PUNCH4NFDI: Particles, Universe, NuClei, and Hadrons for the NFDI

PD Dr. T. Schörner[1] and Prof. Dr. M. Steinmetz[2]

[1]DESY, Notkestr. 85, 22607 Hamburg, *thomas.schoerner@desy.de*
[2]AIP, An der Sternwarte 16, 14482 Potsdam

1 BINDING LETTER OF INTENT

This is the **binding Letter of Intent** of the PUNCH4NFDI NFDI Consortium.

2 FORMAL DETAILS

Planned name of consortium
Particles, Universe, NuClei and Hadrons for the NFDI

Acronym of the planned consortium
PUNCH4NFDI

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

Spokesperson
PD Dr. Thomas Schörner, *thomas.schoerner@desy.de*

Co-applicant institutions and respective co-spokespersons
Bergische Universität Wuppertal, Prof. Dr. Christian Zeitnitz, *zeitnitz@uni-wuppertal.de*
FIAS Frankfurt, PD Dr. Andreas Redelbach, *redelbach@compeng.uni-frankfurt.de*
Forschungszentrum Jülich (FZJ), Prof. Dr. Susanne Pfalzner, *s.pfalzner@fz-juelich.de*
Friedrich-Alexander-Universität Erlangen-Nürnberg, Prof. Dr. Uli Katz, *uli.katz@physik.uni-erlangen.de*
Georg-August-Universität Göttingen, Prof. Dr. Arnulf Quadt, *Arnulf.Quadt@cern.ch*
GSI Helmholtzzentrum für Schwerionenforschung GmbH, Dr. Kilian Schwarz, *k.schwarz@gsi.de*
Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dr. Michael Bussmann, *m.bussmann@hzdr.de*
Hochschule Darmstadt, Prof. Dr. Stefan Rapp, *stefan.rapp@h-da.de*
Hochschule für Technik und Wirtschaft Berlin (HTW, Prof. Dr. Hermann Hessling, *hermann.hessling@htw-berlin.de*
Johannes Gutenberg-Universität Mainz, Prof. Dr. Volker Büscher, *buescher@uni-mainz.de*
Karlsruher Institut für Technologie (KIT), Dr. Andreas Haungs, *andreas.haungs@kit.edu*
Leibniz-Institut für Sonnenphysik (KIS), Dr. Nazaret Bello Gonzalez, *nbello@leibniz-kis.de*

Leibniz-Institut für Astrophysik Potsdam (AIP), Prof. Dr. Matthias Steinmetz, *msteinmetz@aip.de*
Ludwig-Maximilians-Universität München (LMU), Prof. Dr. Joseph Mohr, *joseph.mohr@physik.lmu.de*
Max-Planck-Institut für Kernphysik, Prof. Dr. Jim Hinton, *jim.hinton@mpi-hd.mpg.de*
Max-Planck-Institut für Radioastronomie (MPIfR), Prof. Dr. Michael Kramer, *mkramer@mpifr-bonn.mpg.de*
RWTH Aachen University, Prof. Dr. Alexander Schmidt, *alexander.schmidt@physik.rwth-aachen.de*
Rheinische Friedrich-Wilhelms-Universität Bonn, PD Dr. Philip Bechtle, *bechtle@physik.uni-bonn.de*
Ruprecht-Karls-Universität Heidelberg, Prof. Dr. Stefan Wagner, *s.wagner@lsw.uni-heidelberg.de*
Thüringer Landessternwarte Tautenburg (TLS), Dr. Matthias Hoeft, *hoeft@tls-tautenburg.de*
Technische Universität Darmstadt, Dr. Stefan Typel, *stypel@ikp.tu-darmstadt.de*
Technische Universität Dortmund, Prof. Dr. Kevin Kroeninger, *kevin.kroeninger@cern.ch*
Technische Universität Dresden, Prof. Dr. Arno Straessner, *arno.straessner@tu-dresden.de*
Technische Universität München, Dr. Philipp Eller, *ge54wir@mytum.de*
Universität Bielefeld, Prof. Dr. Dominik Schwarz, *dschwarz@physik.tu-bielefeld.de*
Universität Hamburg (UHH), Jun.-Prof. Dr. Gregor Kasieczka, *gregor.kasieczka@cern.ch*
Universität Regensburg (UR), Dr. Sara Collins, *sara.collins@ur.de*
Universität Siegen, Dr. Carmen Diez Pardos, *diez@hep.physik.uni-siegen.de*
Universität zu Köln, Dr. Jan Mayer, *jan.mayer@ikp.uni-koeln.de*
Westfälische Wilhelms-Universität Münster, Dr. Raimund Vogl, *rvogl@uni-muenster.de*

