Proposal for a consortium in the DFG’s *Nationale Forschungsdaten-Infrastruktur*

PUNCH4NFDI Consortium Proposal

The PUNCH4NFDI Consortium —
Particles, Universe, NuClei & Hadrons for the NFDI

Applicant:
Deutsches Elektronen-Synchrotron

Co-applicants:
Forschungszentrum Jülich, Frankfurt Institute for Advanced Studies, Georg-August-Universität Göttingen, GSI Helmholtzzentrum für Schwerionenforschung GmbH, Hochschule für Technik und Wirtschaft Berlin, Johannes-Gutenberg-Universität Mainz, Karlsruher Institut für Technologie, Leibniz-Institut für Astrophysik, Ludwig-Maximilians-Universität München, Max-Planck-Institut für Radioastronomie, Max-Planck-Institut für Kernphysik, Rheinische Friedrich-Wilhelms-Universität Bonn, Ruprecht-Karls-Universität Heidelberg, Thüringer Landessternwarte, Technische Universität Dresden, Technische Universität Dortmund, Universität Bielefeld, Universität Hamburg, Universität Regensburg

Participants:
Albert-Ludwigs-Universität Freiburg, Europäisches Kernforschungszentrum CERN, Deutsches Luft- und Raumfahrtzentrum, Deutsche Physikalische Gesellschaft, Helmholtz-Zentrum Dresden-Rossendorf, Hochschule Darmstadt, Humboldt-Universität Berlin, Johann Wolfgang Goethe-Universität Frankfurt, Leibniz-Institut für Sonnenphysik, Leibniz-Rechenzentrum, Max Planck Computing and Data Facility, Physikalisch-Technische Bundesanstalt, Ruhr-Universität Bochum, RWTH Aachen University, Technische Informationsbibliothek - Leibniz Information Centre for Science and Technology, Technische Universität Darmstadt, Technische Universität München, Universität Potsdam, Universität Siegen, Universität zu Köln, Verein für datenintensive Radioastronomie e.V., Westfälische Wilhelms-Universität Münster

Contents

1 General Information

Name of consortium [English]

Particles, Universe, NuClei and Hadrons for the NFDI

Name of consortium [German]

Teilchen, Universum, Kerne und Hadronen für die NFDI

Summary of the proposal [English]

PUNCH4NFDI, the consortium of **particle, astroparticle, astro-, hadron and nuclear physics**, covers a substantial fraction of curiosity-driven basic research in physics, and in particular of **data-intense physics at large facilities**. PUNCH4NFDI is a merger of Astro@NFDI and PAHN-PaN. As such, the consortium in its forming process has already gone through a very productive transformative process that might prove representative for the upcoming challenges of the entire NFDI: finding the commonalities in the relevant use cases, agreeing on a common language and methodical approach, and defining and delivering solutions that serve a broader audience – ideally the entire NFDI. The PUNCH community has always been at the forefront of technological developments. In particular, it is a **leading force in "big data" and "open data"** in scientific data management and an avid pursuer and early adopter of the FAIR principles.

Due to new data producers PUNCH science is facing challenges at an entirely new level in terms of data volumes, data complexity, data rates, and data irreversibility. The goal of PUNCH4NFDI is to develop and share solutions for these **data challenges** — which other science fields will also be confronted with in the future. Thus, PUNCH4NFDI already lives FAIR. The main product of PUNCH4NFDI will be a **science data platform** that serves the PUNCH community and the entire NFDI. Technically, the platform will consist of a data lake with storage and cloud computing systems, a data transformation layer allowing the analysis of the data, and a user-friendly interface functioning as a data portal.

PUNCH4NFDI will achieve its objectives through an ambitious work programme. Several **task areas** (TA) address the various aspects of such an advanced layered data management model: TA 2 "Data management" provides solutions for standardised data access and inter-operable storage solutions; it addresses the integration of storage with federated compute resources, and it deals with dynamic and intelligent data handling for advanced workflows. TA 3 deals with the "Data transformations" necessary for the maximum exploitation of scientific data, for the combination of different datasets, and for achieving higher levels of abstraction and thus new scientific insights. TA 4 will provide the "Data portal" which gives access to the underlying knowledge structure. Using the appropriate metadata, it connects the elements of interlinked digital research products that are central elements of the PUNCH knowledge fabric. TA 5 addresses the increasingly important issue of "Data irreversibility" and the problems of data loss, the necessary "real-time" decisions and dynamical archiving capabilities. TA 2–5 are complemented by the TA 6 and 7 on "Synergies & services" and "Training, education, outreach, and citizen science" that address topics of relevance for the entire consortium and for its connections to other consortia, to the entire NFDI, and the public at large.

Summary of the proposal [German]

PUNCH4NFDI, das Konsortium von Teilchenphysik, Astroteilchenphysik, Astrophysik sowie Hadronen- und Kernphysik, steht für einen bedeutsamen Teil erkenntnisgetriebener physikalischer Grundlagenforschung und insbesondere für datenintensive Forschung an großen Forschungsinfrastrukturen. PUNCH4NFDI — hervorgegangen aus Astro@NFDI und PAHN-PaN — ist im Zuge der Fusion durch einen transformativen Prozess gegangen, der beispielhaft für die gesamte NFDI ist: Gemeinsamkeiten in den relevanten *use cases* erarbeiten, eine gemeinsame Sprache sowie einen einheitlichen methodischen Zugang finden sowie Lösungen definieren und bereitstellen, die einem breiten Publikum dienen — idealerweise der gesamten NFDI. Die PUNCH-Community war stets ein technologischer Vorreiter. Insbesondere ist sie im Forschungsdaten-Management führend in den Bereichen ”big data” und ”open data” und hat sich die Umsetzung der FAIR-Prinzipien schon früh auf die Fahnen geschrieben

Neue *data producer* in PUNCH stellen die Community vor ganz neue Herausforderungen in Bezug auf Volumen, Komplexität und Rate sowie die ”Irreversibilität” der erzeugten Daten. Ziel von PUNCH4NFDI ist es, Lösungen für diese Herausforderungen — denen auch andere Wissenschaftsbereiche in der Zukunft begegnen werden — im Lichte der FAIR-Prinzipien zu entwickeln und bereitzustellen. Kernstück der Arbeit von PUNCH4NFDI wird die ***science data platform*** sein, die für PUNCH und die gesamte NFDI zur Verfügung stehen wird. Technisch wird die Plattform über einen ”data lake” mit Speicher- und cloud-basierten Rechenressourcen realisiert; dazu kommen eine Datentransformations-Schicht für die Analyse von Datensätzen sowie eine leicht zu bedienende User-Schnittstelle als Datenportal.

PUNCH4NFDI wird die Realisierung seiner ambitionierten Ziele in verschiedenen ”task areas” angehen, die sich auf die verschiedenen Aspekte eines solchen in Schichten organisierten Datenmanagements beziehen: TA 2 ”Data management” bietet Lösungen für standardisierten Datenzugriff und interoperable Speichermethoden und arbeitet an der Integration von Datenspeichern mit föderierten Rechenressourcen und den entsprechenden dynamischen ”work flows” für die Datenverarbeitung. TA 3 beschäftigt sich mit den ”Data transformations”, die für die optimale wissenschaftliche Ausnutzung von Daten und für die Kombination verschiedener Datensätze (und damit für neue wissenschaftliche Erkenntnisse) relevant sind. TA 4 stellt das benötigte ”Data portal” zur Verfügung und verbindet die zugrundeliegenden digitalen ”research products” mithilfe geeigneter Metadaten. TA 5 behandelt die an Bedeutung gewinnende Herausforderung von Datenirreversibilität und Datenverlust sowie die entsprechenden Echtzeit-Entscheidungen und dynamischen Archivierungen. Die TAs 6 und 7 bieten ”Synergies & services” sowie ”Training, education, outreach & citizen science”, die für PUNCH4NFDI und seine Verbindungen zu anderen Konsortien wie auch für die NFDI und die Gesellschaft insgesamt von Bedeutung sind.

Applicant institution

Applicant institution	Location
Deutsches Elektronen-Synchrotron (DESY)	Hamburg

Name of the consortium spokesperson

Spokesperson	Institution, location
PD Dr. Thomas Schörner-Sadenius	DESY, Hamburg

Co-applicant institutions

Acronym	Co-applicant institution	Location
AIP	Leibniz-Institut für Astrophysik	Potsdam
FIAS	Frankfurt Institute for Advanced Studies	Frankfurt
FZJ	Forschungszentrum Jülich	Jülich
GAU	Georg-August-Universität Göttingen	Göttingen
GSI	Helmholtzzentrum für Schwerionenforschung GmbH	Darmstadt
HTW	Hochschule für Technik und Wirtschaft Berlin	Berlin
JGU	Johannes-Gutenberg-Universität Mainz	Mainz
KIT	Karlsruher Institut für Technologie	Karlsruhe
LMU	Ludwig-Maximilians-Universität München	München
MPIfR	Max-Planck-Institut für Radioastronomie	Bonn
MPIK	Max-Planck-Institut für Kernphysik	Heidelberg
TLS	Thüringer Landessternwarte	Tautenburg
TUDD	Technische Universität Dresden	Dresden
TUDO	Technische Universität Dortmund	Dortmund
UB	Universität Bielefeld	Bielefeld
UoB	Rheinische Friedrich-Wilhelms-Universität Bonn	Bonn
UHD	Ruprecht-Karls-Universität Heidelberg	Heidelberg
UHH	Universität Hamburg	Hamburg
UR	Universität Regensburg	Regensburg

Co-spokespersons[1]

Co-spokesperson	Institution, location	Task area(s)
PD Dr. Philip Bechtle	Rheinische Friedrich-Wilhelms-Universität Bonn	4 (2)
Prof. Dr. Volker Büscher	Johannes-Gutenberg-Universität Mainz	(5)
PD Dr. Sara Collins	Universität Regensburg	(2, 4)
Dr. Andreas Haungs	Karlsruher Institut für Technologie	2, 4 (6, 7)
Prof. Dr. Hermann Heßling	Hochschule für Technik und Wirtschaft Berlin	5, 6
Prof. Dr. Jim Hinton	Max-Planck-Institut für Kernphysik, Heidelberg	6 (3)
Dr. Matthias Hoeft	Thüringer Landessternwarte, Tautenburg	2
JProf. Dr. Gregor Kasieczka	Universität Hamburg	3 (5)
Prof. Dr. Michael Kramer	Max-Planck-Institut für Radioastronomie, Bonn	5, 7 (6)
Prof. Dr. Kevin Kröninger	Technische Universität Dortmund	7 (3)
Prof. Dr. Joseph Mohr	Ludwig-Maximilians-Universität München	3 (6)

[1] Shown in the third column are the TAs with the largest contribution by the co-spokesperson, i.e. TAs lead by him/her, or TAs in which the co-spokesperson is leading a work package. Further contributions of the co-spokespersons to other TAs are given in brackets.

Co-spokespersons (continued)

Co-spokesperson	Institution, location	Task area(s)
Prof. Dr. Susanne Pfalzner	Forschungszentrum Jülich	3 (2,5,6,7)
Prof. Dr. Arnulf Quadt	Georg-August-Universität Göttingen	6, 7
PD Dr. Andreas Redelbach	Frankfurt Institute for Advanced Studies	5
Prof. Dr. Dominik Schwarz	Universität Bielefeld	2, 5 (6, 7)
Dr. Kilian Schwarz	GSI Helmholtzzentrum für Schwerionenforschung GmbH, Darmstadt	2, 4, 6 (7)
Prof. Dr. Matthias Steinmetz	Leibniz-Institut für Astrophysik, Potsdam	1, 4, 6
Prof. Dr. Arno Straessner	Technische Universität Dresden	(5)
Prof. Dr. Stefan Wagner	Rupprecht-Karls-Universität Heidelberg	6, 5 (4)

Participants

Acronym	Participating institution	Location
ALU	Albert-Ludwigs-Universität Freiburg	Freiburg
CERN	Europäisches Kernfornschungszentrum	Geneva, CH
DLR-DW	Deutsches Luft- und Raumfahrtzentrum	Jena
DPG	Deutsche Physikalische Gesellschaft	Bad Honnef
GU	Johann Wolfgang Goethe-Universität Frankfurt	Frankfurt
HDA	Hochschule Darmstadt	Darmstadt
HUB	Humboldt-Universität zu Berlin	Berlin
HZDR	Helmholtz-Zentrum Dresden-Rossendorf	Dresden
KIS	Leibniz-Institut für Sonnenphysik	Freiburg
LRZ	Leibniz-Rechenzentrum	Garching
MPCDF	Max Planck Computing and Data Facility	Garching
PTB	Physikalisch-Technische Bundesanstalt	Braunschweig
RUB	Ruhr-Universität Bochum	Bochum
RWTH	RWTH Aachen University	Aachen
TIB	Technische Informationsbiliothek – Leibniz Information Centre for Science and Technology	Braunschweig
TUDa	Technische Universität Darmstadt	Darmstadt
TUM	Technische Universität München	München
UP	Universität Potsdam	Potsdam
USi	Universität Siegen	Siegen
UzK	Universität zu Köln	Köln
VdR	Verein für datenintensive Radioastronomie e.V.	Jena
WWU	Westfälische Wilhelms-Universität Münster	Münster

Summary of contributions of participants

ALU will contribute to the TA 2 work package (WP) ”Compute4PUNCH” with its expertise to develop the operation, use, and integration of heterogeneous computing resources in particular in the area of monitoring tools. It will support and contribute to TA 7 in the work package ”Development of standardised curricula”. It will provide temporary access to parts of the Worldwide LHC Computing Grid (WLCG) Tier/23 cluster ATLAS-BFG, the High Performance Computing (HPC) cluster NEMO, and the storage system bwSFS.

CERN will contribute via a collaboration on the developments in TA 2 (Data management) and by offering their experience from the open-source project Rucio for large scale scientific data management. CERN will also collaborate on the developments in TA 3 (Data transformation) on the support of heterogeneous resources.

DPG will contribute to TA 6 and TA 7 and foster exchange of concepts, services and activities within the consortium and with other consortia. DPG will try to raise the awareness of proper data management through talks at DPG Spring Meetings, meetings on data management, seminars and mini-symposia of the Young DPG (jDPG), in the working group on ”Information” and in other divisions or working groups. NFDI-related articles in the Physik Journal, a white paper on research data management in the physics curricula and other publications and activities will also raise awareness for proper data-management beyond the scientific community.

DLR-DW will contribute in TA 2 with its experience in data management for data-intensive sciences, such as earth observation and radio astronomy. DLR-DW has gained expertise in the monitoring and performance analysis of large-scale storage systems, ranging from complex tiered storage architectures to fine-grained I/O analyses of data analysis pipelines on modern storage technologies, such as NVMe SSDs. DLR-DW plans to contribute expertise and in-house software tools to evaluate the performance of storage systems.

GU contributes with the expertise on large-scale complex numerical calculations making use of Monte Carlo algorithms and of relativistic hydrodynamics and magnetohydrodynamics in astrophysics and nuclear physics. The know-how on open-source development, work-flows to compare theoretical calculations to experimental data and statistical methods is important for TA 3. GU Frankfurt offers video seminars and a pedagogical talk series on software development to be integrated in TA 7. In TA 7 scientific visualisations of heavy-ion collision and neutron star merger simulations will be provided.

HDA will contribute to education particularly in data science and machine learning (ML) topics (TA 3 and TA 7). It intends to organise joint interdisciplinary workshops and do co-teaching with students of computer science and physics on real experiment data. The HDA will bring in a computer science perspective into the consortium.

HUB will lead WP 3 of TA 5. ”Dynamic archives” are needed in both particle physics and astrophysics to account for missing data that appear as a subset of measurements. This work will extend ongoing projects at HUB which focus on joint analysis of optical, gamma-ray, neutrino and gravitational-wave data streams. Dynamic archives combine these data sources, and will allow searches for anomalous signals which, in turn, update the future behaviour of active real-time programs.

HZDR will contribute in the following places: CASUS/HZDR will provide expertise in many-core programming, performance portable programming for data-intense applications and high performance computing in the field of astrophysics (TA 3).

KIS is currently undertaking strong efforts to further develop its Science Data Centre aiming at becoming a European reference for curation, analysis and access to solar data and a forum for solar-science discussions. KIS will contribute to PUNCH4NFDI in the following places: i) large versioned, tiered data-storage for online archive and data-processing of solar observational data and the production of high-level data products therefrom (TA 2), ii) a compute cluster (1000 cores) and a GPU server for producing high-level data products (TA 4), iii) software, workflows and data pipelines for automated image processing and data reduction (TA 3), iv) data from the German solar observatories and DKIST/NSO data from 2021 (adhering to the NSO data policy). KIS will bring in into the project 1 FTE on TA 2 and TA 4 for metadata development, data curation and dissemination of solar data to the broader community, and 0,5 FTE in TA 3 for high-level solar data products accessible to the broader community.

LRZ will contribute with its expertise in infrastructure, astrophysics simulations, and research data management (RDM). LZR will provide support in application support services (AstroLab), RDM counselling, and services for making large datasets FAIR. LRZ will contribute interface specifications and software components for data display, metadata databases, or dissemination via OAI-PMH, and experience in interface development (WP 2.1, WP 4.2/3). The contribution in TA3.2 is with services like high-level support and code optimisation. LRZ is involved in RDA and other NFDI consortia, and will act as bridge to harmonise RDM policies (TA6).

MPCDF is offering access to the data centre infrastructure, compute resources, and data management skills in close collaboration with the PUNCH4NFDI partners from the Max Planck Society, including those for whom data repositories are or will be hosted at MPCDF (TA5).

PTB will contribute to: (i) semantic representation of numeric factual data by means of controlled vocabularies and matching metadata modules. These will have to be internationally understandable. PTB will use its international contacts, especially to other national metrology institutes; (ii) development of methods for quality assessment of numeric factual data and the corresponding metadata. This includes the assessment and certification of algorithms acting on these data. Harmonising metadata systems, and the development of ontologies for numerical factual data are core topics to be raised with the PUNCH community. PTB will contribute to WP 4.1, 4.2, 6.1, 6.3, 6.4.

RUB will contribute to big data analysis of astronomical survey data (Low Frequency Array (LOFAR), MeerKAT, IceCube, Chrenkov Teleskope Array (CTA), Fermi-LAT, Euclid, Rubin Observatory)) and hadron-physics experiments (BESIII, PANDA) in TA 2; software development for analysis pipelines as well as testing of algorithms with data, and ML applications (classification of astronomical time series and spectra, sky images, photometric redshifts, simulation software, astro-particle theory, hadron spectroscopy, partial-wave analysis with coupled channels) in TA 3 and 6.

RWTH will contribute to TA 2 and TA 4. For TA 2, RWTH is experienced in providing large storage volumes accessible remotely. The focus will be on adaptations and to provide a test facility for standardised data access protocols or interfaces to access data transparently. For TA 4, RWTH will engage with dedicated man-power in a future Gravitational Wave Telescope computing and intends to implement the FAIR principles in the conception of large computing projects.

TIB As a national and international service provider, TIB will contribute its long-time expertise in operating research data services along the life cycle of research data, like the PID-Services DataCite (Digital Object Identifier, DOI) and ORCID. TIB is working on terminology services and automatic semantic annotation of data and will contribute to TA 4 and TA 6. The contributions will come from the existing resources.

TUDa runs a dedicated storage system for research data produced within the Collaborative Research Center 1245. These data are of different origin, type, and size and come from both theory and experiment, and here in particular from the electron accelerator S-DALINAC. They will be used in a demonstrator project to apply and test the tools developed, the design and implementation of digital dynamic research products and their Catalogue, as part of TA 4.

TUM will contribute with its expertise in statistics for data analysis. In particular, TUM will participate with members of the recently founded ”Origins Data Science Laboratory”. The TUM will contribute to PUNCH4NFDI in WP 1 of TA 3 to develop new and parallelised algorithms for MCMC sampling. The contributions will include 0.25 FTE/year for the described work.

UzK will contribute to the development of research products, interfaces, and the data portal of TA 4. In addition, datasets and workflows from small scale nuclear astrophysics and nuclear structure experiments and simulations will be provided; to be published on the PUNCH4NFDI platform early as test and demonstration material.

UP will contribute to TA 4 focusing on design and implementation of an open research catalogue. Based on experience in hosting data from computational astrophysics simulations through the Computational Relativity Collaboration and in the data infrastructure of LIGO and Virgo, UP will support the collection of distributed digital research objects such as open-source databases and research papers in the fields mentioned above.

USi will contribute to the area of automated machine learning processes, especially in the optimisation of the performance and the implementation of machine learning in large data sets, as part of the machine learning WP in TA 3. It will collaborate within TA 7, focusing on the training of physicists in concepts and methods related to data science.

Members of **VdR** are engaged as co-applicants in various TAs. Members of VdR not active as co-applicants will participate by providing support as testers for the quality and the usability of the outcome of PUNCH4NFDI in particular for TA 5 and TA 6.

WWU will provide cloud resources in the WWU Cloud (400 cores, 400 TB) to be used as testbed for the data lake development in TA 2 and to host TA 6 services. WWU will provide support with experience operating Kubernetes in an OpenStack multi cloud environment and operating JupyerHub. WWU will share experience and code for creating Jupyter notebook images with X11/GPU support.

Partners[2]
ALICE Collaboration (ALICE)
ATLAS Collaboration (ATLAS)
Belle II Collaboration (Belle II)
Bergische Universität Wuppertal
CBM Collaboration (CBM)
CMS Collaboration (CMS)

[2] See `https://www.punch4nfdi.de/e113062/e113456/` for support statements from selected partners.

Cherenkov Telescope Array (CTA)
Astronomische Gesellschaft (AG)
European Space Agency (ESA)
European Southern Observatory (ESO)
Grid Computing Centre Karlsruhe (GridKa)
International LOFAR Telescope (ILT)
Komitee für Astroteilchenphysik (KAT)
Komitee für Elementarteilchenphysik (KET)
Komitee für Hadronen- und Kernphysik (KHuK)
LHCb Collaboration
Rubin Observatory Rubin Telescope
Max-Planck-Institut für Astrophysik (MPA), Garching
Max-Planck-Institut für extraterrestrische Physik (MPE), Garching
Max-Planck-Institut für Physik (MPP), München
PANDA Collaboration
Rat deutscher Sternwarten (RdS)
Square Kilometre Array (SKA)

Names and numbers of the DFG review boards that reflect the subject orientation of the proposed consortium
Primary: 32 Physics
Primary: 309 Particles, Nuclei, and Fields
Primary: 311 Astrophysics and Astronomy
Secondary: 308 Optics, Quantum Optics, and Physics of Atoms, Molecules, and Plasmas
Secondary: 310 Statistical Physics, Soft Matter, Biological Physics, Non-linear Dynamics
Secondary: 312 Mathematics
Secondary: 313 Atmospheric Science, Oceanography, and Climate Research
Secondary: 315 Geology and Geodesy
Secondary: 409 Computer Science

2 Scope and objectives

2.1 Research domains or methods addressed by the consortium, specific aim(s)

Research domains of the PUNCH4NFDI *consortium*

The PUNCH4NFDI consortium covers the German community of **particle, astro-, astroparticle, nuclear, and hadron physicists**[3]. The aim of PUNCH physics is to identify the elementary building blocks of matter, the laws governing their interactions, and their impact on the development of the universe and the formation of structures in it.

With about 9,000 scientists with a Ph.D. in Germany, PUNCH physics represents a **significant share of all physics efforts** in Germany, and worldwide. The structures, services, and tools to which PUNCH4NFDI scientists contribute are **genuinely global** and serve several 10.000 scientists. PUNCH4NFDI members are connected to all relevant networks, initiatives and R&D efforts in the field and thus have access to the ”latest and greatest” of all developments. **Collaboration and division of labour** are virtues practised by PUNCH scientists since decades.

PUNCH physics is a **pioneer of data-intense science (”big data”)**. A prime example is the WLCG that serves the LHC experiments at CERN, Geneva. The **concept of ”open data”** has been pursued by PUNCH science since long — as witnessed by the Sloan Digital Sky Survey (SDSS) pioneering the use of relational database technology in astronomy for the Virtual Observatory (VO) showcasing a federated information infrastructure). **Citizen science** methods developed by the PUNCH community (”SETI@HOME”, ”EINSTEIN@HOME”, ”Folding@home” ...) meanwhile have propagated into other fields of science (e.g., GalaxyZoo ⇒ Zooniverse). The PUNCH community has ample experience in the **design, development, and operation of infrastructures** (e.g. WLCG), methods, tools, and services for all levels of data management.

Research methods of the PUNCH4NFDI *consortium*

PUNCH research is, to a large extent, carried out at **large and international facilities**: accelerators and detectors of particle and hadron & nuclear physics, observatories, and telescopes.

PUNCH research is **empirical, data-driven observational science**, involving among other things the observation of the sky, the detection of particle showers in the atmosphere, and the measurement of particle collisions in accelerators. These experiments are accompanied by high-precision **theory predictions, model calculations, and phenomenological simulations** of the relevant processes and their complex interplay. In PUNCH research, many **synergies** between different sub-fields, facilities, methods, and between experiment and theory are exploited, emphasising the **necessity of accessible and interoperable data**. Examples are the comparison of measurements with theory predictions, the combination of data on dark matter from various experiments in particle and astroparticle physics, or the multi-messenger approach in astro- and astroparticle physics. In PUNCH work, methods of machine learning (ML) are not only widely employed but adapted and further developed.

To a large extent, PUNCH science is **discovery science**: Many findings in the field have led to

[3] PUNCH— the ”u” signifies ”universe” and reflects the contributions from astro- and astroparticle physics.

fundamentally new insights and changes of paradigm, and consequently have been rewarded with Nobel prizes (24 since 2001). Similar discoveries can be expected for the future: The questions that PUNCH pursues address the most fundamental topics in the natural science, and consequently many of the current experiments and theoretical investigations might easily lead to game-changing insights. Given this potential impact of PUNCH research, especially the **re-usable character of PUNCH data** is of high importance: Data taken by PUNCH instruments today may well contain information the scientific value of which becomes evident only tomorrow and **may require their re-analysis** and also their combination with other information.

Large parts of PUNCH research in Germany and globally are pursued as **collaborative work in large international organisations** and jointly organised and operated facilities. However, **also ”niche science” and smaller activities** can contribute essential insights, and often have done so. Their integration into the overall landscape, in particular, requires **accessible data and clear interfaces** and may well exploit (so far) hidden potentials.

Besides small-scale and local facilities, data analysis and management in the PUNCH field rely heavily on **large distributed computing and storage centres** (”Tier” centres); the already mentioned WLCG, but also the Virtual Observatory (VO) are important examples. This fact implies a very **tight connection between PUNCH4NFDI members, data providers, and data managers**. It ensures i) technical developments along the needs of the users; ii) a feedback loop between providers, managers, and users; iii) up-to-date information of users; and iv) an optimal basis for PUNCH4NFDI to act as facilitator and multiplicator towards the entire German science system.

PUNCH **data are partly proprietary**, and their access and use is partly limited to members of the data-producing collaborations. More and more data are, however, already **completely open or in the process of becoming so**. In PUNCH science, ”data” (and metadata) exist at very different levels of abstraction: raw collision data in custom formats from particle physics experiments, surveys and monitoring of the sky in astrophysics, reconstructed and calibrated data of heavy-ion collisions in hadron & nuclear physics, cross-section measurements involving theoretical input, to name just a few. For each level, the application of the FAIR principles needs to be verified and extended. To work towards this goal is a central objective of PUNCH4NFDI.

It is characteristic of PUNCH data that they are **without direct commercial value**. Besides, there are **little if any privacy concerns** connected to the data. This makes PUNCH data management and usage potentially simpler than in many other fields of science and opens the possibility to design example solutions with low-threshold access that are easy to try by and port to other science communities. PUNCH4NFDI may therefore **serve as a development platform for data science** in the entire German (and international) landscape.

Specific aims of the PUNCH4NFDI *consortium*

The PUNCH4NFDI consortium pursues several high-level aims related to the NFDI purposes:

1. PUNCH4NFDI aims at **enabling future PUNCH science**: Research in the PUNCH community is confronted with numerous “data challenges” — in terms of data volume, rate, and

complexity — and PUNCH4NFDI has set out to master these challenges.

2. PUNCH4NFDI will **strengthen the implementation of the FAIR principles** in its community, in particular the "open data" approach. The central building block is the setup of the **PUNCH science data platform** (PUNCH-SDP, see objective 1 below). This platform can act as a model for other branches of science, and PUNCH4NFDI — in many ways at the forefront of research in scientific data management — the tool to distribute the community-specific knowledge within the PUNCH sciences and to **share the expertise with the wider science community**. PUNCH4NFDI is dedicated to drive the relevant communication processes within the NFDI. At the same time, PUNCH4NFDI is eager to profit from the experiences of other communities and consortia. Detailed examples of concrete collaborations with other consortia are discussed in section 4 and within the task areas, section 5.
3. PUNCH4NFDI addresses the **"data irreversibility" challenge**, i.e. the fact that in the future only a tiny fraction of the data taken by experiments can be stored in long-term archives and that decisions on what to keep are necessarily based on incomplete information. Generic methods based on ML will be developed that allow to determine and to minimise the loss of information, and to identify in real-time rare anomalies in huge data streams (as possible hints towards new physical effects, see section 5.5). Drastic data reduction by intelligent methods will be indispensable in **green computing** to cope with future power demands.
4. The PUNCH-SDP will also provide **better documentation of analysis workflows**, and it will be an essential step towards enabling **long-term data and workflow preservation**. It will thus improve data quality and the quality of scientific results following the open data paradigm.
5. The PUNCH-SDP and the tools and services necessary for its realisation will facilitate **cross-experiment and cross-community analysis of data** — an important goal of PUNCH science. This furthers the development of science practices that let researchers confidently work with research data "blindly" obtained from a large number of diverse sources, showcasing the potential of this still emerging paradigm.
6. The tools and services that PUNCH4NFDI will develop will **enable "live analysis" and "live peer review"**: Using PUNCH-SDP, in the future, a research product would be associated with the paper under review, and the reviewer can now cross-check the complete implementation of the workflow, wherever needed. This would qualitatively improve the thoroughness of scientific review considerably, and bring it from indirect inference to an interactive learning experience. Conversely, reviewer suggestions on the analysis workflow or the extraction of results can directly be included in improved results.
7. A stated goal when developing metadata schemes is to enable **blind discovery** — that is the discovery of objects or physical effects in data not based on prior knowledge, but purely based on properties of the data.
8. The NFDI will **bundle German efforts** in research data management, and PUNCH4NFDI is eager to harness the ensuing achievements in order to strengthen the German contributions to international endeavours by implementing forefront data management methods.

2.2 Objectives and measuring success

The PUNCH4NFDI consortium concentrates its efforts to several **key objectives** detailed the following. The work programme of all PUNCH4NFDI task areas (TA) and work packages (WP) (section 5) are referring back to these objectives that address and make more concrete the specific aims of the consortium spelt out above (section 2.1). The objectives are also a response to essential needs transpiring from the multitude of use cases (see section 4.1).

