The DAPHNE NFDI Consortium

1 Binding or non-binding letter of intent as advance notification

☒	Binding letter of intent (required as advance notification for proposals in 2019)
☐	Non-binding letter of intent (anticipated submission in 2020)
☐	Non-binding letter of intent (anticipated submission in 2021)

2 Formal details

Planned name of the consortium

DAta for PHoton and Neutron Experiments

Acronym of the planned consortium

DAPHNE

Applicant institution

Deutsches Elektronen-Synchrotron (DESY),
Notkestr. 85
D-22607 Hamburg, Germany

Prof. Dr. Dr. h.c. Helmut Dosch
Head of the DESY directorate

Spokesperson

Dr. Anton Barty
anton.barty@desy.de
DESY Photon Science

3 Objectives, work programme and research environment

Research areas of the proposed consortium

21 (Biology), 22 (Medicine), 23 (Agriculture, Forestry and Vet. Med.), 31 (Chemistry), 32 (Physics), 34 (Geosciences), 42 (Thermal Engineering/ Process Engineering) and 43 (Materials Science and Engineering)

Concise summary of the planned consortium's main objectives and tasks

This proposal is a merger of DAPHNE for X-ray science and NData from the neutron community, centered on the data challenges facing users stemming from significantly increasing data rates and volumes at large scale analytical facilities.

The X-ray and neutron science communities represent a broad area of scientific disciplines with a common need of high level, rapid data analysis and the challenge of implementing research data management. These communities are represented by the KFS and KFN committees who have worked together for many years. Here we have come together to meet the challenges of the digital transformation. The community performs thousands of individual user experiments at central facilities each year across many disciplines using a range of techniques and a diverse set of instrumentation. Individual experiments can produce up to millions of files and in some cases over 700TB data per week depending on experimental configuration. Moreover, the community currently witnesses a fundamental change in both the amount of recorded data and the corresponding data rates triggered by the increase of brightness of the sources themselves (X-ray free-electron laser, high brightness storage rings and new neutron facilities) and by the rapid increase of 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, storing and management previously known only from experiments in areas such as high-energy physics. This applies not only to IT infrastructure for storing data, but also to the development of new algorithms and software concepts in which data is processed at the facility both during and after experiments rather than the model of researchers taking raw data home for later analysis. This revolution in data management requires careful consideration of ethics and legal aspects. To be successful, this transition requires an investment in efficient research data management for the community. This is the challenge addressed by DAPHNE.

DAPHNE brings together the large-scale research facilities in X-ray and neutron science with users representing typical scientific domains to advance the state of data management in the community. The consortium is representative of the broader community of users employing a broad range of X-ray and neutron techniques which is in total more than 5000 scientists throughout Germany. User facilities have traditionally a very close interaction with the user communities driving the scientific and technical developments of the facilities. This interaction and connection between users and facilities is a key element of DAPHNE, since implementing data management requires a joint and coherent approach with the facilities acting as data custodians and the user communities acting as data curators. The consortium consequently comprises the user facilities and users at large scale instruments such as the X-ray sources PETRA III and FLASH at DESY, ESRF in Grenoble, BESSY II at HZB, the European-XFEL, the ELBE Center for High-Power Radiation Sources at HZDR, as well as the Heinz Maier-Leibnitz Zentrum with its neutron source. Included are of course upgrade plans for Petra-IV and BESSY-III.

A federated approach to the DAPHNE consortium will ensure that the needs of the X-ray and neutron science user community are not only met, but that solutions address the needs of the user community. Practical outcomes for the user community will be achieved by working closely with the large user facilities which generate data and increasingly provide storage and cloud computing for users. Solutions and tools generated in DAPHNE will provided to the user community to manage data from lab-based X-ray sources and the institutionally run X-ray sources (DELTA, KARA) as well. We envisage the following organisation of work within DAPHNE:

Task 1: Management

Overall project coordination, definition and execution of the governance structure, implementation of scientific advisory board, etc.

Task 2: Policies, communication and cooperation with other NFDI consortia and international partners

DAPHNE covers a broad variety of scientific communities including biology, condensed matter physics, physics, chemistry, geology, medicine, and material sciences connected through the common use of similar X-ray and neutron methods and data schemes. Task 2 aims to define common data policies, pilot workflows and standardized best practices with the goal to agree on common standards. Task 2 also comprises the cooperation with the other consortia which are connected either by similar scientific questions and/or by issues of data management. Here we will also seek communication with the European user organizations such as ESUO / ENSA and the consortia of facilities such as LEAPS and LENS.