Participants

Albert-Ludwigs-Universität Freiburg, Prof. Dr. Markus Schumacher, *markus.schumacher@physik.uni-freiburg.de*
Deutsche Physikalische Gesellschaft (DPG), Dr. Georg Düchs, *duechs@dpg-physik.de*
Deutsches Luft- und Raumfahrtzentrum (DLR), Marcus Paradies, *marcus.paradies@dlr.de*
Europäisches Kernfornschungszentrum (CERN), Dr. Markus Elsing, *markus.elsing@cern.ch*
Humboldt Universität zu Berlin, Dr. Jakob Nordin, *jnordin@physik.hu-berlin.de*
Johann Wolfgang Goethe-Universität Frankfurt, Prof. Dr. Hannah Elfner, *elfner@itp.uni-frankfurt.de*
Julius-Maximilians-Universität Würzburg, Prof. Dr. Thomas Trefzger, *thomas.trefzger@uni-wuerzburg.de*
Leibniz-Rechenzentrum Garching, Dr. Stephan Hachinger, *hachinger@lrz.de*
Max Planck Computing and Data Facility, Prof. Dr. Erwin Laure, *erwin.laure@mpcdf.mpg.de*
Max-Planck-Institut für Astrophysik, Prof. Dr. Eiichiro Komatsu, *komatsu@MPA-Garching.mpg.de*
Max-Planck-Institut für Extraterrestrische Physik (MPE), Dr. Mara Salvato, *mara@mpe.mpg.de*
Max-Planck-Institut für Physik München, Dr. Oliver Schulz, *oschulz@mpp.mpg.de*
Physikalisch-Technische Bundesanstalt (PTB), Dr. Joern Stenger, *Joern.Stenger@ptb.de*
Ruhr-Universität Bochum, Prof. Dr. Hendrik Hildebrandt, *hendrik@astro.ruhr-uni-bochum.de*
TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek (TIB), Dr. Esther Tobschall, *Esther.Tobschall@tib.eu*
Universität Potsdam, Prof. Dr. Tim Dietrich, *tim.dietrich@uni-potsdam.de*
Verein für datenintensive Radioastronomie (VdR), Prof. Dr. Hermann Hessling, *hermann.hessling@htw-berlin.de*

Further partners

ALICE Collaboration
ATLAS Collaboration
Belle II Collaboration
CMS Collaboration
Cherenkov Telescope Array (CTA)
Astronomische Gesellschaft

European Space Agency (ESA)
European Southern Observatory (ESO)
GridKa
Komitee für Astroteilchenphysik (KAT)
Komitee für Elementarteilchenphysik (KET)
Komitee für Hadronen- und Kernphysik (KHuK)
LHCb Collaboration
Large Synoptic Survey Telescope (LSST)
PANDA Collaboration
Rat deutscher Sternwarten (RdS)
Square Kilometre Array (SKA)

3 OBJECTIVES, WORK PROGRAMME AND RESEARCH ENVIRONMENT

3.1 Research area of the proposed consortium

Primary: 32 Physik
Primary: 309 Teilchen, Kerne, Felder
Primary: 311 Astrophysik und Astronomie
Secondary: 308 Optik, Quantenoptik, und Physik der Atome, Moleküle und Plasmen
Secondary: 310 Statistische Physik, Weiche Materie, Biologische Physik, Nicht-lineare Dynamik
Secondary: 312 Mathematik
Secondary: 313 Atmosphären-, Meeres und Klimaforschung
Secondary: 315 Geophysik and Geodäsie
Secondary: 409 Informatik