Objective 1: Setup of a science data platform (PUNCH-SDP)

The overarching objective of PUNCH4NFDI is the **setup of a PUNCH science data platform** (PUNCH-SDP) serving the PUNCH community, the NFDI, and science and the public at large. The platform will be able to deal with Exabyte data sets and offer the necessary modern tools to handle them efficiently. The PUNCH community is well-known, since decades, for preparing **cutting-edge research data management** methods and for handling **complex workflows for many thousands** of scientists. Now it is time to bundle all these single, often ”island-like” solutions — that often live only inside collaborations — and make them FAIR.

The proposed PUNCH-SDP concept advocates generalising all the various types of **digital objects into a unified abstract scheme of digital research products**. Raw data, metadata, code, graphics, tables, papers — they all shall be treated on equal footing as interlinked research products. The goal of the PUNCH-SDP is to make the relations between these products — and the continuous growth and refinement of knowledge from their interaction — visible, discoverable, and expandable. Of particular interest will, therefore, be the use of links between two or more independent datasets — this will become an act of new research that can easily be published using the established knowledge network. As a unique feature of this knowledge fabric woven by the platform, this linking mechanism also offers the possibility to pursue a research programme further using the bundle of data, parameters, software, analysis tools, paper — the **platform will enable ”live peer review” as well as ”live” and ”continuous” analysis**.

Technically, the PUNCH-SDP will consist of a data lake with storage and cloud computing systems (TA 2), a data transformation layer allowing the data processing and transformations (TA 3), and a user-friendly interface and data portal (T 4). The entire system also needs sophisticated and FAIR data management (addressed in several TAs). In short, the input from all TAs defined later in section 5 — and a whole slew of individual technical and methodical developments — is required to accomplish the platform, which can be considered the ”gist” of the use cases that PUNCH4NFDI has collected and aggregated (see discussion in section 4.1).

The success of the PUNCH-SDP will be measured by the number of users of the platform and their degree of satisfaction with the services provided.

Objective 2: Description of data, metadata, persistent identifiers, analysis workflows and code, simulation, and knowledge

A vital element of the PUNCH-SDP described above is the ability to store, download, run, modify, and re-upload analysis chains, including data, metadata, and simulation and analysis code, including references to statistical tools provided on the PUNCH platform. In order to enable the user to run, modify, and combine different analyses on the platform, a **metadata description of the analysis workflow** will be developed. It will contain information about the abstraction level of the data and simulation, about the analysis steps performed and the necessary software tools, and about the format and content of the analysis output. This new level of metadata description will allow to perform different analyses independently or in a consistently combined way on the same platform, even if they start from different data formats and data abstractions.

An easy-to-use and transparent interface for the user should facilitate finding and using the data, metadata, and simulation and analysis workflows on the platform: The aim is that searches by keywords (for example, types of experiments, objects studied) or analysis DOIs can be performed and the analysis content and workflow, if available, automatically accessed.

Objective 3: Establish an NFDI-wide marketplace for tools, methods, and services

An essential element for the success of the entire NFDI is the **exploitation of synergies and services across different communities**. To this end, PUNCH4NFDI will establish a ”marketplace” that will serve as a **platform for the exchange** of tools, methods, services, and joint development activities. The marketplace will be a point of contact and general exchange within and beyond the PUNCH4NFDI community; it will facilitate the exploitation of synergies between PUNCH4NFDI and its partners and the joint use of services.

The marketplace will be realised within TA 6 ”Synergies & services” (section 5.6). The success of the marketplace will be measurable by the number of fruitful collaborations between PUNCH4NFDI and other consortia, and by the number of services ”traded”.

Objective 4: Integrate* PUNCH4NFDI *into the NFDI and the science landscape

The marketplace is part of the overarching efforts to integrate PUNCH4NFDI and the PUNCH community into the NFDI and the larger German and European structures. The former — **integration into the NFDI** — will be achieved by producing the relevant results that the community requires, as discussed mainly with the PUNCH4NFDI User Committee. Also, the close contact to data providers and data centres of PUNCH4NFDI members via the Infrastructure & Resource Board will be helpful. In short: an **efficient management and governance structure** (sections 3.4 and 5.1) will help PUNCH4NFDI to integrate into the science landscape.

On the European level, the European Open Science Cloud (EOSC) has the mission to unite the relevant funding agencies, politicians, research organisations, and research infrastructures under one single roof. Consequently, IT projects in Europe have to interact with the EOSC and use its defined interfaces. The NFDI will provide the necessary **connectivity of the German research environment** to the European research cloud. The European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE) project aims at addressing the open science challenges shared by ESFRI facilities of the PUNCH4NFDI communities and at connecting them to EOSC, ensuring the proper integration of data and tools. The computing and data management solutions developed and provided by ESCAPE and EOSC will be **propagated by the PUNCH4NFDI consortium** to smaller members of its community and via mutual exchange programmes and active collaborations also to other consortia in the NFDI.

Success can be measured in the way the resulting PUNCH4NFDI science data platform will be linked to ESCAPE and EOSC and in what scale these ideas will be propagated also to smaller PUNCH4NFDI communities and to other consortia within the NFDI.

This objective contributes to the aims of strengthening ”the realisation of the FAIR principles”, sharing ”expertise with the science community”, and ”bundling efforts” in data management.

Objective 5: Educate professionals and non-professionals in data science technologies

It is a concrete objective of the PUNCH4NFDI consortium to **educate scientists at all career levels** and the general public at large in issues related to data sciences and data management. PUNCH4NFDI will provide training opportunities for the physics community on these topics, and it will work towards implementing **data management curricula** in PUNCH studies at universities, or in the natural and technical sciences in general. Society at large will be addressed in **citizen science and outreach projects**, where they will be familiarised with important aspects of data management in the physical sciences. These activities are mainly located in TA 7 ”Training, education, outreach, and citizen science” (section 5.7), with the assistance of tools and services from other TAs, provided in the framework of the data science platform.

Success in achieving these objectives can be judged by studying the numbers of interested participants in events related to PUNCH4NFDI topics, in relevant master classes, in citizen science projects, and in training events. On the longer run, success would manifest in increased numbers of students in data-relevant subjects, scientists with data science backgrounds in leading positions, or bachelor / master / Ph.D. thesis with a focus on PUNCH4NFDI topics.

This objective addresses the aim of sharing ”expertise with the wider science community”.

Objective 6: Comprehensive technical solutions and services

The realisation of PUNCH4NFDI aims in general, and the creation of the PUNCH-SDP in particular, will require numerous detailed technical and methodical solutions, partly community-specific and partly already now of a very generic nature. Details of the related work are described in section 5. The PUNCH4NFDI consortium sees the **transformation of its solutions into services for the entire community** as a central requirement of the entire NFDI initiative (see section 4.4). In particular the PUNCH4NFDI marketplace (objective 1 described above, WP 1 of TA 6, section 5.6) will be the ideal place for promoting and exchanging the ”products” of NFDI consortia. Examples for services that PUNCH4NFDI intends to offer are the PUNCH-SDP, platform, an authentication and authorisation infrastructure (AAI) prototype, or a dynamic disk cache technology for the integration of opportunistic resources.

A generic problem not only for PUNCH4NFDI is a real-time identification of rare and/or unusual signals within huge data streams. This is a fundamental aspect of the **”data irreversibility” challenge** (see TA 5, section 5.5). PUNCH4NFDI will develop scalable solutions based on machine learning and simulations.

The success of the objective will be measured by the success in realising the technical and other solutions advertised in section 5 and by the effectiveness of the marketplace in fostering exchange between different consortia.

3 Consortium

Involvement of PUNCH4NFDI members in other NFDI consortia

The following table represents the best knowledge at the time of writing.

Institution	Other consortia
ALU	DATAPlant
DESY	DAPHNE4NFDI
DLR-DW	NFDI4Earth, NFDI4Ing
DPG	DAPHNE4NFDI, FAIRMat, MaRDI, NFDI4phys
FZJ	DAPHNE4NFDI, DATAPlant, FAIRMAT, NFDI4Agri, NFDI4Earth, NFDI4Ing, NFDIMatWerk, NFDIxCS, Text+
GAU	DAPHNE4NFDI, FAIRMat, KonsortSWD, NFDI4Agri, NFDI4Biodiversity, NFDI4Chem, NFDI4Culture, NFDI4Earth, NFDI4Health, NFDI4Ing, NFDI4Objects, NFDIxCS, NFDI-Neuro, Text+
GU	BERD@NFDI, FAIRMat, KonsortSWD, NFDI4Biodiversity, NFDI4Earth, NFDI4Microbiota, NFDI4Memory, NFDI4Objefts
HZDR	DAPHNE4NFDI
JGU	NFDI4Chem, NFDI4Culture, NFDI4Health, NFDI4Memory, NFDI4Objects, NFDIxCS
KIT	NFDI4Cat, NFDI4Chem, NFDI4Earth, NFDI4Ing, NFDIMatWerk
LMU	BERD@NFDI, KonsortSWD, NFDI4Earth, Text+
LRZ	BERD@NFDI, FAIRMat, GHGA, NFDI4Earth, NFDI4Ing, NFDI4Memory, NFDI Neuroscience, NFDIxCS, Text+
PTB	DAPHNE4NFDI, NFDI4Chem, NFDI4Ing, NFDI4Phys, NFDIMatWerk
TIB	NFDI4Culture, NFDI4Earth, NFDI4Ing
TUDa	NFDI4Earth, NFDI4Ing, NFDIMatWerk, Text+
TUDD	GHGA, NFDI4Agri, NFDI4Earth, NFDI4Ing, NFDIMatWerk, Text+
TUDO	NFDI4Earth, NFDIxCS
TUM	GHGA, NFDI4Earth, NFDI4Ing
UB	NFDI4Agri, NFDI4Biodiversity, NFDI4Health, NFDI4Memory, NFDI4Microbiota
UHD	GHGA, NFDI4Culture
UHH	BERD@NFDI, NFDI4Earth, Text+
UoB	NFDIAgri, NFDI4Earth, NFDI4Health, NFDI4Objects
UR	NFDI4Health
UzK	BERD@NFDI, GHGA, NFDI4Culture, NFDI4Earth, NFDI4Health, Text+
WWU	MaRDI, NFDI4Agri, NFDI4Earth, NFDI4Health

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

The **PUNCH4NFDI consortium** is an effort of the combined German particle, astro-, astroparticle, hadron and nuclear physics community, representing roughly **9,000 scientists with a Ph.D.** In the consortium, all important players in science are present — universities, Leibniz and Max Planck institutes as well as Helmholtz centres. The consortium also enjoys the support of the elected representing bodies of the involved fields of physics: the committees for astroparticle physics (KAT), particle physics (KET), hadron and nuclear physics (KHuK), and the Rat deutscher Sternwarten (RdS). Furthermore, the Deutsche Physikalische Gesellschaft (DPG) and the Astronomische Gesellschaft (AG) support PUNCH4NFDI. PUNCH4NFDI is the result of a **merger of the former Astro@NFDI and PAHN-PaN consortia**; it 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. As a consequence, in its forming process, PUNCH4NFDI has gone through a transformative process, exemplary for the NFDI (see figure 1): finding commonalities in relevant use cases of several sub-disciplines, agreeing on a common language and methodical approach, and defining and delivering solutions that serve a broader audience — ideally the entire NFDI.

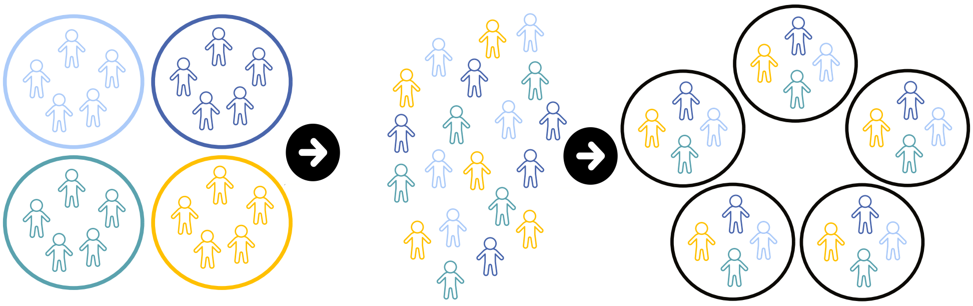

Figure 1: The merger of several sub-communities to the PUNCH4NFDI consortium. See the text for details.

The PUNCH community has since long been a leader in scientific computing and data management; it is experienced in designing, setting up and operating the relevant infrastructures and has acquired unique expertise in the scientific exploitation of data-intense large research infrastructures — especially in the fields of ”big data” and ”open data”, and in the involvement of the public (”citizen science”). This is illustrated with the following few (out of many other) examples:

- The LHC experiments are operating a worldwide computing infrastructure for about 6,000 users. In a distributed data taking, reconstruction, simulation, calibration and analysis system, relying to a large degree on the WLCG specifically created by the PUNCH community for the vast computing challenges of the LHC, on average 1M jobs are run per day by a mixture of centrally organised production jobs and decentrally organised analysis activities. 800k computer cores are tasked with these jobs, 10 PB of data are processed on each single day to cope with the demand of some 300 independent analysers submitting jobs each day. The metadata of the simulation and data can be accessed from experiment-specific web and computing tools. The upgrade to the High-Luminosity LHC will increase the network and computing requirements by more than a factor 10.
- During LHC Run 3, the ALICE experiment will have a data throughput of 3 TB/s from the detector to the online systems. After data reduction in real-time via an online farm of 50k CPU cores and 3k GPUs, 90 GB/s will be stored on disk. Also the FAIR facility experiments will record data with a rate of 1 TB/s and will have accumulated storage requirements of a few 100 PB. The corresponding software framework ALFA is a common development of ALICE and the FairRoot group at GSI.
- In astrophysics, data are commonly made openly accessible via dedicated archives. Long established common formats enable broad use. Some facilities make data obtained for time-domain studies available 24h after recording. Most datasets obtained by leading facilities are downloaded for multiple use — sometimes more than 10,000 times.

- Open access and easily accessible archives facilitated links within and between communities. Scientific exploitation often requires the combination of data obtained by different facilities (e.g. in order to obtain measurements of the same cosmic object obtained in different ranges of the electromagnetic spectrum). Cross-community use has been carried out e.g. by the astroparticle physics experiment H.E.S.S., which in its about 200 publications has used data from more than 80 different observatories and experiments thanks to common formats and open-access archives.
- The LOFAR telescope processes a data rate of 35 TB/h. However, only a reduced data cube can actually be kept in the long-term archive, a third of which (about 20 PB) is located within the PUNCH community at FZJ. The next generation telescope, SKA, will produce about 60 EB/yr with instantaneous data rates that exceed the global internet traffic of today by a factor of a few.
- The access to archival data from different sources has always been crucial for Universe sciences. But it was the advent of ”virtual observatories”, which brought the entire sky at different parts of the electromagnetic spectrum to the fingertips of the astronomers, that led to the successful concept of ”multi-messenger” astronomy. An excellent recent example is the observation and understanding of the first double-neutron-star merger observed first with LIGO. The bright after-glow of this cosmic collision and firework could only identified in an external galaxy by comparing observations with archival data. But virtual observatories go indeed much further. Based on the infrastructure platform originally pioneered for SETI@home, further projects included, for instance, Einstein@home to find gravitational waves. A total of more than 1 million helpers provided an average performance of 6.5 PetaFlop. A similar project, Zooniverse, benefited from the help of more than 2 million volunteers (!) to produce more than 500 million classifications of galaxies.

PUNCH science addresses the deepest questions of nature, it provides **unparalleled fascination for the public**, it educates the brightest minds, and it is a driver of progress and technical development. PUNCH science is continuously operating at the limit of the technologically feasible. In particular, PUNCH’s next generation of facilities will produce unsurpassed amounts of data that cannot be tackled with existing tools and methods, but require **completely new approaches in scientific computing and data management**. PUNCH4NFDI will make indispensable contributions to these new solutions.

PUNCH4NFDI addresses **next-generation data challenges** already today. Current and in particular upcoming experiments are producing data volumes so huge that only a tiny fraction can be kept in long-term storage. Decisions about the rejection of data have to be taken in real-time and, consequently, will be based on incomplete information as there is not enough time for comprehensive analyses. Loss will be inevitable and mostly irreversible. The impact of ”data irreversibility” on the reproducibility of scientific results must be traced and determined.

PUNCH4NFDI is **international**. The involved branches of physics are, to a large extent, globally organised — examples are the global collaborations of the LHC experiments or the telescopes of astroparticle physics like CTA or IceCube, or the satellites like Gaia in astrophysics. In all these contexts, PUNCH4NFDI members take leadership and coordination positions and are actively shaping and executing **global, European and national roadmaps and strategies**.

PUNCH science has **many partners**, most importantly a large number of European and international initiatives, as EOSC, and ESCAPE, and facilities as CERN, ESO, and ESA. There is also a whole slew of important contacts to industrial firms with whom development tasks are shared, as SAP, Google, Hewlett-Packard Enterprise, IBM, Microsoft.

PUNCH4NFDI members are **well connected** with related and complementary **national and international initiatives**, like ErUM-Data or the HEP Software Foundation. The ErUM-Data funding line of the BMBF focuses on the specific needs that are common to users of large facilities. PUNCH4NFDI can build on solutions developed for the ErUM communities and make them available for a broader range of data infrastructure users, for example PUNCH theorists.

In summary, PUNCH4NFDI is ready to face the upcoming challenges, both organisationally — as a leading partner of a global community — and technically — as a driver of new developments in data management.

The following institutions bring in their expertise in the PUNCH4NFDI developments (see also table 2 for the contribution per TA):

DESY is a major contributor to experiments in particle and astroparticle physics and has a worldwide leading theory group. As national laboratory for particle physics, DESY contributes to the leading experiments and efforts worldwide. The laboratory operates compute infrastructures like a WLCG Tier-2 centre for ATLAS, CMS and LHCb, services for the Lattice Data Grid, the CTA Science Data Management Centre, and a major centre for Belle II. The National Analysis Facility, which supports all German particle physics groups, is also located at DESY. Together with the compute facilities, DESY offers ample services and support to the entire community. Staff scientists from DESY are members of the computing management of ATLAS and CMS. The development of the storage middleware dCache is led by DESY, and developers are engaged in the WLCG DOMA effort, in ESCAPE, and in other EOSC related projects. **Key words: big data management, open data, compute centres, services, middleware, real-time computing.**

Leibniz-Institut für Astrophysik Potsdam (AIP) is a major research centre for Astrophysics and major contributor to instrumentation at large ground and space based observatories. AIP has a long tradition in building scientific infrastructure. It lead the AstroGrid-D, and WissGrid with focus on research data management infrastructures. It is co-founder of GAVO and one of the German Virtual Observatory data centres. RDMO, created by AIP, is the most widely used tool in Germany for data management plans and is used by Universities and state-level NFDI initiatives, and also, e.g. by NFDI4Ing. AIP contributes its developments and experience in curating and publishing astrophysical data collections, in development of open-source software, and collaborative research environment. For PUNCH4NFDI, AIP provides an operational infrastructure via a multi-cloud and multi-storage compliant setup using Kubernetes/OpenShift (OKD), containerisation software developed by RedHat. With integrated Gitlab and continuous integration, this is an optimised environment for continuous application development and multi-tenant deployment, notionally for the PUNCH-SDP and its microservices. **Keywords: research data management, collaborative research environments, data publication, data curation,**

Virtual Observatory, grid and cloud infrastructure.

Frankfurt Institute for Advanced Studies (FIAS) is an independent research institute performing cutting-edge research in the areas life science, physics, neuroscience, computer science and systemic risk. In the area of computer science, FIAS focuses particularly on research and development of new computer architectures and algorithms to achieve better energy-efficiency. FIAS has significantly contributed to the ALICE High Level Trigger also including its upgrade towards LHC Run 3 requirements. Additionally, the FIAS group has taken responsibility to develop a large real-time processing farm, the ”First-level Event Selector” designed for real-time data taking of the CBM experiment in Green IT cube at GSI, a hybrid supercomputer with GPU infrastructure. Access to experiment-related systems of these experiments is complemented by the general purpose HPC cluster Goethe-HLR. This cluster system provides more than 18 000 cores connected via EDR-Infiniband. **Key words: energy-efficient high-performance data centres, efficient algorithms to support specialised hardware, real-time data reduction.**

Forschungszentrum Jülich (FZJ) operates supercomputers of the highest performance class, leading-edge storage systems and data analysis resources, and will host the first Exascale compute facility in Europe. FZJ supports computing and data resources users by a comprehensive support group to access and use the various systems in the form of Simulation and Data Labs. As a key European HPC institute, FZJ’s natural focus is TA 3 with its core task of providing tools for simulation (WP 3.2) and processing of data (WP 3.4) including developing advanced machine learning tools (WP 3.3). Another of FZJ’s assets is its specific provision and managing of data storage facilities for the astronomy and astrophysics community in the context of the LOFAR collaboration, which is of high relevance for TA 2 and TA 5 alike especially in the context of MeerKAT and SKA. TA 6 profits from FZJ’s long-term expertise in AAI. FZJ provides routinely general training courses on a variety of topics related to simulations and data handling, which will be extended to cater for the specific needs with PUNCH4NFDI (WP 7.1). **Key words: simulation, machine learning tools, big data management, AAI, compute centres, data storage, concept development, services, training.**

Georg-August-Universität Göttingen (GAU) has a long-standing tradition in computing for the particle physics experiments ATLAS and Belle as well as in teaching and outreach. The institute operates the WLCG Tier-2 grid computing centre GoeGrid. Recently, the ”High Performance Computing in northern Germany - HLRN” moved to Göttingen. In addition, GPU and compute clusters are available for development, research and education. Recently, the ”Campus Institute for Data Science - CIDAS” has been established with a focus on the development on modern data science techniques, in particular machine learning, which also applies to high-throughput online computing using FPGAs. Göttingen staff is involved in the CERN School of Computing and other teaching and education activities. Göttingen has been active in physics outreach for many years (re-development of Kamiokanne, offers ”PiA - Physics in Advent”, initiator of ”physics for refugees”), staff members served for five years in the steering committee of the DPG as outreach coordinator. **Key words: grid computing, compute centres, services, resource provider, service, middleware, real-time computing, outreach, teaching.**

The **Helmholtzzentrum für Schwerionenforschung (GSI)** in Darmstadt operates a worldwide leading accelerator facility for research purposes. Additionally, an international accelerator facility is currently being built in cooperation with international partners for the research with antiprotons and ions (FAIR facility). GSI is a major contributor to the hadron and nuclear physics community. GSI is involved in a leading position in ALICE physics, ALICE detector construction, as well as ALICE software development and computing: the lab and operates the world's largest ALICE Tier-2 centre. An additional service to the German ALICE community is the National Analysis Facility. GSI has experience in the design and development of storage middleware like XRootD and data lakes. Also analysis frameworks like FairRoot are developed at GSI. **key words: big data management, compute centres, services, storage middleware, software development, detector construction.**

Hochschule für Technik und Wirtschaft Berlin (HTW) has long-standing experiences in contributing to frameworks for the management and the curation of large data volumes. It is cooperating since years in this field with important national and international IT service providers and research institutions. HTW is contributing to the development of memory-based computing. The University is a member of the *German Long Wavelength Consortium (GLOW)*, and the *Verein für datenintensive Radioastronomie (VdR)*. **Key words: big data management, memory-based computing.**

Johannes Gutenberg-Universität Mainz (JGU) is a major centre for research on particle physics. It hosts the PRISMA+ Cluster of Excellence with research groups from astroparticle, particle and hadron physics. This includes a strong involvement in the ATLAS experiment at the LHC, the operation of the Mainz Microtron (MAMI) and the construction of the Mainz Energy-Recovering Superconducting Accelerator (MESA). The data centre of JGU is a Tier 2 HPC centre operating a supercomputer with more than 50,000 CPU cores and 2 PFlop/s Linpack performance that is coupled to a 10 PByte Lustre file system and a 2 PByte Ceph archive. **Key words: HPC, big data management, compute centre, scale-out storage, services, real-time computing, machine learning.**

Karlsruher Institut für Technologie (KIT) – The Research University in the Helmholtz Association – is one of the leading German research institutions with a focus on engineering and the natural sciences as well as a university of excellence within the German excellence strategy since 2019. Research, teaching, and innovation are the core tasks of KIT. The research infrastructure includes state-of-the-art laboratory equipment, modern information systems, and large-scale experimental facilities. In the PUNCH research field KIT plays a leading role in astroparticle physics, particularly with the Pierre Auger Observatory for cosmic rays, IceCube for neutrino astronomy and the Karlsruhe Tritium Neutrino (KATRIN) experiment and operates the KASCADE Cosmic-ray Data Center (KCDC). In particle physics KIT has large research groups in CMS and Belle II and is also home of the Tier-1 German data and computing centre GridKa. As one of the largest and best-performing data centres for particle and astroparticle physics worldwide, GridKa is an indispensable large-scale research infrastructure, enabling the successful participation of German physicists in international particle physics collaborations. **Key words: big data management, data storage, services, middleware, open data.**

Ludwig-Maximilian-Universität München (LMU) has one of the leading physics faculties in Europe and brings in the expertise of two research groups. The particle physics group has a strong tradition in distributed computing and the operation of a large Tier 2 computing centre for the ATLAS experiment in the worldwide LHC computing grid. It plays a leading role in the software development of the international Belle II collaboration, provided the chief referee for the computing in the LHC experiments review committee, and currently coordinates the ErUM-Data pilot project for the development of experiment-overarching computing solutions. The cosmology and astronomy group has played a leading role in the initiation of several world leading cosmology-focused multi-wavelength surveys (Dark Energy Survey (DES), South Pole Telescope, eROSITA and Euclid satellite missions) and in the development of new software tools to address data challenges in these projects and others. This group has established the leading Bayesian, hierarchical analysis tools within the context of cosmological analyses of structure formation based observational constraints. Both research groups are co-leading work packages within TA 3 “Data transformations”. LMU Munich is also fostering a close collaboration with Technical University München (TUM). **Key words: data management, distributed computing, collaborative software development and integration, analysis workflows, combination of data, statistical tools**

The **Max-Planck-Institut für Radioastronomie (MPIfR)** is Europe’s leading research centre focusing on radio astronomical observations. It owns and operates Europe’s largest radio telescope, the 100-m Effelsberg telescope, and is a major stakeholder and partner in the ILT, the APEX telescope, SOFIA, global, European and mm-VLBI, Event Horizon Telescope (EHT), MeerKAT and the SKA. Together with South Africa it is extending MeerKAT to ”MK+”. The corresponding research produces data from m- to sub-mm wavelengths and includes, for instance, research on spectroscopy, time-domain and transient sky, and high-resolution imaging, like the EHT, all of which are processed in its own dedicated compute facilities. **Key words: observations, large data volumes, imaging, time-domain astronomy, real-time processing, software development, machine learning, citizen science.**

The **Max-Planck-Institut für Kernphysik (MPIK)** is a major centre for gamma-ray astronomy and astroparticle physics in Germany, strongly engaged in the astronomy projects H.E.S.S., HAWC, CTA and SWGO, and in particle and astroparticle projects including LHCb, LEGEND and XENON. MPIK has experience in the handling of Big Data and is host to the archive of data from the H.E.S.S. gamma-ray observatory, and is a key site for air-shower and detector simulations. MPIK is heavily active in the development of scientific software tools for ongoing experiments and for upcoming observatories, and a partner in the H2020 programme ESCAPE. **Key words: data handling, simulation, software integration / development, open data formats.**

The **Thüringer Landessternwarte (TLS)** carries out research in many fields of observational astronomy, the topics range from the search for exoplanets to the study of cosmic magnetism. The TLS operates the 2-m-Alfred-Jensch-Telescope (AJT) and one of the international LOFAR stations. More than 60 years of operating the AJT have lead to an immense collection of scientific data. The TLS participates in a wide range of consortia to obtain and analyse large amounts of research data, e.g. KESPRINT, GROND, PLATO and the LOFAR Two-Metre Sky Survey (LoTSS). The radio astronomy group at the institute leads the efforts to process the enormous amount of LOFAR data stored at FZJ LOFAR Long Term Archive at the JUWELS supercomputer and contributes to software development of general LOFAR data processing. **Key words: data management, data archives, software integration and development.**

Technische Universität Dresden (TUDD) has a long tradition in experimental nuclear and particle physics with experienced research groups participating in the design, construction and operation of large particle detectors within international collaborations. One research focus is on the particle detector upgrade programme in preparation for the High Luminosity LHC. TU Dresden physicists and personnel provide experience in real-time data processing and modern machine learning techniques as ingredients to novel solutions in real-time data reduction and related challenges. TU Dresden is profiting from a modern high-performance computing centre with full support for associated research groups. Moreover, the TU Dresden will form regional cluster with the HZDR with renowned compentencies in computation intensive physics simulations. **Key words: construction and development of particle detectors, real-time data processing, machine learning, computational radiation physics.**

Technische Universität Dortmund (TUDO) has a long tradition in building, extending and operating particle-physics experiments, in particular the LHCb and ATLAS experiments at the LHC. Members of the different working groups participate in the analysis of large-scale data sets using modern statistical methods and complex models, which is also reflected in a collaborative research center with the faculty of computer science and the faculty of statistics. The NFDI-relevant expertise of the PIs is in the design of real-time components, e.g. the LHCb trigger and BCM, ATLAS and LHCb tracking detectors, as well as in the development of general statistical tools, e.g. the BAT project. The PIs also place an emphasis on teaching and outreach activities, also in an international context, e.g. via an ERASMUS+ strategic partnership. **Key words: statistical tools, real-time computing, teaching, outreach.**

Universität Bielefeld (UB) founded as a research university, is dedicated to transcending boundaries. Its radio astronomy and lattice quantum chromo dynamics (LQCD) groups are part of PUNCH4NFDI. The radio astronomy group is a major partner in LOFAR and D-MeerKAT, operates the compute and storage cluster of the GLOW consortium, and contributes with its expertise gained with the LOFAR Long Term Archive and LOFAR data management software. The lattice QCD group operates a GPU cluster for scientific computing and brings in its expertise in hardware specific code optimisation for large-scale simulations and data analysis on distributed data sets. Both groups are engaged in the Bielefeld Centre for Data Science, which advances data literacy across disciplines. **Key words: data management, data archives, simulations, real-time data, data literacy, citizen science.**

The Ruprecht-Karls-Universität Heidelberg (UHD) hosts one of the largest faculties for physics and the ”Zentrum für Astronomie (ZAH)”, the largest university institution in astrophysics in Germany. It hosts strong research groups in all fields covered by the PUNCH4NFDI consortium, with astronomy and astro-particle physics taking lead responsibility on the PUNCH4NFDI project. ZAH has been a major hub of VO activity for the past ten years, having significantly contributed to many standards from file formats to data access to Registry. Its software package DaCHS has been the basis of the GAVO data centre and operates the default Registry endpoint for tools like TOPCAT or Aladin. DaCHS is deployed worldwide, in particular to support the EPN-TAP protocol used in solar system sciences. The astro(-particle) activities particularly relevant for PUNCH4NFDI further include the involvement in H.E.S.S., CTA, MeerKAT, GAIA, GAVO, and significant involvement in international cooperations for data-management (e.g IVOA). **Key words: data management, services, metadata standards, time-domain studies, statistics.**

University of Hamburg (UHH) is a major centre for particle- and astrophysics, and hosts the federal excellence cluster Quantum Universe. The Observatory of the University has a focus on numerical astrophysics and survey astronomy. The particle physicists have leading roles in a wide range of international experiments. In the past years, particle- and astrophysicists have collaborated on a range of topics in the field of machine learning, e.g. for the automatic classification of sources and events as well as the generation of artificial data]. **Key words: machine learning, HPC, numerical methods, software development.**

The department of physics and astronomy at the **Rheinische Friedrich-Wilhelms-Universität Bonn (UoB)** covers a wide range of scientific areas, including astro, hadron, nuclear, and particle physics. It operates its local electron accelerator and storage ring ELSA, which delivers primary polarised electrons for the hadron spectroscopy experiments Crystal-Barrel and BGO-OD. Bonn is involved in the particle physics experiments LHCb, Belle II and ATLAS, and operates a Tier3 computing centre in the WLCG with a strong focus on developments for distributed and opportunistic high throughput computing, and on analysis preservation. Bonn also contributes significantly to HPC computing with its strong Lattice QCD theory group. On the astrophysics side, Bonn has major stakes in the FYST, Euclid and eROSITA observatory missions, and hosts the German node of the ALMA Regional Center that supports the operation and enhancement of the ALMA observatory. The operational experience of data taking and handling, and of a computing infrastructure for particle experiments and astronomical observatories are an asset that allows to bridge large, international experiments to our local research groups. Bonn physics and astronomy has a strong outreach program, which includes public talks, teacher training, girl's days, the ”Schülerlabor Küstner”, the school project ”Astronomie / vor Ort”, and the ”Physikwerkstatt Rheinland”. **Key words: outreach, distributed and opportunistic computing, computing for small-to-medium scale experiments, analysis preservation strategies.**

Universität Regensburg (UR) is a major centre for lattice QCD worldwide. The team develops HPC hardware (the QPACE 1–4 machines) and software and is involved in the generation, analysis, storage and access of large sets of simulation data. Main activities involve development and optimization of HPC community software for lattice QCD (GRID, Chroma, GPT), data analytics and data handling. The group operates large HPC clusters and small test machines of different architectures and runs a multi-PB storage system as well as smaller parallel file systems, using BeeGFS and GlusterFS. It hosts a copy of the gauge ensembles of the European CLS effort. **Key words: HPC, numerical methods, simulation, (meta)data handling, software integration and development.**

3.2 The consortium within the NFDI

As stated above, PUNCH4NFDI commands **unique expertise in the exploitation of data-intense research infrastructures and in data management**, and in particular in the fields of ”big data” and ”open data”. PUNCH science is already today concerned, in a leading position, with future questions and challenges that other areas of science will face only in a number of years from now. PUNCH4NFDI will thus be a pathfinder for new solutions for the entire NFDI.

