system with external peer review for every experiment conducted. Thus, experiments and data entries are performed on the basis of scientific merit and the data generated has a high scientific value. The beamlines and experimental stations are operated by dedicated beamline personnel, where high-end instrumentation detectors and equipment ensure high quality data from the experiments. The principal philosophy is to record as much quality data as possible during the short and valuable beam-time to maximize the access to the rather expensive large-scale facilities.

For efficient data storage, it is necessary to implement quality control of the data as early as possible in the collection process. To deal with redundant data, data rejection schemes are being implemented, for example when no sample is hit by the beam. This step prevents storing blank data volumes. Another approach is to pre-select and only store part of the data for example during alignment. Such decisions cannot be made easily and require an expert decision or valuable data may not be stored and lost forever. Often the approach is to initially collect everything in the raw data files and then do a first data reduction immediately following clearly defined protocols. These protocols need to be clearly documented and referenced in the metadata. The data is currently stored in a site or even experiment specific architecture during the experiment for access by the users and facility staff. For good quality control, an initial visualisation of the data in a simple reduced form or following a rudimentary analysis is required so that good and bad data sets can be tagged for later use. This step is currently carried out by hand and listed in a logbook or an electronic spreadsheet. The incorporation of this step into the ELN with the associated metadata is an aim of TA1. Moreover, TA1 will provide tools to ensure that the minimum (meta)data requirements are met for each aggregated data set. The repository and findability of good data is dealt with in TA2 while the providing the analysis tools is TA3 required to make such a decision.

For reusability, the metadata provided by the users requires labels or tags such as high quality or useful data. This information is not only valuable for the experimental team during and after the experiment but also later so that the data can be identified for reuse by other groups or confirmed independently. In addition, clearly labelling "bad" data is required where for example the sample was not working properly, was not sufficiently clearly defined or where the experimental control failed. Experimental failures like those described may only appear in the aftermath of the experiment during the in-depth analysis. It is therefore important that this metadata information can be added by the user to the data file in the repository later. A possible first implementation is

to label and identify the high-quality data sets upon publications - such as the PDB has implemented.

The storage facilities need to be robust and safe so that data remains findable and available for the agreed time period. In particular, repositories for processed data need quality control as outlined in TA2. TA1, TA2 and TA3 are all concerned with how to assure the quality control of the data storage services and the choices we make concerning data storage. Deciding which metadata schemata to adopt (TA1, TA2) and how to transport it to the community and beyond (TA4, TA5) will be central to the success of DAPHNE.

4.4.2 Role of users and facilities in metadata capture

The facilities provide metadata information on the level of the instruments i.e. typical beamline parameters such as positions of stepper motors on the instruments, specifications of the detector used, information about monochromators, beam position, x-ray or neutron flux, x-ray pulse energy, x-ray pulse IDs etc.

The user needs to provide metadata information, such as description of samples (sample identifiers, description of production and handling routines, chemical composition, pre-characterisation information etc.), experimental parameters from user supplied set-ups (e.g. optical properties, catalytic performance, monitored in parallel). The user needs to provide also processed data sets and information about the analysis tools used when depositing data into catalogues and repositories.

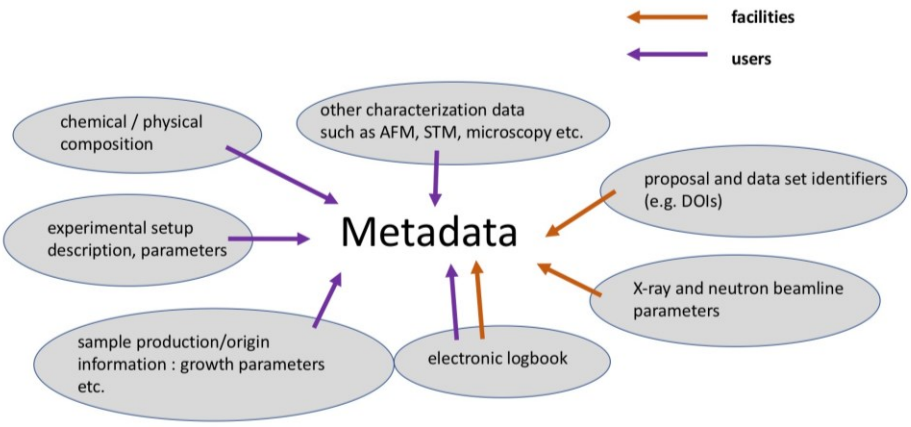

Fig 4.2: Composition of metadata in DAPHNE4NFDI from users and facilities.

The facilities foster and encourage the collaboration with the users through access and data policies which require to contribute metadata information for their experiments.

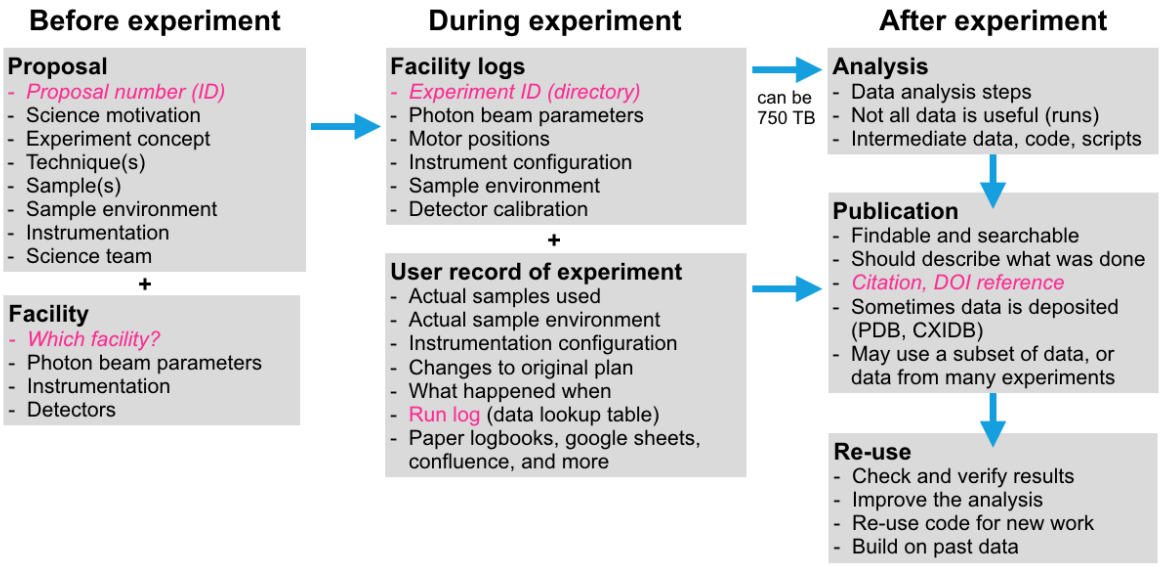

Fig. 4.3: Various points for metadata capture through the experiment cycle for FAIR data.

4.5 Services provided by the consortium

At the large-scale photon and neutron facilities where the data is generated effort is focussed on ensuring that measured data can be saved at an adequate data rate and preserved on disk for a certain period of time. At the majority of instruments, users are responsible for defining the collection of their own metadata, while the facility records the data for the user to take home after which the facility copy is often deleted. Only a few instruments promise to archive centrally all data with curation limited to a list of directories corresponding to proposal IDs.

DAPHNE does not address the saving of raw instrument data nor the transfer of that data to the user's home institution - this service is provided by the facilities as part of their institutional mission.

DAPHNE adds the following services on top of the basic data acquisition services provided by the facilities to enhance the research data ecosystem along the lines of NFDI goals:

- User community driven electronic logbooks (ELNs) linked to and integrated with data collection for the recording of user metadata at the time of measurement in a standardised and searchable manner;
- Science driven definition and curation of (meta)data vocabularies and schemata
- Searchable repositories for processed data, results, intermediate data, and analysis scripts linked to the raw data sources, traceable from publication back to the raw data;

- A curated ecosystem of analysis software developed using professional software development methods available to the entire research community;
- Development and fostering of best practices within the wider research community of facility users.

DAPHNE partners with the facilities to implement and host these services on facility hardware. Indeed, this partnership is highly beneficial as close integration of the DAPHNE services with facility data collection serves the goals of ensuring the long-term viability and impact of services developed by DAPHNE while avoiding unnecessary fragmentation in service provision. It provides a cohesive solution for a diverse community.

The services that are to be developed and promoted in DAPHNE are outlined in section 2.2 and include tools to capture data and metadata efficiently and systematically during experiments, to define and specify suitable metadata schemata for the communities, as well as setting up federated data catalogues and repositories, and the provision of data processing, visualisation and analysis tools. The software infrastructure provided by DAPHNE will be hosted on hardware provided on an in-kind basis by the facilities and made available to all users at no cost.

In summary, DAPHNE4NFDI will pursue /provide

Science driven solutions to formats and standards of metadata by a common approach between the facilities and users to develop side by side these standards (TA1, TA2, TA3). This comprises metadata capture by both facilities **and** users and the development of ELN (TA1, TA2). The user community is fully integrated by means of the KFS and KFN as elected bodies of the user representatives and by active communication with the users (TA4, TA5).

A sustainable analysis software infrastructure for re-use and re-analysis of raw and processed data and for data reduction. Users and facilities will develop, deploy and curate quality assured data analysis software (TA3). Users will have access via a common entry cloud like analysis service (TA4).

Integration of persistent digital identifiers such as DOIs and ORCID references and metadata into the digital workflow (TA1-5) by defining use cases, sample description schemata, and best practice examples together with users and facilities.

Searchable and linked data catalogues and repositories of raw and processed data: Users and facilities will develop, deploy and curate quality assured catalogues and repositories (TA2).

A community platform for photon and neutron users to cover all aspect of research data management and FAIR data principles, including training and education of young researchers.

4.6 Software curation and best practice software development

Professional-level software curation is essential for ensuring the longevity of DAPHNE services as a part of NFDI beyond the project duration.

DAPHNE is backed by scientific computing software development expertise from scientific computing groups at the university partners and especially the facilities who will guide the progress of software development within DAPHNE. For example, DESY plays a leading role in professionally developing the software package dCache for data management [26] including modern, industry-like methods of quality assurance, release management and support. MLZ is similarly the main developer of the NICOS instrument control software [27] which incorporates contributions from both universities and large-scale facilities into highly reliable code. The experience and procedures from these projects such as code peer review, continuous integration-deployment-testing cycles, time-based-releasing and formation of teams responsible for programming and release processes will be adopted by and used within DAPHNE. EMBL will facilitate interaction and transfer of existing knowledge between experts in the field of MX data handling and other disciplines in DAPHNE through workshops, mutual work-visits, and giving advice within individual work areas. A significant portion of the large personnel in-kind contribution from facilities such as DESY, FZJ, HZG, TUM, EMBL and BESSY is provided in the form of software development expertise. At the end of the project, infrastructure developed as a part of DAPHNE will be hosted and further maintained by the facilities ensuring its longevity.

Addressing the question of software development methodology, we anticipate working closely in small teams with scientists from universities and research institutions as well as staff at facility instruments as end users to develop solutions that match their workflows and provide solutions that match their needs with rapid prototype turnaround followed by thorough testing before facility wide deployment. At the same time, large-scale infrastructure is highly complex and relies on many software elements that need to work together, and must not interfere in overall facility operation. This makes it is necessary to carefully define software function, interfaces and interface points so that the different parts work together collectively without any one element disrupting operation of the facility. Additionally, careful interface design and specification is required so that the software is portable across the project and across the many facilities involved in DAPHNE and into the wider community. We therefore see a hybrid model of small teams working closely with the end users in a rapid development cycle combined with strongly coordinated definition of sub-program roles and interfaces coordinated through TA6 with the continuous support of professional scientific software developers from the facilities.

The DAPHNE Consortium involves large scientific computing centres which already today are engaged in software development, provision and maintenance at the highest quality level. As an example of the processes to be used at DAPHNE as well, the software package dCache[26] for

data management, which is developed by DESY in a leading position, will be used here. Modern and industry-typical methods of quality assurance, release management and support are successfully applied here. In detail, this includes the following steps:

- code (peer) review - each code change is reviewed by at least one other developer (dual control principle). If necessary, the changes are adjusted and sent back to review. The code is not checked into the main repository until all participants in the review process have given their consent.
- continuous integration/deployment/testing - Each commit to the main repository triggers a full build-deploy-test cycle. At this stage, the software is fully developed and all defined unit tests and static code analysis have been performed. At the end of the successful build, a package is produced for the target platform. This package is then installed on a test machine and run in pre-production mode with extensive test suites (over 4000 individual tests at dCache). The entire build/test cycle is automated with Jenkins CI.
- Time based releasing. Similar to the dCache project, we will use time-based release procedures where a new major/feature release is produced every three months following feedback from the DAPHNE community. The advantage is that the new functionalities are quickly available and the whole process remains functional at all times. Besides the mentioned releases, versions with long-term support are also provided. Such 'enterprise' versions are intended for stable operation in large laboratories and are often synchronised with the operation of PETRA III and EuXFEL.
- Release team. Although all developers in the team have the same rights, there is always a dedicated person assigned by DAPHNE who is responsible for a software release. This person decides which changes will be added to the productive version. This release manager role is taken over alternately by other members of the team so that everyone is familiar with the release process.