3.2 Summary of the planned consortium's main objectives and task areas

PUNCH4NFDI leads the way towards tackling future data challenges in terms of rate, volume, complexity, reusability, and irreversibility. By its decade-long exploitation of data-intense large research infrastructures, the PUNCH4NFDI consortium has acquired ample and unique expertise in scientific computing and data management, in particular in the areas of ”big data” and ”open data” — expertise that will prove essential for the NFDI.

PUNCH4NFDI is a direct outcome of the NFDI process: The consortium was formed after intensive discussions of common and complementary strengths of two formerly separate consortia — PAHN-PaN and ASTRO@NFDI; PUNCH4NFDI thus has the NFDI ”in its DNA”. PUNCH4NFDI stands for a community of about 9.000 scientists. Its research regularly attracts immense public interest (e.g. black holes, origin of life, CERN and the ”god particle”), inspires public users and citizen scientists (e.g. Einstein@HOME), and is highly recognised (e.g. 24 Nobel prizes in PUNCH-related physics since 2001).

PUNCH4NFDI has set out to facilitate user-oriented and community-overarching data-driven research in the fields of particle, astro-, astroparticle, nuclear & hadron (PUNCH) physics, efficiently employing heterogeneous and federated computing and storage infrastructures. The prime goal of PUNCH4NFDI is to establish a decentralised community-serving science data platform obeying the FAIR data principles, and to integrate it into the NFDI.
The PUNCH4NFDI platform will feature i) tools and infrastructures to provide access to data and computing resources that together will be operated as a federated science cloud, ii) scientific tools and data services necessary to make optimal use of the data and to fully exploit their scientific potential, and iii) all necessary interfaces to allow for multi-directional use, enabling users to store and make their (meta)data, analysis chains, and results available. The PUNCH4NFDI science data platform will provide access to data and metadata. It will also bundle scientific tools, data services, advanced protocols and standards of our communities and integrate these into efficient means for scientific work for PUNCH scientists, the broader physics community, as well as the entire research landscape: i) PUNCH scientists will compare their ideas and tools and test them on their own data as well as on multi-experimental and trans-community data. A necessary precondition for this are standardised metadata and the adherence to the FAIR data principles; ii) scientists of the broader physics community can obtain a more synoptic view of data or can compare their results with other research data; iii) scientists from other communities will employ the tools and services for working on their own or on provided datasets.

PUNCH4NFDI will also lead investigations of and provide solutions for the data irreversibility challenge: the digestion of the upcoming overwhelming data streams will be hampered by resource limitations (time, power, money). These streams will have to be reduced while retaining the essential information.

In building the science data platform and tackling key data challenges, PUNCH4NFDI is working at the forefront of research and development. Members of the consortium actively participate in relevant national, European and international efforts and R&D initiatives such as ErUM-Data, ESCAPE, and EOSC. PUNCH4NFDI thus makes

the results of these efforts accessible to the entire NFDI and connects it to further national and international research infrastructures. Especially the smaller PUNCH communities will benefit from these activities. The impact of PUNCH4NFDI will be significantly leveraged by i) networking efforts across the NFDI and the continuous exploitation of synergies and by ii) intensive training, education, citizen science, and outreach measures.

The PUNCH4NFDI vision sketched above naturally leads to the following task areas (TAs):

TA 2 ”Data management” provides access to data and computing resources for the PUNCH community on a technical level. The task area develops the necessary tools (preferably based on common standards and existing solutions) for accessing and handling data in heterogeneous storage infrastructures. Additionally, the task area will establish methods that will facilitate conjoining these large-scale and to some extent federated storage infrastructures into so-called ”data lakes”. In order to permit the development of a federated science cloud, embedded in the concept of a science data platform, standardised interfaces are required as well as methods to integrate compute resources. The necessary developments can build on top of already well-established tools for distributing jobs as well as data and managing virtualised computing infrastructures.