PUNCH research is international and collaborative. This ensures, on the one hand side, that PUNCH4NFDI with its connections to national, European, and global endeavours, has always the best and most comprehensive sources of new solutions at its hands. On the other hand, PUNCH4NFDI will always be happy to share data, methods, and tools — PUNCH has a sharing culture and intends to bring this spirit into the NFDI.

Also within its own branch of science — physics — PUNCH4NFDI is well organised. The DPG coordinates exchange between all physics-related NFDI consortia, with a view to their complementarity and their complete **coverage of all aspects of the physics landscape**.

Concerning the role of PUNCH4NFDI within the NFDI, the consortium 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. An interdisciplinary view of datasets raises even more data-structure related issues than a cross-experiment combination of data in the PUNCH domain alone. 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 topics. The following numerous contacts and projects have already been established and proven to be productive:

– **MaRDI**: Immediate collaboration is envisaged with the field of mathematics. One topic of collaboration will be the development of common concepts for data integration and annotation with metadata. The micro meta schemes developed by the mathematicians will be investigated in collaboration with TiB. PUNCH4NFDI will also contribute domain-specific knowledge to the MaRDI database, which will then also be added to the MaRDI knowledge graph. PUNCH4NFDI and MaRDI will also explore viable analysis methods and statistical procedures, where the mathematicians can offer an expert service for training as well as guidelines for validated computational workflows to improve the re-usability. PUNCH4NFDI in turn offers a variety of high-statistics datasets for extensive testing and further methodologic development. A special point of mutual interest is the analytic integration of Feynman integrals for scattering processes at high-energy colliders. Within this context, MaRDI can contribute their algorithm database as well as FAIR data formats. MaRDI also creates reference datasets for the training of machine learning tools. The consortium also maintains a repository for these that also PUNCH4NFDI will use and to which it will contribute with own datasets. Collaboration is also foreseen in the optimisation of neuronal networks for specific PUNCH4NFDI use cases. In order to be able to use the PUNCH-SDP, MaRDI plans to provide a use case together with a corresponding research container. Last but not least, MaRDI will support the PUNCH4NFDI activity to contribute to an NFDI-wide AAI infrastructure.

- **NFDI4HPC/NFDIxCS**: In the coming years, PUNCH physics will process Exabyte datasets — the analysis of 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 important topic to cover here is the identifcation and application of standardised interfaces. In this context it is important to ensure the use of FAIR principles across the communities. Intensive work on the definition of standards (DOI infrastructure, domain ontologies) in collaboration with NFDI4xCS is planned. In order to be able to use the available HPC and high-throughput computing (HTC) resources efficiently, it is necessary to continuously optimise 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.
- **NFDI4Earth** and PUNCH4NFDI have the following overlapping research areas: First, Earth science results are important ingredients for studying solar system planets and exoplanets in astronomy. Second, when studying the influence of the sun on the weather systems on Earth, participants of both consortia often use identical detector technologies and sometimes even the same satellites. Third, the operation of IR telescopes is dependent on reliable atmospheric models. These require high-quality meteorological data that in turn are dependent on transmission measurements of Earth's atmosphere in optical wavelength. The latter are again a by-product of astronomical data. Finally, at all PUNCH4NFDI observation sites, information about the environment (temperature, precipitation, air pressure, etc.) is gathered, and it is planned to publish these in more accessible ways. In these areas, it is planned to collaborate with NFDI4Earth in order to ease the exchange and mutual use of data.
- **NFDI4Ing** and PUNCH4NFDI will collaborate on metadata, software architecture, and AAI. Moreover, the exchange of use cases and archetypes may be beneficial for both consortia. One plan for concrete interaction is that PUNCH4NFDI could utilise the terminology services, in close collaration with TIB. In this context also community-overarching metadata concepts will be addressed. Further topics of collaboration are the "Leipzig-Berlin declaration of NFDI cross cutting topics" and the action plan and interoperability catalogue of the EOSC, which includes important topics like FAIR digital objects.
- **NFDI4Microbiota** and PUNCH4NFDI will collaborate by exchanging metadata and in developing metadata standards. The main interest of NFDI4Microbiota is a description of samples and the findability of data in their repositories. For that the consortium has provided a use case to PUNCH4NFDI that will be the basis for common investigations. NFDI4Microbiota is operating a cloud infrastructure for life sciences distributed over several sites in Germany. Collaboration is foreseen in developing automated caching mechanisms for the corresponding cloud storage platform based on PUNCH4NFDI caching technologies. Moreover, AAI and the development of cloud standards are topics of mutual interest.

- **NFDI4Culture**: One important area of mutual interest are legal and cross-cutting topics according to the ”Leipzig-Berlin Declaration”, e.g. the legal aspects of an AAI and FAIR data management. The consortia will also address distributed computing and the way data, the corresponding processing software and subsequent publications are coupled, and they will exchange experience with metadata for describing domain-specific ditigal cultural artefacts. Common work will also go into running NFDI4Culture applications on the PUNCH-SDP and the consortia will engage together in education and training.
- PUNCH4NFDI and **NFDI4Chem** will collaborate on metadata and data reusability, where standards and a unification of community-specific notations are important topics. Both consortia work towards one common AAI solution for the entire NFDI.
- Connections exist also to the consortia **DAPHNE4NFDI and FAIRMat** — the consortia of photon and neutron research and of material sciences. Together with the DPG, these consortia will jointly address general physics-related aspects, for example the education and training of young scientists in the fields of research data management, and the outreach to the general public. More concretely, with DAPHNE4NFDI, PUNCH4NFDI shares use of large-scale research facilities (e.g. at DESY). With FAIRMat, PUNCH4NFDI will address topics of curation, archiving and processing of large data files. All three physics consortia are concerned with the challenges of very large data sets and metadata schemes. Topics of collaboration will comprise the curation, archiving, processing and reduction of very large data sets, and the collection of the relevant metadata schemes. Furthermore topics like search engines, AI tools, image processing, or (statistical) analysis methods will be addressed.

As a very specific contribution to the NFDI’s synergetic character, PUNCH4NFDI will set up, in its TA 6 ”Synergies & services”, a marketplace for ideas, tools, and services — a base for communication and exchange over the entire NFDI. Details can be found in section 5.6.1.

3.3 International networking

The international character of PUNCH science has been alluded to before: PUNCH research is, **to a large extent, global**:

- PUNCH research is guided by international, often global, **roadmaps and strategies** (ICFA, European strategy for particle physics, APPEC, Astronet, ...), in the definition of which PUNCH4NFDI scientists are centrally involved in many places. Not least due to the sheer size of many experiments and facilities in PUNCH science, the field and technical solutions developed in it are necessarily, almost *per definitionem*, international.
- A large share of PUNCH research is taking place at global or international facilities (CERN, ESO, ...), under the guidance of international institutions (ESA, ...), and in international collaborations, often reaching thousands of members from all over the world. Again, **PUNCH4NFDI members take many leading positions** in these endeavours and contribute to the definition of their directions and developments.
- PUNCH4NFDI is, therefore, an integral part of the global and European efforts in PUNCH science, deeply involved in the European and international task sharing in all relevant areas — not least in **scientific computing and data management**.

– PUNCH4NFDI members are also involved, in key positions, in large, subject-specific **international data initiatives** (WLCG, IVOA, ILDG, EOSC, ESCAPE, ...) and in important standardisation working groups (IVOA and others).

3.4 Organisational structure and viability

The consortium structure is displayed in figure 2. Based on the continuous feedback from our TAs and cross-cutting topics, this structure and the corresponding management policies will be updated during the life cycle of the consortium. Section 5.1 on TA 1 describes the various administrative tasks connected to the governance and management of the consortium. It is understood that all PUNCH4NFDI bodies will give themselves bylaws.

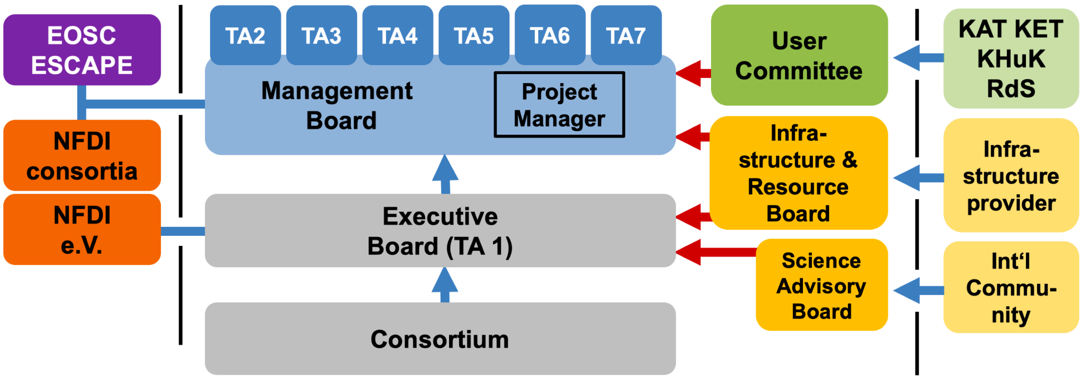

Figure 2: The foreseen governance structure of the PUNCH4NFDI consortium. Red arrows indicate advisory functions. For further details, see the text for details.

On the **executive level**, the PUNCH4NFDI consortium is handled by the *Executive Board* (EB), composed of six board members representing the entire PUNCH4NFDI community and all relevant institutional groups (universities, Helmholtz, Max Planck, Leibniz, ...), plus a *project manager* (PM, see deliverable D-TA1-WP1-3). At the time of writing this proposal, the board members are Andreas Haungs (KIT), Susanne Pfalzner (FZJ), Kilian Schwarz (GSI), Thomas Schörner (spokesperson, DESY), Matthias Steinmetz (AIP), and Stefan Wagner (Heidelberg). The board members will initially be appointed by the consortium in a kick-off workshop (deliverable D-TA1-WP1-2), in which all co-spokespersons of the consortium have one vote.

The EB is responsible for the **overall oversight of the project** in scientific, financial, and outreach terms. Its tasks comprise in particular the appointment of a project manager (PM), the appointment of the TA coordinators and the forming of the Management Board (see below). It is also responsible for the representation of PUNCH4NFDI towards the DFG, the NFDI directorate, and the management of other consortia or initiatives in the NFDI and beyond.

The **scientific organisation** of the consortium is in the hands of the *Management Board* (MB), consisting of the TA leaders and the PM. The task of the MB and in particular of the PM is to organise the consortium internally on the scientific and administrative level. Concretely, the following tasks — among others — will be pursued here (see also section 5.1): scientific coordination and controlling, progress monitoring, resource allocation, outreach organisation, working relations to other consortia (see also sections 5.6 and 5.7 on the TAs 6 ”Synergies & services” and 7 ”Training, education, outreach, and citizen science”).

EB, MB and the PM are supported in their work by administrative and other staff, forming a **consortium office at DESY** (section 5.1 for details on the administrative tasks and the necessary personnel). The office will organise the financial aspects of the consortium and the disbursement of funds to the co-applicants, and it will in general support the PM in his tasks.

In their tasks the EB and the MB are advised by a *Scientific Advisory Board* (SAB), in which international experts from PUNCH physics and from scientific computing in different communities inside and outside of the PUNCH world are assembled. The SAB will meet once a year and will **advise the consortium** and its management on future directions of the consortium. The SAB also reflects the fact that structural efforts in PUNCH scientific computing can only be successful when agreed upon by all international partners from e.g. the large experimental collaborations.

An *Infrastructure & Resource Board* (IRB) will ensure the **close interaction with infrastructure and resource providers** and the correct consideration of their developments and possibilities in the planning of the consortium’s future directions. The IRB will contain representatives of relevant outside infrastructures and data sources (WLCG, EOSC, the HEP Software Foundation, DPHEP, LHC experiments, FAIR facility experiments, Belle II experiment, a major astroparticle experiment or observatory, theory, ErUM-Data initiative, ...).

Finally, a dedicated *User Committee* (UC) will be filled by appointment by the **community representations** of KAT, KET, KHuK, and RdS. The task of the UC, to be carried out in regular meetings with the Management Board, is to ensure the strict orientation of developments in the consortium towards the needs of the PUNCH users. It will in particular give feedback on the services offered by PUNCH4NFDI.

At the time of submission of this proposal, concrete suggestions for the filling of the relevant positions in all boards exist. They will be put forward to the consortium for formal election at the PUNCH4NFDI kick-off meeting mentioned above.

PUNCH4NFDI foresees to integrate the **activities of 20 (co-)applicant institutions and 22 participants**. Measures will be taken to ensure coherence of all activities and actors in the consortium and to avoid a trickling-away of the funds into separate and unconnected efforts. In particular, the various bodies of the consortium will organise **intense communication** among the relevant parties. Furthermore, care will be taken in the hiring process to PUNCH4NFDI positions to carefully select personnel that retains a focus on the NFDI goals in general. Furthermore — and this is a critical point in particular in the field of research data management and scientific computing — PUNCH4NFDI aims at achieving gender equality in the hiring process (see also the detailed discussion of TA 7 in section 5.7).

3.5 Operating model

The PUNCH4NFDI operating model addresses the key aspects for a successful long-term infrastructure, in particular the aspects of the consortium organisation, its development phases, its engineering approach, and its sustainability.

Consortium organisation: The overall organisation and the establishment plan are presented in section 3.4 and in the description of TA 1 ”Management and governance” in section 5.1. At the time of writing of this text, the overall structure of the NFDI e.V. and the integration of the various NFDI consortia was still in flux. Should the development of the NFDI organisation and its entities have sufficiently progressed, the PUNCH4NFDI legal framework will closely follow that of existing NFDI consortia and the recommendations of the NFDI directorate, ideally provided by the NFDI directorate in the form of a template. Alternatively, and possibly as an interim solution, a consortium agreement will be drafted following the rules and procedures of numerous existing scientific cooperation and project agreements, to be signed by all co-applicants and participants before the beginning of the project. All (co-)applicants and participants will work on their own account and according to the requirements of their respective legal status. PUNCH4NFDI will thus not set up a new legal entity for the consortium. In either way, special emphasis is given that **the constraints imposed by the German Fiscal Code and Value Added Tax Act** are appropriately taken into account. For the time being, all (co-)applicants have committed themselves to collaborate according to the organisational structure described in section 3.4 and according to their respective tasks described in section 5. By signing this proposal, the co-applicants and participants have also agreed to enter a long-term agreement with the according commitment to ensure sustainability of the NFDI and its goals.

Decision-making process: Decisions are made by the Management Board (MB), the TA leaders, and the WP leaders — depending on the scope of the engagement. The MB is periodically updated on the progress of the TAs and advises, if necessary, the leader of a TA in possible adjustments of the work. In case of disputes between partners, the lead of an affected TA will try to resolve the dispute. If no consensus can be reached, the MB will take decisions.

Risk management: An important component of the process control is the notification of risks (see section 5. Building a federated data infrastructure is a complex project and has to cope with many risks, whereby some of them may even have an impact on the success of the project as a whole. Events presumably leading to deviations from the delivery of envisaged output have to be noted by the lead of a TA to the MB. It induces mitigation processes, for example re-assigning work or FTEs between partners, and will supervise the progress. An important component of the risk management will be a **mid-term review** that will be conducted by the MB, the SAB and further external experts, where appropriate (deliverable D-TA1-WP3-5, section 5.1). This mid-term review will evaluate the effectiveness of the governance structure and the various TAs. Depending of the outcome of this review, resources may be re-allocated between the different work packages or even TAs.

Development phases: As a matter of principle, PUNCH4NFDI discriminates between the *development phase* of tools and services, their testing by the experts in the consortium (*alpha testing*) and by expert users in the community (*beta testing*). These phases can, eventually, be followed by a *roll-out* to the PUNCH community and the NFDI community at large (see section 4.4). PUNCH4NFDI development can rely on a **decades-long experience in handling large scientific datasets and related services and tools**, and making them available to the expert community as well as to other communities and to the public, where appropriate.

PUNCH4NFDI therefore anticipates that all four project phases basically will co-exist in parallel for different work packages and deliverables.

Owing to the strong engagement of national research centres (Max Planck, Helmholtz, Leibniz) and supercomputer centres (LRZ, FZJ, MPCDF), resources for a **sustained operation** for the beta-testing and, within limitations, for the roll-out phase can be guaranteed, and proper dedicated hardware resources are committed as in-kind contributions to PUNCH4NFDI (see discussion of PUNCH4NFDI in-kind contributions in section **??**). Furthermore, many of the developments by PUNCH4NFDI will enter into internationally coordinated development efforts (e.g. ESCAPE and EOSC), and it is anticipated that some of the outcomes of PUNCH4NFDI will find a long-term home in the data management efforts of international facilities.

Engineering plan: The **basic engineering entity** of PUNCH4NFDI is the layer model, a well-established concept in the computer sciences for organising complex projects (see section 4.1). TA 2 to 4 (sections 5.2–5.4) provide interfaces between two neighbouring layers. TA 5 to 7 are more focused on an overarching approach between layers and TAs, in particular with the intent of dissemination (section 5.6), be it to other communities, be it in the area of education and outreach (section 5.7), in particular in the context of being prepared for the data rates and volumes anticipated for the coming decade. WPs within each TA combine ”process-oriented” and ”result-oriented” approaches.

Material results of PUNCH4NFDI(software, (meta)data, documentation, training materials, etc.) will be made available as **open-source or open-access products** free of charge.

Sustainability: At the time of writing of this proposal, no clear long-term strategy of the NFDI as a sustainable infrastructure has been devised. The establishment of such a strategy is a key ingredient for any planning beyond the project-oriented funding cycle of the NFDI. As explained above, it is nevertheless foreseen that services developed by PUNCH4NFDI can at least initially be perpetuated thanks to the engagement of large research and supercomputing centres. Considering the strong presence of such structures in the PUNCH science domain, a **limited long-term operation model** provided by the aforementioned partners is foreseen to find easier acceptance than a model based on (moderate) user fees. Furthermore, many co-applicants are either themselves data providers or have **strong ties to data providers** in their respective area of expertise and research domain, again helping to root a **long-term service model**. The large national and international facilities, compute centres and research institutions will also be a core backbone for long-term data archiving. PUNCH4NFDI will very actively participate in the NFDI-wide development of concepts and models for a sustainable national research data infrastructure.

4 Research data management strategy

4.1 State of the art and needs analysis

Data and metadata of the PUNCH community — introduction and description of data landscape: **PUNCH data are very diverse** in nature, stemming from very different facilities and experimental approaches. In particle physics and in hadron & nuclear physics, data are very often records of particle interactions in colliders. However, other types of experiments produce different data — e.g. testbeam results, non-collider-based experiments as for dark matter or axion searches, etc. In astroparticle physics, ”pictures” of showers stemming from photons, cosmic rays or neutrinos impinging on an array of sub-detectors are recorded, while in astronomy pictures of the sky in different wavelengths are the focus.

The data of PUNCH physics thus **concern very different natural phenomena**; they cover vastly different length scales and concern processes taking place at very different timescales. Furthermore, the data come in all **different sizes** — from individual photon counts at low repetition rate to the Terabyte per second regime — and at **largely differing abstraction levels**. Due to the latter circumstance, also the metadata of PUNCH physics are very diverse and in addition depend on the abstraction, reconstruction, and analysis levels of the data in question.

Data management in the PUNCH field of science must reflect this **heterogeneity of data**, and it must increasingly consider the problem that PUNCH experiments will be faced with the inability to store all potentially accessible raw data due to the ever growing data rates and volumes (this aspect is considered in TA 5 ”Data irreversibility” in section 5.5). With its science data platform, PUNCH4NFDI intends to accommodate all the varying data types, formats, and sizes.

Classification of PUNCH data: Although PUNCH data are very diverse in character, they can roughly all be categorised into one of the following **classes or abstraction levels**:

- At the lowest level of abstraction, the instruments of PUNCH physics typically produce **raw ”data streams”**: data that are directly delivered by the detector of an experiment or facility or — in the case of simulations — by a simulation algorithm. Frequently, these data streams are too large in volume and therefore cannot be fully preserved owing to storage capacity or bandwidth limitations.
- To cope with these limitations, **”online” or ”real-time” selection** steps close to the hardware, and at timescales ranging from nanoseconds to a few minutes, are applied to the data streams according to a variety of science-specific criteria (so-called ”triggers” in particle physics, ”correlators” in radio astronomy, etc.). This introduces an unavoidable irreversibility and even **loss of information** (see TA 5 ”Data irreversibility”, section 5.5).
- The ”raw data” from the initial data streams that pass through these first levels of data reduction and on-the-fly analysis constitute the **lowest level** that is, at least temporarily, **stored by the experiment**. The raw data are often ”ear-marked” according to the physics purposes for which they have been (pre-)selected — their metadata are enriched with information on the precise trigger / selection criteria that have caused them to be the initially selected.
- In further steps — at more advanced levels of data treatment — raw data are reconstructed (i.e. filled with aggregated and higher-level information) and (partly) analysed, often leading

to **objects that are open for direct physics interpretation** like tracks or energy clusters in experiments of particle or hadron & nuclear physics, or corrected images of the sky.

- At high levels of reconstruction and analysis, particular data (specific collision events in particle physics, certain parts of images in astronomy) might be selected following stringent criteria for very specific purposes (”Higgs candidate events” of the ATLAS experiment at the LHC, or ...). These **classified or categorised data** are a core ingredient of physics publications.
- Finally, for publication purposes, the classified and other ancillary data (e.g. theory calculations or simulations) are used to generate **scientific knowledge** like e.g. cross sections or individual measured quantities like couplings or masses.

How FAIR are PUNCH data? The above classification of PUNCH data — with different and enriched metadata sets at each level — and the underlying life cycle of data in PUNCH physics imply a very diverse set of data at different abstraction levels. Consequently, the **FAIR principles** [1] are **realised to very different degrees**, depending on the sub-community, the abstraction level, the curation status, etc. In short, the degree of compliance with FAIR principles depends on the location of the data collection in the data landscape.

Interestingly, though, the PUNCH community is characterised by a very open and collaborative spirit — **openly sharing tools and data** has a long tradition in the field. An example are data (and their associated metadata) collected by astronomy satellite missions, which regularly are made publicly available. In this environment open data and open science go along well with the community traditions (see section 5.6). In the following, the degree of realisation of the principles is briefly discussed. Figure 3 illustrates the status of the FAIR principles over the four PUNCH sub-communities: particle physics, astroparticle physics, hadron & nuclear physics, and astrophysics.

In **particle physics**, some aspects of the FAIR principles are well realised. The raw data that are gathered by the scientific instruments, the lower-level derived data, and the corresponding simulations are typically only available within the producing collaborations. Within a collaboration, the data and analysis software are freely available for all researchers. Nevertheless, there are efforts ongoing to gradually release more and more (older) data together with the required software to allow re-analyses by non-collaborators. There is also a long-standing tradition to **release higher abstractions of the data (and relevant simulations** (distributions, histograms, likelihoods) electronically in searchable public archives, e.g. `hepdata.net`. These archives are connected with databases of research results such as `inspire.hep.net`. Also, high-level (statistical) analysis and simulation software is freely available.

For **hadron & nuclear physics**, the same holds true as for the particle physics community in many experiments such as the ALICE experiment. The smaller hadron & nuclear physics communities still require significant development in order to fulfil the FAIR principles. The future experiments at the *FAIR* accelerator facility at GSI, finally, have not yet started to collect data.

For **astronomy**, the requirement of sharing and using data collections from archives and sites across the world began long ago and has accelerated in the digital age. For sharing (initially

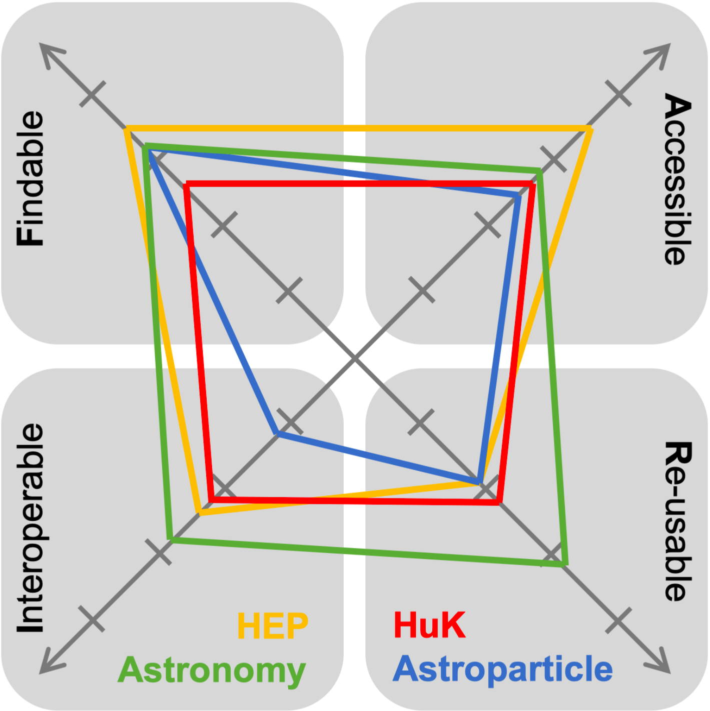

Figure 3: Visualisation of the degree of FAIRness in PUNCH today. See the text for details.

satellite) data, the **FITS format** (with a limited and controlled metadata vocabulary) has been established as a ubiquitous data storage format in all astronomy. The TIFF image format, widely used in public life, is an immediate offspring of FITS. With bigger data collections, the online selection on the one hand side (using databases and SQL) and the notion of a ”multi-wavelength view” on the other started the **Virtual Observatory initative** (2002), whose goal has been to develop standards and protocols for the **exchange and re-use of digital data collections**. Many astronomy specific standards include or extend **common metadata systems** (e.g. OAI-PMH). Its focus has been the **service paradigm**, and one of its main achievements has been the specification of a data access layer with protocols and standards. Especially in optical astronomy, newly generated data collections use at least partially VO recommendations for metadata or data access protocols even with non-public data. Also, the development of open-source tools such as astropy and pyVO have boosted the use of VO protocols. Often, public access to data is restricted only for a short limited period, but the curation of data collections (organised as ”data releases”) often extends over years.

The degree of implementation of the FAIR principles to **astroparticle physics** data lies in between those from particle physics and astronomy. Lower data levels and the associated software packages are findable and accessible for researchers inside the corresponding collaborations. Notable exceptions are e.g. the open KCDC providing public access to raw data and software tools. Final data products in reusable data formats (e.g. FITS) are often connected to the publication through the ADS service and accessible through the collaboration’s websites. The gamma-ray community is in the process of moving towards the usage of open community software (e.g. Gammapy) and data format standards on the calibrated event level (”open gamma-ray data format”), which enables one to publish these data and combine multi-messenger datasets from different observatories in a consistent manner. Observatories under construction (e.g. CTA) will follow FAIR principles in their research data management.

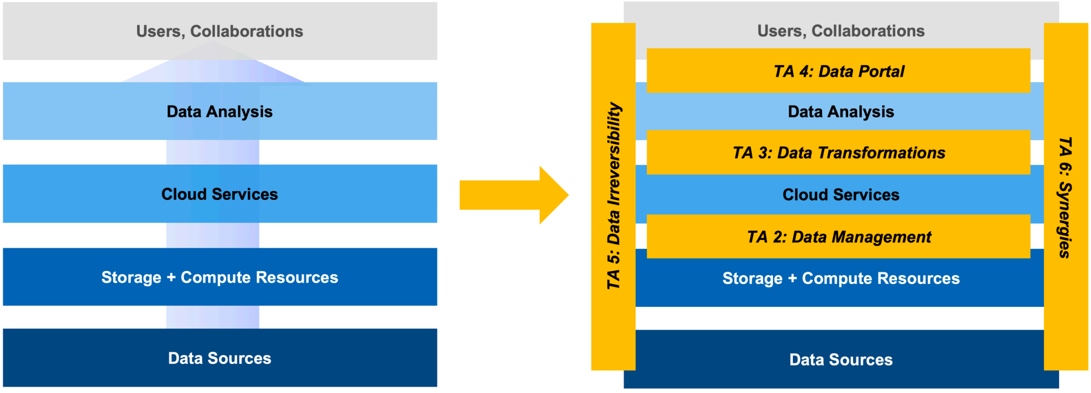

Figure 4: Left: envisaged layered computing architecture of PUNCH4NFDI. Right: assignment of task areas.

The layered model of data management: A classical approach for developing complex IT systems is to design a stack of layers. The layers are independent of one another and interact only with adjacent layers via a defined **set of interfaces**. An essential advantage of this concept is that making changes within a specific layer does not affect the entire architecture.

Given the success of this approach, a **layered architecture** has been adopted by PUNCH4NFDI, see the left-hand side of fig. 4. In the large communities of PUNCH, a two–layered infrastructure is already realised by a separation between high-level ”data analysis” tools and access to low-level ”storage and compute resources”, whereas the infrastructure of smaller communities tends to be rather monolithic.