Task 3: Dissemination and outreach, communication with user communities

For successful implementation, the scientific users of the data need to be strongly involved in all phases of specification and development of tools and methods. Pilot projects are planned to bring user and facilities together. The task further aims on the coordination and communication among the user communities and after successful determination of tools with users on the outreach and dissemination of data management related topics (workshops, schools, etc.). The first measure is to educate the community in the FAIR data principles. This task also develop schemes to implement data management topics into university curricula e.g. through lab courses etc. and aims at communication with existing and planned excellence clusters, collaboration research centres etc. In addition, the topic of IP-sensitive data needs to be discussed.

Task 4: Managing data during experiments

As experiments will be carried out in large scale facilities on individual instruments, this task will address standardization and automatic creation and extraction of relevant instrument metadata and measurement data. In addition, easy-to-use and open source electronic logbook (eg. Jupyter notebook) will be developed, and implemented to capture information from both the facility and the users. This aim is to capture sample metadata and record the experiment protocol and information in an electronic logbook integrated into the instruments at the facilities. Standardisation and common data formats are also promoted. In collaboration with other consortia, a code of ethics and legal aspects for recording experiments will be developed in keeping with FAIR data principles.

Task 5: Data catalogue “Find and reuse data”

DAPHNE will aim to develop a common ontology and standardised metadata database at X-ray and neutron facilities in close collaboration with the relevant user communities (see task 2+3). Linking these metadata with standardized datasets (task 4) in a federated searchable database will allow locating and interpreting data. For some techniques even reference databases are missing (e.g. XANES, XES). The advent of big data analyses necessitates that the catalogue is not only searchable by humans but also machine readable.

Task 6 – Storage, Analysis and Archiving, “Store and Analyse Data”

Data collected at x-ray and neutron facilities are typically subjected to software intensive data reduction and analysis. DAPHNE aims to promote best practice and workflows including tool to handle large datasets within federated infrastructures for providing accessibility and interoperability of the data sets. DAPHNE provides tools for data archiving and preservation including an archive for file formats and software. Tools for remote data analysis will be developed and deployed, as will new algorithms for standardised data analysis workflows including data visualization, data reduction and data filtering.

Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium's objectives

Infrastructure:

Data storage and analysis facilities at the photon science facilities are largely handled in-house. PETRA III and Eu-XFEL take advantage of the DESY computer centre, which is connected to and draws experience from the field of particle physics through operation of a Tier-2 centre within the WLCG for HEP. This infrastructure provides fast disk storage for online analysis, backed by tape storage of data for long term archiving. Remote analysis and data download portals are available through shared resources. Standard data formats and containers such as NeXus help simplify data handling.

In the field of neutron science, data storage and archives are currently operated locally in collaboration with the Leibniz Supercomputing Center (LRZ) which can also provide long term tape storage. The instrument control software is already highly standardized (NICOS) across the instrument suite and it will be further developed for automatic capture of the instrument and sample metadata. NICOS is also being employed at other neutron facilities, such as Paul-Scherrer Institute (PSI) and the European Spallation Source (ESS). Software analysis services are developed within the Scientific Computing group of MLZ, which is active in the NeXus community for standard data formats.

DAPHNE will build on top of this infrastructure and extend the architecture towards cloud and remote computing models in line with the goals of the European Open Science Cloud (EOSC) and associated European projects Photon and Neutron Open Science Cloud (PaNOSC) and EOSC Photon and Neutron Data Services (ExPaNDS), and within the League of European Accelerator-based Photon Sources (LEAPS) and the League of Advanced European Neutron Sources (LENS) community.

DAPHNE will use services and tools from the Physikalische Technische Bundesanstalt (PTB) to ensure data quality, comparability and proper data standards. The PTB is also operating a synchrotron source and one accelerator based neutron source for their missions. DAPHNE will use outreach tools and services from the DPG to ensure communication and interfacing to the German Physics community and to develop modules for university curricula. DAPHNE will utilize library tools, services and knowledge (such as ORCID, DOI, interlinking, semantic annotation etc.) from the TIB for incorporating professional library aspects into the consortium.