These basic principles of software development are to be implemented within the Daphne Consortium. Many partners of the DAPHNE Consortium are actively involved in European projects such as PaNOSC/ExPaNDS, EOSC etc. and contribute to quality assurance and sustainability initiatives in software development. We intend to fully exploit the expertise in software development at the facilities in order to strengthen the code base developed within DAPHNE.

Agile development model of rapid turnaround between prototypes and customers and developers is attractive to responsively meet the needs of the user community. However, the tight integration with established facility infrastructure requires more planning than typical for the agile software model, which can result in chaos when interfaces are not well enough controlled, as found through past experiences in facility software development. At the same time, we do not wish to go down the route of long delays in release cycles driven by the specification-development-deployment

cycle found, for example, in the US military industrial complex. Software development in DAPHNE will end up closer to a DevOps model of deployment, with Continuous Integration and Delivery, followed by deployment and operation/monitoring in production environment. The DevOps approach is a Software Development methodology, which involves Continuous Development, Continuous Testing, Continuous Integration, Continuous Deployment and Continuous Monitoring throughout the entire software development life cycle.

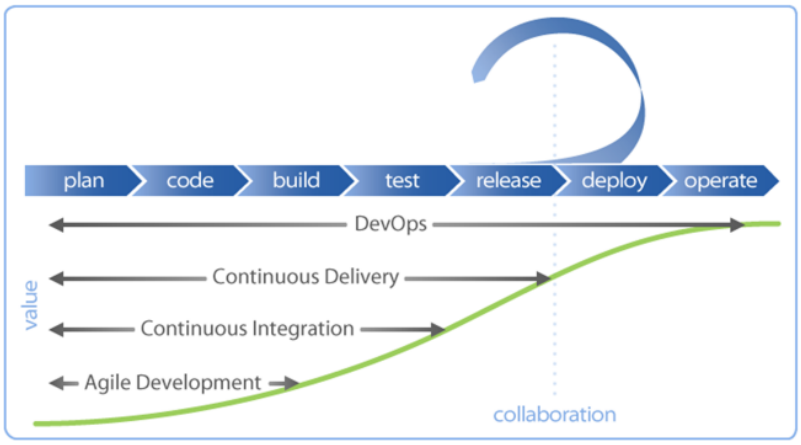

Fig. 4.4: DevOps is about ensuring a closed-loop, fully automated workflow between software development and deployment, increasing the value of the software while strengthen the collaboration.

Software expertise and the role of the scientific IT centres

The user groups participating in DAPHNE4NFDI are by and large experts in data analysis for their own research needs. To a large extend, the daily routine in our community is programming for the data analysis flow. The users develop programs and scripts in a variety of languages depending on their needs (e.g. Python, Jupyter notebooks, Matlab, IDL, or similar tools). These software/scripts and the structure of the specific data flow do strongly depend on the respective science field (such as e.g. crystallography, coherent scattering, spectroscopy, tomography etc.). The idea of DAPHNE4NFDI is to complement this "user" driven science knowledge with the needs and expertise of the large-scale facilities which (a) need to ensure operation of the beamlines, (b) the curation and archiving of the data and (c) do provide expertise and FTE in professional software development from their scientific computing centres. Bringing data experts from the science fields together with the IT experts from the large-scale facilities is the key working principle of DAPHNE4NFDI ensuring

- availability
- professional software development cycles
- science and user driven solutions
- long-term sustainability of data curation, data quality and software quality.

5 Work Programme

The work programme of DAPHNE centres on the following task areas:

Task Area
1: Managing Data Production
2: (Meta)data repositories and catalogues
3: Re-use of data and software
4: Dissemination and outreach
5: External communication and policy
6: Management

5.1 Task Area 1: Managing Data Production

Summary

Managing data production is becoming an increasing challenge for scientists everywhere and in particular in the neutron and x-ray science community where high brilliance sources and high-speed two-dimensional detectors have meant a rapid increase in the quality and quantity of data. To benefit from this opportunity, it is important to capture all data and experiment relevant information in a clearly documented manner and store it so that it can be easily found and understood for use and re-use. This is easily said but requires complex technical solutions, expert involvement and a high level of communication within and across communities to ensure that a sustainable solution can be found and maintained.

Experiments at large-scale neutron and photon facilities take place on individual instruments tailored for diverse applications. This task will address standardisation and automatic creation and extraction of all relevant instrument and sample metadata along with the measurement data and experiment protocol. This is the basis for easy-to-use and open source ELNs that will be evaluated and adapted to capture information from both the facility and the users during the experiments. Moreover, common standards for a minimum (domain) specific vocabulary and metadata schemata and digital object identifier (DOI) will be drafted and discussed within the DAPHNE community, including participating companies (together with TA2-5). The aim is to make the captured and annotated data meaningful, understandable and findable for a large community well beyond DAPHNE. Moreover, we aim to define and curate vocabularies as the basis for automatic access and machine readability for future re-use. Standard data formats will also be promoted. The neutron and photon community frequently (but not uniquely nor uniformly) employs common data standards based on the NeXus format, however the suitability of HDF5 for high data rates needs to be established and possible solutions that mitigate shortcomings need to be

identified. The same applies for the validation of standard scalable solutions for high data rates. The data production and collection addressed in this task will adhere to the code of ethics and legal aspects for recording experiments in keeping with the FAIR principles and data policies that are to be developed in collaboration with other consortia (TA4, TA5).

The objectives of TA1 are therefore grouped in the following measures: (1.1) metadata and vocabulary specification, (1.2) online logbook integration, (1.3) automatic metadata capture and (1.4) standard high performance data formats. Funds are foreseen to set up, test and operate the services at the facilities. The first three measures address the common aim of the DAPHNE community to bring together the currently discrete information on sample and instrument metadata into a well annotated data aggregation before and during the experiment, for ease of analysis, publication and deposition in data repositories as well as future re-use. In view of the very different communities using the large-scale facilities (and their different beamlines with varying characteristics), this task requires close communication within the DAPHNE community and beyond. For a successful and long-lasting implementation, coordination within the international community is also essential. DAPHNE brings together users and operators of the beamlines at the facilities and key companies. This implies that besides the close communication on common metadata schemata and standards any technical implementation will be a joint undertaking to ensure that both the requirements of the users and the facilities are considered and met. After the initial test phase among the DAPHNE project partners, the solutions developed will be available to a wider community at the beamlines for further testing and evaluation. On the one hand, the facilities will focus on providing generic technical solutions for ELNs and interfaces for automatic data capture provided by their considerable in-house contribution to the project. On the other hand, in collaboration, the participating user groups from different scientific domains will drive the implementation and evaluation at selected beamlines. Due to this constant interaction, evaluation and re-iteration initially within DAPHNE but following the first test phase also with the wider community, we foresee that DAPHNE will contribute significantly to establish and promote common standards for domain specific vocabulary and metadata schemata that will then be automatically part of any data aggregation that is produced at the large-scale facilities.

Setting standard high performance data formats together with providing common metadata schemata will increase the usability of experimental data and enhance the interoperability of analysis programs. Detectors and instrument control software currently record measurements and relevant metadata in a multitude of different formats. Cooperation and information exchange with the relevant companies will help us attain a common standard file format. Currently, due to variation in data formats, similar experiments executed at different instruments (and different facilities) have rich variety of formats and analysis tools even when a similar approach is pursued in their analysis. This hampers both a timely analysis and the accessibility for infrequent users.

Introducing standardised data formats with a well-aligned inner structure and vocabulary will be of great benefit to the user community especially in cases where data collected at different experiments is to be analysed together.

The objectives will be attended by the action of the cooperating university groups in close collaboration with the beamline scientists and computing staff at the facilities. There will be a number of scientific use cases where experienced user groups within DAPHNE will test the developed tools. There will be close interaction with the related industrial partners and the other NFDI consortia (e.g., FAIRmat, NFDI-MatWerk, NFDI4ING, NFDI4Chem) to attain common standards and ensure that common solutions are found across the techniques.

5.1.1 Measure 1.1: Specification of domain specific metadata schemata and vocabulary

Metadata specifications and standards are essential for data interoperability and reuse. Collaboration and the development of community consensus is needed. Therefore, this measure is common across all tasks of the DAPHNE community. The task also aims to contribute to the NeXus standard and validate and adapt any application definition to identified requirements, as well as to communicate the findings to the NeXus international advisory committee (NIAC, one member of the NIAC is from HZB)), as well as to commercial companies providing detectors and their software. All partners in TA1 will contribute to the specification for the annotation of data in ELNs, the automatic capture and the entries for standardised data formats.

Deliverable	Description
1.1.1	Initial white paper on proposed metadata schemata and domain specific vocabulary (together with TA2, TA4, TA5, to be updated annually during the project)
1.1.2	Common metadata collection implemented at key experiments across facilities
1.1.3	Publication defining metadata schemata and domain specific vocabulary (together with TA2, TA4, TA5) – ongoing

5.1.2 Measure 1.2: Online logbooks for extended metadata capture

Here, we will first evaluate existing notebook solutions with the aim to identify the needs for the community. An effective electronic experiment logbook (ELN) should provide a central, accurate record of the experiment. In the simplest form, this includes recording instrumental and experimental parameters, metadata and data file listings. For an effective FAIR data management system and a full experiment description, an ELN should also include unique DOI (or PID) links to sample description, sample preparation, previous and in-experiment characterisation. During

the experiment, visualisation of the collected data and preliminary analysis will be incorporated in the ELN providing a full digital record of the experiment. To achieve this, the ELN should be fully integrated into the instrument control log to collect relevant instrumental metadata and should allow for text, sketches and graphic input from scientists participating in the experiment locally and remotely. Further, for our diverse communities often working in multidisciplinary teams, important considerations are that multiple users can make parallel entries including graphics locally and remotely and that it is flexible and easy to use. Such an ELN must meet legal requirements, including time stamping, authorship and complete version control. Within DAPHNE, beamline operators and users will work together on standards and implementation. A close collaboration with other consortia will ensure that solutions are feasible across a broad range of scientific communities. DAPHNE has the advantage that providing an ELN service integrated at facility experiments not only aids FAIR data management but is also a powerful education tool for scientists using facilities. By using a modular approach, it is the aim to have many common modules at different instruments and across facilities and in the wider scientific community.

There has been a rapid progress in the field of ELNs in the last year and many possible solutions have been tested by the DAPHNE participants and the facilities. These include cloud shared documents (CAU) and more developed ELNs such as at ISPyB from DLS at EMBL, ICAT+ at ESRF (tested at DESY). In the first part of the project, available solutions for ELNs will be further evaluated and discussed among the DAPHNE community with emphasis on domain specific needs, the minimum entries and metadata standards as well as technical requirements for interfaces to the instruments. After a decision has been made, first implementation at selected beamlines will be started and tested within the DAPHNE project. These prototype implementations will cover a range of different scientific domains and techniques. They will also cover the interfaces to the instrument to add and retrieve relevant metadata that is automatically captured at the beamlines. The technical solutions, the domain specific implementation as well as the evaluation of the functionality will be presented and discussed among the DAPHNE community – which will act as a continuous advisory panel (TA4, TA5). After the initial test phase is successful, the ELN should be made available for the wider community for use at the beamlines and eventually deployed facility wide.

This measure can be subdivided in the more technical aspects to provide the online logbooks at the beamlines and home institutions. Solutions at the beamline rely on the support at the facilities both in technical infrastructure support, as well as from the local beamline scientist. This support constitutes a significant own contribution of the facilities. Initial tests and the domain specific evaluation will be carried out at selected beamlines in a joint effort between the operators and the users of the beamlines, tests are foreseen at various stations at DESY / PETRA III as well as neutron instruments at MLZ.

Deliverable	Description
1.2.1	Draft ELN specification document
1.2.2.	Evaluation of existing ELN solutions – decision on technical solution to be tested and/or adapted
1.2.3.	Prototype ELN available at selected beamlines incorporated in facility hardware
1.2.4.	Final version developed, deployed, documented and archived including tutorials

5.1.3 Measure 1.3: Instrument and sample data capture

In this measure, tools for the automatic capture of instrument and sample data are to be developed to facilitate the efficient creation of (machine-readable) data packages adhering to the FAIR principles, guided by the OAIS reference model /ISO standard 14721:2012.[28] All along the workflow from the experimental proposal through sample preparation and the data collection itself (meta)data are created and need to be gathered by a data collector at the facilities, in particular the actual physical instrument and sample data during the experiment. The information captured along this workflow belong to different categories with different requirements of metadata schemata, such as bibliographical metadata (guided by e.g. DataCite[29] schemata or the Dublin core initiative[30]) and data preservation information (e.g. following Premis standards[31]) to ensure sustainable accessibility in data catalogues.