TA 3 ”Data transformations” focuses on algorithms and methods for data analysis and exploitation that are broadly used in PUNCH. Many of these have clear applications outside their original domain, such as i) statistical tools, e.g. fitting complex models with many parameters to huge datasets in a resource-efficient way; ii) numerical methods and simulations, especially on large, heterogeneous compute grids; iii) development of robust methods for the automated design and optimisation of machine learning models; iv) joint analysis across multiple and complex datasets, allowing scientists to exploit the full potential of data by combining information from different sources.

TA 4 ”Data science portal” will provide the technology and the reference implementation for the access layer to the FAIR-compliant science data platform. This will help users from the PUNCH and NFDI communities to build sustainable decentralised science platforms for data analyses in their respective fields. The access layer will combine FAIR access to data, metadata, analysis code and analysis tools, and allows PUNCH workflows to be handled. It builds upon many developments, often already in active use, and it employs a variety of secure and standardised interfaces, environments for working with software and data, and APIs to access the platform. Such a layer requires carefully designed and developed metadata, protocols and standards.

TA 5 ”Data irreversibility” addresses upcoming challenges related to the processing and archiving of extremely large real-time data streams as they arise in the next generation of data-intense facilities, in particular the resulting impact on the scientific interpretation of the data and reproducibility of the results. Concepts and methods such as dynamic filtering processes will be developed that permit real-time data selection without human intervention. The goal is to establish dynamic archives, which also quantify the information loss and level of irreversibility and import relevant characteristics into the metadata. The algorithms, concepts, and methods developed here can also serve as blueprints for future NFDI applications in other fields and for the society at large as information technology enters the “Internet-Of-Things” era with its immense data volumes and power consumption demands.

TA 6 ”Synergies and services” targets cross-cutting activities that foster a mutual exchange of concepts and developments among the PUNCH community as well as with other consortia and the NFDI in general. Synergies are often closely related to the common use of services being provided either to subsets of the community or to the entire NFDI. Special emphasis is placed on a set of core topics (i.e. open data and metadata, big data management, and authentication and authorisation infrastructure). A marketplace will manage the exchange of concepts and solutions within PUNCH4NFDI and with other consortia.

TA 7 ”Education, training, outreach and citizen science” will train the professional physics community on data science and management methods, and educate and engage the society at large. Special focus is given to the structural challenges related to gender equality. The different target groups have a wide range of interests with diverse backgrounds, to which measures within four main areas will be tailored: i) training of experts in advanced data science methods, and career counselling; ii) support, development, and provision of training resources; iii) support of data science methods and infrastructure for public education and outreach; iv) utilisation of diverse

public computing and digital communication resources to engage citizens in active science, through citizen science initiatives (e.g. Einstein@Home, Zooniverse). TA7 is essential for the sustainability of PUNCH4NFDI.

TA 1 "Governance" bundles all necessary administrative tasks and workflows. It maintains interfaces to other consortia and to the NFDI. An *Executive Board* with representatives from the PUNCH community will be in charge of the oversight of PUNCH4NFDI. A *Management Board* will include task area and work package (WP) leaders to coordinate and monitor progress and resource allocation / usage. An external *Science Advisory Committee* of physicists and computer scientists will provide advice on ongoing and future developments that need to be addressed in the evolving NFDI. An elected *User Committee* ensures feedback and input from the user side regarding the scientific directions of the project and the services offered. An *Infrastructure Control Board* will be responsible for synchronising the top-level requirements and deliverables with the national and international data providers.

All methods and services developed in PUNCH4NFDI will be made available to the entire scientific community as plugins via the science data platform.

3.3 Proposed use of existing infrastructures, tools and services

The existing infrastructures in the PUNCH field of science fulfill four functions: data production, data analysis, long-term availability, and data sharing and publication. Together the infrastructures form the backbone of the digital landscape. All functions are accompanied by usually community-built tools and services. Increasingly often, these tools make use of the capabilities of commodity software.