To cope with the increasingly heterogeneous infrastructures within the computing centres, it is suggested to provide further layers, see e.g. [2, 3]. Likewise, PUNCH4NFDI introduces a layer in between called ”data transformations‘’, which consists of a set of algorithms and workflows that interface with the portal layer above and the data management layer below, enabling ease of use within the complex computing environment described above. Realising a three–layered infrastructure is challenging and needs close cooperation with international efforts in this area. Smaller communities are typically using frameworks with special demands that can best be met by a rather thin layer in-between, i.e. by an effectively two-layered architecture.

From use cases to deliverables and task areas: The PUNCH4NFDI consortium has taken great care to evaluate the current and future needs of its communities and to tailor its activities accordingly (see fig. 5):

– In a first step, **use cases** in the field of scientific computing and data management were **collected from the entire community**. This has resulted in a lot of feedback; all in all, close to 100 use cases were contributed, giving a very detailed picture of what PUNCH4NFDI

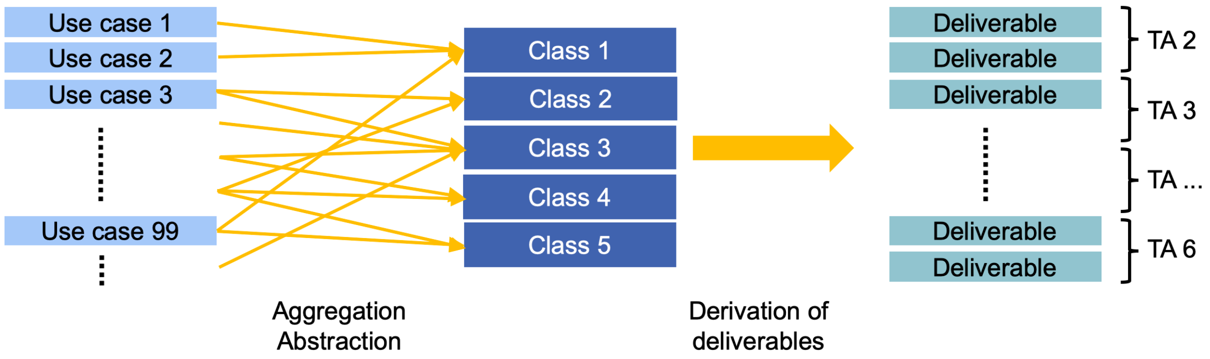

Figure 5: The process from use cases to deliverables. See the text for details.

either considers as typical workflows or wishes to see realised as such[4]. This investment of effort ensures the **scientific relevance** of the activities planned by PUNCH4NFDI as well as their added value.

- The **use cases typically contain descriptions** of the intended process or project, the addressed and involved sub-communities or activities, the required expertise, the current status of the fulfilment of the use case plus currently available assets in terms of tools, expertise, etc., detailed specifications (necessary developments, required resources like software, hardware, databases, services, necessary interfaces, ...), and potential applications outside of the PUNCH community.
- In a second step, a dedicated working group within PUNCH4NFDI scrutinised the use cases and carried out an **aggregation and abstraction process** by identifying conceptual and technical commonalities among the use cases and defining shared underlying principles. This process led to the definition of five **use case classes** described below.
- The **actions necessary** to fulfil the use cases and use case classes have been further broken down and cast into accessible task definitions with associated schedules, which in the following are called **”deliverables”**.
- In a final step, the deliverables have been localised in the conceptual ”layer” model, fig. 4). This leads to a natural **clustering of deliverables** at certain points or interfaces in the model, implying a bundling of closely related deliverables into work packages and, finally — at a higher level — into **task areas**. The TAs defined in this way are indicated in yellow in the right hand side of the figure.
- The TAs together with their associated WPs define the work programme of PUNCH4NFDI this will be discussed in detailed in section 5.

The layers indicated in blue–grey in figure 4 form the infrastructure ”backbone”; they are to a large extent fixed externally, mostly by internationally defined community practices. PUNCH4NFDI with its TAs, WPs, and deliverables, provides the package that aims at optimally harnessing the available resources and infrastructures in this environment.

Use case classes: In the following, the five use case classes are briefly introduced. The

[4] See `https://www.punch4nfdi.de` for a complete list of PUNCH4NFDI use cases.

connection of the so defined use cases classes to the PUNCH4NFDI TAs is discussed below. Their connection to the FAIR principles is the subject of subsection 4.3.

Use case class 1: Validating and publishing scientific data collections

Scientific data collections produced by individuals or science collaborations are more valuable when made available to the broader community in a manner consistent with the FAIR principles. Doing so requires efficient tools for making the data accessible through standard protocols for selection and retrieval. Many data collections published by their collaborations need integrative work. Other data collections need assistance in curating and publishing. This includes metadata management, minting digital object identifiers (DOIs), and integration into larger collections according to FAIR standards. Furthermore, infrastructure is needed to create workflows that can be used for vetting these data collections for publication and subsequent analysis. The availability of these data collections would then enable new cross-experiment/cross-collaboration data sharing, leading ultimately to new scientific discoveries.

PUNCH4NFDI will **build the PUNCH-SDP** to offer services that improve the usability of the data collections by translating and bridging different metadata domains within the consortium and with all of NFDI (TA 4 and TA 6). Selection of suitable data for scientific work and finding and applying scientific tools developed by TA 3 will be enabled. For data providers, PUNCH4NFDI offers assistance to publish their data using FAIR standards. This also is part of the TA 6 work of promoting the use of PUNCH4NFDI data collections by all natural sciences.

Use case class 2: Analysing local or distributed data sets

The analysis of distributed or local data sets constitutes a core element of the scientific enterprise. These analyses are enabled or made more efficient through the deployment of a data management system that enables data finding, efficient data movement, and the execution of compute jobs to transform the data into science-ready data products or final results. Experience shows that it is crucial for users to be able to control and monitor their analyses through an **easy-to-use portal interface**. Coverage of the use cases of this class requires the support of **three types of analyses**: i) small to intermediate datasets are downloaded from an archive node for analysis on a local compute platform; ii) analyses where the local compute resources are not sufficient require moving the data to a large-scale, shared compute resource (HPC/HTC) for analysis; iii) extremely large datasets — where data movement becomes excessively costly — adopt the so-called *code-to-data* approach where the analysis job is performed at one or more archive nodes, each with an associated high-bandwidth coupling to powerful computing resources. The workflows required for these analyses will often consist of a complex hierarchy of jobs and, as demonstrated in the contributed user stories on the PUNCH4NFDI web site, will require domain-specific data access tools and a range of transformations, including forefront statistical, numerical and machine learning methods.

By constructing the PUNCH4NFDI data portal, the consortium will not only enable this use case class for our PUNCH communities but will also provide a working solution for other communities to adopt and extend. In particular, the tools developed in TA 2 (data management), TA 3 (data transformations), TA 4 (data portal) together with the TA 6 services all come together to enable a broad and flexible solution that will enable the achievement of the majority (if not all) of the user stories within use case class 2.

Use case class 3: Executing and analysing numerical simulations
Scientific understanding requires an underlying theory that explains or interpret the origin of the observed phenomena. Within the PUNCH community, the theoretical contribution is provided in the form of numerical simulations that often produce large quantities of data. Since the production of these data is **computationally expensive**, it is highly desirable that data production be optimised and data maintained for re-examination and analysis by others. Currently this is the exception rather than the rule because — unlike for the experimental/observational side of these research areas — simulations are often performed by small groups or even individual researchers that do not command the necessary resources for providing these services. Therefore, additional structures and tools are needed to support the optimised use of the simulation data and to enable the analysis of these simulations through common sets of tools. Besides, the **direct accessibility of simulation data** is essential for a high-quality comparison between experimental/observational and theoretical results.

PUNCH4NFDI will provide several actions to fundamentally improve the current situation. Namely, it will optimise the **performance of the most commonly used codes** in the PUNCH field (TA 3). PUNCH4NFDI will also provide easy-to-use tools that allow the publication of data directly from the compute platforms (TA 2). This also requires the development of metadata standards (TA 4). Once this is achieved, the experiences gained by PUNCH4NFDI can be transported to other research fields via the TA 6 marketplace.

Use case class 4: Community-overarching data challenges
This use case class addresses community-overarching combined data analysis. For many scientific questions, data from various experiments are being joined to probe for physical effects. A large amount of research data of different origin and without common structure is produced by experimental and theoretical groups that cover very diverse topics. Often there are specialised experiments recording precision data for specific properties. These data are often useful in further scientific studies and applications, but in many cases are not available within a FAIR context. The main challenge in realising this use case is then to make such preprocessed datasets available in a format that will enable a common analysis. This can be achieved through the introduction of **structured data management tools**. This requires the development of highly federated storage technologies that meet the required scales and provide high bandwidth connections to the compute resources. Currently every experimental collaboration maintains their own specialised code. Combinations of datasets are only performed within the collaborations. The methods to perform combined fits across experiment/collaboration boundaries have been developed in principle, but the field is currently slowed down by the lack of universal access to the required datasets across communities. A cross-experiment analysis with combined datasets would allow one to reduce systematic uncertainties and has the potential to provide new physical insights. The main ingredient is an interface that allows experimental collaborations to **export their data in accessible formats**. The most general and useful format would be that of an unbinned likelihood, but in some cases the joint analyses would require direct joint analysis of the

lower-level data products. In both scenarios, the physics model has to be formulated in a way that allows common fit parameters across the global analysis. This use case class is addressed by TAs 3, 4, and 6.

Use case class 5: Real-time challenges, data irreversibility
In experiments of high-energy physics and upcoming astronomical observatories, data are taken at rates much too high to be stored in long-term archives. Use case class 5 addresses the challenge of extracting, in real-time, the tiny subset of ”interesting data” out of huge data streams in an automated way. The tools and methods currently used are not sufficient to cope with future demands. Firstly, the **data rates are increasing** even further. Secondly, the size of the outcome of single measurements may be as large as one Petabyte, and these ”data monsters” cannot be analysed in a reasonable amount of time by traditional computing systems. Thirdly, a continued increase of data volumes will result in a drastic increase of energy consumption: by 2030, approx. 20 % of the **worldwide power consumption** will be due to IT needs (according to a study published in *Nature* in 2018) [4]. The information content that can be extracted from most of the archived data is very small (”stored & forgotten”). A crucial challenge of overarching relevance is to identify the subset of data that are ”of interest” during their creation as well as by re-analysing archived data and to do this in an automated way. The use cases of this class require the exploration and development of methods for coping with challenges arising from **”data irreversibility”**. Decisions on what part of the incoming data streams need to be rejected have to be taken in real-time, and the ensuing unavoidable loss of information is mostly not reversible. The use cases cover different parts of the *dynamic life cycle model* (see figure 9 that indicates the interconnections between the work packages of TA 5). The subset of use cases that is realised in PUNCH4NFDI will provide interfaces that are eventually compatible with the requirements of TA 2 and TA 3.

Use case classes and task areas: Figure 6 summarises the intricate connections between use case classes and the work programme in the form of TAs. Use case class 1 (publishing scientific data collections) primarily uses the PUNCH4NFDI data portal (TA 4). Use case class 2 (analysing distributed data sets) requires massive compute resources and therefore primarily addresses ”Data management” and ”Data transformations” (TA 3 and 4). Use case class 3 (simulations) also requires ”Data transformation” (TA 3) because of the simulation work package therein. Use case class 4 (overarching data sets) requires ”Data portal” (TA 4) and especially to the work package ”Interfaces”, with further impacts on other TAs, including especially the work package ”Methods for analyses across datasets” of TA 3. Last but not least, use case class 5 (real-time challenges) is mainly addressed by task area ”Data irreversibility” (TA 5).

The colour layout of the figure visualises the **correlation between TAs and use case classes**. A 1-to-1 mapping cannot be expected as the topics addressed in the use case classes are based on existing or envisaged broad workflows covering many aspects of PUNCH4NFDI work. It is therefore fair to say that the structure of the TAs is compatible with the overall intention of the use case classes and thus reflects the general needs of the PUNCH community.

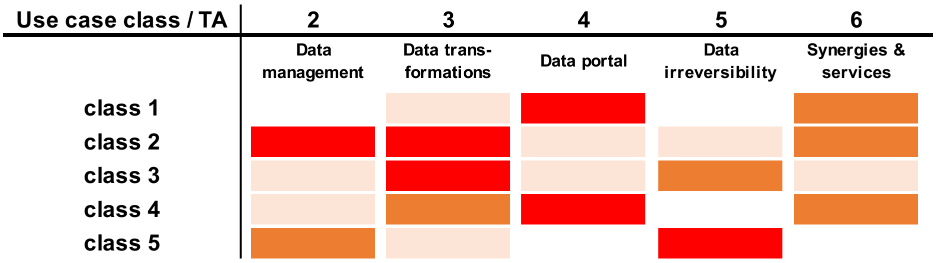

Figure 6: Overview of the connection of use case classes with TAs. Darker colours indicate stronger connections.

4.2 Metadata standards

The data landscape of PUNCH is very diverse, and even more so are the systems used to standardise and record metadata. Many advanced community- or experiment-specific approaches exist, albeit with only limited scope. The NFDI process offers a chance to take a fresh look at issues of **metadata and their use** within the PUNCH community. It quickly becomes apparent that there is **no common understanding of the term ”metadata”**. For a productive discussion of metadata standards, the following **distinction** shall be used:

- ”Data” in the PUNCH context are a collection of data points, typically numbers.
- ”Metadata” are expressions that assign meaning to the data points of the data — a data point without an explanation (i.e. without metadata) is meaningless and cannot be considered valid scientific data.
- A ”data collection” is a collection of data points with a set of descriptive metadata.

The generation of data collections in PUNCH physics has been described in detail in section 4.1. The data considered almost always come in two types: the signal data and ancillary data describing environment, observing conditions for observations, or initial conditions and parameters etc. for simulations as the starting point. Both **signal and ancillary data** are data collections, i.e. data points with metadata. The used metadata can either be taken from a fully developed ontological system or are very specific descriptions of data points for one instrument, and they are not always in ”machine readable” format. Each PUNCH sub-discipline (particle, astro-, astroparticle, hadron & nuclear physics) has developed its own nomenclatures and systems of metadata best-suited for their domain. However, one very generic system of metadata — the physical units — is used by all, but is not sufficient for the purpose of providing FAIR data collections.

The **access to the relevant data** collections of the PUNCH community — and the role of metadata in this process — can take different forms, each offering its own problems and limitations:

- First, on the file level there are data formats into which metadata can be incorporated: The FITS file format from astronomy, for example, combines metadata (with a formal metadata system) and data points. Some limitations of FITS have led to the suggestion of a newer format (ASDF). In particle physics, the ROOT format is frequently used; the ROOT file internal structure resembles a UNIX file system, its use being tightly bound to the ROOT software system. Both file formats — FITS and ROOT — share the problem that accessing their metadata requires reading the file first.

- Secondly, metadata may be encoded on the storage system level in various ways (via a metadata or attribute store, but even e.g. on pure file systems via file naming schemes). The ”data lake” (which operates on the file system level and improves the findability and the accessibility of data collections) also provides access to these metadata, which are encoded using the file system, or in object stores. The data lake is chiefly part of the solution for one of the ”big data” problems: Most of the PUNCH data collections are too big to efficiently move across the network for individual use alone.
- Thirdly, metadata (and processing or context information) can be tied to and implicitly be built into the access software for the data collection in question.

Some communites have already separated data and metadata, creating searchable metadata stores.

One of the main challenges for improving the FAIRness of PUNCH4NFDI data collections lies in finding ways to separate out sufficient metadata for the selection and transformation between the various data collections to use these in a common context. But this is only one part, and it is addressed in each of the TA 2–4 in specific ways. A further aspect of developing suitable metadata standards can be best described by the **data curation continuum** [5], which in PUNCH4NFDI has three broad segments:

- segment (1) is the generation of data collections, which may include initial data reduction (e.g. electronic filtering within the instrument or artefact removal);
- segment (2) is the validation and analysis process in a collaboration; and
- segment (3) is the publication of data collections for use by a broader range of scientists.

Within segment (1), the directly involved instrument builders are usually providing the data collections with metadata tied to the specific instrument. For collaborative use, segment (2), transformations of the data collections take place and a common understanding for the content of the data, its limitations and descriptions within the collaboration is established. This implies extended metadata usage, and implicit knowledge is built into analysis tools. By making the resulting data from analysis accessible to scientists outside the collaborations — segment (3) —, metadata need to be mapped to systems used by others, and implicit processes need to be made explicit. For segment (3), the FAIR principles are the best guidelines. For segments (1) and (2), longer timescales need are expected for the realisation of data FAIRness, since this process will only work by back-propagation of good solutions towards acceptance by the huge collaborations creating most data collections.

PUNCH4NFDI takes a pragmatic approach and will **develop metadata mappings and transformation routines** that satisfy the requirements of the use cases. The main focus is on achieving working solutions for using the existing and future data collections in new ways. To this end, some additional metadata sets with improved standardisation are required to describe results of processing data sets, including e.g. provenance information. For these, as much as possible of the available common metadata systems will be used. This will lead to **extensions or adaptations of metadata schemes** within PUNCH4NFDI , working with terminology services and using expertise from TIB and common efforts within NFDI.

The use of DOIs for identifying *publicly accessible data collections* has begun in all scientific disciplines including PUNCH. DOIs provide a unified means to connect data to publications, but only provide a very abstract layer of content characterisation and PUNCH4NFDI promotes their usage to improve findability of data collections. For this, PUNCH4NFDI will work out solutions that provide back-pointers to the richer metadata collections of PUNCH data to enable discovery mechanisms, e.g. to search and discover data collections using physical parameters (see TA 6, WP 3) . The resolving mechanisms will be supported through the services of the data portal. On the other hand, defining a common set of DOI metadata [6] for published data collections of PUNCH connects with all disciplines and supports one of the main goals of the NFDI.

4.3 Implementation of FAIR principles and data quality assurance

Findability, **A**ccessibility, **I**nteroperability, and **R**eusability of digital datasets are a cornerstone of the PUNCH4NFDI efforts. Besides, their machine-actionability with no or minimal human intervention is a central aim. Based on considerable, although often quite heterogeneous, experience in FAIR in PUNCH sciences (see section 4.1 and figure 3), PUNCH4NFDI strives to achieve maximum compliance with the FAIR principles.

Findability: Data and metadata need to be easy to find, both for humans and for computers (requiring machine-readable metadata for automatic discovery). Several schemes for **connecting data and scientific publications** exist already in the community. Examples are ADS (including hints towards the used data collections and software), or the hep-inspire portal. Close partnerships between PUNCH4NFDI and specific journals/publishers have been established, and together, guidelines for data publication will be designed (D-TA6-WP3-1 and D-TA4-WP1-4).

Findability of data crucially depends on the underlying curation processes. Currently, the degree to which published data associated with publications are findable varies tremendously. For example, there are ongoing efforts within IVOA to incorporate DOIs into their registry of data services to make it easier to efficiently find the respective metadata. However, for particle, hadron & nuclear physics, only a fraction of the data associated with publications are published on portals, despite the fact that the underlying infrastructure often is available (e.g. the classic grid frameworks used in particle physics are all able to schedule jobs according to data locality).

The future PUNCH user's **starting point** for finding data will be **the PUNCH data portal**. Several measures will establish the envisioned functionality: digital research products will be enriched with descriptive metadata (D-TA4-WP1-3) and followed by dynamic ingestion and curation processes (D-TA4-WP1-2). New metadata schemes will be developed and existing metadata standards will be extended (D-TA4-WP2-1/2). D-TA4-WP3-2 integrates these achievements into the data portal. In data processing, analysis algorithms can be matched automatically to data but this requires the development of data-locality aware scheduling in overlay batch systems in the PUNCH4NFDI federated science cloud (D-TA2-WP2-7). Besides, new and improved data placement and replication mechanisms in data lakes (D-TA2-WP3-4) and improved monitoring systems for federated storage infrastructures (D-TA2-WP3-1) are needed.

These developments go hand in hand with the definition of the functionality of dynamic digital research products and their interfaces (D-TA4-WP1).

TA 6 ”Synergies & services” carries these developments over to offer the complimentary services for open data archives (D-TA6-WP5-1). Numerous **collaborations on metadata** have been arranged with other NFDI consortia (MaRDI, NFDI4Ing, NFDI4Microbiota, NFDI4Culture — see section 3.2). The use of terminology services from TIB or the HMC will ensure compliance with NFDI developments.

Enhancing the FAIRness of PUNCH datasets for public accessible data collections requires considerable efforts for assessing and registering the published data collections and their metadata systems. While several data providers (particle physics data centres, astronomy data centres, etc.) use DOIs, the **DOI metadata** often have very different depth and utility. PUNCH4NFDI will work towards standard agreements on metadata for DOI, including eligible entries to use for drilling down into the data collections using their specific (or adapted) metadata.

Accessibility: A key requirement of the PUNCH-SDP is data availability, i.e. fast and simple access to the relevant data and associated metadata in order to perform the intended research tasks. In a second step, researchers from other science areas and the interested public should also have data access — public data for public money! Accessibility thus relies on well-defined authentication and authorisation mechanisms and standardised, open and free, universally implementable communication protocols. These give access to the relevant metadata on the analysis process together with the processed data. Ideally, FAIR data should be retrievable by anyone with a computer and an internet connection, using a well-defined protocol — if the user is authorised — also restricted-access data can be FAIR!

To improve accessibility, PUNCH4NFDI will develop **authentication and authorisation interfaces** (AAI) across many network domains (D-TA6-WP2-1). For cloud-based environments, advanced AAI (such as the token-based authentication for WLCG) are evolved towards compatibility with the requirements of communities that use the OAuth2.0 protocol. TAs 2 and 5 will devise standardised protocols for data access and for dynamic archival and time-critical feedback to data. PUNCH4NFDI will benefit from its extensive experiences in developing standardised protocols in the context of virtual observatories. Furthermore, the prototype data lake implementation (D-TA2-WP1-1) will be optimised for data accessibility.

Interoperability: PUNCH data need to be integrated with external data sources and must interact with a broad array of applications and workflows for analysis, storage and processing. Incredibly complex challenges arise from experiment-overarching analyses and the combination of complementary datasets; such use cases challenge the interoperability, flexibility, and scalability of any system and need to be tested and demonstrated on a large scale.

PUNCH facilities implement a broad array of computing and storage models, which makes the need for interoperability even more pressing. Indeed, interoperability is a prerequisite for an overarching science platform as envisaged by PUNCH4NFDI. While the overall layout of workflows for data access and analysis are quite similar across the entire PUNCH field, many concepts are unknown in other fields of science. Even within PUNCH science, deviating approaches to interoperability are taken. Beyond the PUNCH community borders, the degree of interoperability decreases further. PUNCH4NFDI will focus efforts and collaborate with partners like TIB and other NFDI consortia.

In particular, PUNCH4NFDI will develop **interfaces that allow the combined analysis of datasets** from different data sources and experiments (see D-TA4-WP3-3, WP 3.4 in TA 3) and the publication of research products together with stored data and interoperable analysis workflows (D-TA4-WP4-2). In general, PUNCH4NFDI will improve interoperability following the model of the Virtual Observatory: defining and implementing interfaces and protocols for data publication, data selection and retrieval.

Re-usability: Scientific progress is increasingly relying on **data archives** — without access to (published) data collections, little scientific work is feasible. In astronomy in 2018, for example, more than 50 % of publications based on Hubble Space Telescope data actually used the HST archival data rather than new observations. Consequently, the pressure is high (also from the funding agencies) to make data publicly available and to make them adhere to common standards. This is already the case for many published datasets — albeit with a large variance.

Having a clear policy for **providing licenses** is crucial for data re-use. Data re-use requires data and software to be released under clear conditions for both humans and machines. As a service developer and provider, PUNCH4NFDI aims for a **consistent application of open licences** like CCO waivers for data and documentation, and open-source licences (GNU/GPL, BDT, Apache2.0 being the most important ones) for software.

For massive data streams the situation is even more challenging because **new concepts of metadata** need to be developed to **ensure reproducibility of scientific results** under these conditions (TA 5). Rich metadata will have to provide information on how a given dataset has been created, calibrated, and reduced. Provenance information should be readable by both machines and humans. A **provenance standard** for PUNCH data must be established to be compliant with the W3C provenance approach. Currently, however, there are only pilot implementations of this standard, e.g. the IVOA provenance standard [7] or the lattice grid (e.g. D-TA6-WP3-1). For an existing data collection made public by PUNCH4NFDI, collaboration with the data provider will ensure that it meets a basic set of requirements, including a defined DOI minting process. The data quality assurance of existing data collections has to be left to their producers. However, PUNCH4NFDI will work towards an improved use of the DOI metadata.

For the PUNCH-SDP, research products will be defined that contain pointers to all data, parameters, software and other necessary information to allow unpacking and re-execution of its content through the data portal. They will enhance re-usability and ensure a **high level of quality assurance** for all components of the PUNCH science research products. Deliverable D-TA4-WP4-2 focuses on the publication of research product examples with stored data and interoperable analysis workflows. In this context, also the reference guide for publishing software (D-TA6-WP3-2) and the example repository for open software development (D-TA6-WP4-2) will significantly contribute to enhance re-usability.

The goal of PUNCH4NFDI is to harmonise and extend the FAIR status across the entire field.

Such an effort will have to address **education and training activities** as described in TA 7, so that scientists will become better acquainted with FAIR data management.

Table 1 provides an overview of how the PUNCH4NFDI TAs and WPs address the FAIR principles[5].

Table 1: Relation of FAIR principles and PUNCH4NFDI TAs and WPs

	Explanatory text	**TA.WP**
Findable		
F1	(Meta)data are assigned a globally unique and persistent identifier	2.1, 3.2, 3.4, 4.2, 6.3
F2	Data are described with rich metadata	2.1, 4.2, 5.1, 6.3
F3	Metadata clearly and explicitly include the identifier of the data they describe	4.2, 4.4, 6.3
F4	(Meta)data are registered or indexed in a searchable resource	4.4, 2.1, 6.3
Accessible		
A1	(Meta)data are retrievable by their identifier using a standardised communications protocol	4.4, 2.1, 6.3
A1.1	Protocol is open, free, and universally implementable	2.1, 4.2
A1.2	The protocol allows for an authentication and authorisation procedure, where necessary	2.1, 4.3, 6.2
A2	Metadata are accessible, even when the data are no longer available	2.1, 4.4, 5.1, 6.3
Interoperable		
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language	4.2, 2.1, 6.3, 6.5
I2	(Meta)data use vocabularies that are FAIR	6.3
I3	(Meta)data include qualified references to other (meta)data	4.2, 5.1, 6.3, 6.5
Reusable		
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes	4.2, 6.3, 5.1
R1.1	(Meta)data are released with a clear and accessible data usage license	6.3, 4.4
R1.2	(Meta)data are associated with detailed provenance	2.1, 2.3, 3.2, 3.4, 4.3, 6.3, 5.1
R1.3	(Meta)data meet domain-relevant standards	3.2, 3.4, 4.2, 6.3

4.4 Services provided by the consortium

As a result of the cooperative work in the different TAs, PUNCH4NFDI will provide several **large-scale distributed services for the processing and management of scientific data** to the users (see also sections 3.1, 3.2, 4.1). These services will support the users in making their data available, accessing data provided by other researchers, extracting information from the data, and transforming data efficiently. Performing these tasks will be facilitated by the availability of highly informative metadata. The ample experience of the PUNCH community in operating services is a strong point. Specifically, PUNCH4NFDI provides the following services (in brackets, the related deliverables are given that are described in section 5):

– **PUNCH central infrastructure** (see section 2.2): Central to PUNCH4NFDI is comfortable

[5] See the definition at https://www.go-fair.org/fair-principles/.

access for the users to all the provided services, offered via the **PUNCH-SDP**. All other services mentioned below are an integral part of this single, dedicated platform. Through the marketplace and the **data portal** these services will be published, including the AAI prototype (work packages WP 4.1–3, 6.2).

- **Data access and management services:** Access to PUNCH4NFDI open data archives such as the open data of CERN's LHC experiments and the IVOA (D-TA6-WP5-1), but also smaller data collections will be offered in the **data lake**. Dynamic disk cache technology for the integration of opportunistic storage resources (D-TA6-WP5-2) will be provided (D-TA2-WP2-6). Data processing will be supported by FTS and Rucio services for evaluation purposes (D-TA6-WP5-10). Specifically, corresponding reference guides will support data producers in publishing their data and software (D-TA6-WP3-1, D-TA6-WP3-2).
- **Analysis software and data irreversibility services:** For the extraction of high-level information, tools can be found in the PUNCH **software repository** (D-TA6-WP4-2). Among others, this contains data analysis and simulation routines optimised for multi-GPU systems (D-TA3-WP2-1), a framework for AutoML on scientific data (D-TA3-WP3-1), and a framework for conversion/reading of data for combined analyses on heterogeneous systems (D-TA3-WP4-1). For real-time applications, algorithms optimised for sorting, hardware-specific clustering and pattern recognition (D-TA5-WP2-5), and methods for transforming dynamical archive queries into dynamic filters (and vice versa, D-TA5-WP3-2) will be developed, as well as a generalised toolkit for predictive maintenance and anomaly detection (D-TA5-WP5-3).
- **Metadata services:** The possibility to build from the research products of PUNCH4NFDI a dynamic **knowledge fabric** is the next step. The definitions of their functionality and interfaces (D-TA4-WP1-4), prototype metadata schemes, data formats and published and interoperable digital research products using the full range of PUNCH4NFDI services (D-TA4-WP4-4) (D-TA4-WP2-1) are provided. The corresponding catalogue technology (D-TA4-WP1) will be accessible for all interested users through the PUNCH-SDP.
- **Computing and storage resource services:** For efficient data management and analysis, IT resources as provided by **Compute4PUNCH** are needed. Therefore, interfaces to existing infrastructures are provided, for example, to the supercomputer HLRN (D-TA6-WP5-4). Additionally, small fractions of community-specific resources jointly managed by the COBalD/TARDIS compute resource management software framework will be made available to PUNCH4NFDI and beyond for data analysis (D-TA6-WP5-5). Finally, prototyping interactive analysis via multi-cloud resources will be facilitated for interested users (D-TA6-WP5-5 and D-TA2-WP3-4).

Future work of the consortium will lead to more services that can be added to the portfolio. Many of the relevant services are transferable to other domains and can thus serve as a **blueprint for other communities** that are dealing with similar challenges. TA 6 "Synergies & services" has been established to create **clearly defined communication channels** and strengthen the collaborative spirit inside PUNCH4NFDI and towards the NFDI and its other consortia (section 5.6). In particular, the PUNCH4NFDI ”marketplace” (WP 6.1) will be instrumental in this regard.

PUNCH4NFDI has defined a workflow for offering services: Each TA will identify deliverables that have the potential to become services. After accomplishing a deliverable in question, it will undergo a **prototype/alpha-testing phase** together with an extended circle of developers and dedicated expert users. In the following **beta-testing period**, additionally, the TA 6 (marketplace) together with interested users from the target community are involved. In case of a successful evaluation, potential additional development work identified in the evaluation process will be carried out by the involved TAs. Services that can already now be offered by PUNCH4NFDI — e.g. such services that are already in use by the community — will undergo the same evaluation steps at a later stage.