The aims of this measure are to define a common workflow for (meta)data capture and to provide the technical tools where needed. Most importantly, together with measure 4.1, common metadata schemata and defined, domain specific vocabularies are to be promoted and included in the automatically captured data in a joint effort between the users and operators of the beamlines. The facilities have tools in place to gather mainly instrument, sample environment and bibliographic data, e.g. based on NeXuS filewriters and metadata composers to be used at DESY and ESS. At MLZ, the in-house instrument control software NICOS (also to be employed by PSI and ESS) includes data collector software. Furthermore, efforts to standardise the interface to external sample environments (SECoP-protoco[32]) are already ongoing, one result of the SINE2020 cooperation. The coherent capture of metadata will be extended to the sample, e.g. by establishing suitable standardised interfaces for sample identification (including the promotion of persistent identifiers PIDs) and sample data capture. Again, the required entries will be domain specific and DAPHNE aims to provide a platform to promote common workflows and common standards for metadata collection at beamlines. Some communities have already established sample data schemata and sample data capture tools (e.g. ISPyB in protein crystallography,

LaMMB: Sample and Measurement Database, at HZB, SampleDB at FZJ), but the majority of beamlines still relies on a free form entry of the sample description by the user.

Together with measure 4.1, the available tools for sample metadata capture and domain specific requirements considered and a draft specification of the requirements will be presented and discussed within DAPHNE. Moreover, close interaction to other ongoing metadata initiatives such as the Helmholtz Metadata Collaboration (HMC) or ExPaDNS will be pursued. Starting from selected beamlines specific use cases (listed above) will be implemented within the project that then will be made available for the wider community.

This measure is mainly situated at the beamline operators as it involves the connection to the instrument control system as well as data storage at the large-scale facilities. In close collaboration with measures 1.1 and 1.3, this measure aims to provide the interfaces for automatic data capture and data aggregation to provide high quality data sets adhering to the FAIR principles.

Deliverable	Description
1.3.1	Prototype interface available for automatic metadata capture at selected beamlines (user input)
1.3.2	Define common aspects across facilities
1.3.3	Final version developed, deployed and documented

5.1.4 Measure 1.4: High performance data format standards

In this measure, high performance data formats and instrument description files will be promoted and implemented at selected beamlines at the facilities providing data. Combining an open and high performance file format with automatic metadata collection at the beamline enables the setup of software assisted experimental chains where the data is reused in successive steps of the experiment and during the analysis, e.g. employing different software from different users/developers/facilities interacting through the common file format. The use of standardised file formats could therefore be a key for interoperability across different facilities and users.

Due to the variety of different instruments and research topics at photon and neutron sources, there will be no forced ‘one-fits-all’ solution; instead, common solutions are envisioned according to instrument classes, wherever possible, which are suitable for the user community. Together with TA4 and TA5, the current state of data standardisation with respect to instrument description, metadata capture and data formats will be collected and discussed, in order to define the needs of the user communities. In the photon and neutron community, NeXus and openPMD based on HDF5 are frequently – but not uniquely – used to store raw and processed data. From these

examples, standardised data formats are to be deployed successively at the instruments and to be included in the experimental workflows. The task aims to further promote and develop open data formats that are capable to handle large data rates and event-based data collection. Any standardisation has to ensure that the requirements for common analysis programs are considered, and suitable interfaces and converters have to be in place. Again, these efforts are to be coordinated across facilities and in an international context.

Deliverable	Description
1.4.1	Draft specification for high data rate high performance data formats and review existing solutions
1.4.2	Prototype high rate data format available at selected beamlines
1.4.3	Prototype standard data format available at selected beamlines
1.4.4	Final version developed, deployed and documented

5.2 Task Area 2: (Meta)data repositories and catalogues

The goal of this task is to obtain a federated, searchable, and interlinked repository and catalogue

Summary

The goal of this task is to establish federated, searchable, and interlinked data repositories and catalogues at all participating neutron and synchrotron radiation sources – working towards FAIR data for the whole neutron and x-ray community. Well curated searchable data repositories will increase the findability and reusability of data. On the other hand, the link to the provenience of the deposited entry (i.e. the raw data), the record of all intermediate processing and analysis steps from the raw data to the final result and the detailed description of the sample will increase transparency and thus quality and trustworthiness of the deposited data sets.

- of raw (meta)data (provided by TA1)
- and processed (meta)data (initially provided by members of this TA)
- the scripts and programs needed for processing the (meta)data (provided by TA3)
- with a link to a run-time environment for the software (provided by TA3).

The task will mainly address the implementation of raw (meta)data repositories (at facilities where not yet in place) and integration of processed and published data. This will necessitate the standardisation of data and metadata as well as the link to software repositories. In order to make the interconnection of federated repositories that relate to different stages of the analysis process up to the publication, well defined interfaces and common standards for the used vocabulary need

to be established. While the technical implementation of such repositories and catalogues is driven at the large-scale facilities (including work conduced within the European Projects PaNOSC and ExPaNDS), the actual specifications and implementation for different scientific communities using synchrotron and neutron techniques is far less advanced. A notable exception is crystallography with the Protein Data Bank (PDB),[4] well established in the MX community, which has deposited over 160 000 structures up to now, and the crystallography open database[33] for non-biopolymers (more than 460000 entries). A similar repository exists for structures determined by cryo-electron microscopy[34]. These repositories do not archive the raw data, but rather start from intermediate data sets.

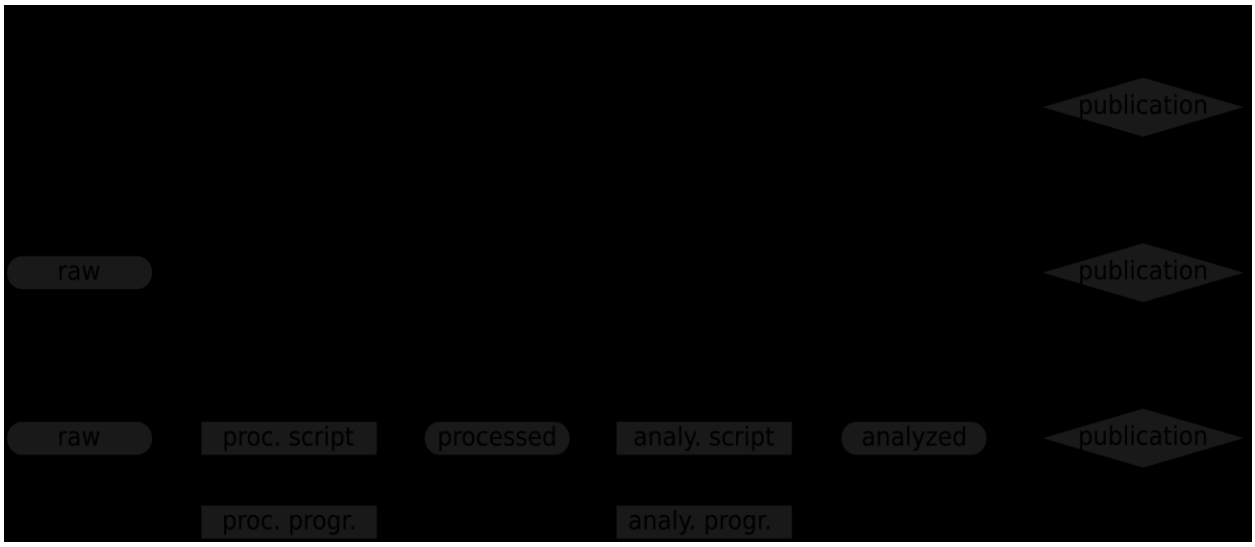

Fig. 5.1: Current workflow for data publication and the vision that is the aim of DAPHNE4NFDI. The data repositories have to (i) store all raw data sets and allows the user to extract the data for their experiment, and that makes the data publicly available after an embargo period, (ii) store and display the complete workflow between raw data sets and publications, making the complete path of data processing transparent and reproducible; and (iii) create an archive of processed and reduced data corresponding to published results.

The Coherent X-ray Imaging Databank (CXIDB)[20] is still comparatively small and far less structured than the PDB, but has gained traction in the coherent imaging community as a platform for sharing processed data, reference data sets, and in some cases the raw data and code/scripts required to process it (depending on the enthusiasm of the user). However, for many fields, easily accessible reference databases are still missing, or suffer from insufficient details on the sample and technique, or inconsistent data formats, e.g. the Farrel Lytle database[35] or the XAFS spectra library at the University of Chicago[36]. In some parts, the user community has also started to use commercial platforms such as figshare.com[37]. Therefore, there is a clear user-driven demand for data exchange, but co-ordination between the different user groups is currently lacking.

Here, DAPHNE aims to make a significant contribution in the establishment of community standards for vocabulary and data repository entries and – in collaboration with the facilities – in the establishment of well curated data repositories. While TA1 of DAPHNE is mainly addressing

the automated capture of (meta)data as early as possible in the data generation process, TA2 is dedicated to the 'Findable', 'Accessible' and 'Re-useable' aspect of the FAIR data policy. Common standards will be further promoted in TA3 where well defined metadata definitions and entries in any data set will facilitate the use and provision of data processing and analysis tools, in such a way that it is both accessible for humans as well as machine readable.

5.2.1 Measure 2.1: Repository and catalogue roll-out and development

'Data sets' collected at large-scale facilities typically consist of a data aggregation that is generated during one experimental campaign: i.e. all measurements that are taken during the duration of the beamtime allocated via the (peer reviewed) proposal system. This data aggregation includes the actual raw data files that store the detector outputs, but also metadata information on bibliographic data of the group conducting the experiment, automatically logged metadata on instrument settings as well as the electronic logbook of the users containing information on the sample and the running of the experiment (cf. TA1).

Several institutions in DAPHNE have already a data catalogue for all conducted experiments at the facility such as EuXFEL (MyMdC), HZB (ICAT), and HZDR (RODARE). A notable exception is MLZ, which aims to install SciCat,[38] that is currently developed by ESS, MAX IV, and PSI. These different facility-based data catalogues are already being transformed into a federated catalogue by defining and implementing a common API to search across all of the individual ones. These catalogues also have interfaces to connect to trans-national infrastructures such as OpenAIRE[39] and EUDAT[40]. While each facility will operate its own repository and catalogue for the data collected at its instruments, the user experience will be a single point of entry.

This task aims to ensure that all participating neutron and x-ray facilities will at the end of this project have a (meta)data repository and be part of a searchable and cross-linked data catalogue of raw and processed data that can also contain the scripts and programs needed for processing the data. The user-driven approach of DAPHNE will ensure that the different programs installed at the facilities actually answer the need of the scientific community. Within the framework of DAPHNE, MLZ will therefore establish a data repository/catalogue, adhering to the community standards and to be included in the federated catalogue, including the assignment of DOIs.

Moreover, DAPHNE aims to support the development of the federated data catalogues, mainly through the further specification of relevant entries (and thus search parameters) from various scientific fields, as well as through the establishment and promotion of a well-defined vocabulary (measure 2.2). TA4 and 5 will initiate the community-wide discussion on data policies, thus addressing questions of ownership, access rights and embargo periods relevant for data

catalogues that contain unpublished data. The raw data sets from the instruments will be ingested automatically (in collaboration with TA1).

It will have to be assured that bidirectional cross-links between all steps of the data processing and analysis workflow are implemented in all participating repositories/catalogues. These links must not only be able to point to different entities in the same repository, also compatibility with third-party repositories has to be provided. These third-party repositories could contain for example additional lab-based measurements or calculations and can be provided either by large-scale facilities or other interested players in the field. The aim is to be able to generate a complete PID graph, linking all involved entities together similarly to persons in a social network. Interaction with the European FREYA project[41] will be ensured.

Deliverable	Description
2.1.1	Comparison of different repository/catalogue systems and recommendations which one to adopt at participating facilities or third-party contributors
2.1.2	Prototype implementation of repository at MLZ
2.1.3	Prototype implementation of catalogue at MLZ with DOI minting
2.1.4	Integrate bidirectional links, also with the possibility to refer to third-party repositories
2.1.5	Improvement of repository / catalogue systems based on user feedback (repeated continuously)
2.1.6	Bidirectional links throughout the whole workflow implemented
2.1.7	Prototype ingestion of automatic metadata capture at selected beamlines (with measure 1.3.1)
2.1.8	Repository and catalogue at all participating facilities available

5.2.2 Measure 2.2: (Meta)data standardisation and sample identification

Both, data and metadata, have to be standardised to a certain degree across user communities, instruments, and facilities. They have to be named in a way that the user community expects it, but ideally, their meaning is also unambiguously clear for a wider community. In addition to clear and unique names, this information has to be stored in a standardised, machine-readable way.