On the data production side, the work in PUNCH4NFDI is mainly concentrated on ESFRI and other large research infrastructures and facilities (e.g. CERN, FAIR@GSI, ESO) and large international endeavours (nuclear and particle physics experiments, astronomy and astroparticle physics observatories, satellite missions). Furthermore, there are many smaller-scale and partly national infrastructures and experiments at university level. Data-intense theoretical computations and simulations, often carried out at high-performance computer centres, complement this picture. All in all, a very diverse array of facilities of all sizes gather scientific data that are exploited by the PUNCH community. The data volumes produced by these facilities grow faster than the IT innovation cycle (Moore's law). The necessary reduction of data (data loss, irreversibility challenge) as well as the need for cross-experiment, cross-community, cross-science access are the main drivers behind the PUNCH4NFDI efforts.

The data harvested by the data producers are processed in different ways. In high-energy physics, data are mostly digested by large data and computing centres (e.g. at DESY, GSI, FZJ, MPCDF, and KIT). With the worldwide LHC computing grid (WLCG), PUNCH physics as a leader in big data science has created a network of computing centres distributed over the entire globe. The WLCG is the largest network of such Tier centres — developed and operated by physicists, and serving of the order of 10.000 scientists. Data analysis in astronomy is so far mostly conducted on hardware operated by infrastructure providers or in medium-sized or smaller compute facilities, although future facilities will increasingly rely on large data and compute centres similar to the HEP community. In general, there is an increasing tendency to make data public after a proprietary period. With the International Virtual Observatory an environment has been established to promote interconnecting data archives and to allow reuse of data for new scientific applications. PUNCH4NFDI will cooperate with these facilities and projects, enhancing the efficient navigation of this multi-faceted landscape towards implementation of the full FAIR principles and to adapt these concepts to other science areas. The diversity of data sources, the increasing data volumes in many branches of PUNCH science, and emerging new data analysis techniques require new data analysis structures. Keywords are federated data lakes or the matching of data repositories with sufficient compute power.

Different data management and data analysis workflows apply, depending on the user community and facility. A two-layered architecture of a low-level data-storage layer and a high-level analysis layer is, however, evolving as a common model. In order to cope with the increasingly heterogeneous infrastructures within the computing centres, a third layer in-between will extend this architecture, employing virtualisation techniques ("cloud computing"). Ideally,

from the user's perspective, it is not necessary to know about the complicated details of data storage and access; these are being taken care of by tools for data storage and transfer, for data management and for job distribution. This abstraction gives users the freedom to concentrate on developing their analysis frameworks and individual analyses — a task that is again assisted by numerous community-wide tools and services.

PUNCH4NFDI will improve tools for making its data FAIR. Starting from available common formats and other standards and protocols, this requires interfaces and extensions to widely used tools that allow the collection of sufficient metadata. PUNCH4NFDI will promote public access to the data, building on existing initiatives in the research field such as CERN Open Data in high-energy physics, GAVO in astronomy, or KCDC for astroparticle physics, setting common standards and creating a common platform for the access. For published data, the use of persistent identifiers is one basic requirement. PUNCH4NFDI will work with IVOA, and TIB/DataCite, and GeRDI towards deployment and efficient use of such services, thereby joining forces with all other NFDI consortia.

PUNCH research is, to a very large extent, international, and the efforts of PUNCH4NFDI can not sensibly be considered without considering many important European and international partners and infrastructures. PUNCH4NFDI members are involved in European and global activities for data management, e.g. in EOSC or ESCAPE. The consortium will thus be able to harness international developments for the German science landscape.

3.4 Interfaces to other proposed NFDI consortia

The PUNCH4NFDI communities cooperate intensively with international partners and in international collaborations in all their activities, and not least in the field of scientific computing and data management.