Common data formats:

X-ray/neutron data are diverse. Data captured by area detectors is commonly but not exclusively stored in HDF5 format. The Hierarchical Data Format (HDF) itself is a de facto standard for storing binary data and is used by various scientific communities and companies. NeXus combines the hierarchical data format (HDF5) with well defined, experiment-specific metadata schemata, but does not capture all metadata. Most major application frameworks support the HDF standard, and thus NeXus, since NeXus is fully HDF-compliant. The NeXus community comprises members of the consortium, is continuously developing to meet the needs of experiments at photon, neutron or muon facilities.

DAPHNE, together with the facilities, will develop common protocols to allow harmonisation and interoperability across the community consistent with the FAIR principles.

Data processing / data analysis methodologies

Data processing and data analysis methodologies in X-ray and neutron science are quite diverse due to the different scientific fields involved, requiring a flexible approach for data technologies such as data containers or data processing/analysis software. The 'one size fits all' top-down model of particle physics is inappropriate to the needs of this community. At the same time there is currently a transition of responsibility for data processing and analysis from the user to the facility due to the large volumes of data collected on ever faster

timescales. This requires the development of automated analysis pipelines for standard methods such as tomography, small angle scattering or diffraction.
Making such tools accessible to the community and managing the transition is one of the central tasks in DAPHNE.

Interfaces to other proposed NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration
Interfaces and overlap to other consortia have been discussed on a pre-NFDI workshop in Berlin on May 6 with FAIRMAT, NFDI4CHEM, NFDI4ING and NFDI4MSE and with NFDI4CAT in the following week. During the workshop we identified the scientific fields to be addressed by the different NFDI consortia, the structure of the potential user community and data producers, challenges and potential fields of synergies.

FAIRMAT
A detailed agreement for collaboration has been issued with FAIRMAT. While the specific needs, tasks and challenges for both consortia in terms of data management are quite different we agreed on an extended collaboration with the following tasks:

- Define a science driven common pilot project to identify the needs and interfaces in terms of data structures, data exchange and metadata between the two consortia
- Organize a common DAPHNE-FAIRMAT workshop in 2021
- Define and develop translating software of the electronic logbooks to facilitate data exchange.

NFDI4CHEM, NFDI4MSE, NFDI4ING
In the pre-NFDI meeting we identified common tasks and topics with the three consortia (such as e.g. streamlining metadata, e-logbooks) and agreed on collaborating. Developing the details of the envisioned collaborations, however, will take some more time than for the other consortia and is subject to the full proposal.

NFDI4CAT
The characterization of data plays and important role in catalysis and, hence, X-ray and neutron data conducted operando are relevant for NFDI4CAT. It is furthermore an excellent platform for interaction with industry and thus outreach. Joint development of data formats also for correlative techniques with joint workshops are planned.

PAHN-PaN and Astro-NFDI
Driven by the BMBF ErUM Data initiative the eight user communities organized in the ErUM field conceived a coordinated action plan for the next ten years. This action plan comprises a catalogue of measures for advancing the digital transformation in the field of ErUM data. In this context extensive discussions have been held between the X-ray/neutron science community and astrophysics, particle and hadron physics. In the field of data management we clearly identified the need for science driven solutions and avoiding forcing one-fits-all solutions to the different communities which will ultimately not be accepted by the scientists in their daily work. However, despite the different science fields we also identified overlap and mutual interest in the area of management of metadata, open data and handling large data sets etc. Here, the X-ray and neutron community identified needs in their field and would benefit from the experience and best-practice examples of the particle and astrophysics community.
Within the ErUM data process a cross community platform "partnership for digitalization" has been proposed for funding. This partnership should ensure collaboration also in terms of wider aspects of digitalization such as hardware, modern data analysis, outreach, web-interfaces etc. Depending on the (to developed) structure of the NFDI and funding levels available for ErUM data, we see the NFDI consortia of DAPHNE, PAHN-PaN and Astro-NFDI strongly linked to the cross-community actions in ErUM data.

NFDI4PHYS

DAPHNE agreed on collaborating with the Deutsche Physikalische Gemeinschaft (DPG), Technische Informationsbibliothek (TIB) and PTB as partners in our consortia - all of them are also organized in the NFDI4PHYS consortium. The partners can uniquely contribute to DAPHNE in the fields of outreach, library services and standardization and metrology services of X-ray and neutron data which will benefit the whole NFDI.