While harmonization of metadata schemata and vocabulary is a common objective across all tasks of DAPHNE, measure 2.2 puts special emphasis on both the sample description and the definition of (known) metadata schemata and their relations. While the identification of humans (ORCID), publications/software/instruments (DOI), or institutions (GRID) is fairly straight-forward, the measurement metadata and the sample persistent identifier (PID) are at the same time crucial

and difficult to implement and will be the focus of this measure. Tight collaboration with other NFDI consortia and trans-national efforts will be necessary and, e.g., KIT will provide the link to NFDI4CAT.

The sample PID refers to the actual piece that was measured. It is necessary that it can also be used, for example, in the case of samples that show a time-dependence, be it ageing or reaction to stimuli during an experiment. Persistent identifiers (PID) of samples that are developed by other NFDI consortia (e.g., FAIRmat, NFDI-MatWerk) have to be accommodated by allowing a link to their database. Following this link (implemented in measure 2.1), the user obtains access to further measurements or the preparation protocol of this particular sample. For samples that do not have a PID yet, we aim to implement a procedure based on IGSN[42]. Originally developed for geological samples (International Geo Sample Number), IGSN has extended to all kinds of physical samples and is designed to provide an unambiguous globally unique persistent identifier for them. It requires some adjustments to the samples often used in neutron and x-ray experiments, such as the provision of a field for the manufacturer. It will then facilitate the location, identification, and citation of physical samples used in research.

The sample metadata (and other quantities) will not only have to specify the particular piece that was measured, but also the type of sample (e.g., Fe_3O_4) so that links to other quantities on this class of sample, such as a theoretical calculation of their band structure, can be established. In order to work with different existing metadata schemas (e.g. the IUPAC International chemical identifier (InChI)[43], known schemas have to be registered and mappings between these schemas must be created. This effort will build on recently implemented registries for semantic assets, such as Taxonda[44] or the NIST schema registry.[45]

Not only the metadata concerning the sample have different naming conventions in different experiments or communities. An essential part of this measure is therefore also to find metadata definitions to describe metadata. These metadata schemata and ontologies are describing the meaning and structure of the metadata. In their machine-readable form, mappings can be developed between metadata schemata and between ontologies in order to allow (automatic) conversions. The long-term aim here is not to develop a separate language for sample description in DAPHNE, but rather to co-ordinate with the other NFDI consortia and to register standards and mapping between standards.

Deliverable	Description
2.2.1	White paper on metadata definition and known metadata schemata (updated annually)
2.2.2	Adaption of IGSN sample identifiers to the needs of the DAPHNE4NFDI sub-communities
2.2.3	Prototype implementation of PIDs, also for time-varying samples
2.2.4	Registration of metadata schemata (with delivery 1.1.3)
2.2.5	Possibility to either use PIDs developed here or by another NFDI consortia established
2.2.6	Comprehensive documentation of metadata schemata and data formats used at the participating LSF

5.2.3 Measure 2.3 Insertion of additional (meta)data into repository/catalogue at a later stage

(Meta)data that have to be stored in a repository/catalogue will not only be the metadata and raw data known at the time of the measurement (which are covered by TA1). Additional (meta)data will be generated by the user at a later stage. These include among others: processed data sets, analysed data sets, connected publications, automated and manual quality checks/curation, additional home laboratory based sample characterisation, or simulation results. These additional (meta)data can either be inserted directly into the facility-based repository or via a link to a third-party repository. This choice has to be unimportant to the user, the additional (meta)data have to be included in the catalogues irrespective of their location.

This measure aims to open the facility-based repositories and provide clear interfaces to third-party repositories for such additional (meta)data. Part of these additional data will also be the processing and analysis scripts and programs provided by TA3. It is envisaged that these third-party repositories will usually group around certain scientific sub-communities. Successful examples that will have to be reachable from the DAPHNE catalogue are the PDB[4] and the Coherent X-ray Imaging Data Bank.[20] Whether a sub-community decides to set up their own repository or use the ones provided by the large-scale facility will also depend on the size of their data files: When the amount of data is very large, it is beneficial to bring the data analysis to the data and use the facility-based resources. Sub-communities with smaller data sets might prefer to download the data from the facility to do the analysis elsewhere and to ingest the results into their own repository.

This measure consists of both the technical aspect to interact with or potentially host the community repositories as well as the specification of the community required metadata standards, initial population of the repositories, and data curation aspects. The technical part will mainly be addressed and supported by the facilities, and therefore it will draw and benefit from the efforts on the international level that is part of the EU-funded projects PaNOSC and ExPaNDS. The technical implementation includes also the interface for access both for humans and the API for automated machine access. Well-defined interfaces for access of humans and machines onto the (meta)data are of paramount importance. While human access will be facilitated through a web frontend, a well-defined API will be the key to ensure interoperability of different repositories. Both access methods will be open to the public so that external data harvesting services can get easy access to the data, too. A direct link to the run-time environments for software developed in TA3 will be established. With the advent of algorithms based on big data, providing machine-readable access to the data via this API should open up future possibilities to interact with the data in ways that we can't even imagine today. The standards developed in measure 2.2 are the basis that the (meta)data are machine-readable. This is also important for the link to TA3, where data reduction and analysis programs want to access the (meta)data as well as for machine learning applications.

The incorporation of community repositories will start with several use cases in the areas of x-ray absorption spectroscopy, tomography, diffraction (wide and small angle) and spectroscopy (quasi-elastic neutron scattering and XPCS) as well as time dependent measurements. We will follow the principles of rapid prototyping where the complete workflow from raw data creation to publication is tested as quickly as possible to allow an agile development of the processes. In order to do so, existing already processed data will be ad-hoc inserted into the repositories without waiting for new raw data to be measured and processed. This makes it possible to test the "downstream" part of the project immediately without waiting 1-2 years until fresh raw data have made their way down to this stage.

The DAPHNE sub-community of X-ray absorption spectroscopy, including x-ray emission and further photon-in/out techniques, is probably the most expectant one regarding data repositories and catalogues: In contrast to other scattering or tomography measurements, these data are often evaluated by comparing them to previously measured or calculated reference spectra that are stored in a repository. This sub-community is therefore acutely aware of the high requirements concerning documentation from their experience with previous databases and they have processed reference data already, which can be injected into a repository. They have also established best practices how to search metadata and have clear ideas how they want to organise quality assurance and reliability of metadata and data, which is a challenging but important topic, especially for reference data. An automatic assessment will be established (together with TA3) and a manual peer review possibility will be implemented that will be saved

in the metadata. Users will be able to easily judge the quality and the usability of each data set by looking at these quality markers.

In the case of the other use cases, harmonisation of common metadata standards is far less advanced and there is no common repository for processed and published data as of now. Here, DAPHNE will transfer the experience from the x-ray absorption spectroscopy community as well as the coherent x-ray tomography to initiate further community repositories in the above areas. The use cases will first be implemented and tested within the DAPHNE community (together with TA4 and 5).

Deliverable	Description
2.3.1	Prototype implementation of user interface
2.3.2	Preliminary specification for use cases minimum metadata
2.3.3	Prototype implementation of reference database for XAS/EXAFS
2.3.4	Provide write access for processing/analysis software to deposit results (with measure 3.2.3)
2.3.5	Mechanism for quality assurance developed
2.3.6	Prototype implementations of reference databases for the other use cases
2.3.7	Link to run-time environment for software provided (with measure 3.1.5)
2.3.8	Deployment of community database

5.2.4 Measure 2.4: Search data

The ability to search data rather than just metadata (in everyday terms: search within the contents of books rather than just author, year of publication etc.) is a quantum leap. In the particular case of x-ray absorption spectra, the search for metadata (e.g., the Pt L3-edge in a powder sample of PtO_2) and subsequent manual comparison of the deposited spectrum to the newly acquired data would be replaced by submitting a new spectrum to the search engine and obtaining “this spectrum is similar to the Pt L3-edge in a powder sample of PtO_2” as an answer. With the advent of machine learning algorithms, this pattern recognition search is becoming more and more powerful – while still leaving room for more traditional approaches. Direct access to indexed data rather than metadata is the focus in this task area while the techniques that make use of this access, be it machine learning based or not, will be the focus of TA3.

In fact, searching the data should become a very powerful tool across the whole DAPHNE community once the repositories contain enough well-documented data sets. Many large-scale

facilities face the challenge that the number of “power users” is steadily declining and most users perform an experiment at the facility as just one of many characterisation techniques of their sample. These groups lack the experience of the power users and may not exploit the data completely, or even mistake experimental artefacts for physical features. To a certain extent, the highly experienced group member who would have spotted these things in the past can now be substituted by machine learning algorithms. The user measures a certain set of data, operates on the data, uploads the result to the pattern recognition system and receives a list of similar-looking data, together with the linked data analyses and publications.

Since many scattering experiments are evaluated using very abstract analytical models, such a comparison can provide useful input for the data analysis, even if the samples have nothing in common apart from the shape of the precipitates, for example.

Deliverables	Description
2.4.1	Prototype implementation of a machine learning algorithm to find similar data in a training set
2.4.2	Using the XAS/EXAFS reference database (measure 2.3.3) as training set
2.4.3	Using the data repository of a LSF as training set
2.4.4	Demonstrator for pattern recognition for several use cases
2.4.5	Repository search tool based on pattern recognition

5.3 Task Area 3: Infrastructure for data and software reuse

Summary

Discussion about FAIR principles typically focuses on the data. However, FAIR software and data analysis tools to make sense of that data is also a critical part of the research data ecosystem. This need is particularly acute given the increasing complexity of scientific data analysis in x-ray and neutron data. Task area 3 is concerned with making scientific data analysis tools and software FAIR alongside with the data itself. DAPHNE4NFDI achieves this by engaging with selected power user groups - the scientific 'influencers' of the photon and neutron community - to develop findable and repeatable data analysis tools and foster best practices in open research software development. We achieve this by focusing on particular use cases of high benefit to the community of researchers.

Modern detectors at state-of-the-art photon and neutron sources provide a flood of data which easily exceeds the ability of individual researchers to process it. The flood of new data opens new opportunities such as high-resolution structure determination and the ability to study dynamic rather than static structures. However, keeping pace with increasing data rates requires robust data processing pipelines, real-time or near real-time data processing, and experiment-specific data compression and filtering (triggering) in order to efficiently conduct and control experiments. The data analysis challenges involved in analysing that data are significant and in many cases beyond the individual capabilities of reasonable sized research groups. Facilities are increasingly taking care of the core infrastructure for data processing and analysis so that users can take home scientific results or reduced data sets instead of an ever-increasing number of hard disks filled with raw data. For example, at facilities such as EuXFEL and Petra IV, the data sets created are of the order of petabytes and it is not realistically feasible to take or transfer the data to the home laboratory.

In the context of data infrastructure, findable data is of little value without findable and repeatable software tools with which one can analyse that data. Put another way, properly curated software for the user community and the promotion of best practice in software development are both a critical part of the research data ecosystem.

Currently, the situation in photon and neutron science today is relatively inhomogeneous (or even chaotic) – while some fields such as protein crystallography have a well-developed software ecosystem, the approach of propagating scripts written by students or postdocs remains common in other areas. The latter approach has never been sustainable, nor compatible with findable and reusable data and software. Properly curated and released software is an essential part of making new methods repeatable and new techniques useful to the community as a whole. DAPHNE

recognises this in the context of NFDI with this work package devoted to curating software tools alongside the data.

Since much expertise in method-specific data analysis lies in the user community, we will address this need by working closely with key ‘power user’ groups in the community to improve and deploy properly curated data analysis software on centralised infrastructure in order to make new techniques available to all researchers. Engaging the ‘power user’ groups in making their analysis software available to others in a properly curated manner will both improve repeatability and make the techniques available to others.

This demands interaction with a broad user base within the user community, first targeting applications which will have the greatest impact. A further benefit of engaging directly with the user community is the promotion of best practice in software deployment and reusability within the wider user community.

Beyond analysis software itself, analysing large and complicated photon and neutron data sets is an essential and non-trivial process for knowledge extraction which can extend for weeks or months after the experiment is completed. Combined with the increasing size of data sets, this means that the ability for remote data analysis is essential in order to make measured data useable for research. Users must be able to effectively analyse data from their home institutes even if that data is stored at the facility. We therefore need to develop remote data analysis services suited to the photon and neutron community data analysis workflows as a part of the project, so that the data can be exploited once it is found. This task is closely linked to the development of user-driven analysis software so that the services provided match the needs and workflow of the user community.