PUNCH4NFDI is a merger of the former PAHN-PaN and ASTRO@NFDI consortia. The consortium was formed following the insight that both communities share numerous issues (e.g. massively growing data volumes) and at the same time have complementary strengths so that a merged consortium could optimally exploit synergies and make an almost irresistible offer to the NFDI. It is in this spirit and with this merging experience in mind that PUNCH4NFDI is pursuing collaboration with other NFDI consortia.

PUNCH4NFDI has ties to FAIRmat and DAPHNE, whose data management shows structural similarities to those at an observatory or accelerator and will, in due time, also face significant challenges. Moreover, the PUNCH communities are naturally linked to the DAPHNE consortium via the BMBF's ErUM-Data initiative.

PUNCH4NFDI fosters numerous collaborations with many different disciplines. These cooperations range from physics over informatics, mathematics, earth science and engineering to fields that develop advanced image processing technologies, e.g. for medical or biological applications, and to genetics. To give an example: NFDI4Earth has challenges in bringing together data recorded at different wavelengths that are very similar to those encountered in astronomy. However, an interdisciplinary view of datasets raises even more data-structure related issues than the cross-experiment combination of data in the PUNCH domain. NFDI provides the ideal context to jointly address such structural and other data management related issues. PUNCH4NFDI pictures itself as a driving force in this respect and will boost joint activities and an early set of actions that promote the identification of common problems. In particular, task area 6 "Synergies and Services" will implement a marketplace that will provide a platform for the exchange of services, ideas, and joint development activities between PUNCH4NFDI and other consortia of the NFDI. The following contacts and projects have already been established:

- Immediate collaboration is envisaged with the field of mathematics (i.e. the consortium MaRDI, the Mathematical Research Data Initiative). The consortia will e.g. develop common concepts for data integration and for annotation with metadata, and they will explore viable analysis methods and statistical procedures. PUNCH4NFDI also offers a variety of high-statistics datasets for extensive testing and further methodologic development.
- PUNCH4NFDI, MaRDI and other consortia have common interests in terms of scientific software. One important aspect to follow up is the question of sustainability of software: Code needs to be developed, maintained and supported by scientific software experts over long time scales.

- In the coming years, PUNCH physics will process Exabyte datasets — which requires the inclusion of opportunistic resources from the national HPC computer centres. It is planned to establish a direct communication channel to PRACE (Partnership for Advanced Computing in Europe) and to GCS (Gauss Centres for Supercomputing) and GA (Gauss Allianz) via NFDIxCS. An effective use of the HPC and HTC resources requires constant optimisation of PUNCH4NFDI applications and workflows. Therefore, the HPC subgroup of NFDI4xCS will provide the necessary logging information that cannot be collected at the application level. Conversely, PUNCH4NFDI will give feedback on the necessary data and metadata, and on the use of NFDIxCS applications. Another collaboration topic with NFDI4xCS is the definition of standards (DOI infrastructure, domain ontologies).
- The German Human Genome-Phenome Archive (GHGA) deals with rapidly increasing data volumes in genomic and medical research. GHGA aims to address a similar challenge as TA 5 of PUNCH4NFDI: analyses over long time periods require a shift in paradigm — from processor-centric to memory-based computing.
- One crucial challenge for the NFDI is the harmonisation of metadata schemes among different consortia and domains. In this respect, PUNCH4NFDI intends to strongly collaborate e.g. with the NFDI4Ing consortium and the Metadata4Ing group formed in its context, where a contact has already been established through LRZ and PTB that are involved in both initiatives. PUNCH4NFDI will also actively follow the NFDI-wide metadata workshops launched by this group.

4 CROSS-CUTTING TOPICS

4.1 Relevant cross-cutting topics

In the past decades, fundamental physics is per se increasingly engaged in new synergetic approaches. A prominent example is the field of astroparticle physics that brings together research methods from astronomy and high-energy physics, not only in the area of data management and analysis.