DAPHNE will promote local interfaces to other consortia. Synergies can effectively promoted when consortia have active involvement in more than one consortium: in the case of DAPHNE for example Kiel University (DAPHNE, NFDI4Objects, NFDI4CHEM and FAIRmat in the first round), Karlsruhe Institute of Technology (DAPHNE, NFDI4CHEM, NFDI4CAT, FAIRmat, Helmholtz-program “Information”), Siegen University(DAPHNE, PAHN-PaN). Here we plan internal meetings twice a year in order to develop a cohesive approach to explicit cross links. For example in Kiel: curriculum development; a common approach to Metadata and a university internal policy on FAIR data handling.

4 Cross-cutting topics

Cross-cutting topics relevant for DAPHNE and that need to be designed and developed by several or all NFDI consortia

Making data FAIR beyond single consortia

DAPHNE can serve the x/n user communities and DAPHNE interfaces to related consortia in the natural sciences. However, making data FAIR between other consortia and even between the disciplines is a considerable challenge that needs to be addressed by the whole NFDI. Interoperability, storage, archiving, definition and use of complex metadata structures beyond a single discipline need to be addressed by the NFDI as a whole.

Training, outreach and design of the NFDI

University curricula are dominated by classical domain specific tasks and modules. While the awareness for the need of FAIR data is increasing, bringing these topics into university curricula is a considerable challenge which is directly linked to the success of the whole NFDI idea.

The NFDI needs to develop science driven solutions - otherwise the developed tools will simply not been used by the scientist - but NFDI also has to assure a broader impact: the question of how to organize the interface areas of the consortia is highly relevant for DAPHNE.

Storage and archiving of data, hardware

While hardware is no topic for funding in the published framework of the NFDI, in reality the question where and how the data is stored and how to access, copy and work with the data in technical terms needs to be addressed by the whole NFDI.

Analysis of big data

The analysis of data is also not a topic within the framework of the NFDI but the large amounts of data that will be provided by NFDI also implies that for many disciplines new ways of analyzing data (e.g. machine learning and others) will be required. The interface of the NFDI to other initiatives in Germany (e.g. ErUM Data, Artificial intelligence funding by BMBF, etc.) needs to be developed by NFDI as well.

We identified the following cross-cutting topics where our consortium could contribute to the overall NFDI.

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 scientific fields in the natural sciences but still have well defined data providers determined by the large scale facilities involved. Thus we see the potential that DAPHNE can quite efficiently develop coherent standards for metadata, data management and best practice examples which are accepted/of interest for the broader user community and therefore disseminating it into a broader context of the NFDI. DAPHNE has also considerable experience in handling and managing large data sets. DAPHNE reaches into and is of interest for smaller university groups in a wide range of activities within the natural sciences. DAPHNE is also representing users from industry.

Helping to implement data policies/data structures by structuring of user communities

A large part of the NFDI is about communication. Communication between data users and data providers is needed for establishing e.g. data policies such as open data or defining metadata standards. DAPHNE is well prepared to tackle this problem due to the structured user organization and the coherence of 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 of defining data policies and organizing a larger community.

Implementation of FAIR principles on a European level

The German user organizations KFS/KFN active in DAPHNE and the X-ray and neutron facilities are embedded in a larger European context with well-organized structures. User organizations with democratically elected representatives comprise over 30.000 users organized in the European user organizations (ESUO/ENSA). DAPHNE has close ties to ESUO/ENSA and LEAPS/LENS and it thus ideally suited to actively participate in European initiatives such as e.g. PANOSC or EXPANDS. Solving FAIR and challenges such as metadata creation and curation on a European level is a realistic goal for the X-ray and neutron community. We therefore envision that DAPHNE can contribute considerably to the international/European aspect of the whole NFDI.

Education/training

The aspect of education and training of future scientists and generating awareness for data management matters is of considerable importance for DAPHNE. Bringing these topics into university curricula in the natural sciences is a considerable challenge. Here, DAPHNE will, with the help of both the DPG but also by cooperation with universities and data science schools, establish courses at universities. Our broad user community in combination with the science driven needs of the DAPHNE user groups will help to generate enough (science driven) interest for advancing university curricula. Here, the context to the ErUM data funding is also of importance in which we have foreseen a tenure track university program for data scientist.