Finally, there is a logical incentive to reduce needs for raw data storage by performing on-the-fly data reduction and triggering so that only detector events containing potentially useful data are retained. Alternatively, data can be processed to reduced small data sets containing only the information of interest in near real time. Examples include retaining only radial profiles for SAXS/WAXS (although not always applicable), or only detector readouts containing crystal diffraction in the case of serial crystallography. Basic and generic intermediate analysis steps are performed as soon as the data is collected, retaining only the information which is needed for subsequent sample-specific analysis. The potential savings in storage space could be significant, with triggered or reduced data sets being 10% or less than the raw data size. While facilities can provide the mechanisms to do this, the analysis strategies for determining what reduced data is appropriate is domain-specific and can only be developed by experts in the user community. After all, it is the domain experts in the user community who know what data reduction steps need to be performed to produce a result, and thus which immediate data reduction may be appropriate and acceptable to the user community. For a community not accustomed to selecting data on the

fly, approaching this task through the user community is essential in order to identify the cases where real-time data reduction does not compromise result quality, to build confidence that valuable data is not lost in the process, and in the long run to gain acceptance of the practice in the community.

Task Area 3 (Data and software reuse) addresses this through the following measures:

1. Curate and deploy data analysis software for key techniques for use on cloud like analysis services at the facilities so that ordinary users can repeat and benefit from the work of ‘power users’ in establishing new measurement techniques. Move software towards event- and run-based analysis workflows from container file formats for the analysis of big data sets. Perform on-the-fly calculation of reduced ‘small data’ and near-real-time processing for the rapid reduction of large data sets to more convenient sizes for users to take home for further analysis – including vetoing and data compression to reduce the volumes of measured data to workable levels for individual researchers.
2. Work with the facilities to develop and deploy tools for remote (and local) data analysis, data access, data visualisation, and data retrieval suitable for x/n workflows,supporting and ensuring the compatibility to the various experimental techniques.
3. Operation of services.

We focus on the following use cases in order to manage the project scope. These use cases (1) generate large data volumes suited to on-site processing and data reduction, (2) have a community of application-based users who would like to treat the experiment as a measurement with a professionally curated data analysis pipeline, and (3) are arranged around existing power user ‘influencers’ willing to develop data analysis software pipelines along the vision of DAPHNE:

1. Tomography techniques including full field, ptychographic and fluorescence tomography;
2. EXAFS and related spectroscopies;
3. SAXS/WAXS/SANS workflows;
4. Serial crystallography; and
5. Interfacing research data to machine learning methods.

These techniques are largely reliant on user community developed software for analysis. Unlike other scientific disciplines, there are essentially no commercial products available for data analysis from photon and neutron facilities. There is a large community of researchers who make use of x/n data but who are not expert in software development and have no desire to write their own code, yet results are dependent on custom codes written by other research groups that are in general poorly curated. Making this software FAIR is an integral part of research data management. DAPHNE addresses this need by targeting the user groups writing the software and fostering its deployment for other researchers within facility infrastructure.

Firstly, software engineering aspects of existing user-developed software analysis need to be improved. The development of academic software within the user community generally suffers from (i) a lack of scientists who have the required training in software engineering or a lack of software engineers embedded in academic research groups, and (ii) a lack of funding to support such activities. For most research grants, the key metric are publications, and the quality of software that is developed to achieve these publications is of secondary interest. Software development itself is thus naturally not encouraged and takes a secondary role. This is the right decision for each individual research group but prevents the community from benefiting from the developed software more widely. The NFDI gives us the possibility of starting to address this missing gap by specifically funding the development of analysis software infrastructure. The key point of measure 3.1 is to fund the substantial additional effort required for development, curation and deployment of key analysis software which is often overlooked in the academic process of pursuing publications. Note that this task is separate from the researching of new methods, which is already funded by other means.

Secondly, the software needs to be integrated at the facilities (see measure 3.2). Here again the NFDI is unique in enabling funding of both academic and facility groups to work in a structured way on common analysis software infrastructure for the benefit of the science community.

5.3.1 Measure 3.1: Analysis software infrastructure: strengthening and sharing user tools

Measure 3.1 improves analysis software in developed for targeted use cases by 'power users' for their own research to enable flexible curated and traceable deployment of the software at the facilities. The deployed analysis tools – developments by the user community but ultimately provided to other users at the facilities (see measure 3.2) – will enable ordinary users to use these state-of-the-art software tools to support the analysis of their experiment. By developing, curating and sharing the best tools, which are currently only accessible to the authors in the expert groups or through collaboration with these groups, as facility services we create a data analysis infrastructure that benefits the whole community.

We have identified the following use cases as focus areas for DAPHNE:

1. **Tomography techniques** including full field, ptychographic and fluorescence tomography are routinely desired for analytic measurements of samples ranging from materials science to biology and cultural heritage. The software for data analysis is developed by a handful of power user groups and requires specialist knowledge to use. Currently data analysis is a significant bottleneck and barrier to entry to users, limiting the FAIRness of archived data. We will address this by deploying standard, optimised workflows on facility or cloud infrastructure which can be cross-linked to measured data.

2. **EXAFS and related spectroscopies** rely on databases of calculated and measured spectra for determination of local chemical composition, and the ability to use that in spectroscopic workflows. There exists a user community interested in reprocessing data but who are not expert in the specific software packages developed by user groups. This problem will be addressed by deploying standard, optimised workflows on facility or cloud infrastructure.
3. **SAXS/WAXS/SANS experiments** rely on azimuthal integration which can reduce data sets volumes by several orders of magnitude. Related to this (although not from a slightl different field) are time-resolved quick-EXAFS data that is usually highly oversampled, with e.g. several hundred measurements per eV and thus allows on-line smoothing and data reduction instantaneously. Another application is hit-finding, indexing and integration for serial crystallography experiments: once the data has been processed, the resulting data set is orders of magnitude smaller. This is also a very effective strategy to reduce data: once the initial processing has been carried out, the original raw data do not have to be kept. Alternatively, raw data can be moved directly to tape storage for long term archiving rather than being retained on expensive disk-based storage.
4. **Serial crystallography** relies on measuring protein microcrystals delivered into the beam in a stream, but can generate vast quantities of data – over 100 TB per day. The sheer data volume overwhelms many research groups. Many of the recorded data frames do not contain crystal diffraction and need not be retained for analysis, while data can also be pre-processed and reduced on the fly, reducing data volumes to manageable levels.
5. **Interfacing research data to machine learning methods.** An important topic in the context of Big Data is machine learning (ML)/artificial intelligence (AI). While DAPHNE is of course not aiming to replace ML/AI projects funded by other programs, one of DAPHNE's goals is to *connect* to ML/AI projects professionally and efficiently in order to help exploiting the full potential of ML/AI. A key ingredient in ML/AI approaches to data analysis in the area of x-rays and neutrons is the availability of suitable data sets for *training*, which is at the heart of machine learning. Providing the data in a transparent and searchable form and format is of key importance, and the ML/AI efforts will benefit from this greatly. *DAPHNE will thus serve as an interface* between the experimental hardware of the photon and neutron facilities and their users producing large quantities of high-quality data on the one hand and ML/AI projects on the other. An example for the latter is the Tübingen Cluster of Excellence "Machine Learning in Science"[46] with involvement by University and Max-Planck groups, which is at the forefront of ML/AI research, but occasionally lacking an efficient connection to the right database. There are, of course, numerous other efforts internationally, not all efficiently connected, and not all connecting suitably the potential of ML / AI with experts for algorithms to the experimental data.

Clearly, synthetic data are not sufficient to replace experimental data. In the preparation of the DAPHNE proposal we have already been interacting actively with the ML/AI community and identified this interface as one of the bottlenecks for their progress, as well as strategies for solution. DAPHNE is committed to providing suitable access to the data.

6. **Multi-dimensional data-treatment and refinement of neutron tof-diffractometers with large-area detectors.** Many latest generation neutron diffractometers both at spallation and reactor sources are measuring in time-of-flight mode *and also* utilize large-area detectors up to 10 sr (e.g., POWGEN and GEM and in future POWTEX and DREAM). The most-often tailor-made, He^3-free detectors produce multi-dimensional event-data. We as 'power users' develop data-treatment methods and algorithms to best exploit this multi-dimensionality by avoiding to mix or integrate over data regimes with differing peak-shapes or peak-broadness. This approach allows to exploit the maximum instrument performance. Within DAPHNE, instrument and software specific implementations of these methods will be translated to a general and free open-source library.

This measure funds positions within power user groups at the university research groups to improve their tools for use in the community, we make their investment and expertise available to all facility users. Measure 3.1 includes improvement of the software engineering aspects, improving robustness and providing missing features. Resources for developing user software for deployment will go largely to university groups, as this is where effort is needed to interface the user software with for community benefit. However, the work will be done in close interaction with the facilities requiring dedicated supporting resources for development, support, and operation of services (see measure 3.2). The experience of MX, where the software ecosystem is relatively mature and well developed, will act as a role model for the other fields. Here, the practical experience from EMBL will play an invaluable role.

An important aspect of DAPHNE is improving longevity and sustainability by using software engineering practices in the improvement of these tools, such as version control, automatic unit and system tests, and continuous integration. The expertise is available in some groups and facilities, and will be shared through appropriate training. By deploying staff with these skills in the user groups, that knowledge will spread and underpin future developments of the user groups and the universities they are embedded in, which is an essential step in making data- and software-driven methods FAIR. Bringing funded groups together regularly promotes the sharing of expertise, resources, and distribution of know-how.

We will further improve software documentation to make it accessible to and effective for a wider audience. All too often, what seems trivial to the person who has developed the software may be

impossible for other users to guess. In addition to somewhat standard documentation for all features of a tool, we will develop for each tool an introduction to the science question that the tool addresses (to help make the tool findable: which problem does it solve), an introduction to usage of the tool for beginners, and a collection of tutorials for typical use cases. Parts of this will be made available as video recordings. This task cross-links to education and outreach.

For each of the software units supported in this measure, we anticipate the following internal development schedule and internal deliverables to be monitored and tracked as a part of TA6 (project management):

Deliverable	Description
3.1.x.1	Service/software specification document (to be revised annually during the project)
3.1.x.2	Appointment of the hired staff and commencement of employment, team formation and organisation
3.1.x.3	Adoption of development and testing standards
3.1.x.4	Prototype services/software deployed, tests by “power users”
3.1.x.5	Revised services/software available for use on facility hardware for use by all users with full documentation
3.1.x.6	Final version developed, deployed, documented and archived including tutorials

This model will be repeated for each of the six software development task supported (labeled ‘x’ in the table above). Project management plays a critical role in coordination of the elements in this measure including project oversight and harmonization of effort through the diverse user community represented by DAPHNE. Developing power users as role models for software development is an intrinsic part of one of DAPHNE’s goals to educate our science community.

5.3.2 Measure 3.2: Analysis software infrastructure: integrating and hosting user tools at facilities

The deployment of user software on facility infrastructure for the wider community requires a greater level of care compared to user groups who develop software solely for their own experiments. To work towards FAIR data, all tools must be versioned and such information must be stored together with outputs produced by it. Aspects to consider include provision of data and metadata to the software, selection of meaningful and robust analysis parameters in the absence of the experts who developed the tool, documentation of the tool to a level that it does not need expert input to use it, and reliable deployment of the tool at the facility within the existing facility-specific environments.

The user tool software needs to be able to digest the facility data, for example through file-based data input or streaming of the data over network ports. This requirement would be trivially solved if we had data standards that captured all required data and metadata, but we are not yet in that luxurious position. As part of this proposal, we push forward the development of such standards and file formats. This is very important for interoperability and needs to be conducted in collaboration with the user community. However, this is a difficult task that needs time and long-term investment. We cannot wait for this effort to be completed for this project and need to accept somewhat facility-specific data conventions for data in the meantime; some of the metadata that is required to run the tool may not be universally available either and facility-specific solutions need to be found.

As a result of this initiative, facilities will host an increasing number of such user tools for a wide range of experiment analysis. Self-contained and up-to-date documentation, addressing people who have not used the tool before need to be developed in order for these services to be useful to a wide range of users. Additionally, facility specific example data sets need to be available that users can use and analyse using shared software (ideally already before their first experiment at the LSF) and ideally before attempting to use the tool on their own data. In addition to being a key part of an effective tutorial, such data sets and example scripts provide system tests and deployment tests for the software at the facility. Indeed, such datasets make ideal unit tests for software development.

Analysis infrastructure developed in measure 3.1 needs to run on appropriate services at the facilities. Generally, these can be used remotely, for example through secure shell logins (ssh) and so-called X-forwarding, or other technologies that allow having a graphical user interface in a browser, or allow remote access to a Jupyter notebook. While these access methods provide the required functionality for remote data analysis in some form, the access tends to be facility-specific.

The deployment of shared user software within given facility computing environments also needs to be addressed. Historically, this is done in whatever operating system and software environment the facility services provide. High Performance Computing centres use packaging strategies such as containerisation or modules to be able to offer a variety of analysis tools, often with conflicting library requirements. To unlock efficiency savings and move towards more reproducible computation and analysis, we propose to explore the deployment of tools through containers (such as Singularity and Docker). The benefits of this approach include that the software can be installed inside the container in whatever operating system works best for the software and with an arbitrary set of support libraries, and the same container can then be shared across facilities or even be used by multiple users within the same institution. This will reduce deployment efforts and take us closer to a world of interoperable tools. However, container technology is fairly new in science, and particularly for the analysis of large experimental data sets. Some solutions may

introduce security risks (for example Docker), while other, more secure approaches, result in difficulties accessing the raw data. While any additional complexity from calling a command from inside a container can be hidden from the user by experts at facilities, this is less clear for use of the containers outside facilities.