In order to prepare for data flows from upcoming observatories and large experimental facilities with Exabyte-scale data volumes, the PUNCH community has engaged in a considerable number of across-the field cooperations: Data management procedures for massive data rates have been addressed since the advent of the LHC and LOFAR, often by providing open software solutions to the research community. PUNCH4NFDI intends to use the opportunities provided by the NFDI to carry out joint activities, addressing the increasingly demanding requirements in data management in the forthcoming decade and sharing our experience with other researchers. Likewise, PUNCH4NFDI anticipates to profit from close interaction with other disciplines.

The idea of cross-cutting topics is to agree, within the PUNCH4NFDI consortium and with other consortia, on synergetic tasks in the pursuit of potentially generalisable solutions for upcoming challenges. With several consortia (e.g. MaRDI, NFDIxCS), first joint tasks have been identified. In any case, all connections to neighbouring NFDI consortia will serve as basic threads in the enlarged NFDI network. The exploration and exploitation of cross-cutting topics aims at designing future cross-NFDI services and benefits. PUNCH4NFDI will actively pursue, presumably together with the NFDI Directorate and other consortia, a programme of topical workshops that are open to all NFDI consortia and related initiatives.

Synergies with other NFDI consortia will be managed via an intensively moderated exchange programme, for which the PUNCH4NFDI marketplace will serve as a fruitful platform for generating software and services.

Examples for cross-cutting topics include:

- *Provenance and metadata*: While classical metadata describing static data sets are sufficient to characterise the setup of a measurement for very diverse sets of data and data generating facilities in the PUNCH community, they prove to be insufficient for optimum use throughout the whole life cycle of a dataset. Challenges emerge for the combination of simulation and observational data, or for processing of streaming data. While these challenges

arise as domain-specific problems, they are also known in other fields like geosciences or climate research. Simulations of measurements provide essential metadata for data curation and data reuse but already today these metadata may exceed the volume and complexity of the data themselves, forming new challenges to data management. Simulations of physical processes that are matched against measured data generate datasets in their own right whose metadata need to incorporate all information required to associate the simulated datasets with measured data. Metadata shall provide links to all other elements of the data life cycle that is connected to specific datasets, in particular provenance. These include documentation about the motivation of measurements (proposals), changes in the data-taking procedure as a result of changing parameters (dynamic filters), and connections between various datasets in data re-use. The connection between dataset(s) and scientific articles is also a part of data FAIRness. The development of metadata standards also includes activities e.g. in the ”Semantics working group” of IVOA.

- *Converters between internationally established data formats*: While our communities have established common data formats for very diverse sub-communities, used globally by all data providers, cooperation with other communities that have their own well established formats requires conversion tools between widely used data formats. The creation of such converters will be relevant across the entire NFDI community.
- *Data management plans, tools and policies*: PUNCH4NFDI will provide tools to assist collaborations, organisations, and projects in gathering and retaining information about the data flow throughout the lifecycle of their datasets, thus facilitating the establishment of coherent data management plans. These tools will improve the data management processes across all PUNCH communities. Special emphasize will be given to transfer the acquired knowhow and solutions to other consortia and their respective domains.
- *Legal barriers*: From former cross disciplinary and collaborative projects it is well known that an exchange of resources across borders (institutional or regional) is difficult to organise, also in many more formal aspects (legal constraints, constraints by the funding agencies, national borders). This is a topic that needs to be addressed within the framework of the NFDI.
- *Training*: A central aim of the NFDI will be to establish expertise at an advanced level not only in the development of data infrastructures, but also in training and support, from student assistants to project scientists. PUNCH4NFDI will organise training events that are tailored to the needs of the PUNCH4NFDI community.
- *Education and teaching*: The education of students and early-stage researchers is another central aim of the PUNCH4NFDI consortium. The goal is to provide and improve proficiency in NFDI-related themes and thus to enhance career prospects. PUNCH4NFDI will provide basic educational resources for university-level teaching that will also be offered to other consortia. This also calls for the integration of topics related to research data management into university curricula, preferably in form of a commonly recognised core curriculum. Such an integration will profit neighbouring, if not all, NFDI initiatives.
- *Citizen science and outreach*: PUNCH4NFDI is committed to outreach activities and will address a wide audience via science communication by promoting the NFDI initiative and the PUNCH science case in schools. In view of the gender imbalance in the natural sciences, in particular with respect to the number of first-year students, PUNCH4NFDI will take active measures to address female pupils. Furthermore, citizen science projects building on top of the science data platform will facilitate public participation in PUNCH physics.
- The *German Physical Society (DPG)* is a participant of the PUNCH4NFDI consortium and will help in providing solutions and services offered by the PUNCH community to all other communities within the entire field of physics.
- On 15 June 2020, the *Leipzig–Berlin declaration* on cross cutting topics and infrastructure development in the NFDI has been published. In this declaration relevant cross cutting topics and means to address them are being identified. PUNCH4NFDI is also supporting this activity.