Services offered by PUNCH4NFDI need to be maintained and documented, and **user support** provided. For this purpose, a sufficient amount of personnel is foreseen in the marketplace and the corresponding work packages of TA 6. Also the TA which initially developed the service will delegate personnel to this task. TA 7 will provide training in the use of the newly developed services in the PUNCH Young Academy (section 5.7) Some of the services offered by the consortium will require **dedicated hardware resources**. These will be taken to some extent from in-kind contributions of PUNCH4NFDI members, especially for the needs of the consortium itself.

5 Work Programme

This section lays out the detailed work programme of the PUNCH4NFDI project in terms of its task areas (TAs) and the relevant work packages (WPs) and how it relates to the overall objectives of PUNCH4NFDI (see section 2.2). Responsible persons for each measure as well as the participating partners are given. Significant achievements and tangible results aimed for during the project are collectively called ”deliverables” and are enumerated to allow for easy scientific controlling. Where meaningful, and possible, dates for the achievement of deliverables are also given. Figure 7 on the next page gives a rough visualisation of the time dependence of the deliverables. The TAs and WPs are also summarised in table 2. The relations of the WPs to other NFDI consortia (rightmost column ”Transfer”) are explained further below in the text.

Table 2: Overview of the task areas, work packages, and responsible scientists. Responsible persons that are not co-spokespersons are given in italics (details on the PUNCH4NFDI web page `https://www.punch4nfdi.de`).

Work package	Responsible	Transfer
TA 1 ”Management and governance”: T. Schörner, M. Steinmetz		
WP 1.1: Management setup	T. Schörner	
WP 1.2: Financial administration	*project manager (N.N.)*	
WP 1.3: Controlling and reporting	*project manager (N.N.)*	
WP 1.4: Consortium organisation	*project manager (N.N.)*	yes
TA 2 ”Data management”: M. Hoeft, *C. Wissing*		
WP 2.1: Standardised access to data and metadata	*C. Wissing*, D. Schwarz	yes
WP 2.2: Compute4PUNCH	M. Hoeft, *M. Giffels*	yes
WP 2.3: Automatisation and optimisation of big data management workflows	K. Schwarz, *A. Brinkmann*	
TA 3 ”Data transformations”: *M. Brueggen, T. Kuhr*		
WP 3.1: Statistical methods	K. Kröninger, J. Mohr	yes
WP 3.2: Numerical methods and simulations	*F. Karsch*, S. Pfalzner	yes
WP 3.3: Machine learning methods	*M. Brüggen*, G. Kasieczka	yes
WP 3.4: Methods for analyses across datasets	*T. Kuhr*, J. Mohr	
TA 4 ”Data portal”: P. Bechtle, *H. Enke*		
WP 4.1: Digital (dynamic) research products and their catalogue	*G. Maier*, *S. Neubert*	yes
WP 4.2: Mapping and collating metadata	A. Haungs, *C. Urbach*	yes
WP 4.3: Implementation of interfaces	*A. Geiser*, *O. Kaczmarek*	
WP 4.4: Build and operate the science data portal	*H. Enke*	yes
TA 5 ”Data irreversibility”: M. Kramer, A. Redelbach		
WP 5.1: Implications for discovery potential and reproducibility	D. Schwarz, S. Wagner	
WP 5.2: Dynamic filtering	*T. Ferber*, M. Kramer	yes
WP 5.3: Dynamic archiving	*J. Nordin*, *L. Spitler*	yes
WP 5.4: Scaling workflows	A. Redelbach, H. Heßling	yes
WP 5.5: Evaluation and validation of instrument response ...	*B. Spaan*, *L. Spitler*	
TA 6 ”Synergies & services”: K. Schwarz, S. Wagner		
WP 6.1: Marketplace	S. Wagner, K. Schwarz	yes

WP 6.2: Authorisation and authentication infrastructure	K. Schwarz, *D. Mallmann*	yes
WP 6.3: FAIRness	S. Wagner, H. Heßling	yes
WP 6.4: Open-source data analysis tools	*G. Duckeck*, J. Hinton	yes
WP 6.5: Services in big data management	M. Steinmetz, A. Quadt	yes
TA 7 "Training, education, outreach, citizen science": *F. Bertoldi*, K. Kröninger		
WP 7.1: Training of scientists - Young Academy	A. Quadt, S. Pfalzner	yes
WP 7.2: Education of students	K. Kröninger	yes
WP 7.3: Public outreach	*F. Bertoldi*	yes
WP 7.4: Support for citizen science	M. Kramer, A. Haungs	yes

Explanation of relevance to other consortia

Results obtained in the various work programs will yield tools and services for research data management that will be beneficial for many other NFDI initiatives. TA 6 shall facilitate interaction and exchange between work packages of PUNCH4NFDI and other consortia. Discussions with initiatives listed in section 3.2 suggest that most of the deliverables of the PUNCH4NFDI WPs will be used in other initiatives.

Sustainability, Risks and mitigation

The description of each measure summarises status and goals of the corresponding actions. The associated risks and the risk mitigation strategy is explicitly mentioned. There are, however, a number of aspects common to many if not all task areas. These are addressed in the following:

- *Risk:* Recruiting. With the start of PUNCH4NFDI a considerable number of trained experts in data science will need to be employed, in competition with other consortia and a very competitive "market". *Mitigation:* TA 7 addresses this very point, as it will initially enhance the training of talents at PUNCH4NFDI institutions, later on, PUNCH4NFDI contributes to educating the next generation of experts.
- *Risk:* A shortage of storage or compute resources might occur in a specific WP. *Mitigation:* Storage/computing resources are pooled so that members of the affected WP can use with high priority the shared resources.

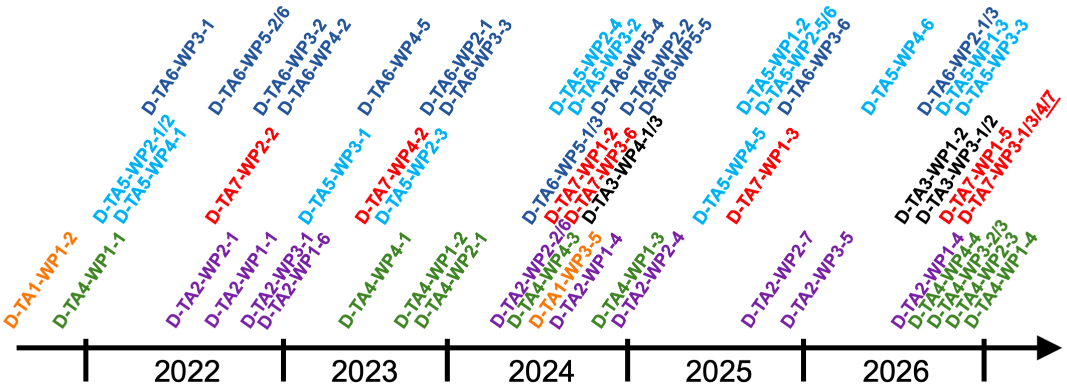

Figure 7: Overview of the deliverables of PUNCH4NFDI. Only the more important deliverables are indicated. Different colours indicate different task areas.

- *Risk:* Unforeseen technical difficulties might delay the provision of PUNCH-SDP. *Mitigation:* In this case already finished products and services are first offered on individual smaller platforms and later incorporated inPUNCH-SDP. This way, finished products can be made available early on for the user.

PUNCH4NFDI will develop a considerable number of services, for the PUNCH community as well as for other fields. With the end of the project funding, these services will have to be handed over to the responsibility of other entities.

Sustainability is a significant challenge for the NFDI as a whole, and solution strategies will heavily rely on the overall NFDI sustainability strategy, which is currently being developed. PUNCH4NFDI has access to a whole slew of external computing facilities and is strongly involved in national and international research infrastructures. Furthermore, several supercomputing centres participate in PUNCH4NFDI. The strong involvement of these centres will help to establish at least interim solutions for PUNCH4NFDI services until a sustainable NFDI structure has been established.

5.1 Task area 1: Management and governance

The PUNCH4NFDI consortium involves over 40 co-applicant and participating institutions and maintains close ties with another large number of, mostly international, infrastructures and resource providers. Close connections have already been formed to numerous other current and future NFDI consortia, and their number is expected to grow further as the NFDI as a whole is built up. The consortium will also command significant funds — both as requested from the NFDI and in the form of in-kind contributions (personnel, dedicated hardware resources). Consequently, many **managerial, financial and communicative tasks** are to be handled by the consortium. These are bundled in the four work packages of task area 1 ”Management and governance” that will be introduced below, together with their deliverables and milestones. The related budget request is summarised in a separate subsection.

The work programme in TA 1 is carried out mainly by personnel from DESY, with own contributions from GSI and KIT for community support in the fields of hadron & nuclear physics and astro- and astroparticle physics, respectively. DESY, as national laboratory for particle physics, has ample experience in managing large projects technically, administratively, and financially.

5.1.1 Work package 1.1: Management setup

The ”Management setup” work package of the task area will be the main responsible for setting up the necessary **structures for managing the consortium**. Consequently, most tasks of this work package have to be fulfilled at a very early stage, to a large extent even before the official start of funding (currently assumed to be 1 October 2021). Since at these early stages, the project manager of the consortium will most probably not yet be selected, many of these tasks will be handled by the current Executive Board and its spokesperson[6].

[6] Details of the PUNCH4NFDI governance are explained in section 3.4.

As a starting point for the official collaboration of all PUNCH4NFDI partner institutions, **memoranda of understanding** and contracts regulating the scientific and financial relations between the institutions need to be setup (D-TA1-WP1-1). The PUNCH4NFDI partners shall then, in a **"kick-off workshop"**, **establish the Executive Board** as the main governing body of the consortium board, and appoint a consortium spokesperson from the Executive Board (D-TA1-WP1-2). This can either be done by confirming the current group of people (see section 3.4) or by deciding to elect members of the board.

The Executive Board will then appoint the Management Board by either confirming the currently acting task area coordinators (see table 2) or appointing new ones. The Executive Board will also select a project manager (D-TA1-WP3-3), who in turn will set up the consortium office at DESY. Finally, the **User Committee** (UC), **Scientific Advisory Board** (SAB), and **Infrastructure & Resource Board** (IRB) will be composed by inviting the relevant external bodies (KAT, KET, KHuK, RdS), the relevant infrastructure providers, and the entire national and international communities, respectively, to nominate representatives (D-TA1-WP1-4).

***Deliverables**:*

- **D-TA1-WP1-1 (30 Sep 2021):** Memoranda of understanding and contracts.
- **D-TA1-WP1-2 (31 Oct 2021):** Kick-off workshop; establishing of Executive Board (EB).
- **D-TA1-WP1-3 (30 Nov 2021):** Setup of Management Board (MB); employment of project manager (PM); setup of consortium office at DESY.
- **D-TA1-WP1-4 (30 Sep 2021):** Establishing of User Committee (UC), Scientific Advisory Board (SAB), and Infrastructure & Resource Board (IRB).

5.1.2 Work package 1.2: Financial administration

In terms of **financial management**, the PUNCH4NFDI consortium can rely on the well-established and experienced DESY administration structures: the DESY administration is used to managing large national and international projects. Funds have to be transferred, reporting numbers have to be collected, and reports on the expenditure of funds have to be prepared. The two deliverables express this and will be tackled by DESY personnel.

***Deliverables**:*

- **D-TA1-WP2-1:** Organisation of financial relations among the partners.
- **D-TA1-WP2-2:** Financial reporting (collection of numbers, report preparation).

5.1.3 Work package 1.3: Controlling and reporting

The third work package is concerned with the **controlling and reporting** of the consortium's work. Here, the relevant monitoring steps are performed, mostly by the project manager, and the necessary reports are prepared, including the results of the financial administration from work package 1.2 (see above):

The PUNCH4NFDI work programme consists of a **set of deliverables**, the fulfilment of which documents the success of the consortium. Therefore, also as a basis for the regular reports to be published by the consortium, the project manager will closely follow up on each of the deliverables across all task areas and make sure that progress of the consortium's work is correctly monitored (D-TA1-WP3-1).

The PUNCH4NFDI partners **pledge significant own personnel and in-kind contributions** in the form of hardware or services. These resources are an integral part of the consortium's work programme; they are thus vital to the success of the consortium's planning. Therefore, great care will be taken by the project manager to ensure that sufficient resources are adequately delivered by the partner institutions (D-TA1-WP3-2).

The DFG and the NFDI will require **regular reports** on both the spending of funds and the achievements of the consortium. The preparation of these reports is another task of the project manager (D-TA1-WP3-3). Similarly, the project manager will be responsible for the preparation of an annual report for the consortium members and beyond (D-TA1-WP3-4).

Finally, in a project running over, initially, five years in a very dynamical environment such as the NFDI, care has to be taken not only that long-term goals are stringently pursued, but also that new developments and achievements are adequately considered in the planning of next steps and a re-arrangement of the long-term goals. To ensure this, the PUNCH4NFDI consortium foresees the **organisation of a mid-term review** (D-TA1-WP3-5), with the involvement of the UC, SAB, and IRB, and additional external experts. The review results will feed back on the orientation of the consortium in the final phase of the first 5 year period, but even more so on the setup of the consortium for a potential second 5-year funding period from 2026 onward.

Deliverables:

- **D-TA1-WP3-1:** Deliverable monitoring.
- **D-TA1-WP3-2:** Monitoring of own and in-kind contributions.
- **D-TA1-WP3-3:** DFG / NFDI reporting.
- **D-TA1-WP3-4:** Annual PUNCH4NFDI report.
- **D-TA1-WP3-5 (30 June 2024):** Internal mid-term review.

5.1.4 Work package 1.4: Consortium organisation

The day-to-day organisation of the consortium is handled in WP 1.4. In a first task (D-TA1-WP4-1), the project manager will ensure the **organisation of the meetings of PUNCH4NFDI bodies** (MB, UC, SAB, IRB).

The PM is also responsible for maintaining close ties to the management of other NFDI consortia, and he or she will actively promote the **integration of PUNCH4NFDI into the NFDI** with the strong support of this activities by the PUNCH4NFDI executive board (D-TA1-WP4-2).

The PM will also be responsible, in collaboration with colleagues mainly from the task areas 6 ”Synergies & services” and 7 ”Training, education, outreach, and citizen science”, for the **organisation of all kinds of events**: training events, workshops, outreach events, contacts to other consortia and the NFDI, etc. (D-TA1-WP4-3).

Finally, this work package maintains the **PUNCH4NFDI communication channels** inside the consortium and to the interested public (D-TA1-WP4-4). Elements of this are the PUNCH4NFDI web site `https://www.punch4nfdi.de`, the regular publication of a PUNCH4NFDI newsletter, and the organisation of social media.

Deliverables:

- **D-TA1-WP4-1:** Organisation of PUNCH4NFDI management meetings.
- **D-TA1-WP4-2:** Continuous integration of PUNCH4NFDI into the NFDI.
- **D-TA1-WP4-3:** Event organisation.
- **D-TA1-WP4-4:** Maintenance of PUNCH4NFDI web site, newsletter, and social media.

5.1.5 Sustainability, risks, and mitigation

The governance and management structures envisaged by TA 1 are well embedded into the overall German PUNCH community and are intended to last beyond the initial 5-year period of the consortium.

Sub-optimal management is a risk that could put the overall project at danger and thus requires an efficient early warning system. Mitigation: As a large organisation tasked with the management of numerous large projects, DESY offers ample know-how to assist the setup and management of PUNCH4NFDI even in crisis situations. Furthermore, the advisory structure (EB, UC, IRB, SAB) and the foreseen midterm review will support the project in identifying management issues early on.

5.2 Task area 2: Data management

This TA primarily addresses the integration and, where necessary, the development of **modern middleware components** (objective 6). As such it provides building blocks for the PUNCH-SDP (objective 1), while being well aligned with international research and development efforts (objective 4).

Many scientific goals of the PUNCH community can only be achieved if the most immense amounts of data, originating from experiments, observations or simulations, are managed and provided to the community for scientific exploitation. It is imperative for this field of fundamental sciences to advance the technologies and methods to **cope with the increasing data volumes** generated with the state-of-the-art research infrastructures. Moreover, it needs to be ensured that a community as broad as possible can access and analyse the obtained research data to maximise the scientific return.

A large part of the PUNCH community, especially the extensive collaborations that process immense amounts of data, are **already familiar with operating and using large, distributed data storage and compute infrastructures**. Presently, the largest globally distributed scientific IT infrastructure is the WLCG, which consists of more than 160 data centres worldwide, providing mostly dedicated resources for this particular community. As of today, about 800k CPU cores are running at any given time, and over 500 PB disk space as well as close to

800 PB archival (tape) storage are managed dynamically.

In the second half of the 2020s, when the High-Luminosity LHC (HL-LHC) will start, the disk requirements are expected to reach Exabyte scales. At roughly the same time, the SKA will produce a similar data volume, and the experiments at the FAIR facility will require storage capacities of several 100 PB. These, but also other, experimental and theoretical **PUNCH research efforts require extreme data volumes and computing capacities** that can only be fulfilled by making use of additional, non-community-specific resources such as HPC computers and clouds including special hardware such as GPUs, that are currently not part of WLCG and can not efficiently be utilised. Tearing down computational barriers between communities, on the other hand, also allows a sharing of resources. In particular, it also lowers hurdles for smaller-scale experiments or theory groups to exploit the sophisticated large-scale computing infrastructure pushed forward by extensive experimental collaborations.

Astronomy has a long tradition in surveying the sky, carefully documenting observations, and making the measurements available to the community and the general public. There are many examples in astronomy where projects have pioneered ways to store the large data collections obtained from systematic surveys and to **provide open access to the data**, as e.g. done by the SDSS. A generation of new, large telescopes, e.g. LOFAR, MeerKAT, the Rubin Observatory, and the SKA, require new concepts for data processing, storage, and access. LOFAR is already driving the transition by having its data stored in a long-term archive (LTA) distributed over several HPC centres. LOFAR data are processed on supercomputers at these sites.

Modern, large-scale simulations in astrophysics or lattice QCD pose a challenge not only on the algorithmic side, but also in terms of production, storage, analysis, visualisation and FAIR handling of data volumes beyond the multi-PB scale (see also TA 3). The analysis of the resulting simulation data nearly requires the same amount of computing resources as its production. Its enrichment with metadata, persistent identifiers, and FAIR access is a key topic of PUNCH4NFDI. The **strong similarities in the strategy for extracting information from simulations, experimental data and observations** can be leveraged to build a standardised data management workflow.

The PUNCH community has a long tradition in **developing middleware for distributed computing**. Over roughly the last decade very powerful cloud middleware has become available as open-source software. The PUNCH community started early on to operate and integrate such solutions into the existing distributed infrastructures. The increasing usage of industry-standard protocols will allow the construction of the backbone for PUNCH-SDP and the tools that will be developed in The solutions will be generic and applicable also for other data-intense sciences.

Based on their experience and expertise relevant to topics addressed in TA 2 this TA will be handled by PUNCH4NFDI partners DESY (WP 2.1, WP 2.2, WP 2.3), FZJ (WP 2.1, WP 2.2), GAU (WP 2.2), GSI (WP 2.1, WP 2.2, WP 2.3), JGU (WP 2.1, WP 2.2, WP 2.3), KIT (WP 2.2, WP 2.2,WP 2.3), TLS (WP 2.1, WP 2.2), TUDO (WP 2.3), UB (WP 2.1, WP 2.3), UoB (WP 2.1, WP 2.2), and UR (WP 2.1, WP 2.2).

The work will be supported by the following participants of PUNCH4NFDI which contribute with the specific expertise: ALU (WP 2.2), CERN (WP 2.3), DLR-DW (WP 2.1, WP 2.2), KIS (WP 2.1). LRZ (WP 2.1, WP 2.2), RUB (WP 2.1, WP 2.2), and RWTH (WP 2.1), and WWU (WP 2.3)

5.2.1 Work package 2.1: Standardised access to data and metadata

This work package addresses the objectives 1, 4 and 6. It provides standardised access to diverse and **federated storage** resources, in the following referred to as ”data lake”[7][2], in line with international community standards (e.g. AAI), and it supports the services offered via TA 6.

Status: The landscape of storage technologies and data management tools used by the PUNCH community ranges from very simple setups to very advanced distributed installations. Smaller research groups often cannot afford to invest vast resources into their data infrastructure. The data are usually stored at the host laboratory of the experiment or the observatory. To access the data local accounts need to be negotiated. Many setups **lack a systematic catalogue** that describes the contents of the files and the relation to the actual measurements. On the other end of the spectrum, there is the WLCG. Various levels of advancement can be found in between the extremes. For example the Belle II experiment uses middleware that was originally developed in the context of LHC experiments, and the lattice community uses parts of WLCG middleware to enable community access to so-called ”gauge ensembles”.

The astronomical community has a **long tradition of operating open data archives**. In contrast to particle physics, however, these come without attached computing resources. An example is the CDS, which hosts a large number of astrophysical catalogues that can be openly accessed. It is also available to smaller research groups for uploading and sharing their data. However, the most recent and next generations of astronomical instruments, like LOFAR, 4MOST, the Rubin Observatory, or the SKA, reach data rates and volumes that do not allow to follow the same strategy. The LOFAR LTA is currently the largest astronomical data archive ($\sim$50 PB, mainly on tapes) worldwide. Individual measurement sets have sizes of tens of TB and need compute clusters with high-memory nodes at the archive sites. This makes them hard to access, and it is expensive to move the raw data.

Numerical simulations play a major role in the PUNCH community. Examples are gauge ensembles in lattice QCD simulations and big datasets arising from cosmological N-body simulations. Typical large-scale simulations produce massive amounts of data and metadata and make up for a significant part of the workload on Germany’s HPC machines. Corresponding storage and archival solutions mostly exist at HPC centres. However, solutions to make the **massive, hard-to-access and hard-to-transfer output data** FAIR are still missing. FAIR data access and data management plans are more and more actively requested by scientists.

Goals: An infrastructure with standardised data access is needed to ease or enable analysis of **distributed datasets** (use case class 2). Hence archives and storage instances will be made **interoperable** leveraging established standard protocols. This functionality is a prerequisite to

[7] Here, ”data lake” refers to a collection of geographically distributed and diverse storage centres, operated and accessed as a single entity. This data lake definition differs from others in IT technolology (e.g. on `wikipedia.org`).

include the (dynamical) archives in the data flux of large experimental setups (use case class 5). Various metadata systems are already in production use. Their APIs need to be assessed such that they can be exposed to community-overarching services like the PUNCH-SDP.

Work programme: The PUNCH community has very ambitious plans with its flagship projects, requiring the recording of huge amounts of data. The community, including members from the PUNCH4NFDI consortium, addresses the challenges for example in the ESCAPE project and the DOMA initiative by WLCG. Members of PUNCH4NFDI are actively involved in these **international activities**. The ongoing research is targeted at the needs of the international collaborations. Within PUNCH4NFDI, significant effort will be spent to employ these developments for the national needs and provide state-of-the-art solutions for local experiments at smaller accelerators that are hosted at universities.

Based on ongoing developments, prototypes (D-TA2-WP1-1) will be provided to PUNCH4NFDI to allow the necessary integration work. That requires compatibility with existing and evolving systems from the international collaborations and the possibility to interconnect with infrastructure constructed within PUNCH4NFDI. The WLCG is presently converging to support two protocols for wide-area data transfer, WebDAV, the **standard protocol** for file transfers over the internet, and XRootD, which is a protocol very specific to the HEP and HuK communities. The readiness for data transfers using these protocols will be established with the prototypes.

AAI is of overarching interest for the consortium and beyond, consequently it is also addressed in WP 6.2. In TA 2 considerable amount of work is expected to setup the recommended AAI implementation and its validation in the complex environment of the various communities.

The data structures used by the PUNCH community are often complex and depend on the data source. A number of data sources will be addressed in an early stage to make them available to the PUNCH-SDP via the PUNCH data lake. The identified pilot data sources are the DES [8], GLOW-LOFAR pulsar data [9], LoTSS [10], data from one experiment at ELSA (e.g. [11]) at UoB, data from experiments at MAMI/MESA at JGU, and S-DALINAC at TUDa (D-TA2-WP1-2).

Later, more data from the list in table 3 will be added, based on the experiences gained with the pilot data (D-TA2-WP1-4). Priorities will be negotiated particularly with the User Committee (UC) and matched to available resources in coordination with the Infrastructure & Resource Board (IRB). The variety of data structures is caused by the nature of the scientific instruments, usually unique custom-designed detectors, observatories or satellites. That is also reflected in the data, metadata and corresponding software. Any approach to **harmonise the access** to those data and metadata can only be arranged on top of the existing systems and in collaboration with international partners. Solutions will be found case by case (see also WP 4.1 & WP 4.3).

Metadata handling plays an important role at various levels of the PUNCH-SDP. At the lower levels, metadata concern technical and administrative information for accessing data or for provenance tracking. Additional metadata are needed to define, describe, and track computing steps that (re-)generate data in more complex workflows. Higher levels provide information on the scientific content, like physics parameters or publications. Thus, many instances of interoperable metadata catalogues for different and partially community-specific metadata will eventually be needed within a federated system like the PUNCH-SDP (see also WP 4.2 & WP 6.3). **Interoperable metadata catalogues** with a standard web interface and for freely configurable metadata schemes are used in the lattice data grid (D-TA2-WP1-3). Astrophysical and general simulation data at HPC centres need to be enriched with metadata in a similar way. The aim is that the data, often too large to be moved to other places, can remain in place while being made FAIR, with persistent identifiers assigned and data search engines being aware of the data products.

Existing metadata catalogue solutions shall be extended and harmonised within PUNCH4NFDI in order to accomplish this task. While the technical development of the flexible metadata catalogue is subject of this TA, design of new schemes and corresponding interfaces for various PUNCH4NFDI applications is foreseen in WP 4.2 & WP 4.3.

***Deliverables**:*

- **D-TA2-WP1-1 (30 Sep 2022):** Prototype data lake setup at DESY and GSI.
- **D-TA2-WP1-2:** Integration of initial data resources:
  (31 Mar 2023) Radio astronomy resources (GLOW facilities).
  (30 Jun 2023) Hadron physics resources (experiments at MAMI & MESA, at ELSA and S-DALINAC).
- **D-TA2-WP1-3 (31 Dec 2023)** Metadata catalogue reference implementation for LQCD.
- **D-TA2-WP1-4:** Further integration of data resources:
  (30 Sep 2024) Astronomical and HEP data, including access to external resources.
  (31 Mar 2026) Nuclear physics resources (FAIR facility).
- **D-TA2-WP1-5 (30 Sep 2026):** Documentation of PUNCH data lake storage particularly for TA 6 services.
- **D-TA2-WP1-6 (31 Dec 2022):** Start of support of PUNCH data lake storage.

5.2.2 Work package 2.2: Compute4PUNCH

This WP addresses in particular objective 1 by providing the necessary **federated compute infrastructure** for the PUNCH-SDP, it contributes to objective 3 by developing tools that can be offered at the marketplace, it addresses objective 4 by integrating the infrastructure into European efforts, and it addresses objective 6 by making the solutions sufficiently generic.

Status: Large amounts of data in the PUNCH community are publicly accessible, in astronomy for instance via VO and CDS. The **archives** in astronomy are **rarely connected to computing** resources, hence limiting advanced analyses. For large datasets, this effectively restricts their use to users having access to appropriately large processing resources. In the particle physics community, computing is available via the Grid and heavily used. However, grid computing often **lacks efficiency and requires expert-level knowledge**, as indicated by the fact that the resources are rarely used by smaller groups.

The astronomical community has well established methods and repositories for publishing data. Individual projects commonly set up own infrastructures with interfaces mainly in accord with VO standards. This allows to automatise searching and retrieving data of interest, implying **workflows based on data retrieval and local analysis** carried out on compute resources of the user. For data-intense projects a **paradigm shift** is necessary and **code-to-data workflows** need to be realised to allow a wide range of users the exploitation of the data. For instance, LOFAR has established data centres which combine storage and computing. However, the services can currently only offered to a small number of users. For upcoming data-intense projects, as the SKA and the Rubin Observatory, it is even more imperative to develop **data centres providing data access and computing for general users** to exploit the scientific potential of the data. For the SKA, for instance, the EU-funded AENEAS consortium has developed recommendations for a federated service management of SKA regional centres.

The future data-driven scientific flagship projects like the HL-LHC, the FAIR facility, and the SKA will pose enormous challenges on data analysis infrastructures. One promising approach to tackle those challenges is to supplement community-specific resources by **temporarily available**, additional resources. However, including a variety of different resources not anymore fully controlled by community-specific policies (concerning e.g. operating system, hardware and software environments) leads to an increased **heterogeneity of the overall system**.

Beside data obtained from experiments and observations, also **data from large, sophisticated simulations** have an enormous scientific potential. An increasing number of these simulations has been made public, e.g. the Magneticum simulation. The full potential of the simulation products can only be exploited when complex data analysises can be carried.

Access to data will be provided by tools developed in WP 2.1 and can be **enhanced by establishing opportunistic data caches** close to the processing resources. The PUNCH community can profit in this case from existing developments at GSI, KIT and the JGU like the XRootD plugin based cache system or ad-hoc file systems.

The **integration of compute resources** into a common infrastructure can be accomplished by using the `COBalD/TARDIS` software framework developed at KIT in the context of the ErUM Data-IDT project. `COBalD/TARDIS` allows for a dynamic provisioning of compute resources from various providers using cloud and batch system interfaces and enables a transparent integration of those resources into a common infrastructure. **Access to compute resources** in this infrastructure can be enabled utilising established technologies like JupyterHub, grid compute elements, or traditional login nodes to cover the broad needs of the PUNCH community. In addition, established products like modern **container technologies** (e.g. Singularity) and the CERN Virtual Machine File System (CVMFS) can be adopted and utilised to provision various operating systems and to deploy the mandatory software components.

Goals: In order to ensure an **efficient utilisation**, to increase the operational effectiveness as well as to provide **uniform access** to the compute resources of various providers like HTC and HPC centres as well as cloud providers, it is reasonable to dynamically and transparently integrate those heterogeneous resources into **one common federated compute infrastructure**. This is how PUNCH4NFDI aims for ensuring that the PUNCH community can fully exploit the scientific potential of the upcoming data-intense experiments, observations and simulations.

To this end, PUNCH4NFDI will test, **combine and develop tools** to set up such a federated infrastructure (see figure 8) including available and suitable compute resources as well as their interconnection to data storages and archives.