In this task, we aim to go beyond this facility-specific access and start to develop the next generation infrastructure by developing a portal-based user access interface. We will develop tools and portals for remote (and local) data analysis, data access, data visualisation, and data retrieval suited to neutron and photon science use cases. We will focus on neutron science, as the data analysis community and software landscape for neutron science is more well-established and advanced than in photon science, and the ambitious creation of such a portal needs to be based on well-defined experiment and data classification and standards. The portal will provide an overview of available data sets/publications, and associated analysis tools available to operate on that data. By working closely with end users with domain expertise, we will ensure this is matched to scientific workflows.

Resources for developing data analysis services to support the user analysis code will go largely to the facilities, as this is where effort is needed to integrate, develop and deploy services to meet community needs. Funding both the facility and user groups as part of one project enables them to work together effectively towards a common goal in a project with coordination across disciplines and facilities, averting the development of island solutions.

Deliverable	Description
3.2.1	Common service specification document (to be revised annually during the project)
3.2.2	Prototype portal-based user access interface
3.2.3	Prototype services/software developed by ‘power users’ deployed on facility infrastructure
3.2.4	Final versions of ‘power users’ deployed on facility infrastructure, documented and archived including tutorials

5.3.3 Measure 3.3: Operation of services

This measure provides resources for the operation and maintenance of the IT infrastructure on which user-community software will be made available and run. This task is operations-focused ensuring the running of backend infrastructure is working as required and is separate from the development tasks listed above. Resources are required at both photon and neutron facilities.

Deliverable	Description
3.3.1	Annual report on service provision at each member facility

5.3.4 Funding profile for Task Area 3

This task is by definition widely distributed among consortium members by virtue of its engagement with user groups in software development and curation. The task area will be led by Technical University of Munich, Deutsches Elektronen-Synchrotron, European XFEL, the University of Kiel and University of Tübingen with substantial contributions from University of Wuppertal, University of Siegen, RWTH Aachen University, Helmholtz-Zentrum Geesthacht, University of Göttingen, and the European Molecular Biology Laboratory, and the Helmholtz-Zentrum Dresden-Rossendorf.

5.4 Task Area 4: Dissemination and outreach

Summary

One of the key challenges to advance towards a common data management culture and infrastructure is to address appropriately the needs of individual sub-communities and techniques, the requirements from university users, institutions and industry, and to agree on (possibly different) specifications for processing their data. The goal of this task is to serve as communication platform among the DAPHNE4NFDI community to discuss, evaluate and specify the requirements for the common data management infrastructure as well as disseminate the project results within DAPHNE4NFDI.

This task area aims at disseminating the results obtained in the various project areas, particularly, of TA1-3. At the same time, it aims to implement a common platform for the community to discuss their needs and refine the data management strategies. The specific character of DAPHNE, which acts as a consortium arranged around the user communities of photon and neutron large-scale facilities, requires that this is also reflected in the dissemination of DAPHNE's work. Serving a wide range of scientific areas, the specific needs and cultures of these sub-communities will be considered. Going beyond DAPHNE itself, dissemination and outreach require a joint effort with all topically related NFDI consortia in combination with the efforts in TA5. The advantage of establishing infrastructures at large-scale facilities (LSFs) is that the widely distributed communities work naturally together. Even though the neutron and photon user communities already display a high degree of cooperation, the awareness of synergies and standards in data infrastructures needs to be further improved. In many cases reference databases (RDB) or repositories for processed data and analysis scripts need to be established (see TA2). The goal of this task is to serve as communication platform among the DAPHNE community to discuss, evaluate and disseminate both the requirements for the data management infrastructure as well

as the project results. To involve the scientific users in all phases of specification and development of tools and methods, use cases (UC) will play an important role, i.e. they will test the developments along the whole data workflow in a collaboration between facility users and operators. Results of these use cases will be reported within DAPHNE, but also presented to users of different communities or techniques. This also supports creating synergies and to push cooperation. Main tools to attain these goals are workshops, project webpage and regular newsletter, schools, surveys and web-based discussions.

The outreach activities will be directed towards:

● Scientific researchers to demonstrate the impact of data management tools on high quality research. This will be targeted via specific projects and initiatives including collaborative research centres (SFBs), DFG priority programs (SPP), graduate schools, and the DFG excellence initiative (ExIni))

● Universities, lecturers and teachers to include data management topics into the curricula

● General Public to increase awareness of the economic and socio-cultural impact

The main focus in TA4 will be the organisation of workshops covering topics such as data management and analysis, standards in metadata vocabulary and collection, establishment and extension of (reference) databases, or the use of electronic logbooks (ELN). Workshops should be organised annually as DAPHNE consortium meetings, in addition to opportunities at user meetings or in collaboration with DFG-funded SFBs, graduate schools and SPPs. It is important to link the activities to both the mathematics and informatics communities that are already quite engaged with the photon and neutron science communities. Thereby, the already established links between universities and LSFs as well as links to the activities within the Helmholtz Association concerning data management topics (such as HIP, HAICU, HMC, see section 3.1) will help. Finally, we aim to transfer the knowledge gained into university curricula. The activities will be spread internationally via sessions on research data management at the LSF conferences, transported to the young scientists via PhD and research schools. Moreover, the activities will be coordinated and disseminated via the newly established DAPHNE homepage and DAPHNE portal. Outreach also implies public relations, presentation for the general public and pupils, as well as events for academia-industry networking.

This will be achieved by six measures listed below:

4.1. DAPHNE homepage and webportal

4.2 Dissemination within the DAPHNE consortium

4.3 Promotion of infrastructure within DAPHNE communities and beyond

4.4 Dissemination to PhD students and postdocs and implementation of data management topics into University Curricula

4.5 Outreach to society and industry

4.6 Long term outreach and dissemination

5.4.1 Measure 4.1.: DAPHNE4NFDI homepage and webportal

Objective: Establish a common point of entry for information and services connected to DAPHNE4NFDI activities.

At the moment, the webpage www.daphne-nfdi.de informs briefly about our initiative now. At the project start, information on DAPHNE, the FAIR data principle within the consortium and the contact persons will be presented. The webpage shall inform about current activities of DAPHNE, including announcements of workshops, conferences etc. DAPHNE related publications and project results, such as white papers or policy recommendations, will be accessible from the website. The portal will show the current status on standards for electronic logbooks, proposed metadata schemata, or high performance data formats as developed in TA1. An important part of the DAPHNE portal will be the links to repositories, data catalogues, reference databases (cf. TA3), software catalogues for data analysis tools and data reduction (cf. TA3). Note that further deliverables related to this measure are reported with other measures, e.g. dissemination to society via webpage.

Deliverable	Description
4.1.1	Setup of webpage with news section, General information and announcements on workshops, courses, conferences, etc.
4.1.2	Functional setup of webportal with link to different actions
4.1.3	Distribution of newsletter

5.4.2 Measure 4.2: Dissemination within the DAPHNE4NFDI consortium

Objective: Organisation of seminars, tutorials and courses on data management and remote data analysis, initiate topical workshops to foster common activities across tasks (such as metadata standards and vocabularies for repositories, data collection or software). Communication within DAPHNE4NFDI to promote data management and FAIR data principles.

The success of DAPHNE will depend on finding the most appropriate (meta)data formats, processing schemes, repositories and software tools. Although common standards and processing schemes are in place in parts of the, it is not yet easily transferable beyond those

community boundaries. The objectives of this measure are thus twofold. On the one hand, this measure aims to increase awareness and knowledge about existing standards and experiences from sub-communities, on the other hand, workshops and discussion forums will be used to jointly work towards widely accepted standards and best practices that are appropriate and applicable for DAPHNE and beyond. This is the basis for the success of all of NFDI.

At first, we will bring together the ideas of the NFDI, those of the LSF and those of the users. In this way, facility users will not only benefit by recording their data at LSF but also perceive the advantage of building up a national research data infrastructure, to host, archive, and exchange data, data analysis tools and protocols. The obvious advantages such as easy access for comparison with other samples that show similar scattering, microscopic or spectroscopic features, or developing tools together are already realised in some areas (e.g. EXAFS, tomography, or neutron spectroscopy). Extending to this to larger communities will improve accessibility and support high quality research. Similar standards and harmonised data formats will also facilitate the application of correlative techniques. Awareness of previous data sets for similar samples/experiments enable optimised data acquisition procedures and improved analysis tools. The latter applies in particular to users facing big data (e.g. time-resolved x/n tomography, time resolved studies at EuXFEL, or the upcoming ESS, etc.) to whom solutions for automatised metadata storage appears important.

Deliverable	Description
4.2.1	Organisation of kick-off/annual meetings, initiate actions at satellite meetings at LSFs, surveys of user needs
4.2.2	Establish integration in satellite meetings at neutron and photon radiation user meetings
4.2.3	Establish connections to existing or obvious topical clusters within DAPHNE4NFDI (e.g. RDB, experiment/theory, data analysis)
4.2.4	Set up of discussion forum for the areas: (A) setting up/running an experiment, (B) storage – data format and metadata, (C) evaluating, (D) reporting and publishing

5.4.3 Measure 4.3: Promotion of infrastructure within DAPHNE4NFDI communities and beyond

Objective: Promotion of the realization of the national research data infrastructure in DAHNE4NFDI and related communities

The tools for managing data production (ELNs, data storage, metadata), data catalogues and RDB as well as the provision of data analysis tools are developed in TA1-3, respectively. Use cases will be in place to test, evaluate and refine the developed tools and data processing chains along all relevant aspects within DAPHNE. Regular updates on the progress of the developed services and demonstration of the use cases will be disseminated and discussed with the wider

community, e.g. via newsletters to the KFN/KFS members, via the website or in dedicated workshops (e.g. with DPG, DBG, DGK, DECHEAMA, etc. as well as SFBs, SPPs or Ex-INIs). These will include e.g. Research Data Management sessions (a) as satellite meetings at neutron and synchrotron radiation user meetings, (b) at the national conferences like the SNI-conference, the German neutron scattering conference, conferences of DGK etc., (c) at international conferences like the European Powder Diffraction Conference, the international Small-Angle Scattering Conference, the International Conference on X-Ray Absorption Fine Structure (XAFS2022), the International Conference on Synchrotron Radiation Instrumentation (SRI), or European/International Neutron Scattering Conferences. The collaboration with other NFDI consortia is part of TA5.

Deliverable	Description
4.3.1	Distribution of information via newsletters, distributed to KFS and KFN members
4.3.2	Symposia, workshops and training on use cases, developed infrastructures and their extension to other areas
4.3.3	Special scientific symposia with partners (PTB, DPG, GdCh, DGK, DECHEMA,…) and other NFDI-consortia

5.4.4 Measure 4.4: Dissemination to PhD students and postdocs and implement Data Management Topics into University Curricula

Objective: Increase awareness of data management topics during education (Master and PhD level) and train the next generation of researchers.

For a successful outcome of the NFDI/DAPHNE initiatives, it will be important to develop schemes how to implement data management topics into university curricula and how to educate the next generation of doctoral researchers. With the increasing pace of the digital transformation, students on master and PhD level also need to be educated to manage the challenges in this area in order to be successful in their respective scientific field. We aim to achieve it for DAPHNE-related topics starting from a seminar where the young generation meets experts from physics, mathematics, IT and applied sciences (chemistry, biology, materials science) – e.g. by organising a WE Heraeus seminar. In addition, a Research Data School in photon and neutron science will be established, complemented by webinars (“Ringvorlesung”) with speakers from some seed universities (CAU, U Siegen, KIT) and LSFs as well as contribution to existing PhD schools (e.g. RACIRI, MATREC/MATRAC, European powder diffraction school). The long-term aims is to use the DAPHNE platform together with organisations such as DPG and DBG to discuss the implementation into university curricula. To support this, we aim to provide lecture material that can be used by universities. Moreover, we will pursue the dissemination via existing PhD-schools and initiate/promote networking possibilities for PhD students (network DAPHNE-VISION)

interested in data processing and organised beyond the sub-communities. This network (DAPHNE-VISION) will be supported by the DAPHNE core team.

