1st Conference on Research Data Infrastructure

Natural Sciences

https://doi.org/10.52825/CoRDI.v1i.293

Published: 07 Sept. 2023

The DAPHNE4NFDI and PUNCH4NFDI Consortia in the NFDI

L. Amelung[1] [https://orcid.org/0000-0002-8137-1553], A. Barty[1],
B. Murphy[2] [https://orcid.org/0000-0002-1354-2381], C. Schneide[1] [https://orcid.org/0000-0003-1024-6875],
A. Schneidewind[3] [https://orcid.org/0000-0002-7239-9888], and
T. Schoerner[1] [https://orcid.org/0000-0002-7213-0352]

[1]Deutsches Elektronen-Synchrotron, Hamburg, Germany

[2]Christian-Albrechts Universitaet zu Kiel, Germany

[3]Forschungszentrum Juelich, Germany

Abstract: The DAPHNE4NFDI and PUNCH4NFDI consortia represent the large scale facilities in the German physical sciences community. Work in DAPHNE4NFDI and PUNCH4NFDI is characterised by the use of large-scale research infrastructures – reactors, light sources, accelerators, telescopes, observatories, satellites – that serve international research communities of up to several thousand users and produce data in the terrabyte, often petabyte and in future exabyte range. The communities will have to master massive challenges in data management, building on and extending their leadership in “big data” management, distributed computing, multi-user management, and data loss / data irreversibility issues.

Keywords: DAPHNE4NFDI, PUNCH4NFDI

1 First section

The DAPHNE4NFDI and PUNCH4NFDI consortia represent the large scale facilities in the German physical sciences community. Together they comprise 64 co-applicant and participant institutions – universities, Helmholtz Research Centres, Leibniz Institutes, Max Planck Institutes – and close to 15,000 researchers in the related physics and natural science communities.

DAPHNE4NFDI – “DAta from PHoton and Neutron Experiments” – focuses on data management for experiments using large scale photon and neutron research facilities. The properties of photons and neutrons allow us to probe the structure of matter to find out where atoms are and how they move - in solids, liquids and thin films even at very low temperatures or high pressures, penetrate through thick materials to obtain 3D structures, and map spatial chemical distributions. This makes it possible, for example, to see the tiniest cracks and pores in a turbine blade, to find small amounts of impurities in a semiconductor, the chemical states of catalysts or batteries in operando, or to determine the overall structure of a protein molecule or virus down to position of individual atoms. Photon and neutron research therefore includes a diverse range of

scientific disciplines. Individual instruments can generate over 1PB of data per day, and each facility hosts multiple instruments. The large amount of experimental data generated at high data rates presents us with substantial challenges: The data is often user-specific, as a wide variety of software is used for experimental control, data collection and data analysis. Therefore, it is currently not easy to share the data. There is a great need for digital tools to capture the data and meta data, curate the storage and provide high-level data analysis so that the data is reusable. Significant expert knowledge is required in order to use the data. DAPHNE4NFDI addresses the research data lifecycle, aiming to make data from photon and neutron experiments accessible to non specialists thereby making scientific work more efficient and gaining more knowledge from the data collected by others.

PUNCH4NFDI – "Particles, Universe, NuClei and Hadrons" – is the NFDI consortium of particle, astro-, astroparticle, hadron and nuclear physics. PUNCH physics addresses the fundamental constituents of matter and their interactions, as well as their role for the development of the largest structures in the universe - stars and galaxies. The achievements of PUNCH science range from the discovery of the Higgs boson over the installation of a 1 cubic kilometre particle detector for neutrino detection in the antarctic ice to the detection of the quark-gluon plasma in heavy-ion collisions and the first picture ever of the black hole at the heart of the Milky Way. The prime goal of PUNCH4NFDI is the setup of a federated and FAIR science data platform, offering the infrastructures and interfaces necessary for the access to and use of data, analysis workflows and computing resources of the involved communities and beyond. PUNCH4NFDI also offers tools for the efficient scientific exploitation of research data.

Work in DAPHNE4NFDI and PUNCH4NFDI is characterised by the use of large-scale research infrastructures – reactors, light sources, accelerators, telescopes, observatories, satellites – that serve international research communities of up to several thousand users and produce data in the TB, often PB and in future EB range. New and upcoming facilities like the High-Luminosity LHC (HL-LHC), the Square Kilometre Array (SKA), the Einstein Telescope (ET), PETRA IV, or FAIR (Facility for Antiproton and Ion Research in Europe) will soon produce data with unprecedented rates, volumes and complexities, compounded by the increasing prevelence of hybrid multi-modal experiments. The communities will have to master massive challenges in data management, building on and extending their leadership in "big data" management, distributed computing, multi-user management, and data loss / data irreversibility issues. The increasing demands for FAIR and open data and science now pose additional challenges that also drive the developments in the two consortia.

There are also differences between the two consortia, stemming from the different work organisation and scientific methods involved. In particular, the separation between "facilities" and "users": In DAPHNE4NFDI users often come to a given facility for a limited beam time, bring their individual sample, and take data at pre-installed experimental set-ups that they have not built themselves. Facility users are often expert in their own domain, but non-expert in photon science data analysis. In PUNCH4NFDI it is much more common that the scientists build their own, sometimes massive, detectors and operate it and exploit its data at the same time, sometimes over very long time-scales (several decades e.g. at the LHC), in large international collaborations that live equally long. Here, users and providers typically coincide. Nevertheless, there are also significant differences in approaches within the consortia between e.g. particle physics and astronomy, or smaller and larger enterprises.

Due to these differences, also the immediate challenges and goals of the two consortia partly differ in some aspects: For DAPHNE4NFDI, the following main tasks have been identified:

- Improve metadata capture through consistent definitions and 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;
- 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;
- 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.
- Develop education and outreach programs to export the knowledge and standards developed within DAPHN4NFDI.

PUNCH4NFDI has defined the following high-level goals:

- An integrated prototype package of dynamic digital research products, a science data platform, and the Compute4PUNCH compute and storage resources (Storage4PUNCH) coupled with single-sign-on (AAI). This will, for example, enable the collaborative creation and re-use of dynamic research products including analysis and simulation workflows based on a selected range of tools and required software environments. The milestone will enable the collaboration within and among research groups on the development of analysis workflows, and the combination of research products within the supported range of formats and tools.
- Data irreversibility solutions – a new kind of science in the age of too-large-to-be-stored data streams: Irreversible data reduction and compression based on real-time decisions are a prerequisite for future discovery science in the PUNCH fields and will become more and more important in other branches of science.
- Rolling out the PUNCH4NFDI outreach and education programme: It is important to share the expertise with the wider science community.

Despite these differences, a close and lively cooperation between the two consortia exists. An assessment of synergies and the exchange of information on tools or developments takes place on regular basis.

2 Second section

Data availability statement

The submission is not based on data.

Author contributions

Conceptualization: AB, BM, AS and TS; Project Administration: LA, AB, CS and TS; Writing original draft: TS; Writing review and editing: LA, AB, BM, CS, AS and TS.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the consortia PUNCH4NFDI and DAPHNE4NFDI in the context of the work of the NFDI e.V. The consortia are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project numbers 460248186 (PUNCH4NFDI) and 460248799 (DAPHNE4NFDI).