4.2 Potential contributions from PUNCH4NFDI

The PUNCH4NFDI community has decade-long experience with the development and operation of distributed services to process and manage large-scale scientific data, many of which can also be applied in other domains. Members of the community are actively involved in national and international programmes that aim at providing generic services and tools for scientific computing, and they adopt them to the needs of current and future facilities, e.g. in the framework of EOSC and ESCAPE. The PUNCH developments can serve as blueprints for other communities who will face similar challenges in the future. Such services can thus constitute a major synergetic contribution of PUNCH4NFDI to the entire NFDI.

Special emphasis is given to a set of core topics: i) open data and metadata; ii) data curation and archival access; iii) the development and deployment of corresponding standards; iv) tools and methods for data-intense work and for the provision of large-scale distributed computing as a service, including visualisation, statistical treatment, ontology, and outreach in data science, as well as an overarching authentication and authorisation infrastructure.

In many of the aforementioned activities, cooperations with other disciplines — and thus to other NFDI consortia — have emerged, in particular in the fields of natural sciences, engineering and life sciences. Specific examples include:

- *Open data* is a well-developed approach in astrophysics that is of interest for the NFDI community at large. Reuse of data — possibly after a proprietary period — generates additional science products and allows to address new scientific questions. Various modes of data access need to be established, in particular during the proprietary period. These modes may mirror data access management problems in many other fields, e.g. those that deal with personalised or commercial data resulting in restricted data access rights. Open data thus strongly connects to the FAIR principles, which opens the possibility for other NFDI consortia to test their algorithms on the wealth of available PUNCH4NFDI data.
- A common *authorisation and authentication infrastructure (AAI)* will be essential for the NFDI. PUNCH4NFDI will contribute to the set up of an NFDI-wide infrastructure.
- *Metadata* are key to re-usability and of fundamental importance for the NFDI at large. PUNCH4NFDI offers expertise in the development of metadata standards, in the curation of datasets, and in quality control. The challenge is the provision of format interfaces between different highly developed data and metadata formats that have emerged as standards in internationally closely connected research fields. Benefitting from the intensive groundwork in internationally coordinated and established processes, PUNCH4NFDI can take a leading role in many aspects of metadata and provide solutions for other partners in the NFDI.
- *Big data management and distributed computing* is the daily business in PUNCH. Corresponding tools, preferably based on open standards, will be offered by PUNCH4NFDI. Using these methods, the data and computing centres connected to the consortium will be combined into a federated science cloud. This will serve as an important starting point for corresponding services for the entire NFDI.
- *On-the-fly reduction* of Exabyte-scale data streams results in data irreversibility and data loss. This irreversibility will challenge many data-intense branches of science and technology. The optimum choice of ”data acquisition” settings requires ”on-the-fly” decisions in the feedback loops between data gathering and storing (dynamical archives). PUNCH4NFDI will contribute key technical, methodical, and algorithmic expertise to this branch of big data management.
- *Connecting data and simulation* for the purposes of data calibration and curation is an essential task to which PUNCH4NFDI will contribute decisive data management procedures.
- *Deep learning methods* are nowadays commonly used for science analysis in PUNCH physics. The consortium will actively share its expertise in AI application with the entire NFDI.