Deliverable	Description
4.4.1	Establish data management as topic in existing schools (RACIRI, MATRAC, PDS, HERCULES, BOMBANNES, etc.)
4.4.2	Research data school in photon and neutron science, seed for young scientist's network
4.4.3	Ringvorlesung (webinars) and once a Heraeus seminar (2nd or 3rd year)
4.4.4	Preparation of lecture material for use in teaching at universities
4.4.5	Preparation of hands-on material for training at LSFs
4.4.6	Develop the young scientist network DAPHNE-VISION

5.4.5 Measure 4.5: Outreach to Society, Application and Industry

Objective: Inform public and strengthen the German innovation and industry, general lectures, workshops with other societies and industry

We aim to disseminate the objectives and results of DAPHNE to the wider public to raise awareness of the importance, the challenges and opportunities of the digitization for the DAPHNE community and beyond. Beside the general public, special emphasis will be put on industry- and commercially oriented areas such as catalysis, engineering, life sciences or health which will benefit and might contribute to the overall NFDI. Moreover, graduates form the DAPHNE community will be educated/trained in efficient data processing. Finally, increased efficiency in use of measured data justifies the use of expensive, tax-paid facilities for indispensable scientific information.

Deliverable	Description
4.5.1	Section on webportal to address industry and general public
4.5.2	DAPHNE4NFDI participation at events like facility open days, "Nacht der Wissenschaften", "Bunte Nacht der Digitalisierung", Dies Academicus, etc., children's university ("Kinderuniversitat"), "Maustag", general events for scientists and industry like "Scientists meet Scientists", VDI expert's Forum, DECHEMA/GeCatS day and related events
4.5.3	Initiate discussion on IP rights and further aspects relevant for commercially relevant areas
4.5.4	Position results in print/online media

5.4.6 Measure 4.6: Sustainable dissemination infrastructures

Objective: Sustain further development of the created infrastructures

To sustain the success of DAPHNE, networking and information flow between participating communities, users and LSFs, DAPHNE and NFDI need to be installed in a way of continuously working instances and workflows. This naturally will entail outreach to partners, other communities and the public. The native interest of the users on data processing and management combined with the responsibility of KFN/KFS for DAPHNE will guarantee the sustainable link to NFDI and continuous work on all DAPHNE duties. In addition, the aims and efforts of DAPHNE need to be coordinated with the IT infrastructure at the LSFs. DAPHNE aims to lay the foundations for a sustainable network between partners, use case teams, KFN/KFS and facilities as a driving force of data management and process development within the communities. Webpage, social media presences and their interactive tools will be maintained on long-term scale, with the goal to be accepted and adopted by the users. Workshops, schools and frequent meetings need to be organised in recurrent manner (to be known and expected). The user platforms for data management and processing tools will be installed at the LSF, and the network of young scientists (DAPHNE-VISION) will become an independent instance to continue as an integrated part of the NFDI infrastructures.

Deliverable	Description
4.6.1	Data management as a resort in KFN/KFS with support of DAPHNE-VISION
4.6.2	Support DAPHNE-VISION

5.5 Task Area 5: External communication and policy

Summary

Commonly accepted data policies need to be formulated, aligned and harmonised with the stakeholders of DAPHNE4NFDI on both the national and on a European level.

TA5 seeks to establish and define common data policies by continuous communication with the user and facilities. Interaction, exchange and cross-consortia activities with other NFDI consortia need to be initiated, formulated and executed. TA5 will establish collaborations and networking with NFDI consortia. The definition of pilot workflows and use cases, pilot science projects and procedures for data processability that are relevant to the different communities involved are also part of this task. DAPHNE4NFDI will collaborate, network and coordinate with European partners.

DAPHNE covers a broad variety of scientific communities including biology, condensed matter physics, physics, chemistry, geology, medicine and materials science, which are connected by their common use of similar photon and neutron methods and data schemes. TA5 aims to define common data policies, use cases and pilot workflows and standardised best practices with the aim of agreeing upon common standards. TA5 also encompasses cooperation with the other NFDI consortia, which are connected either by similar scientific questions and/or by issues of data management. Here we will also seek to coordinate and communicate with European user organisations, such as ESUO and ENSA, and the consortia of facilities, such as LEAPS and LENS. Participation in European projects (PaNOSC, etc.) that are closely linked to DAPHNE is envisaged, through partners or representatives of DAPHNE who are already active as observers.

This task area contains the following three measures:

5.1 Define common data policies in DAPHNE: best practice, realistic limitations, FAIR principles, legal aspects such as GDPR, logbook archiving and data retention requirements

5.2 Cooperate and network with other NFDI consortia

5.3 Collaborate and coordinate with European partners such as LEAPS, LENS, ESUO, ENSA, PaNOSC, EOSC, ExPaNDS.

5.5.1 Measure 5.1: Adoption of common data policies

We seek to move towards common data policies for the German photon and neutron community, including the facilities. This policy needs to be aligned with users, with facilities, with efforts on a European level and with the overall NFDI process. DAPHNE will participate in framing data policies at the centres and in initiatives such as PaNOSC. In this section, the members of DAPHNE will work out a best practice example for a common x-ray and neutron data policy. The broader goal is a (European) data policy that is accepted by users and facilities independently of the beamtime granted for a specific facility or beamline.

At the moment data policies are defined by the individual user facilities and users must agree to whatever data policy is applied when they accept access to beamtime. This often includes general principles, such as accepting rules for beamtime allocation, definitions of ownership, curation and legal aspects specific to the respective countries. A very good example is the policy framework defined in the European PaNOSC project.[47] Here, acting partners such as facilities, users, principal investigators, and experimental teams are properly defined as well as the terms raw data, metadata, data catalogues etc. Access to raw data and the associated metadata is defined, including an embargo period of 3 years during which access is restricted to the experimental team. Rights and responsibilities of the PIs are defined and specified, as well as ownership of and access to results and curation of results including responsibilities of the different partners involved. Finally, best practice for metadata capture and storage of results for IP-sensitive data will be defined.

An understanding of the implications of FAIR principles for the DAPHNE community is required. We will generate a white paper on FAIR principles and their consequences both for users and for facilities, which will provide input for the other tasks within DAPHNE. These efforts will be aligned with broader European efforts.

Finally, we must gain an understanding of the implications of the legal aspects of data management for DAPHNE, for both non-profit university and industrial users, especially with regard to open data, such as logbooks and metadata. Legal aspects must be discussed on an NFDI-wide level.

Deliverables Measure 5.1

Deliverable	Description
5.1.1	Organise DAPHNE4NFDI and community-wide workshops for defining and discussing data policies. Generate a DAPHNE4NFDI data policy white paper.
5.1.2	Contribute and participate, as DAPHNE4NFDI, in European activities such as PaNOSC, LEAPS, LENS and work together with users in ESUO and ENSA.
5.1.3	Organise workshops for discussing data policies within the NFDI, especially with consortia within our cross-community activities.
5.1.4	Organise joint workshops with BMBF ErUM Data partnership for digitization of the BMBF.
5.1.5	Organise workshops/sessions, e.g. during user conferences, for addressing the implications of FAIR principles for both users and facilities (see also TA4).
5.1.6	Discuss and align strategies on a European level with ESUO, ENSA, LEAPS, LENS and PaNOSC.
5.1.7	Generate a white paper on FAIR principles within DAPHNE4NFDI.
5.1.8	Implement the results of the FAIR white paper into the TAs of DAPHNE4NFDI.
5.1.9	Organise workshops with typical industrial users and university users to discuss legal aspects, such as 10-year storage, privacy rights etc.
5.1.10	Generate a white paper on the legal aspects of DAPHNE4NFDI.
5.1.11	Organise workshops on an NFDI level for addressing legal aspects of the entire NFDI.

5.5.2 Measure 5.2: Cooperation and networking with other NFDI consortia

This measure defines the scope on DAPHNE in relation to interfaces, overlap, complementarity, differences and synergies between DAPHNE and other consortia. The NFDI consortia FAIRmat, MaRDI, NFDI-MatWerk, NFDI4CHEM, NFDI4CAT, and NFDI4ING provide or will provide data and services which are of interest for DAPHNE. This comprises data from DAPHNE related fields such as chemistry, biophysics, catalysis or the material sciences but also simulations and AI/ML tools. The consortium PUNCH4NFDI is connected to DAPHNE via the use of large-scale research facilities (e.g. DESY) and the challenge of coping with very large data sets and meta-data schemes.

In this measure, we will define and identify common needs and wishes for the ability to process raw data sets online and for the availability and re-usability of processed data. Data repositories need to be set up which provide easy (single-point) access to the processed data sets documented in scientific publications. A searchable database is needed as well. This will allow the large sets of data generated by the photon and neutron community to be used by the other

consortia as well. At the same time, DAPHNE would benefit tremendously from being able to access simulation data sets and other complementary data produced, e.g. by FAIRmat, MaRDI, NFDI-MatWerk, NFDI4CHEM, NFDI4CAT and the other consortia. Theoretical data sets should also be compared directly with experimental data - or else they can also be used as a learning input for AI measures, as provided by Math4NFDI. This scientific boundary is where we see the greatest potential for extending the NFDI beyond our own user community.

However, we are fully aware of the difficulties in cross-community work in the area of data and data formats mainly caused by a lack of scientific motivation for establishing a sustainable long-term collaboration across disciplinary boundaries. We will mitigate this risk by defining science-driven use cases (or pilot projects) with other consortia, which ought to work out concrete steps and procedures for re-use and exchange of databased on prototypical scientific use cases. The aim of these projects is to establish and foster collaboration between the actors in the various consortia and to identify the most urgent and pressing problems for FAIR data within NFDI. Typical examples of common science projects comprise

- **FAIRmat**: simulations/theory in which theoretical results are complemented and compared with x-ray/neutron results etc., as well as sample identification, e.g. persistent sample identifiers (PID) to connect home laboratory and large-scale facility experiments.
- **NFDI4CAT**: correlation of structure and functionality in catalysis (e.g. nanoscale particles)
- **NFDI4CHEM**: self-assembly processes of nano structures, battery cycles, x-ray spectroscopy of electronic levels, crystallography
- **NFDI-MatWerk**: micro and nanostructure of technically relevant materials
- **NFDI4ING**: strain and stress in load bearing materials (e.g. train wheels, car industry etc.)
- **MaRDI**: PB of high quality x-ray data for correlation analysis, training, event analysis etc., to be essential for testing of AI codes

In addition, we envision to initiate a close partnership with **PUNCH4NFDI** with regards to topics of

- curation, archiving, processing and reduction of very large data sets
- collection of meta-data schemes for very large data sets.

Based on these projects, we will work out best practice examples and define procedures for data processability that are relevant to the different communities involved.

Beyond the measures described above, DAPHNE is committed to support the overarching goal of the NFDI by addressing cross-consortia issues. In agreement with the Berlin declaration, we are pursuing the following common objectives in the NFDI:

Collaborative governance and general framework:

- Common vision of the NFDI, long-term foresight and common strategic planning
- Governance & sustainability
- Cultural exchange aka reputation, publication/funding policies and credit systems
- Policy advice, consultation and outreach with respect to the NFDI
- International visibility and networking of the NFDI
- Human resource management, recruitment, development

and for the community (user) involvement:

- User-driven adaptive development of the NFDI (cross-domain use cases)
- Training, undergraduate and graduate education, professional development
- Stimulating a cultural shift of users and providers towards FAIR and open research data

Deliverables of Measure 5.2

Deliverable	Description
5.2.1	Organise workshops between funded NFDI consortia with the goal of establishing common examples, policies, best practice and guides for metadata catalogues, data repositories, online electronic logbooks, etc.
5.2.2	Identify potential power user groups which are active in the field of DAPHNE4NFDI and interested counterparts in other NFDI consortia for forming science driven use cases and pilot workflows
5.2.3	Identify and execute common use cases between DAPHNE4NFDI and the consortia **FAIRmat, NFDI4CHEM, NFDI4CAT, MaRDI, NFDI-MatWerk, NFDI4ING**, in the field of data exchange and data processing between the different communities
5.2.4	Workshop and meetings with **PUNCH4NFDI**. Develop strategy for large volume data handling in NFDI. Update regularly.
5.2.5	Organise user workshops for identifying common needs in terms of data policies and accessibility of raw and processed data in NFDI.
5.2.6	Establish a cross-topic platform for regular meetings and discussions between the consortia
5.2.7	Establish an inter-consortium working group addressing cross-cutting topics. The leadership of a such working group needs to be organised, but can only be defined after funding decisions have been made.

5.5.3 Measure 5.3: Cooperation with European partners such as LEAPS, LENS, ESUO, ENSA, PaNOSC, EOSC, ExPaNDs, ORSO, ISPyB consortium

On the European stage, efforts need to be coordinated and coherently staged with the user facilities organised in LEAPS (photon science) and LENS (neutron sciences) and the user organisations ESUO and ENSA. These four organisations are the European stakeholders of the photon and neutron community, together representing 25 facilities and 30,000 users. The German user community organised within the KFS and KFN represents a considerable fraction of this

European network. The projects PaNOSC and ExPaNDs that are currently active on a European level display a strong overlap with the objectives and goals of DAPHNE. These projects have been initiated and are driven by the large-scale facilities only, and are funded by European programs such as Horizon2020. In contrast to these European programs, DAPHNE is strongly user-driven and will complement the European efforts of the facilities. DAPHNE brings in the user community and will help to specify the needs for data analysis software and prepare the infrastructure for the storage and re-use of data from the facilities on a national level, taking into account the international aspect of many data sets. In a European context, it is also vital to ensure that network connectivity is enhanced and cheap (if not free), to make sure that setting up the EOSC is not impeded by the bandwidth between LSFs, universities, and data centres.