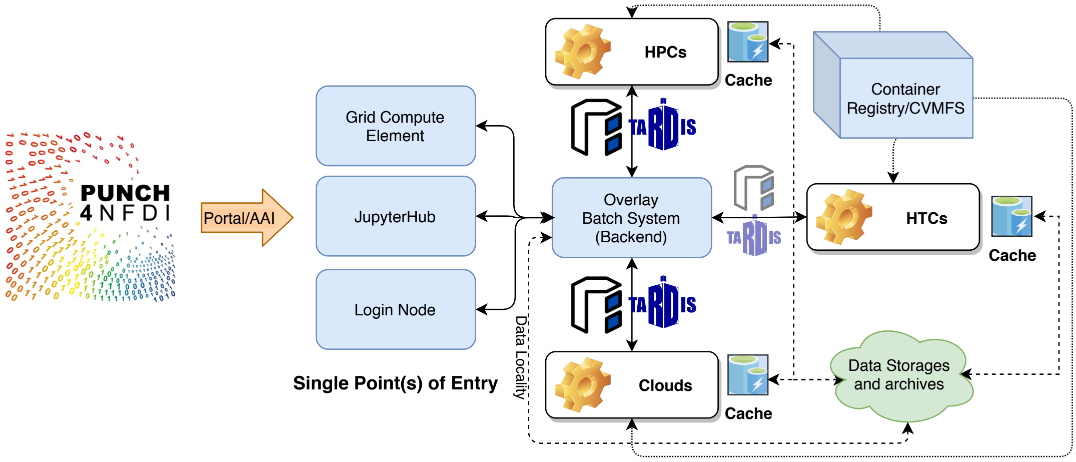

Figure 8: Compute4PUNCH — federated compute infrastructure. Components in light-blue rectangles are developed within this work package. Data storage and archives (green cloud) are developed in WP 2.1. The orange arrow is contributed by TA 4.

Work programme: To achieve the goals, the WP plans to proceed as follows: First, design and construct a federated compute infrastructure, named Compute4PUNCH, for the PUNCH community based on established solutions. For example the `COBalD/TARDIS` **resource manager**, to enable a dynamic and transparent integration of available heterogeneous compute resources (HPC, HTC and cloud resource) into one **common overlay batch system** acting as a back-end to the resources.

To access the Compute4PUNCH infrastructure, **different entry points** will be realised. Developments around well-established technologies such as JupyterHub, grid compute elements and traditional login nodes will provide very flexible entry points to the entire infrastructure, which is allowing the community to tackle a wide range of scientific questions.

All mandatory operating systems and software environments for the community-specific tasks will be provided using **modern container technologies** via a container registry (D-TA2-WP2-8). Initially starting with developing a prototype registry based on CVMFS ensuring an efficient distribution of the available images (D-TA2-WP2-3).

The entry points together with the container registry will be used by other TAs to offer even more complex workflows to the community.

Opportunistic caches deployed close to processing resources utilising semi-permanent or temporarily available storage will be used to enhance the access to the input data (D-TA2-WP2-6). Such cache systems will be made available by PUNCH4NFDI consortium members and will be integrated with the data storage infrastructure developed in WP 2.3 as well as the overlay batch system to further improve the efficiency of the infrastructure by enabling **data locality aware scheduling** (D-TA2-WP2-7). Those caches need to be systematically tested concerning performance and efficiency.

This infrastructure will **allow to execute complex workflows** on the data collections made available in WP 2.1 using hardware resources made available to the PUNCH4NFDI consortium. First, PUNCH4NFDI will provide a **demonstrator** Compute4PUNCH that allows to test e.g. the integration of opportunistic cache resources (D-TA2-WP2-1). As second step, the consortium will extend the Compute4PUNCH infrastructure to permit **integration of the variety of compute resources** available for PUNCH4NFDI, ranging from compute clusters to HPC and HTC infrastructures (D-TA2-WP2-4), see table **??**. The goal is to integrate a majority of the large variety of compute, storage resources and data collections provided by the PUNCH4NFDI consortium members from the start into Compute4PUNCH covering about 90 % of the use-cases (D-TA2-WP2-2). A small portion of the resources for special use cases (like lattice QCD) or resources lacking internet connectivity can be made available using traditional access methods.

***Deliverables**:*

- **D-TA2-WP2-1 (30 Jun 2022):** Demonstrator for federated compute infrastructure Compute4PUNCH.
- **D-TA2-WP2-2 (30 Jun 2024):** Adaption of Compute4PUNCH for domain specific large data collection (LOFAR, MeerKAT, CERN open data).
- **D-TA2-WP2-3 (30 Jun 2022):** Prototype for container registry.
- **D-TA2-WP2-4 (31 Dec 2024):** Integration of a variety of compute resources available in PUNCH4NFDI into Compute4PUNCH.
- **D-TA2-WP2-5 (30 Jun 2023):** Realisation of entry points as JupyterHub and batch system.
- **D-TA2-WP2-6 (30 Jun 2024):** Integration of opportunistic cache systems into Compute4PUNCH, and testing.
- **D-TA2-WP2-7 (30 Sep 2025):** Data-locality aware scheduling available in the overlay batch system.
- **D-TA2-WP2-8 (30 Sep 2026):** Fully capable container registry.

5.2.3 Work package 2.3: Automation and optimisation of big data workflows

This work packages tackles objective 1 and optimises the PUNCH-SDP by introducing quality of service (QoS) and **intelligent workflow techniques**. It supports use case class 2 by optimising data placement in and data movements across large-scale computing resources for huge distributed datasets. WP 2.3 also extends workflows and code (objective 2), and derived technical solutions will be part of the marketplace (objective 3). WP 2.3 will be closely aligned to national and international efforts (objective 4), e.g. through ESCAPE.

Status: Many data pools accessed by PUNCH users already exceed a capacity of 100 PB, while growing rapidly. The CERN community has shown that it is not desirable to store such big repositories only within a single site, as this either requires that the hosting site is also (completely) responsible for the overall compute infrastructure or that data always have to be moved from the hosting site to peers. Instead, data and storage should be distributed over multiple sites. Resulting **distributed data lakes** should support replication of data objects and should enable client applications to find the nearest copy of a data object within a unified namespace.

Data can be stored on arbitrary storage elements like tape, disk, or in memory, and grid sites can manage their data pools either as object storage, file systems, or databases. In particular the nuclear & particle physics community inside PUNCH4NFDI has developed open-source tools like dCache, XRootD, or StoRM to connect data centres with data repositories, while frameworks like Rucio [12] or Dynafed [13] combine these pools and expose them under a unified namespace. The astronomical community faces similar challenges as the particle physics community but is not yet as advanced in adopting modern data management tools; data locations are often still statically managed and determined independent from workflows.

Goals: In order to cope with the future challenges of the research field, the storage capacity of individual data lakes, the capability to work across data lakes, and to move data between them have to increase by an order of magnitude. It is necessary to enhance the existing tools to enable them to efficiently use scale-out storage environments. This can be achieved by collecting information about the underlying storage infrastructures and the intended access patterns to the stored data, which then can be used to optimise data placement and data movements, while also steering workflows. WP 2.3 therefore focuses on two main aspects:

– Increase the capabilities of PUNCH data lakes to **intelligently place and move data** by analysing monitoring information and information about future access patterns
– Extend usage scenarios for PUNCH workflows to **include multi-cloud environments**

The tasks in WP 2.3 to add intelligence to the management of data lakes will be closely aligned with international data management initiatives like the ESCAPE Data Infrastructure for Open Science (DIOS) project that is aiming to provide a scalable federated data infrastructure to manage Exabyte-scale data volumes. Several applicants of the PUNCH4NFDI proposal are actively driving the development of data lakes within ESCAPE.

Work programme: Intelligence can only be added to the PUNCH4NFDI data lake infrastructure by including **monitoring information** into the decision-making process of the applied tools (D-TA2-WP3-1). One example is a coupling between underlying object storage solutions, like Ceph, and grid tools like Rucio. Ceph provides a rich set of monitoring information that can be fed into the policy engine of Rucio. Rucio can then use this information to steer data placement within Ceph. These policies can then e.g. enforce the placement of hot data objects on an SSD tier and the placement of cold data on tape. Nevertheless, PUNCH4NFDI will use this example to also show the limitations of today's tool sets on different technological levels.

First of all, there is still a mismatch between the capabilities of individual tools and the requirements to build distributed storage infrastructures that also enable fine-grained QoS. Data objects from huge PUNCH experiments are typically placed in a single storage pool, and IO rate limits to these storage pools should be based on dedicated data streams between hosting sites. Underlying technologies, like Ceph, often only allow rate limits for individual devices and have to be extended to also include individual jobs. Furthermore, it is necessary to develop **interfaces between backend storage and frontend Grid tools**. WP 2.3 will therefore develop techniques to couple QoS between storage systems and Grid tools (D-TA2-WP3-2).

Secondly, QoS only exists inside the current grid tool sets and in present data lake developments as a concept and first implementations are expected as results of ongoing projects in the context of WLCG DOMA and ESCAPE. Strategies how to use QoS within the data management of experiments are in a very early state. However, good **concepts for data placement**, data replication, and traffic steering would be beneficial to drive down the immense costs for storage that are coming along during HL-LHC and other data-intense endeavours. WP 2.3 will therefore implement QoS techniques for major PUNCH4NFDI data archives (D-TA2-WP3-3).

Finally, it is important to use **forecasts about future access patterns** to proactively migrate or replicate data or jobs. This information is (partly) available in workflow descriptions and can especially help workflows that rely on all input data being present before a computation can start. Computational problems can then either be divided to only cover what is present at a single location, jobs can be transferred, or data can be proactively prefetched. WP 2.3 will therefore include information from workflow descriptions to improve data management.

Exabyte datasets also require new algorithms and protocols to scale concerning their management. An example is the design of distributed file catalogues. Each site participating in a data lake, for example, today needs to be aware of all participating storage elements and data caches to correctly locate objects and file copies. As a consequence, file catalogues do not scale well with the number of participants. An alternative is to use **distributed hashing** that refers clients to the correct metadata information even in case that this information is not locally available. Distributed hashing has shown to scale to (nearly) arbitrary sizes in peer-to-peer computing, while keeping the resource requirements of each participant extremely low. WP 2.3 will investigate how such hashing strategies can be applied to data lakes and exemplary be combined with tools like Rucio or Dynafed (D-TA2-WP3-5).

New data lake technologies can also help to improve the design of workflows. An example is the introduction of an event interface, which can help to trigger workflow executions depending on the creation, modification, or deletion of files. Example implementations of event interfaces are the OpenIO object store, AWS Lambda, and Google's cloud functions. Event interfaces are currently mostly used in serverless computing, where individual object modifications trigger small computational tasks within a container. WP 2.3 will investigate how these concepts can be transferred to PUNCH4NFDI to build **dynamic workflows**.

The design of workflows on an adaptive storage substrate builds a direct link between the astrophysics and the nuclear & particle physics community. The current momentum to adopt workflow languages like the Common Workflow Language in the astrophysics community can be used to specify tool sets that are jointly developed by all PUNCH4NFDI partners and that can then also be exported to further NFDI communities (D-TA2-WP3-6). WP 2.3 will work on workflow definitions that consider data locations and dependencies between processing steps, and then intelligently run the workflow distributed within a multi-cloud environment. This work package will therefore, e.g. extend the JupyterHub deployment from WP 2.3 to support dedicated access to the PUNCH data lake (D-TA2-WP3-4). Resulting notebooks can be run on multiple sites, close to the data and may further include research centre specific extensions, while they will also support intelligent data lake features like prefetching and replication. The resulting

JupyterHub prototype will be integrated into use-case specific notebook images in TA 6.

***Deliverables**:*

- **D-TA2-WP3-1 (31 Dec 2022):** Data lake monitoring infrastructure prototype.
- **D-TA2-WP3-2 (30 Jun 2023):** Middleware and backend storage jointly supporting QoS.
- **D-TA2-WP3-3 (30 Sep 2025):** QoS example implementation for a major HEP experiment.
- **D-TA2-WP3-4 (30 Sep 2025):** Prototype multi-cloud workflows using intelligent data placement.
- **D-TA2-WP3-5 (31 Dec 2025):** Hash table based data placement and replication mechanisms.
- **D-TA2-WP3-6 (31 Dec 2025):** Pilot of astronomy workflows using PUNCH4NFDI data lake features.

5.2.4 Sustainability, risks, and mitigation

The German PUNCH community is part of a larger global community that develops or selects its computing infrastructures, tools and processes in joint collaboration. Members of PUNCH4NFDI are well represented in the relevant international bodies like WLCG Operations Coordination, the management of the large international experiments, or various advisory committees. Furthermore, there are active contributions to ongoing international projects, e.g. EOSC Hub, EOSCsecretariat.eu, EOSC Pillar, EOSC Synergy, CS3MESH4EOSC and ESCAPE. This way, it is ensured, that PUNCH4NFDI developments are well aligned with the user community needs and other development projects of the community.

Difficulties in the inclusion of data sources due to political and legal barriers or unforeseen issues with APIs of existing services could eventually lead to delays. Therefore a significant number of diverse pilot data sources have been identified to ensure availability of data for the development of the PUNCH-SDP.

5.3 Task area 3: Data transformations

This TA contributes to the specific aims 1, 2, 4, and 5 and addresses a crucial component of **objective 1** of the PUNCH4NFDI consortium: the transformations of data in pursuit of scientific insights. The prime aim of PUNCH4NFDI is to maximise the exploitation of research data by providing a suitable data infrastructure to the user. Efficient numerical tools are an essential part of this endeavour, as they enable the user to extract high-level information from large datasets. Essential elements for meeting this goal are tools that extract higher-level information from large datasets, enabling one to study them directly or combine them with constraints from other datasets to draw novel scientific conclusions. The required tools have to be deployable on heterogeneous computing resources and have to be able to handle also the situation of the data being distributed and partitioned. The outputs of these transformations and the libraries containing the algorithms must conform to FAIR principles.

TA 3 will ensure that the developed tools define the new **state-of-the-art** and are **well-integrated** into the scientist's environment and easily usable by a broad community via the data portal developed in TA 4. Tools requiring access to storage and compute resources will make use of the solutions provided by TA 2. Many of the tools developed in TA 3 have potential applications in other NFDI consortia and will be made available via the marketplace offered by TA 6.

WP 3.1 connects to use case classes 2, 3 and 4 (especially the analysis of extensive datasets and the formulation of likelihoods). It focuses on the integration and further development of **tools for statistical analyses** that work efficiently in the limit of the large datasets and complex models of the PUNCH community. An initial goal is to ensure that PUNCH-SDP users find the necessary statistical toolkits available when they build or integrate workflows for data analysis. In parallel, the WP 3.1 effort will centre on the development of cross-community tools that can be deployed within (and beyond) the PUNCH community.

WP 3.2 addresses the challenge of optimally using heterogeneous and evolving architectures for **numerical methods and simulations**. A service for use case classes 2 and 3 will be provided by optimising performance-critical codes/algorithms in data analysis and simulation software used by the PUNCH community. The optimisation will be first performed for a specific compute environment and afterwards transported to others building on this experience. These algorithms/codes will be hosted in the software repository made available via the data portal developed in TA 4 and maintained by TA 6.

While **machine learning** (ML) has proven a powerful tool for the analysis of data, the adjustment of ML algorithms to specific use cases remains a tedious and time-consuming issue. By exploiting the large variety of PUNCH datasets, the consortium will automate the optimisation process and make the tools scale for large datasets in WP 3.3. In both aspects, the developed tools are important to ensure the reproducibility of scientific results because ML-results from large datasets are very difficult or expensive to reproduce.

Some insights can only be gained, or discoveries made, by **analysing multiple datasets** (identified by means provided by TA 4) in parallel, which is one of the key motivations for building an NFDI. Based on the analysis of use cases in classes 4, 2, and 1, tools will be developed in WP 3.4 to enable or facilitate such analyses that become increasingly difficult with the rise in data volumes and their distribution across multiple storage nodes. This includes solutions to deal with different data formats, to manage complex and scalable workflows, and to make data available in the easily combinable format of likelihoods. Well implemented, tested, and curated template workflows will increase the efficiency of researchers as well as the quality and FAIRness of their results.

This TA combines the **expertise on data transformations** from multiple institutes as detailed in section 3.1 to develop adequate solutions for a broad scientific community. The TA will be handled by PUNCH4NFDI partners FZJ (WP 3.2, WP 3.3, WP 3.4), GAU (WP 3.3), JGU (WP 3.3), LMU (WP 3.1, WP 3.4), MPIK (WP 3.1, WP 3.4), TUDO (WP 3.1), UB (WP 3.2, WP 3.4), UHH (WP 3.2, WP 3.3), and UR (WP 3.2).

The work will be supported by the following participants of PUNCH4NFDI that contribute their specific expertise: GU (WP 3.2), HZDR (WP 3.2), KIS (WP 3.4). LRZ (WP 3.2), TUM (WP 3.1), and USi (WP 3.3).

5.3.1 Work package 3.1: Statistical methods

Status: Big data analytics comprises a variety of statistical procedures. While statistical inference of allowed parameter ranges or hypothesis tests for several concrete models are often based on the principle of maximum likelihood or Bayesian inference, classification and regression are often performed with the help of machine learning techniques. In all of these applications, a potentially **large number of free parameters** must be tested against **very large datasets**. Any inference of parameters requires careful treatment of the statistical properties of the underlying data. Within the PUNCH community, statistical methods are applied at all levels of data analysis, from low-level event-based data to higher-level image or multidimensional array-based data, up to the highest level combination of pre-computed likelihoods for cross-dataset analyses. In many fields within the PUNCH community such as particle physics, astroparticle physics, astronomy, and cosmology, such methods or a combination thereof are applied, and often those methods are general enough to be shared between sub-communities.

The statistical analysis of large datasets using complex models is a central element in almost all scientific fields and industrial tasks. Thus it is strongly connected to needs within and outside the PUNCH4NFDI consortium. There exist already many statistical tools (e.g. ROOT [14], R [15], STAN [16]) that can function as starting points. WP 3 will build the **next generation of tools** for analysing large datasets. These tools will fulfil the core needs of PUNCH users building domain-specific workflows for deployment within the PUNCH-SDP, but also be general enough to be shared with other scientific communities.

Goals: Activities in this WP will provide the NFDI consortia with the **statistical tools needed for the analysis of large data sets and complex models**. This will include a focus on the further development of the promising statistical tool BAT.jl that is community-overarching together with a focus on making available a broader range and developing further a subset of — in some cases — more community-specific statistical methods and tools in the PUNCH-SDP environment to support the development of analysis workflows that draw upon these algorithms. These methods will be packaged in containers for deployment within a heterogeneous computing environment in close collaboration with TA 6 WP 6.4 (section 5.6.4).

Work programme: PUNCH4NFDI will address statistical inference in the light of large datasets and highly parallel computing by focusing on extensions of the **Bayesian Analysis Toolkit** (BAT) [17], BAT.jl [18], which is written in the novel Julia programming language [19] and that has the potential to meet the requested requirements (D-TA3-WP1-1). It allows for a fast evaluation of the statistical models and the execution of computational procedures, and it is not constrained to a specific scripting language in which the statistical models have to be expressed. The focus will be on optimising the toolkit for application to large datasets using massive parallelisation and the ability to handle extraordinarily complex models. The BAT.jl project has already been well established by the BAT.jl developer teams in Munich and Dortmund.

In the context of the development of BAT.jl, PUNCH4NFDI will focus on new strategies to efficiently access large, and possibly distributed, datasets. This includes the consideration of constraints from **running numerical algorithms in a highly parallelised way**. Recently, a work on the integration of existing samples to estimate integral values [20] has been published, which lays the groundwork for further development of parallelised sampling via space partitioning [21]. Besides, this technique will also enable the sampling from models with discrete parameters as well as the ability to efficiently sample from more complicated models, e.g. hierarchical models that feature ”distributions of distributions”. All new developments pursued within PUNCH4NFDI will be implemented in collaboration with the BAT.jl developers, and the full BAT.jl toolkit will be made available within the PUNCH-SDP environment.

PUNCH4NFDI will also develop and integrate a **broader set of statistical methods** that are required to analyse data within the PUNCH4NFDI sub-communities (D-TA3-WP1-2). In a first phase, the consortium will participate in the community survey of open-source software planned in TA 6 (**D-TA6-WP4-1**, section 5.6.4) to identify crucial existing sets of tools that can be made available within the PUNCH-SDP environment immediately. PUNCH4NFDI will support the integration of the identified tools into the containers that will be organised in TA 6 and will be made available to users through the TA 4 data portal interfaces. This approach ensures that community users will be able to quickly build workflows for analyses of datasets from across the whole consortium. In a second phase, PUNCH4NFDI will extend and generalise a subset of the most promising community-specific tools to better support combined analyses across multiple, heterogeneous datasets. A particular focus of these continued developments will be to ensure the ability to apply these tools efficiently in the limit of huge and heterogeneous datasets and complex models.

Within the astronomy and cosmology context, these tools will include methods for the study of the statistical properties of populations (e.g. correlation functions, power spectra, mass functions) and the prediction of these properties from first principles given a set of cosmological parameters [22]. Within the astroparticle physics context, these tools will include Gammapy, which is a package that already allows for a combined maximum likelihood analysis of multi-instrument, event-based astronomical data. PUNCH4NFDI will extend the functionality of the tools to include the possibility to export and provide likelihoods according to standards developed in WP 3.4. Within the nuclear physics context, PUNCH4NFDI plans to incorporate work toward a multi-parameter Bayesian analysis for heavy-ion collisions including Gaussian process emulation and Markov-Chain-Monte-Carlo method (MCMC).

Deliverables:

- **D-TA3-WP1-1 (30 Sep 2026):** Statistical inference in the limit of large datasets and highly parallel computing.
- **D-TA3-WP1-2 (30 Sep 2026):** Integration of a broad set of statistical methods; further development of a subset of methods into a common set of cross-community tools.

5.3.2 Work package 3.2: Numerical methods and simulations

Status: The analysis of large datasets and the simulations performed to generate data often require the use of HPC resources and software packages that make the best use of these resources. Although many of the tools are open access and employed by extended communities, currently the crucial, performance-critical sets of algorithms must typically be **optimised** by the individual user for application **on particular hardware**. This code optimisation is a very time-consuming process, and the task needs to be performed again and again whenever the hardware environments of the leading HPC centres change, which happens at frequent intervals. This situation is often challenging for the users as they are usually not trained in state-of-the-art code optimisation techniques. Thus valuable research time is lost by the researchers having to deal with these technical issues.

Goals: The main aim of this work package is to provide **libraries of numerical tools** to the user. The repository of libraries will contain the most commonly applied algorithms used on HPC machines in Germany. It will differ from existing code repositories in that these codes are optimised for the HPC platforms in Germany and are adapted whenever hardware changes occur. This repository will be accessible via the data portal. As most of the computing time used at the HPC computer centres by the PUNCH community can be attributed to about twenty different code bases, optimising these codes is an achievable goal. The performance-critical parts of the analysis software will be encapsulated in low-level libraries that address the needs of users in the PUNCH research community and are tuned for the heterogeneous compute environments.

Work programme: Various research communities are involved in the PUNCH4NFDI consortium using different codes and algorithms. Despite this diversity, there are **common challenges in the development, improvement and optimisation of basic, time-consuming routines** that need to be addressed in order to achieve acceptable sustained performance in large-scale simulations and data analysis campaigns performed in HPC environments.

PUNCH4NFDI will concentrate on the development and optimisation of software and provide support for basic **performance-critical routines** entering data analysis and simulation. The consortium will support multi-GPU systems, heterogeneous compute clusters, and upcoming processor generations. Interfaces to existing and new libraries developed in PUNCH4NFDI will be provided, and data analysis and simulation tools for upcoming, next-generation hardware paradigms will be prepared. This measure will also include providing options in the codes that ease the publication of the research data, including the metadata. The PUNCH community will profit from an optimised work environment in data production and analysis that will be made available to the users through the data portal.

Crucial parts of the data analysis and simulation software corresponding to **well-known bottlenecks** in the research fields covered by the PUNCH community, such as multi-grid or deflated conjugate gradient solvers in QCD and Riemann solvers and gravity solvers in astrophysics, require continuous attention. They need to be improved and optimised for use on new compute hardware. The focus in this WP will be on data-/compute-heavy algorithms and algorithmic/technical aspects of scientific reproducibility (resiliency, uncertainties).

To best meet the needs of the community, a **user-oriented choice of codes** will be made that shall be further developed and optimised. In the first stage, the codes will be optimised that are used by the largest number of users at the HPC centres. In the second stage, PUNCH4NFDI will also include the most resource-intensive codes. From both measures all users will benefit because the overall available resources will be more efficiently used. The optimised data production tools will include large-scale simulations and analysis software packages used in astrophysics and LQCD, as well as hydrodynamical simulations performed e.g.' in the cosmology, high-energy astrophysics and nuclear physics communities. In astrophysics, the focus will be on state-of-the-art N-body and hydrodynamical codes. In QCD applications, the focus will be on solvers for large sparse matrices and improved discretisation schemes for solving Hamiltonian equations of motions on four-dimensional space-time lattices.

The challenges posed by data distributed over **many CPUs and accelerators**, such as GPUs, will be addressed by developing optimised tools for memory management and communication. PUNCH4NFDI will focus on current and upcoming hardware architectures including single instruction multiple data processing as well as multi-GPU systems. The efficient performance of software packages also depends critically on the optimisation of data layouts over multiple GPUs as well as optimised memory access. The corresponding parameters need to be tuned in the compute-intensive parts of analysis and simulation codes. Furthermore, PUNCH4NFDI will develop strategies that dynamically balance the workload on heterogeneous computing systems, where the same application shares different types of computing resources. Work-sharing and load-balancing strategies will be implemented through libraries provided by the project. Such optimisations shall be performed by auto tuners and can potentially be supported by machine-learning methods.

The first set of libraries and tools will become available in 2023 and can be linked to the first version of the data portal developed in TA 4. The following deliverables provide users of use case class 3 with essential optimised codes for their data production, handling and analysis.

Deliverables:

- **D-TA3-WP2-1 (30 Sep 2026)**: Optimisation of performance-critical routines entering data analysis and simulation software on GPU systems, heterogeneous compute clusters and upcoming processor generations.
- **D-TA3-WP2-2 (30 Sep 2026)**: Provision of data-/compute-heavy algorithms with a focus on algorithmic/technical aspects of scientific reproducibility (resiliency, uncertainties).

5.3.3 Work package 3.3: Machine learning

Status: Analyses of the increasingly large experimental, observational, and simulated datasets produced in PUNCH can benefit tremendously from analysis via **modern machine learning methods**. Other scientific domains are moving on a similar trajectory, and research in PUNCH anticipates many problems and opportunities that will also be encountered eventually by other research communities.

PUNCH science is experiencing a transformation where new methods based on artificial intelligence and deep learning are changing the way scientific data are analysed. Due to (1) the large amount of high-quality synthetic data, (2) a deep understanding of uncertainties, and (3) meaningful multi-scale structured data, PUNCH constitutes a front runner domain in this process. However, currently much work is needlessly duplicated as independent research groups implement deep learning workflows for themselves. A typical scientific machine learning pipeline using artificial neural networks starts with data in an idiosyncratic format, specific to the measurement device or simulation environment. These data must first be converted into a file format (such as hdf5) amenable to machine learning. In a next step, salient features are engineered, extracted, and represented using a structure suitable for the problem at hand. Such structures can be, for example, two-dimensional matrices (e.g. images), time series, unorganised arrays, sets, or graphs. Based on the data structure, features, and objectives, a network architecture and training method are identified. The training is iterated multiple times and the tunable parameters of the architecture and learning method — so-called hyperparameters — are optimised until satisfactory performance is reached. This **technically complicated process** is time- and resource-consuming, and currently needs to be implemented and run manually by researchers whenever a deep learning model is to be trained.

ML algorithms will face new challenges as the **data volume** increases dramatically. This is expected in particle physics with the upcoming HL-LHC, in astronomy, which enters an era of survey science of extreme resolution (angular, temporal, spectral), and in other areas of science. While for some time the paucity of training data was an issue, now the limiting factor is the inability of ML algorithms to use all the available data. As a result, a new field in machine learning has emerged: large-scale learning [23]. Another challenge related to ML on large datasets is the application of FAIR principles, because ML-results from large datasets are very difficult or expensive to reproduce.

Goals: This project addresses two specific challenges of overarching importance relevant for the successful and wide-spread use of machine learning. The first deliverable focuses on **automated machine learning** (AutoML), and the second deliverable provides methods to extend machine learning to **very large datasets**. Naturally, both aspects will closely integrated in the common data portal as plug-ins and greatly increase the scientific value of collected data.

Work programme: PUNCH4NFDI will develop a system where users can interactively provide the training target as well as additional specifications such as especially salient features, relevant structures and splits in test, training, and validation data. The feature engineering and extraction is then **automatised**. Using this information, an automatic scan of models and hyperparameters is initiated and the best model is returned to the user. PUNCH data often come with complex interrelations of measurements by individual sensors. Therefore, one of the most promising data representation are graph neural networks. The optimisation of hyperparameters will be based on available and established Bayesian methods. Building on the most successful models, a library of starting points for use in **transfer learning** will be implemented.

Based on available benchmark and reference datasets, the performance of the best model produced by AutoML can continually be monitored and compared to hand-crafted approaches. Further improvements will come from optimisations of the automated system, including hyper-hyperparameter tuning and architectures using state-of-the-art learning to learn (meta-learning) techniques. The variance of predictions will further be reduced by ensembling over different models. Meanwhile, due to the **high complexity and size of PUNCH datasets** — e.g. observations by particle detectors ATLAS, CMS, FASER, and astronomical observatories LOFAR, SKA, Rubin Observatory, the developed AutoML algorithms can be useful for a range of other scientific applications. PUNCH4NFDI will increase the diversity of datasets and study the robustness and performance of the AutoML framework by using the library of benchmark datasets curated by MaRDI.

In summary, TA 3 proposes an automated system for the training and optimisation of neural networks for scientific decision making (D-TA3-WP3-1). This **AutoML** approach will greatly increase the usefulness and efficiency of scientific data as state-of-the-art learning algorithms can automatically be trained with little to no prior knowledge.

Furthermore, PUNCH4NFDI will develop concrete solutions for performing ML on large datasets (D-TA3-WP3-2) using two main approaches: **partitioning** the data and **parallelisation**. TA 3 will investigate building individual classifiers on particular subsets of data (this is sometimes called ensemble learning) in order to enhance the accuracy of a single classifier. Here, the most common method constructs the set of classifiers by training each one on different subsets of data. Afterwards, the classifier predictions are combined in a concrete way defined by the ensemble algorithm.

Alternatively, in cases that all data is needed for learning, a suitable parallelisation to enable learning distributed over many processors (CPU cores or GPUs) is needed. Modern deep learning frameworks already provide the ability to distribute batches over different processors, but a careful study of the relevant trade-offs for PUNCH datasets is needed. In cases where data is split into subsets the degrees of overlap can matter. It may also be important whether the data is disjoint or not. PUNCH4NFDI will study optimal approaches in terms of scalability and speed, for different hardware architectures and inter-processor communication requirements.

A frequent problem is the detection and classification of **extended objects** in large datasets that are stored and processed over many nodes. ML algorithms that are good at determining such objects in unpartitioned datasets usually fail when the objects extend across ”partitions” in the data. TA 3 will develop clustering algorithms that are able to identify extended objects objects across borders that keep their classifications intact. This task is relevant for any kind of extended objects in any kind of space (spatial, spectral, temporal). This problem of pattern recognition in partitioned datasets will also occur in TA 5, but for real-time data streams and other NFDI consortia. Examples are extended objects like clouds in geo-satellite data or cancer cell recognition in medicine. Based on the newly gained knowledge of cross partition pattern recognition, the applications will be later on extended to other application areas.

Finally, the techniques developed for machine learning on partitioned datasets will be **integrated** in the data portal and especially provide a foundation of the AutoML framework.

Deliverables:

- **D-TA3-WP3-1 (30 Sep 2026):** Framework for AutoML on scientific data based on the PUNCH domain.
- **D-TA3-WP3-2 (30 Sep 2026):** Tools and solutions for distributed learning using very large datasets.