PaNOSC: The Photon and Neutron Open Science Cloud project, PaNOSC, to which DAPHNE partners such as EuXFEL, ESRF, ILL, and ESS are contributing, with the goal of building on existing local metadata catalogues and data repositories to provide federated services for making data easily findable, accessible, interoperable, and re-usable (FAIR). PaNOSC will also develop and provide data analysis services. The services will include notebooks (Jupyter-based), remote desktop applications and containers or virtual machines (VMs). These services will be provided locally by the LSFs for their users (especially when they are on site or when the volume of data is too large to be exported); the same services should also be available on the EOSC for general use. The data analysis services will offer the data, the software, the IT capacity and the necessary scientific support, all as a single user experience. All these services should be fully integrated into the EOSC catalogue, in terms of discovery, accessibility, and user authentication/authorisation, SLA, accounting etc. The PaNOSC cluster will also help introduce a new data culture to the user community – via training at each site, and workshops on scientific data management and publishing practices. Best practices in data stewardship will be shared with other laboratories within the photon and neutron community and other clusters. Experiences, trials and results will be shared openly through publications and meetings. The positive experience of implementing an open data policy and connecting data and services to the EOSC will help convince other x/n institutes still struggling with adopting the FAIR principles.

ExPaNDS: The ambition of the EOSC Photon and Neutron Data Services (ExPaNDS) is to enrich the EOSC with data management services and to coordinate activities to enable national photon and neutron large-scale facilities to make the bulk of their data ‘open’, following FAIR principles, and to harmonise their efforts to make their data catalogues and data analysis services accessible through the EOSC, thereby enabling them to be shared uniformly. EOSC currently provides a range of services that needs to be adapted to the ever-increasing requirements of scientific experiments carried out at various x/n LSFs. It is essential that these services become

standardised, interoperable and integrated to fully exploit the scientific opportunities at x/n LSFs. ExPaNDS therefore seeks to enable EOSC services and provide coherent FAIR data services to the scientific users of x/n LSFs, connect x/n LSFs through a platform of data catalogues and analysis services through the EOSC for users from LSFs, universities, industry etc., gather feedback and cooperate with the EOSC governance bodies to improve the EOSC, and develop standard relationships with and connections between scientific publications, x/n scientific data sets, experimental reports, instruments and authors (via ORCID).

DAPHNE will liaise via EMBL with the well-established ISPyB consortium[24] on concepts for sample tracking and experiment reporting in use on various MX beamlines since the early 2000's. Open Reflectometry Standards Organisation (ORSO) [48] is a recently formed word-wide collaborative network of scientists interested in developing common file formats, analysis tools and reference data for the neutron and x-ray reflectivity community. There will be close interactions with the use case.

The objective is to integrate DAPHNE in a European context with the goal of achieving a European solution, by cooperating with European partners and strengthening the user component of European projects.

Deliverables Measure 5.3

Deliverable	Description
5.3.1	Establish collaborations with European projects via workshops and discussions
5.3.2	Approach the user organisation ESUO and ENSAs to organise a user-driven participation in European data projects
5.3.3	Generate a European user white paper together with ESUO and ENSA
5.3.4	Organise workshops and meetings and participate in the data activities of the European photon and neutron facilities with the goal to harmonise data policies of x/n facilities on a European level
5.3.5	Establish a common working platform between DAPHNE4NFDI/NFDI and European projects
5.3.6	Work with LEAPS, LENS, ESUO, ENSA, ORSO and NFDI on technical questions for the EOSC, such as bandwidth costs etc.

5.6 Task Area 6: Management

Task Area 6 is concerned with logistics of running the DAPHNE project and will be based at the host institution, DESY. This includes everyday tasks such as administration, project coordination, bookkeeping, tracking of progress, reporting, and making sure things happen. It also provides mechanisms for managing change and risk across the entire project. The distributed nature of DAPHNE calls for an effective management unit responsible for planning, coordinating and controlling project activities, ensuring transparent decision making, balancing competing priorities, and tracking resource usage relative to planned deliverables. Additionally, project management assumes responsibility for coordinating professional software development across the project.

We plan to hire a project manager, and experienced staff from DESY administrative and financial teams will support the project as required during the course of the project. **Additionally, software development management will be provided by way of in-kind resources to coordinate professional software practices through the consortium.** The software development manager will be embedded in DESY IT and bring professional software development expertise from other large software projects at DESY into the DAPHNE project.

Task area 6:

1. Manages consortium activities to maximise project impact, implementing the organisational structure and appropriately and ensuring the long-term legacy of the project
2. Manages funding and resource adjustments where necessary, react to changing circumstances including managing risks and exploiting new opportunities
3. Ensures compliance with contractual obligations and NFDI requirements
4. Ensures professional software development through the project.

These objectives are implemented through the following measures:

5.6.1 Measure 6.1: Project initiation, establishing legal, financial and compliance

This section will:

- Ensure the legal and financial operation of the consortium
- Negotiate legal agreements with all partners, provide legal framework for distributing funds
- Set up of governance structure and advisory boards, and coordination of hiring of personnel

- Provide financial oversight, tracking and reporting of progress, internal communication and quality assurance
- Communicate with the DFG

Deliverable	Description
6.1.1	Kick off meeting
6.1.2	Establishment of legal and financial agreements with partners, distribution of funds

5.6.2 Measure 6.2: Day to day management of operations

This task manages the timely and precise execution of the work plan in a highly distributed project, including the flow of information between participants, organises regular project meetings, and ensures overall project coordination. The central project office will act as a central point of contact for coordinating with other NFDI communities and with members of the KFS and KFN. Monthly video conferences and an annual conference will be organised.

This measure is responsible for tracking expenditure at participating institutions to ensure that expenditure tracks with the project plan, will define rules for professional software development within other task areas, and verify that standards, procedures and metrics are defined, evaluated and applied across the project. Capturing and reviewing all actual information about actual costs (financial and human resources efforts) and comparing it with the planned forecasts as well as reporting to DFG.

Deliverable	Description
6.2.1	Annual reporting to DFG

5.6.3 Measure 6.3: Professionalising software development

This measure promotes and implements software quality assurance processes for ensuring the necessary quality criteria, indicators, and tests that are necessary to ensure high quality software components and to foster professional software development culture within the collaboration. This enables DAPHNE to support the complete software lifecycle management process, including source code management, continuous integration and delivery, continuous deployment, and ongoing monitoring and assessment of the services provided. The software development management task links strongly to outreach and education so as to foster best practice in the community. This measure is an overarching topic across all over DAPHNE, especially TA1-3. Therefore, it is placed in TA6 that will take over the organisational lead. Implementation will rely on the resources of TA1-3 as well as the in-kind contributions of the co-applicant institutions (IT centres).

Deliverable	Description
6.3.1	Development of software test bed
6.3.2	Deployment of software development infrastructure

5.6.4 Risk Management

Managing risk is an essential part of the project management. For example, we foresee the following management risks associated with implementation of DAPHNE:

1: Coordinating effort across multiple institutions runs the risk of a loss of focus. We address through regular meetings of all funded participants at the coordination level. Likelihood: Low. Potential impact: Medium

2: Hiring skilled staff on DFG rates in the IT labour market area is challenging. Computer scientists and software engineers with experience in infrastructure development and web services are in very high demand. Competing for staff will be challenging, especially at the pay rates we can offer. Fixed term contract limits due to 6 year limit in the *Wissenschaftszeitvertragsgesetz* additionally complicates the hiring process. We aim to mitigate this by allowing adequate time between funding and start of project activities to attract good staff. The current COVID environment may help with availability of skilled staff. Likelihood: Medium-high. Potential impact: Medium

3: Uncoordinated and incompatible software development could delay near term implementation and limit long-term impact. This proposal has a large component related to the development of services, software, APIs, containerised tools and workflows in photon and neutron science experiments, as well as their interaction with infrastructures and services in the NFDI as a whole and beyond (such as the EOSC). We are aware of typical risks in such developments, which are delayed developments, unstable and unsuitable software and the inability of partners to make concerted and well- accepted decisions. We aim to mitigate this by fostering best software practices as a condition of funding. Likelihood: Low. Potential impact: Medium

On the subject of risk management, it is necessary to enable DAPHNE to react to new and unforeseen developments in the national RDM landscape. To this end project management may redeploy FTE annually and modify work plans in a cost neutral manner, in consultation with NFDI management at DFG, to react to new developments in the research community. A possible scenario, for example, is the successful funding of parallel projects which would benefit from integration into DAPHNE, or parallel developments which can be exploited by way of synergy. We will regularly review the need for cost-neutral project extensions or rebalancing of effort to integrate novel and unforeseen developments during the course of the project.

6 Appendices

6.1 Bibliography and list of references

[1] https://www.sni-portal.de

[2] https://www.bmbf.de/de/erforschung-von-universum-und-materie---das-rahmenprogramm-erum-4388.html

[3] https://www.sni-portal.de/de/Dateien/challenges-and-opportunities-of-digital-transformation-in-fundamental-research-on-universe-and-matter

[4] https://www.rcsb.org/

[5] https://creativecommons.org/licenses/by/4.0/deed.de

[6] Data set collected on 8-Aug: 3600 10 ms frames of 7 Mbytes collected in 36 seconds. Total compressed data volume ~ 25 GBytes. Reduced data volume ~ 20 Mbytes.

[7] Note that model construction and refinement usually take place in the home-lab, i.e. is decoupled from and not controllable by the facility where the data were collected.

[8] Bernstein-1971: Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. Journal of molecular biology, 112(3), 535–542. https://doi.org/10.1016/s0022-2836(77)80200-3.

[9] IUCr-1989: Commission on Biological Macromolecules. (1989). Acta Cryst. A45, 658. doi:10.1107/S0108767389007695.

[10] Baker-1996: Baker et al. (1996) Crystallographic Data Position, Nature 379:202.

[11] Nature-1998: Campbell: New policy for structure data. Nature 394:105 (1998)

[12] Science-1998: Bloom: Policy change. Science 281, 175 (1998)

[13] PNAS-1998: Cozzarelli: New policy on release of structural coordinates. PNAS 95:iii (1998)

[14] Berman-2014: Berman, H. M., Kleywegt, G. J., Nakamura, H., & Markley, J. L. (2014). The Protein Data Bank archive as an open data resource. Journal of computer-aided molecular design, 28(10), 1009–1014. https://doi.org/10.1007/s10822-014-9770-y

[15] DDDWG-2017: https://www.iucr.org/resources/data/dddwg/final-report.

[16] https://zenodo.org/

[17] https://www.ebi.ac.uk/biostudies/

[18] https://data.sbgrid.org/

[19] https://proteindiffraction.org

[20] https: www.cxidb.org

[21] B. Ravel, J.R. Hester, V.A. Sole, M. Newville, J. Synchrotron Rad. 19 (2012) 869-874, DOI: 10.1107/S0909049512036886, K. Asakura, H. Abe, M. Kimura, J. Synchrotron Rad. 25 (2018) 967-971, DOI: 10.1107/S1600577518006963

[22] http://ixs.iit.edu/database/data/Farrel_Lytle_data/

[23] https://xaslib.xrayabsorption.org/elem

[24] https://ispyb.github.io/ISPyB/

[25] See also: https://www.sni-portal.de/de/nachrichten/ergebnisse-der-kfs-umfrage.

[26] https://www.dcache.org/

[27] https://nicos-controls.org/

[28] https://public.ccsds.org/pubs/650x0m2.pdf

[29] https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf

[30] http://www.dublincore.org/

[31] http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

[32] https://www.sine2020.eu/about/the-road-to-the-ess/secop.html

[33] https://www.crystallography.net/

[34] https://wwwdev.ebi.ac.uk/pdbe/emdb/emdb_schema3

[35] http://ixs.iit.edu/database/data/Farrel_Lytle_data/

[36] https://xaslib.xrayabsorption.org/elem/

[37] https://figshare.com/articles/Magnetoelastic_hybrid_excitations_in_CeAuAl3/7803092/2

[38] https://scicatproject.github.io/

[39] https://www.openaire.eu/

[40] https://www.eudat.eu/services/b2find

[41] https://www.project-freya.eu/en/pid-graph/the-pid-graph

[42] https://github.com/IGSN/metadata/tree/master/description

[43] https://www.inchi-trust.org/

[44] https://emmc.info/taxonda-dashboard/

[45] https://schemas.nist.gov/

[46] https://uni-tuebingen.de/en/research/core-research/cluster-of-excellence-machine-learning/home/

[47] https://www.panosc.eu/wp-content/uploads/2019/05/PaN-data-D2.1_PolicyFramework.pdf

[48] www.reflectometry.org