5.3.4 Work package 3.4: Methods for analyses across datasets

Status: Scientific insight is often gained by **jointly analysing multiple datasets** (see use case class 4 in section 4.1). A typical example in particle physics is that the interpretation of a dataset of measured events requires the processing of datasets of simulated events of the signal and various background processes to determine signal efficiencies and background compositions and shapes. The multi-messenger approach in astrophysics exploits data about the same object measured with different wavelengths, particles, or gravitational waves. Finally, results from multiple experiments are often combined at a high level to obtain the best possible parameter constraints. As has been demonstrated for example in Ref. [24], such joint analyses often lead to dramatic improvements over what any single experiment can achieve through the process of degeneracy breaking. The current technical solutions for analysing multiple datasets are usually ad-hoc and fragile, addressing only very specific use cases.

Goals: While TA 4 addresses the issue of identifying suitable datasets and TA 2 provides the means to execute code on large, distributed datasets, the goal of this WP is to add **generic and scalable tools** to enable joint analyses of multiple datasets.

One challenge in meeting this goal is that datasets may have **different formats**. To analyse them together a common format has to be identified and suitable conversion methods or dataset readers must be available and included in the workflow.

The definition of **workflows** in the context of multi-dataset analyses is another challenge TA 3 will address in this work package. The inclusion of format converters is just one example where multiple steps have to be combined in the right order. Many science cases require multiple steps with specific dependencies, in particular if data at a low abstraction level (and huge size) are processed. The steps and dependencies must be represented in a way that is understandable by humans and computers. It should be stressed that it is essential for a FAIR data infrastructure to capture and document workflows. Only if workflows are known and recorded can they be reproduced. By making standard workflow steps and even complete workflows available via the data portal, this WP can ensure that they provide proper data and metadata outputs.

While the joint analysis of datasets at a low abstraction level often requires domain-specific algorithms and detailed knowledge about the data structure, a frequently used format for the combination of results at a high abstraction level is the **likelihood**. This may not always exploit the full potential of the low-level data, but it often provides sufficient information to those who do not have the means or desire to carry out a low-level analysis. The challenge here is to make likelihoods available in a FAIR way. In this work package, a standardised way of publishing likelihoods via the data portal developed in TA 4 will be developed.

Work programme: A prerequisite for resolving the heterogeneous data format problem is to ensure that the datasets have appropriate metadata describing their format such that appropriate dataset conversion or reading tools can be automatically selected. Specific examples of **conversion/reading methods** will be implemented for selected LQCD, astrophysics and other data formats. These tools will be deployable within heterogeneous computing environments. The developed framework will be transparent and easy to apply for the user when analysing multiple datasets automatically, including the required converters/reads in the workflow (D-TA3-WP4-1). The framework will be integrated into the data portal implemented by TA 4.

Another element is the development of a **workflow management system** that meets the requirements of the PUNCH communities (D-TA3-WP4-2). This system will make it easy for the scientist to define workflows for joint dataset analysis and test them quickly on local resources. The efficient execution of workflows on large, distributed datasets using the tools developed in TA 2, should then be a simple switch of a configuration parameter, made easy through the data portal interface developed in TA 4. PUNCH4NFDI will review existing tools for workflow management, adjust them for the relevant use cases, and integrate them in the PUNCH-SDP. Furthermore, common workflows or workflow steps will be identified and standardised solutions or templates offered that work efficiently on large datasets. With this a library of **template workflows** for typical PUNCH data analyses will be built from which scientists can choose solutions similar to their own use case and easily adjust them to the specific scientific question they wish to investigate. The metadata based descriptions of workflows will be employed within the TA 4 developed data portal to enable users to configure their workflows and associated datasets for an analysis.

To develop a standardised way of publishing likelihoods (D-TA3-WP4-3) we will exploit collaboration expertise with likelihood representation and in obtaining world averages from multiple measurements in particle physics [25] and cosmology [26]. Likelihoods may be available in a functional form, a histogram, a MCMC, or calculated on demand directly from the data. In addition to providing a **common interface** for the likelihood representations, one has to address the issue of parameter definitions. Likelihoods can only be combined if common parameters can be reliably identified. Therefore, PUNCH4NFDI will implement a **catalogue of likelihood parameters** that will allow scientists to publish their results with clearly defined and commonly understood dependencies. As likelihoods are often the input or output of statistical methods this work will be done in close collaboration with WP 3.1 (section 5.3.1).

Deliverables:

- **D-TA3-WP4-1 (31 Dec 2024):** Framework for conversion/reading of data for combined analyses; implementation of selected conversion/reading methods on heterogeneous systems.
- **D-TA3-WP4-2 (30 Sep 2026):** Tools to define, test, and execute scalable workflows; library of template workflows.
- **D-TA3-WP4-3 (31 Dec 2024):** Standard interface for the publication of likelihoods, including a catalogue for the definition of common parameters.

5.3.5 Sustainability, risks, and mitigation.

Most contributors are used to working on **long-term projects** within large and distributed collaborations. As the partners in the TA have **extensive experience in software development**, they are well aware that sustainability is a crucial aspect for the success of their projects. Some partners have responsibilities for libraries they developed and have demonstrated in the past that they are capable of providing the required **support**. Due to large science and computing centres playing a central role in our collaboration, the **technical infrastructure** for operating the developed services can be provided in a sustainable way.

As a large fraction of the work in TA 3 consists of software development, a key risk is not being able to fill the positions with enough **qualified people** quickly enough. PUNCH4NFDI plan to mitigate this by targeted advertisement and by grooming people for the key qualifications. Given the **large number of contributors**, the risk of having access to an insufficient diversity of datasets, hardware architectures, or workflow examples is very low.

5.4 Task area 4: Data portal

Data-driven research produces a wide range of **digital products**, such as data in various states of processing from raw data to distributions and derived statistical descriptions, simulations, technical documentation, research papers, figures, data tables, but also software, models, data analysis workflows, and computational runtime environments used to execute the analyses. While it is common practice to make scientific publications open access, only parts of the data sources are being made available as open data through mostly domain-specific repositories. Several platforms for science data handling already exist within PUNCH science, with varying degrees of implementation of the FAIR principles, and varying degrees of completeness and homogeneity/heterogeneity of the data. This includes platforms from ESCAPE, the EOSC, GAVO, the CERN open data portal [27], HepData [28], KCDC, ILDG and the VO [29]. In many cases, every data set, especially when dealing with data from different experiments, needs its own dedicated treatment of software and computing environment. While the majority of the simulation and statistical analysis code of the PUNCH community is open source, it is not accessible or searchable in a common way.

In the PUNCH community, as in many other organisations, currently, there is no overarching systematic way to link the various products of research amongst each other, access and interoperate with them, and reuse them. A combined low-level analysis of data from different sources, if at all, is thus only possible with considerable special efforts.

Therefore, the main product of TA 4 is the PUNCH4NFDI **data portal** as central element and main external access point for the PUNCH and broader communities to the **PUNCH4NFDI science data platform, PUNCH-SDP**. Drawing on experience from the development of the EOSC and other portal services, the data portal provides access to a unique knowledge fabric connecting the elements of each interlinked digital research product (RP), and especially each interoperable digital dynamic research product (DRP), which are developed in this TA. **An RP is a collection of interconnected research components on the PUNCH-SDP**, ranging from a publication to the associated metadata, simulation, data and workflow description. **A DRP is a *dynamic* RP, which is enhanced with executable code and machine-readable representation of the workflow**. The portal contains the ability to search, access, and understand the metadata associated with the RPs, be it publications, data, simulation and research procedures, and will contain a set of interfaces for users and knowledge providers. **The knowledge fabric formed by the interconnections within and between the (D)RPs and the representation of the analysis workflow in the DRPs sets the PUNCH4NFDI data portal apart from existing portal services.**

The data portal represents the entry point to the PUNCH-SDP, which is objective 1 as described in section 2.1. It is built on the tradition of collaborative research on scales from small to extremely large collaborations, which the PUNCH community follows since decades. **The data portal will be a crucial step to leverage the international PUNCH community computing expertise and infrastructure (as outlined in section 3.1) for FAIR science**. Currently, the PUNCH communities use this knowledge within their often large collaborations. PUNCH4NFDI aims to bring this cutting-edge technology into the realm of open science for PUNCH and beyond. **The future data portal is a central part of the majority of use-cases described in section 4.1**, most notably the use case classes 1, 2 and 4. It is also a central part of the FAIR research data management strategy defined in section 4. Its complete functionality will be provided as a service under the management of TA 6 (see sections 4.4 and 5.6) for PUNCH and other communities. It builds on the data management developments of TA 2 (section 5.2) and incorporates the data transformation tool developments from TA 3 (section 5.3).

The FAIR principles and the data portal: The data portal will provide interfaces for the user to **find** (D)RPs from existing databases (e.g. from experimental collaboration listings, inspirehep, DOIs, the Virtual Observatory Registry) and via metadata queries. The latter will allow to find and select RPs based on their physics content. The data portal will allow the user to **access** all elements of the (D)RPs, including data in different forms of processing and abstraction, metadata, simulation, statistical information, and code. It will allow the user to access the workflow of the original DRP and alter it. The persistence of the DRPs on the PUNCH-SDP will allow to link information from contemporary and past experiments in a new way.

The data portal will enable the user to **interoperate** with all elements of the (D)RP, download information from and/or execute computing tasks. It will allow uploading original (D)RPs from individual researchers or research groups, or collaborations, and it will expand on the **reusability** of research by enabling users to re-upload DRPs in an expanded, combined or modified form.

The (D)RPs on the data portal will be able to incorporate sources for all possible elements of the (D)RP from **outside repositories** (e.g. HepData), decentralised or centralised data repositories from large and small scale experiments, and data storage and databases specific for the PUNCH-SDP. **It will also provide a means of analysis and data preservation partially independent from the computing infrastructure of the data providers**.

The full functionality of the data portal can **allow for a combination of results from different (D)RPs across all of PUNCH, which automatically creates new connections within the**

PUNCH-SDP knowledge fabric. Such a capability of interconnecting DRPs is vital for the cross-community efforts of the use case class 4 described in section 4.1.

The Development of the data portal is a response to the enormous complexity of established distributed workflows and (meta)data structures in PUNCH science. To deal with this challenge the *depth over breadth* approach will be applied in all WPs: The minimal viable data portal will focus on a flexible design (microservice architecture), but start with implementing the full functionality first on selected workflows openly available within PUNCH4NFDI. With additional interest from other parts of the research community, the portal will then be expanded.

Digital (dynamic) research products and their catalogue. WP 4.1 will define and implement the (D)RP and the searchable catalogue technology on which it operates. Each (D)RP will be the core element of the contents in the data portal and be a dynamic, linkable, interactive and ultimately executable representation of a RP. Depending on the data provider, the (D)RP can consist of information entirely provided by the PUNCH4NFDI infrastructure, or it can link a large variety of existing outside sources of publication repositories and catalogues, data, simulation and metadata, and software repositories.

Mapping and collating metadata. The PUNCH community has a long history of complex and large data and simulation sets with a rich and often implicit structure of metadata. W P4.2 will take steps to make more and more of the metadata of these data structures available in the PUNCH-SDP, by leveraging existing tools to handle data and metadata from all subfields of PUNCH. One core element will be to enrich the current metadata with overarching structures needed for the (D)RPs.

Implementation of interfaces. WP 4.3 concentrates on the interfaces and permission tracking necessary to operate the interactive data portal. This includes access control, the monitoring and accounting of the use of computing resources, the infrastructure for combined analyses, e.g. on different (D)RPs, the improved provenance tracking, and the interfaces to statistical and analysis tools developed in TA 3. Within TA 4, where needed, the latter also involves the development of dataset- and/or experiment-specific interfaces for data transformations enabling data from very heterogeneous sources to be used in or with such standard TA 3 tools.

Build and operate the data portal. WP 4.4 realises the operation of the data portal. Web- and API-based interfaces for users and data providers, the operation of the catalogue services developed in WP 4.1, the access to databases operated outside of PUNCH4NFDI, and the operation of the access to the data lakes and computing resources of TA 2 are facilitated through WP 4.4. In co-operation with TA 6, WP 4.4 is responsible for turning the data portal into a service operated for and/or operated by other communities.

Based on their experience and expertise relevant to topics addressed in TA 4, this TA will be handled by PUNCH4NFDI partners AIP (WP 4.1, WP 4.2, WP 4.3, WP 4.4), DESY (WP 4.1, WP 4.2, WP 4.3, WP 4.4), GSI (WP 4.1, WP 4.2, WP 4.4), KIT (WP 4.1, WP 4.2, WP 4.4), MPIK (WP 4.3), UB (WP 4.2, WP 4.3), UHD (WP 4.2, WP 4.3), UoB (WP 4.1, WP 4.2, WP 4.3, WP 4.4), and UR (WP 4.2, WP 4.3).

The work will be supported by the following participants of PUNCH4NFDI which contribute with their specific expertise: KIS (WP 4.2, WP 4.4), LRZ (WP 4.2, WP 4.3), PTB (WP 4.1, WP 4.3), RWTH (WP 4.4), TUDa (WP 4.1), UP (WP 4.1), and UzK (WP 4.1, WP 4.3, WP 4.4).

5.4.1 Work package 4.1: Digital dynamic research products and their catalogue

Status: The PUNCH community curates and has access to a large number of repositories of different form and type of access (e.g. repositories for data, simulation, software, or publications; and services like github or dockerhub, which are provided by external providers). Smaller experiments and observatories may provide access to their data through institutional web pages. **No community-wide machine-readable research catalogue is currently available**. **The connection of data, metadata, software, and workflows is generally not consistently preserved**, although there are such activities in the community (e.g. on the phenomenological level in the form of GAMBIT [30], MasterCode [31], HEPfit [32], or HiggsBounds/HiggsSignals [33], or on the experimental results level in the form of REANA [34] and the CERN analysis preservation portal [35]). These works use similar approaches on dynamic research products as described below, but are generally restricted to specific use cases, have no overarching common logic of connecting different research elements, and, in case of the phenomenological projects, are only based on high abstraction levels of data.

Goals: WP 4.1 will design and implement an **open research catalogue** to find and access digital research products from across the PUNCH community in a unified way through the PUNCH-SDP. A central innovation will be the definition and inclusion of **dynamic research products** in the catalogue. A DRP can be any machine-executable program, such as physics models, functions of data and parameters or even complete data analysis workflows, interconnected with data, metadata, and simulation. **The catalogue will provide entry points to search, discover, access, and interlink static and dynamic RPs** using services developed by TA 2, TA 3 and TA 4. **Thus, initially unconnected data sources will be made inter-linkable to allow for execution of innovative use cases and improved re-useabillity**. The research catalogue of the data portal will allow to interlink currently separated data silos of different size across the PUNCH community (from those of individual researchers, small experiments and observatories, to large collaborations) using a unified access layer to metadata which complies with the FAIR principles (TA 4 WP 2). **The research catalogue will enable researchers to publish DRPs in a state that allows for re-execution of the code** on local facilities as well as on the science platform (TA 4 WP 4).

Work programme: The design and setup of a database for digital research products is central to this work package. A **central database for PUNCH4NFDI** will be developed, which can run on distributed instances and allows to access RPs from outside sources (D-TA4-WP1-1). The catalogue will incorporate all necessary metadata from available document-, data-, code-repositories and registries across the PUNCH community. Entries in the catalogue will be described and organised using a common standard. It will list where (in potentially distributed sources) the RP is hosted, who is responsible for curating it, and how to access the data. It will provide short descriptions and tagging and allow indexing and discovery.

The assembly of a collection of existing RPs available in the PUNCH community and their **ingestion into the open research catalogue** through a submission process will demonstrate the functionality of above measures (D-TA4-WP1-2). This includes the development of best practices for data publication in collaboration with TA 6.

The third main measure for this work package is **data annotation, linking, and enrichment** of RPs with descriptive metadata (D-TA4-WP1-3). This annotation scheme will form the bases for automatic processing and access through restful APIs provided by TA 4 WP 3. Experience in the PUNCH community on this topic exist for example from the implementation of the KASCADE Cosmic-Ray Data Centre (KCDC) [36].

The **development and definition of DRPs** includes the establishment of functionality of how to logically connect all their elements. This comprises a unified description of interfaces and resources needed to execute a DRP for a successful analysis preservation policy (D-TA4-WP1-4). **External containers and science platform products will be supported** as well as developments in the ESCAPE [37] and ARCHIVER [38] initiatives. PUNCH4NFDI will provide services to access, modify, and execute DRPs in cloud environments accessible to PUNCH4NFDI. The implementation will be based on experiences with containerised workflow preservation (Docker and Singularity containers; e.g. using experience gained on REANA [34]) and will allow for continuous integration during the development of DRPs.

Deliverables:

- **D-TA4-WP1-1 (31 Dec 2022):** Design and implementation of the PUNCH DRP database on the infrastructure provided by TA 2.
- **D-TA4-WP1-2 (31 Dec 2023):** Dynamic ingestion and curation processes of selected existing RPs from different PUNCH subcommunities; demonstration and routine processing.
- **D-TA4-WP1-3 (31 Dec 2024):** Enrichment of DRPs with descriptive metadata; demonstrator project to allow user access to SFB 1245 and astroparticle legacy data.
- **D-TA4-WP1-4 (30 Sep 2026):** Definition of the functionality of DRPs and their interfaces; prototype demonstrators and consistent integration for running them on the PUNCH-SDP; continuously integrate needs of PUNCH community.

5.4.2 Work package 4.2: Mapping and collating metadata

Status: In most of the current community data management plans (DMPs), a central item is keeping the metadata in a searchable and resilient catalogue, but the implementation is based on very different metadata models. For instance, the LQCD community developed an international metadata standard [39] 15 years ago and has a running metadata service, but only for a single layer of data, namely the raw Monte Carlo data. Other examples are the metadata schemes used at various platforms like the VO, the CERN open data portal [27], KCDC, EUDAT, EOSC, zenodo [40], or Rucio [12]. Related to publications, in high-energy physics INSPIRE-HEP provides the literature database and uses the open-source software Invenio [41] for creating and managing literature databases and digital libraries. In astronomy the astrophysics data system [42] delivers similar services.

The various available metadata schemes in the PUNCH community are currently not compatible, nor is the underlying ontology. There exist, however, **approaches to unify the access to data, metadata, and software in parts of the community, like for instance the CERN open data project, or the IVOA in astronomy**. In the data management plans of all the large-scale facilities there is a requirement to describe the data in order to be able to perform selection processes for the analyses as well as comparisons with the corresponding simulations. In most cases, these data descriptions are done using customised metadata models not ready for standardisation or even for general use. Experiment-specific software tools then implicitly include the relationship between metadata and data structures. This high degree of integration is a challenge, especially considering the often complex data structures of both PUNCH experimental data and simulations. Part of this challenge lies in the area of the specific interfaces that must be provided by WP 3.

Goals: The data portal to be developed in TA 4 requires **different layers of searchable and accessible (meta)data in all communities**. Such layers can be defined for instance for raw (experimental or Monte Carlo) data, for different levels of analysis data and for publications. Consequently, the first goal is to define such layers for the different communities in an as flexible as possible manner, such as to make it possible to insert new layers at a later stage. In general, these layers can be envisaged as a directed graph with some data transformation attached to each edge. The top level will be the publication layer, while the lowest layer will be the raw data.

For backward compatibility reasons there cannot be a single standard metadata scheme for each layer, in particular for the lower lying ones. Therefore, **interfaces have to be defined for already existing schemes that allow one to incorporate them into and search the corresponding metadata services from the data portal**. However, the existing metadata schemes have to be extended in a minimal way to include information about the other layers of metadata, the cross-layer information. This will require discussions with the metadata working groups in the corresponding international communities. An alternative approach is a separate catalogue providing the information about the inter-layer connections. Both approaches will be investigated and an informed decision will be taken.

Work programme: In close collaboration with WP 6.3, TA 4 will define a **prototype metadata scheme** which covers the various data formats from the PUNCH community. This will include also a transformation of the different data (and metadata) formats in a prototype data format (D-TA4-WP2-1). This is of particular importance for sub-communities with less developed standards, for which also suitable prototype data formats will be provided.

It will be essential to **provide feedback to the data providers regarding their data format, metadata schemes and data curation**. This process to define **layers of metadata** fitting the entire PUNCH-SDP requires a close connection to the work in WP 1 and WP 3 of the task area (D-TA4-WP2-2). Due to the diversity of the community, TA 4 needs to work out extensions to (existing) registries and data discovery services to incorporate cross-layer information, possibly accompanied by the definition of a metadata catalogue. An appropriate interface to use the new schemes have to be provided (WP 3, WP 4).

For an efficient and sustainable use of the data portal, TA 4 has to define and standardise formats and services to annotate the **cross-layer connections with information about the data transformations** and/or workflows (D-TA4-WP2-3). This will be implemented and tested together with WP 4 and worked out in close connection to TA 6 in order to lead to a general service of PUNCH4NFDI. Formats need to be developed to store the cross-layer information as well as attach the software stack required to perform the corresponding data transformation.

An important part of **interoperability** is to make sure that the different terminologies of the PUNCH community are understandable for members of all communities, which represents a challenge. Compatibility or **machine-readable maps between vocabularies** could be long term goals to cope with this challenge. Until then very precise metadata and cross-layer information can provide a bridge. Data curation needs to be provided by the communities themselves, which only have the relevant knowledge of the data. Depending on the needs of these communities a combined service for several sub-communities might be reasonable.

Deliverables:

- **D-TA4-WP2-1 (31 Dec 2023):** Provide prototype metadata scheme and prototype data format.
- **D-TA4-WP2-2 (31 Dec 2025):** Define metadata layers and minimal extensions for existing metadata standards.
- **D-TA4-WP2-3 (30 Sep 2026):** Define cross-layer connections for transformations of digital objects.

5.4.3 Work package 4.3: Implementation of interfaces

Status: As listed in the introduction, several portals for science data handling already exist within PUNCH, with varying degrees of implementation of the FAIR principles, and varying degrees of completeness and homogeneity/heterogeneity of the data. Different access control and accounting policies apply and no interfaces between different data repositories exist.

So far no platform exists that allows the discovery, access and use of the content of these and other platforms in a unified way. In many cases, every dataset, especially when dealing with data from different experiments, needs its own dedicated treatment of software and computing environment. A combined (at or below the level of published data) analysis of data from different sources, if at all, is thus only possible with considerable special effort for each dataset.

Goals: This WP develops **interfaces** of the science platform **for the access control, accounting, and monitoring**, and it will provide the infrastructure for combined analyses on different datasets. Improvements of **interfaces to statistical and analysis tools will be incorporated and interfaces for improved provenance tracking and metadata** implemented.

Integrating these interfaces in the data portal (WP 4), will lead to a data portal for users as **single entry point**, not only to search for data and access individual data as well as software tools, but allowing for combined analyses of data from different sources with interfaces to optimal tools, including input from TA 2, 3 and 6.

Many data analysis tasks, especially for combined data analyses on different datasets, require

large amounts of computing resources on **heterogeneous compute environments**. In the course of such an analysis, new data and metadata are generated. Besides providing the user access to the data, analysis software and computing resources, **provenance tracking and the documentation of workflows** is an important aspect here, e.g. in mapping and linking different data sources, software and computing resources.

Many datasets, even when dealing with conceptually very similar data, use logical data formats, variable naming schemes, schemes for interrelations between the different parts of the data, software versions and even operating systems that are mutually incompatible and, for older data, sometimes even obsolete. To remedy this, interfaces need to be provided that **transform the content and format of the data into a state that can be handled by the tools and methods** supported by the platform.

Work programme: This work package will focus on the implementation of interfaces for the functionality of the data portal respecting the FAIR principles. This includes interfaces for access management to the central platform, its decentralised branches and external resources linked into the data portal. Furthermore, interfaces to the content and format of initially heterogeneous data both inside and outside the platform, are built to enable their homogeneous treatment with common tools provided by other task areas and work packages. It will strongly interact with the task areas and work packages which provide and implement the platform, which provide and implement the metadata handling, and which will provide tools that will be integrated into these interfaces or for which these interfaces will serve as an input. **Interfaces to access data and metadata from the external portals mentioned above will be developed**. These resources are currently inhomogeneous and fragmented, thus a common access scheme, e.g. a RESTful interface, has to be developed and implemented for each resource (D-TA4-WP3-1). This includes (partly data-specific) conversion for the integration into the PUNCH-SDP and AAI to access restricted entries and will be coordinated with TA 3 and 6.

As required by the user access to distributed data and non-central computing environments, AAI will be integrated into the data and metadata access mechanisms in the user interface (D-TA4-WP3-2). Access to the necessary analysis tools, for instance provided by TA 3, will be provided in terms of an API. Different policies will be supported and implemented in the services and user interfaces. Within the PUNCH community this will include the standards and services provided by TA 2 and 6. Especially for combined and distributed data analyses an **improved provenance tracking in combination with a workflow documentation** will be integrated in the platform, which is an important aspect for data and analysis preservation. Automatic **status monitoring and consistency checks of metadata and other services** like web services, databases as well as data and metadata file catalogues will be developed and implemented, which is important for a sustainable data portal.

Interfaces needed to **enable multi-dataset/experiment analysis** with the tools provided by other work packages and task areas, including TA 3, will be developed (D-TA4-WP3-3). The focus is on key interfaces for dataset-specific aspects of software environment, content and format, whenever generic tools or interfaces can not be applied directly. Many of these will be based on already existing test prototypes. Others will be freshly developed according to user needs.

Deliverables:

- **D-TA4-WP3-1 (30 Sep 2026):** Technical interfaces to external resources.
- **D-TA4-WP3-2 (30 Sep 2026):** Integration of platform services and interfaces.
- **D-TA4-WP3-3 (30 Sep 2026):** Interfaces allowing combined analysis of data sets from different sources and experiments.

5.4.4 Work package 4.4: Build and operate the data portal

Status: Data portals exist in many variants in the PUNCH community. They are almost always built for data collections from an instrument or experiment, and sometimes they are an assembly of heterogeneous data collections. These — notionally in astronomy — are often based on common curation and publication standards, built on paradigms of scientists individually accessing (and selecting from) widely distributed data collections for analyses on their workstations. There are evident weaknesses in these approaches, and these constitute one of the drivers for forming PUNCH4NFDI. Some institutes or communities have started to address the issues, e.g. the KCDC, or the SciServer at SDSS, and MPE, and AIP using CoCalc-based [43] execution environments. In other disciplines, portal solutions built on CKAN and similar software are used for managing in-house data collections.

Goals: The goal is the instantiation of the data portal, which encompasses many functions: It is the access point for users of the PUNCH4NFDI services and facilities ranging from **working with a DRP, to uploading own data for combining with other data collections using offered tools, to searching for and selecting from data collections with certain physical properties** and even taking these out of the data portal environment, if the selection is of suitable size. The user might have to authenticate for accessing some resources or data (e.g. being a member of a group working on a paper), and the user's account carries authorisation to execute on precious compute resources.

Work programme: To support these interactions, an explorative web front-end is provided. This includes login facilities, components to access metadata and data selection, and the interactive use of DRP. The latter requires the implementation of a publicly accessible execution environment to unpack and work with the DRP. This needs to be carefully designed, as some of the PUNCH4NFDI resources will be consumed here.

Aside from an **interactive front-end**, the access to services via APIs based on standard protocols through scripting, like the TAP service used by astronomy data providers, requires another set of components. All user interactions with public accessible data and software is traditionally considered part of the provisions for the publication in PUNCH territory. But if a user authenticates and consumes resources, the monitoring of their resource consumption will form a component of the data portal.

The developed metadata and transformation services of WP 2 and 3 are essential parts of the data portal construction. Since the WP expects heterogeneous implementations with build-on elements developed within the PUNCH community, the portal is seen as **orchestrated microservices**, where the APIs enable the construction of different workflows. WP 2 establishes facilities to connect and assemble heterogeneous data collections, whereas WP 3 enables the execution of tasks on these data, both making use of developments and service components from TA 2 and 3. **The data portal enables the orchestration of such tasks**, and is a container for results. In turn, these need storage space and other facilities, such as support for describing processing and result with proper metadata. Whereas WP 1 defines the features of such products, the creation of DRPs by various users and **the management of the DRP** falls into the operation of the portal.

For the orchestration to work, tools from large software development endeavours are employed: all developments make use of a GitLab hub with Continuous Integration (CI), enabling well founded decisions when to roll out a new service component for public use. This will be another important part of the data portal which needs to be established early on. But this tool is not only for the internal development process, it is also a means to enhance the **reproducibility of results** for portal users, especially if they want to create a DRP, and also for the general goal of **re-usability**.

It can also be a means to make results of the PUNCH community publicly available, but the decision whether to use a data portal-hosted GitLab, or migrate implementations to a central NFDI code repository is to be worked out by TA 6 in interaction with other NFDI consortia. The WP 4 is building and operating the science data portal and provides access to the machinery and environment for the many services that are built by PUNCH4NFDI.

Within the use case collection, the creation of operational centres for astronomical data, also for astroparticle data has been specified, and data centres have offered their data collections for inclusion into the PUNCH environment. And more, some not yet published data collections from various sources were offered for curation, publication and use. WP 4 will work with WP 2 and 3 and TA 2 on **concepts to incorporate these data collections**.

Deliverables:

- **D-TA4-WP4-1 (30 June 2023):** Working prototype data portal (web interface) to access and use the platform.
- **D-TA4-WP4-2 (30 June 2024):** Published research product examples with stored data and interoperable analysis workflows.
- **D-TA4-WP4-3 (30 Sep 2026):** Feature-complete data portal service.
- **D-TA4-WP4-4 (30 Sep 2026):** Published and interoperable (dynamic) research products using the full range of services, including combined analysis of datasets from different sources and experiments.

5.4.5 Sustainability, risks and mitigation

The following **sustainability** strategy is envisaged: Catalogues, registries, and other services of TA 4 need resources available of long timescales. During the funding period of PUNCH4NFDI, the resources will be available. Beyond that, the platform can become a part of an encompassing NFDI platform.

The following risks and risk mitigation strategies are considered:

Community adoption and resources: Success of the data portal depends on its adoption throughout the PUNCH community. Since it folds several still separate components into a new product, it adds complexity. This has to be counterbalanced by ease of use. The implementation will start with various publicly available data collections and open-source software to assess its viability and include the wider community early in the testing and development process in a **depth over breadth strategy**. The wide community input into the use cases (serving as initial guides for design of the data portal) and the participation of PUNCH4NFDI scientists in all major experiments with German participation ensure broad access. Workshops, training, and the buildup of synergies with national and international efforts for FAIR research will play a key role for its success (together with TA 6 and TA 7).

Upstream risks, competing and overlapping activities: A subset of the TA 4 efforts are paralleled by other organisations and efforts, especially by European projects (EOSC, ESCAPE, ARCHIVER, CERN open data). This risk is mitigated by parts of the PUNCH4NFDI consortium being a part of these European efforts, co-operating with these, evaluating their products and using available solutions for parts of the TA 4 work.

5.5 Task area 5: Data irreversibility

The rapid increase in both data rates and data complexity leads to several **vital challenges soon to be seen throughout society** as we enter the ”Internet-Of-Things” era, where large sets of ”sensors” will transmit data upon which autonomous ”actors” will react. However, the substantial increase of **power consumption for storage solutions**, e.g. cloud computing, requires the investigation of resource-optimised data sets with maximal relevance and minimal redundancy. **Decisions will need to be made, often in real-time and without human intervention**, which information to keep or how to compress it with calculable loss [44]. **Loss will be inevitable and mostly irreversible**, while off-line analyses or emerging additional information will feed back and dictate modifications of the on-line processes (“dynamic filtering”). The decision process of rules and methods for the extraction of pertinent information out of huge data streams in real-time will need to be updated frequently and captured as important metadata. Hence, **the impact of the information loss** must be traced and gauged in order to allow drawing adequate conclusions from archives, which will no longer be static but **dynamic entities**.

Nearly all fields within the PUNCH community face crucial aspects of these challenges already today; they are unavoidable on an even much larger scale for the upcoming generation of experiments. **Using cutting-edge architectures of data processing, this TA 5 will study selected real-life applications, comparing set-ups where experiments can be controlled (particle physics) or merely observed as given (astrophysics).** In both cases, efficient real-time algorithms and methods need to be developed and studied, including ML and methods to estimate the degree of information loss and the resulting impact on the scientific interpretation of data and reproducibility of results.

Successful work in TA 5 will form the **basis for future discoveries also beyond PUNCH4NFDI**, since comparable requirements will also be essential for other consortia — as for future commercial and civil projects. Hence, the problem is not only essential but also very timely. While PUNCH4NFDI provides a unique opportunity to establish the required concepts, TA 5 cannot solve the diversity of challenges that data irreversibility will cause. But PUNCH4NFDI can focus on one specific task that is common to many problems, namely the **identification of rare and/or abnormal signals in fast huge data streams**. Similar challenges may occur, for instance, in the identification of disease markers in medical data or glitches in sensor data of smart cities. Hence, the TA's activities will be linked with those of the marketplace in TA 6 (WP 1). TA 5 will provide key contributions to objective 6 among the PUNCH4NFDI key **objectives** (section 2.2), by enabling and evaluating future real-time data reduction mechanisms.

While TA 2–4 describe how to improve existing data handling, some methods may not be applicable anymore when instruments produce **scientific data that exceed in size all that humankind has collected so far in its entire history**. This challenge will require conceptually different solutions. Exchanges with TA 2–4 in well-defined interfaces are essential. The structure of TA 5 and its tightly interconnected work packages is designed along with a **dynamic data life-cycle model** (see figure 9). Concepts developed in TA 5 are directly related to the initial selection of raw data **including the consistent description of the data reduction process in metadata**. All of these aspects are of key importance for developments of the future PUNCH-SDP.

WP 1 ”Implications for discovery potential and reproducibility” will explore the tension and interplay between reproducibility of filtered and refined data and its implications for the discovery potential. Formulating the problem with help of astronomy demonstrators, this WP will explore possible solutions and define curation criteria and structural requirements on metadata in light of the FAIR principles.

A large focus of TA 5 is on the actual implementation of efficient real-time filtering. Making use of large data streams accessible to PUNCH (e.g. LHC, Belle II, MeerKAT), **WP 2 ”Dynamic filtering”** will study a number of analogous challenges in the real-time selection and processing of data extracted from huge data streams.

The focus will be on how to discard irrelevant information with both minimal time budget and ability to describe the decision process in metadata that become part of the data itself. Since information is irreversibly lost, filtering of various forms of noise and irrelevant signal background deserves special attention, in particular for discovering un-

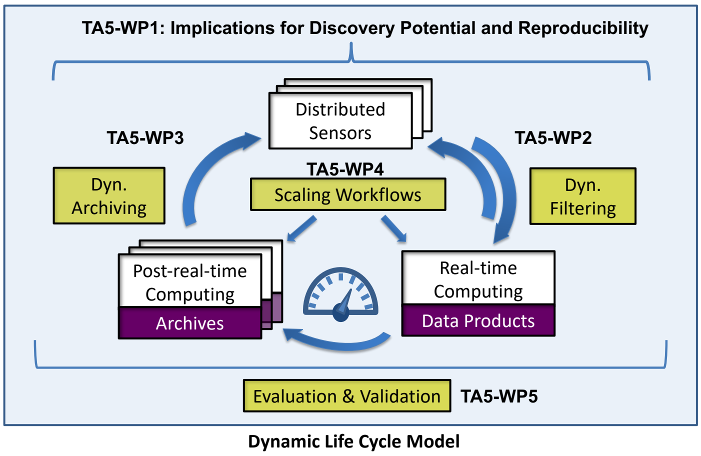

Figure 9: Dynamic life cycle model. See text for details.

usual or rare signals.

The aspects of dynamic archiving are studied in **WP 3 ”Dynamic archiving”** using astronomical data.
WP 3 aims to develop the concept of ”dynamic” archives and to provide methods for accessing these. The focus is on developing generic concepts for workflows that can be used to optimise feedback between online and offline analysis. A test case will be set up to demonstrate a complete dynamic filter and archive feedback loop.

WP 4 ”Scaling workflows” aims to identify technology solutions for scaling the ”online” and ”offline” workflows in WP 2 and 3, respectively. PUNCH4NFDI will study the optimal use of resources, which may require hardware and software to be designed together to create embedded systems, and address the problem of analysing single data sets with huge volumes (”data monsters”) utilizing complex workflows.

The **”evaluation and validation of instrument response and characteristics”** studied in **WP 5** is crucial to validate the success of the measures implemented in WP 1–4. The goal of this WP is the development of tools for predictive maintenance process control, based on unsupervised machine learning algorithms, and to devise strategies to ensure the quality of the data even in the presence of variable, unpredictable background noise.

The work planned in TA 5 will create synergies not only for the PUNCH4NFDI partners but also for other NFDI consortia facing similar challenges shortly.

TA 5 focuses on all aspects related to **use case class 5**: Real-time challenges, data irreversibility as described in section 4.1. A set of highly efficient and robust algorithms required for real-time data analysis including also tools for the near real-time detection of anomalies will be developed as **services**, see also section 4.4. To optimise the concepts developed in TA 5, regular feedback from and discussions with the other TAs and user community are necessary and will take place within the structure described in section 3.

Based on their experience to topics addressed in TA 5 (for participants see section 1, for co-applicants see section 3.1), this TA will be handled by PUNCH4NFDI partners DESY (WP 5.2, WP 5.3), FIAS (WP 5.2, 5.4), FZJ (WP 5.1, 5.4), HTW (WP 5.1, 5.3, 5.4), JGU (WP 5.2, 5.4), MPIfR (WP 5.1, 5.2, 5.3, 5.5), TUDO (WP 5.2, 5.4, 5.5), TUDD (WP 5.2, 5.5), UB (WP 5.1, 5.3) and UHD (WP 5.1, 5.2, 5.5).

5.5.1 WP 1: Implications for discovery potential and reproducibility

Status: Reproducibility is one of the basic principles of modern science and is also heavily discussed in the context of social, economic, and medical studies. The known ”knowns” are established by reproducing experiments, simulations, and theoretical reasoning. The **basis of reproducibility is a complete description of an experiment (metadata) with a complete record of its outcome (data)**. In reality neither data nor metadata are ever complete, but one can aim at covering the relevant parameter space densely. Compared to other branches of science, this is feasible for physics research, where one is able to control many experimental aspects relevant to a given question at hand. However, this is not the case for astronomical observations where the observed object, the universe, is literally out of control. We can (partially) overcome this problem by turning to statistically large samples of objects. For example, we map the cosmic large-scale structure based on the observed spatial distribution of billions of galaxies from large sky surveys, creating data rates and sets of challenging size.

In order to cope with this situation, PUNCH science relies on data models, for which **a known data structure of the ”unknowns” is assumed, according to which filters are built** that reduce data rates and volumes. Recent examples awarded with Nobel prizes that used strategies based on filters are the discovery of the Higgs particle [45, 46] or the observation of gravitational waves from coalescing black holes [47]. But as these models make assumptions on the state of the world, **researchers blind themselves towards potentially exciting discoveries that are unexpected and thus anomalous to the model at hand**.

The search for anomalies can currently be described as a ”fishing trip in today’s data swamp”, where one is unaware of what one is looking for, but follows some intuition and runs some random (consistency) tests on the data. WeIt is expected that the chances for breakthrough science based on ”fishing trips” is decreasing with the increasing amount of data and especially with the increasing amount of most finely filtered and refined data. From time to time new phenomena are discovered by this method, like the mysterious fast radio bursts (FRBs [48]) that have been found in archival data. **Incomplete archives holding only data refined to highest quality according to some world model might hinder new discoveries.**

Goals: WP 1 plans to identify and evaluate **methods and tools that can recover unexpected science (the unknown ”unknowns”) from incomplete data**, be it on- or offline, and its implication on the potential to make unexpected discoveries. For that the WP will test several metrics (information entropy is an obvious one) that could evaluate the information content of raw and filtered data and investigate how sparse a dataset could become before losing any chance to say that the data have been reproduced and that unexpected features can still be discovered. The work in WP 1 aims at finding a measure for the minimal amount of information in the filtered data that would allow for an unexpected discovery, that is it would clearly stick out of the instrumental noise. As an initial step, prototype algorithms will be developed for a general broad usage to install alarm systems for cases that require human intervention to vet unidentified signals.

Measures: After formulating the problems and possible implications for discovery of phenomena events, strategies and concepts will be developed to deal with them (D-TA5-WP1-1). This includes **curation criteria and the structural requirements on metadata in light of FAIR principles** (D-TA5-WP1-2). An important role is played by the comparison of data (real data and simulations) with other data and theories, **in the presence of irreversible information loss**. We will develop a general protocol on capturing the decisions made by, and status of, real-time systems (D-TA5-WP1-3). The results are interfaced with WP 2, 3 and 6. These studies will allow to develop strategies to ensure reproducibility of results. The WP will also coordinate work on real-life examples and the ability to scale those examples (WP 2–5), the interfacing with TA 2–4 and other consortia (TA 6).

***Deliverables**:*

- **D-TA5-WP1-1 (30 Sep 2023):** Report on impact of on-line filtering on discovery potential.
- **D-TA5-WP1-2 (30 Sep 2025):** Report on impact of on-line filtering on FAIR principles.
- **D-TA5-WP1-3 (30 Sep 2026):**General protocol on capturing the decisions made by, and status of, real-time sensors demonstrated on an astronomical example.

5.5.2 WP 2: Dynamic filtering

Status: The processing and selection of detector data at the earliest stages is a crucial aspect of the future of real-time data analysis, as summarised e.g. in [49]. In particle physics, usually, defined triggers are used for selection, but many proposed experiments are planning to run in a trigger-less mode, when complete raw data are read out in order to **allow for searches for yet unknown, rare and exotic physics signatures**. Similarly, in the case of astronomical instruments, data streams are now continuously processed in their standard way (e.g. to produce images), while other systems are designed to tap into the data streams in order to **simultaneously look for unknown or rare signals using machine learning** [50]. The identification of rare or even unknown signals requires, on the one hand, a comparison of data with prior knowledge (via dynamic archives, see WP 3). On the other hand, once a trigger is raised, a fraction of data (e.g. kept within a ring buffer) needs to be subjected to other pipelines. In astronomy this will **enable follow-up observations with additional instruments** in a ”multi-messenger” approach [51] (see use case class 5), where the required speed of the processing is set by scientific (e.g. expected duration of an ”after-glow”) and technical conditions (e.g. physical size of a ring buffer). Generally, fast real-time clustering and pattern recognition algorithms are also needed to identify and separate background from physically interesting events.

In particle physics, Belle II will achieve world-record luminosities with very high beam backgrounds. Potential future luminosity upgrades will require ultra-fast clustering, cluster-splitting, and photon identification utilising energy, timing, and pulse-shape discrimination information on hardware trigger levels. In astronomy, the **data rates of SKA will exceed the global internet traffic by a factor of a few**. Using the MeerKAT, we can prepare for this situation and work on methods to identify rare signals (such as FRBs) and to provide a full analysis of the data (e.g. to obtain the source position) before they disappear.

Goals: WP 2 aims to develop methods to identify and select signatures of specific and rare signal events (D-TA5-WP2-1, D-TA5-WP2-2, D-TA5-WP2-3). We will **probe a range of technology solutions**, from using high-end FPGAs (with input bandwidths of Tbit/s per device and several million programmable logic elements) to a complex mixture of hard-, firm- and software solutions over central or distributed HPC clusters. We will investigate if we can achieve a **paradigm shift** in identifying basically all known processes and signals, so that only the remaining unknown events with generic anomalous signatures are selected. In particle physics this will allow almost model-independent selections of ”beyond standard model” physics, while in astrophysics previously unknown phenomena can be discovered, like in the case of FRBs. We want to demonstrate not only the similarity in the underlying technical challenges caused by huge data streams, but we also want to show that generic solutions can address a common scientific goal, here identifying the nature of ”dark matter” (see use case class 5). These plans require the **very fast processing of a high-dimensional feature space**, so that implementations will be based on a varying set of solutions, on the hardware level (FPGAs, GPUs, CPUs), or on the algorithmic level based on advanced machine learning methods implemented in software. The constraints given by real-time systems require a **focus on algorithmic performance, highly efficient usage of hardware resources, and latency**.

Measures: On the shortest timescales, FPGAs provide efficient solutions for data reduction and pattern recognition tasks when very high bandwidth, parallel processing and short latency are required. These tasks can be performed using deep learning methods, e.g. convolutional and recurrent neural networks. But FPGAs are not the typical target hardware for the solutions provided by deep learning tools like PyTorch, TensorFlow and others. It is planned to develop a **generic tool that converts trained neural networks into efficient high-level synthesis (HLS) and VHDL firmware implementations** (D-TA5-WP2-4). Initially, network implementations in firmware will be pursued, leading to a baseline version of a tool for selected network architectures. In parallel, common interfaces of the deep learning output to HLS design tools will be identified and developed. Moreover, the results may be combined with existing tools, like HLS4ML. Finally, the performance of different implementations shall be compared and optimised with respect to applications at LHC, HL-LHC, Belle II, and possibly future collider experiments (D-TA5-WP2-5, D-TA5-WP6).

Exploring solutions less bound to specific hardware architecture, PUNCH4NFDI will use real-time MeerKAT data to develop **algorithmic methods implementing low-false-alarm-rate detection schemes** (D-TA5-WP2-3). In collaboration with WP 5, the developed methods will overcome the challenge that man-made background signals partly show time-varying features of the looked-for astrophysical signals. The results will demonstrate advantages of using a dynamic filtering architecture that is able to accept feedback from archival signal analysis studied in WP 3 (D-TA5-WP3-3). This part will also adopt results of WP 1, ensuring that a constantly modified real-time processing is correspondingly reflected in the metadata passed on to the archives (D-TA5-WP2-1).

Deliverables:

- **D-TA5-WP2-1 (31 Mar 2022):** Curation & metadata schemes for dynamic filtering.
- **D-TA5-WP2-2 (31 Mar 2022):** Strategy concept for identifying highly complex (multi-parametric) signals in huge data streams.
- **D-TA5-WP2-3 (30 Sep 2023):** Test environment for identifying highly complex (multi-parametric) signals in huge data streams using MeerKAT data.
- **D-TA5-WP2-4 (30 Sep 2024):** Generic tool to convert trained neural networks into efficient HLS/VHDL FPGA firmware optimised for a real-time, low-latency environment.
- **D-TA5-WP2-5 (30 Sep 2025):** Algorithms for massively parallel real-time sorting, clustering and pattern recognition on specialised hardware.
- **D-TA5-WP2-6 (30 Sep 2025):** Algorithms and Machine Learning methods for filtering and selecting relevant transient/anomalous signals.
- **D-TA5-WP2-7 (30 Sep 2026):** Pipeline for anomalous signal detection with low false-alarm probability for multi-messenger follow-up.

5.5.3 WP 3: Dynamic archiving

Status: The aspect of **dynamic archiving is best studied with state-of-the-art astronomical data** due to the large ground work already done and on which also TA 2–4 build their activities.

Those archives rest on the principle of recording all measurements done by an observatory, often together with metadata describing how the measurement was done, as well as some ancillary (e.g. weather conditions) or derived quantities (e.g. energy or classification). Examples of archives include Gaia, the Mikulski Archive for Space Telescopes, and the SDSS. They are functioning well, but carry fundamental limitations: They are **built on the assumption that (1) all data can indeed be recorded (complete archives)**, (2) a sensor archive can be developed in isolation (static, no relation to external actions), and (3) final archives are published as subsequent releases, often containing re-calibrated and/or updated measurements compared with previous versions. A critical task for archives is to **continuously monitor the data quality and (re)examine the causal inference of any dynamic filters** in use (see WP 2). This is particularly necessary as the scientific question may change over time, which might require a complete re–evaluation of the archived data and to implement corresponding new online filters or to adapt the entire online analysis. Such **connections between real-time and archive analysis has led to several breakthroughs in astronomy**, with the most recent textbook example being the already mentioned FRBs, which was discovered in 2007 by analysing archived data from 2001. This discovery allowed dynamic filters to be readjusted such that FRBs are now regularly discovered in real-time. Indeed, real-time filters are today updated between observations based on updated findings in stored data (see use case class 5). This shifts and modifies the distinction between real-time and offline processing and archiving, where the **transition point will be changing depending on the available online compute power and/or the challenge being addressed**. However, while examples like the discovery of FRBs show the potential for linking real-time and archive analysis, there are currently no methods for doing so systematically.

Goals: The WP aims to develop the **concept of dynamic archives and provide methods for accessing these**. Hence, it focuses on developing **generic concepts for workflows that this feedback between online and offline analysis**. We take into account that dynamic archives intrinsically **incorporate the concept of incompleteness**, relations to other sensors (e.g. ”did this measurement trigger a reaction or was it obtained due to another action?”) and evolution in time (i.e. the interpretation/existence of an observation depends on when it is queried). The goal will be to create an **iterative system which makes it easy for scientists to probe large, heterogeneous real-time datasets for previously unknown or rare events**. Such a realisation of dynamic archives, will allow ”what if” questions to be asked: Based on an existing set of combined dynamic archives, what would the result have been of a potential real-time filter and reaction network? Which events would have been selected, and how well can one say this considering the missing information in the current archive? Dynamic archives are thus directly related to dynamic filters (WP 2) in that a query to a dynamic archive can be directly related to a filter applied to (a set of) real-time data streams.

Measures: This WP will provide methods for more general and explicit links between dynamic filters and archives to **facilitate changes to dynamic filters on short time scales** (D-TA5-WP3-1). In collaboration with WP 1 and 2, we will develop a general, FAIR-compatible protocol for how a sensor can archive the criteria by which events were discarded, why a particular event was chosen and what the sensor sensitivity was at a given instance of time (e.g. direction, sensitivity).

Work will be carried out in collaboration with WP 2–4. In the next step, methods will be developed by which one or more dynamic archives can be jointly queried to return potential triggers as well as an estimate of how well the query could be answered (D-TA5-WP3-2). Such general **queries are likely to rely on modern statistical and machine learning techniques**. Methods for transforming a dynamic archive query into a dynamic filter will be tested by transforming dynamic archives into *simulated real-time* streams to which dynamic filters (WP 2) can be directly applied, and by the use of real data streams from MeerKAT. The result is the development of a **complete dynamic filter and dynamic feedback loop** (D-TA5-WP3-3).

***Deliverables**:*

– **D-TA5-WP3-1 (31 Mar 2023):** Methods by which one or more dynamic archives can be jointly queried and interpreted in the presence of information loss.
– **D-TA5-WP3-2 (30 Sep 2024):** Methods for transforming a dynamic archive query into a dynamic filter (and vice versa).
– **D-TA5-WP3-3 (30 Sep 2026):** Complete dynamic filter / archive feedback loop.

5.5.4 WP 4: Scaling workflows

Status: Scaling workflows is often done by parallel computing on HPC resources, but in order to address our dynamic requirements for huge data streams and volumes, complex workflows need to be developed. The workflows in WP 2 and 3 are derived for today's real-life examples. In order to apply them also to the next generation of experiments, massive scaling of these workflows is required. In order to **scale dynamic filtering** workflows, in-depth studies of the performances of the developed software may require developments of dedicated hardware layouts for different detector setups. **Scaling** workflows for **dynamic archives** means to be able to deal with single data items (e.g. high-resolution image from the SKA) as large as one Petabyte. Such **”data monster”** cannot be analysed reasonably fast on traditional computing architectures. The relatively small throughput rate when reading data from disks is a serious bottleneck (memory–wall problem). Here, memory–based computing offers a change in paradigm from the current processor–centric architecture to a memory–based architecture to make data monster available for dynamic archiving. One attempt is Gen-Z, an open ”memory-semantic” protocol and the basis of a memory-based computing prototype at HP Labs (with 100 TB main memory). HTW ported the big data framework *Thrill* to the prototype and performed first scaling tests. Another issue is the efficient usage of storage for processing data of different sizes. In many cases optimising the access to memory for files stored on different local computing clusters would significantly reduce latencies and could, at the same time, reduce energy needs. In particular, disk caches or buffers can hide latencies of the network from the client and correspondingly also from users. This is most relevant for cases of significant data reuse, where caches may also reduce wide-area network bandwidth requirements.

Goals: WP 3 aims at **identifying possible technology solutions** (algorithms/hardware/architecture) and requirements for scaling the workflows in WP 2 and 3. Interfacing with WP 2–4, standardised protocols for dynamic archival and time-critical feedback to data gathering instruments (dynamic filtering) will be developed with a **focus on green computing**. In particular, we want to investigate different strategies for caching files relevant for repeated access. By caching merely what is used, the disk is only used for storing files of the working set rather than the entire file. This strategy can indeed save significant amounts of storage space and, potentially, energy. In order to evaluate the efficiencies of different strategies we plan to study the XCache (XRootD Caching Proxy) method, which allows to setup a ”cache” server which saves frequently-accessed files at a specified ”origin” to local storage (D-TA5-WP4-3). A **prototype for a framework integrating and connecting software components for real-time data taking** and data processing shall be provided to implement algorithms and different topologies in a flexible and efficient way. Tools should monitor, profile and debug the data flow. The complexity of the internal structure should be hidden as much as possible from the user and at the same time allow to implement complex processing graphs.

Measures: Algorithms, workflows and architecture required for **scaling dynamic filtering** will be optimised. We will optimise load balancing by deriving decisions, where to execute a certain algorithm on the current and predicted load of the nearby CPUs and GPUs. We investigate a possible hard/software co-design based on the new type of CPUs which include FPGA or GPU features (see [52] and [53, 54] for applications in LHCb and ALICE experiments, respectively) (D-TA5-WP4-5). The load-balancing algorithms will be augmented to include this new hardware to dynamically optimise the performance of this new compute platform (D-TA5-WP4-6). In order to **scale dynamic archiving** we will optimise offline algorithms, including investigating the optimal transition between real-time and archiving processing using aspects of the astronomy examples in WP 3. We will integrate an analysis of the data Monster on memory-based computing platforms into the *dynamic data life cycle model* (D-TA5-WP4-1/2). We combine this with an evaluation of the efficiencies of different caching strategies for processing a set of relevant files (see also D-TA6-WP5-2). We will provide concepts for analysing huge data streams in parallel by complex iterative workflows as they are typical in radio astronomy. Existing workflows in astronomy will be reorganised such that the communication volume increases only sub-linear with the number of nodes (D-TA5-WP4-7).

Deliverables:

– **D-TA5-WP4-1 (31 Mar 2022):** Porting common off-line packages (e.g. CASA) to a memory-based computing prototype to prepare analysis of “data monster”.

– **D-TA5-WP4-2 (30 Jun 2024):**Standard software (e.g. CASA) compatible with Gen–Z.

– **D-TA5-WP4-3 (31 Mar 2025):** Caching strategies for processing a set of benchmark files with the evaluated efficiencies and latencies.

– **D-TA5-WP4-4 (30 Jun 2025):** Porting CASA to a HPC platform with appropriate scaling.

– **D-TA5-WP4-5 (30 Jun 2025):** Concepts for the optimisation of the hard/software co-design for CPUs which include GPU or FPGA features.

– **D-TA5-WP4-6 (30 Jun 2026):**Efficient real-time data processing framework.

– **D-TA5-WP4-7 (30 Jun 2026):** Scaled feedback interfaces between off-line software (e.g. CASA) and (selected) real-time processes using MeerKAT data.

5.5.5 WP 5: Evaluation and validation of instrument response & characteristics

Status: Data **quality assurance is of utmost importance**. In terms of FAIR, this means that the data need to be complete, self-explanatory, and accurate. While this is **challenging in the era of data irreversibility**, it is crucial to continuously monitor the detector performance, aiming to identify deviations from standard performance as quickly as possible to ensure the correct functioning of the dynamic filters. This challenge is **particularly severe when variable and uncontrollable background noise is present**. Examples in astronomy are the interference caused by signals from radar, mobiles or WIFI, or the disturbances caused by sun-light reflections from satellites (e.g. Space-X). In such cases, the usual **criteria or algorithms for triggering methods may not be applicable**, and a highly dynamical (and usually unpredictable) signal **background can ruin all data if the filters are static and unchecked**. Similarly, the application of improved calibration constants to measurements or the correction of certain data sets are typically performed in a re-processing stage well after the data are recorded. More refined tools are obviously needed to achieve an online calibration of the experiment synchronously to the data taking, allowing a full real-time reconstruction of the events.

Goals: The goal is the development of **tools for predictive maintenance process control**, based on unsupervised machine learning algorithms, and to devise strategies to **ensure the quality of the data even in the presence of variable, unpredictable background noise**. Typical monitoring tools are the comparison of histograms, the inspection of snap-shot images for unusual patterns or the scanning of time series for unusual statistics. If data rates are huge, more sophisticated tools need to be developed, as even simple monitoring is challenging. We want to derive solutions using real-life experiments in particle physics and radio astronomy, On the one hand, a real-time energy reconstruction of calorimeters is required, which directly feeds into the feature extraction and event filter systems. Deviations of detector signals from nominal need to be detected and automatically corrected. On the other hand, we study the impact of the exposure to terrestrial and satellite interference signals, respectively.

Measures: We will develop methods to predict possible deviations or failures of detector performance based on numerous parameters, external and internal to the experiment (D-TA5-WP1/2). The result will be a **tool for predictive maintenance and process control, based on unsupervised machine learning algorithms, and methods for anomaly detection**, and changes in calibration constants or dynamic filtering settings (D-TA5-WP5-3/4). The subsequent generalisation towards the inclusion of the entire detector system is a basis for transferring the methods to other PUNCH experiments as well as to other fields of science and industry.

Deliverables:

- **D-TA5-WP5-1 (30 Sep 2024):** Development and implementation of machine learning prototypes for anomaly detection, predictive maintenance and process control.
- **D-TA5-WP5-2 (30 Sep 2024):** Interference recognition and mitigation schemes for transient discovery leading to a “robust” triggering system for multi-messenger follow-up.
- **D-TA5-WP5-3 (30 Sep 2026):**Expansion of the concept to a generalized toolkit for predictive

maintenance and anomaly detection.

- **D-TA5-WP5-4 (30 Sep 2026):** Evaluation of false-alarm rates and improvements via machine learning, dynamic queries, on-line feedback and modification of archive metadata

5.5.6 Sustainability, risks, and mitigations

TA 5 provides answers to challenges for the immediate future of PUNCH experiments by looking at specific examples. Without its templates, many experiments will focus on its own individual solution to deal with data irreversibilty. The joint cross-community developments will allow the optimisation and generalisation of already existing code to avoid duplicating work. There is a risk that non-optimal or failing solutions will lead to a loss of scientific data and/or wasted compute power. WP 5 of the TA is designed to mitigate this. Miscommunication with other TAs or failure in some deliverable will be felt most severely only in the next funding cycles but remains a matter of concern. A common risk is the late employment of personnel, technical problems in the context of hardware, software, firmware or high-level design tools. As means of mitigation of the potential risks, it is foreseen to regularly get feedback from the user community. Regular meetings within TA 5, and with the other TA leaders in the Management Board, are essential to ensure enduring mitigation.

5.6 Task area 6: Synergies and services

This task area addresses the PUNCH objectives 1, 3, 4, and 6. Collaboration in general and especially in the fields of (big) data management and computing is at the heart of activities in the communities joined in the PUNCH4NFDI consortium. Collaboration between members of the consortium ranges from the joint work of a small number of individuals on very specific scientific or technical questions up to large-scale cooperation in large international facilities like the LHC, Gaia, H.E.S.S., or the FAIR facility, or in the truly global enterprises like WLCG, ACS or the IVOA. Many of these forms of collaboration have been active for more than a decade. By their nature, scientific work and hence also collaborative efforts are intimately related to efficient **data management from observation/measurement up to publications and long-term archives**. The close international cooperation and the very diverse set of partners enabled a fabric of synergies in data management that has been very beneficial (e.g. the global use of standard data formats in sub-disciplines of PUNCH4NFDI). However, collaboration and cooperation do not stop at the boundaries of the PUNCH community. Strong ties to neighbouring communities such as e.g. software development, engineering, mathematics, or high-performance computing exist already. With growing demands on **computing power, storage, bandwidth, and AI/ML solutions** in other fields of science, further collaboration will emerge and increase the benefits of synergies to be realised. Already in the course of establishing the PUNCH4NFDI consortium and defining its contents and work programme, other NFDI consortia were contacted to identify common challenges. **Task area 6 "Synergies & services"** will support and strengthen this collaborative and synergetic approach. It targets cross-cutting activities that foster an exchange of concepts and developments among the PUNCH4NFDI 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 (AAI)**). A marketplace will manage and trigger the exchange of concepts and solutions within PUNCH4NFDI and with other consortia.

While TA 2 addresses data management, TA 3 advances data transformation, TA 4 defines a data portal, and TA 5 investigates the data irreversibility challenge, TA 6 will promote the advances made in all other TAs in the PUNCH community and the NFDI in general. Researchers will find technical information in the marketplace about **setup and usage of data lakes** as well as **easy-to-use entry points for working with grid, HPC and cloud resources** and also **public entry points to open data and a software platform**.

Based on their experience and expertise relevant to topics addressed in TA 6 (see section 1 for participants, for co-applicants see section 3.1 and table 2), this TA will be handled by PUNCH4NFDI partners AIP (WP 6.1, 6.3, 6.5), DESY (WP 6.1, 6.5), FZJ (WP 6.1, 6.2), GAU (WP 6.5), GSI (WP 6.1, 6.2, 6.3, 6.5), HTW (WP 6.3, 6.5), KIT (WP 6.2), LMU (WP 6.4), MPIfR (WP 6.3), MPIK (WP 6.3, 6.4), UB (WP 6.1, 6.2), UHD (WP 6.1, 6.3, 6.4, 6.5), and WWU (WP 6.5). The work will be supported by the following participants who contribute with the specific expertise mentioned in section 3.1: DPG (WP 6.1), LRZ (WP 6.1, 6.3), PTB (WP 6.1, 6.3, 6.4), TIB (WP 6.1, 6.2, 6.3), and VdR (WP 6.1, 6.3, 6.5).

5.6.1 Work package 6.1: Marketplace

This work package addresses the PUNCH objectives 3, 4, and 6. The different communities that coordinate their efforts within PUNCH4NFDI have developed many aspects of data management within their specific fields, are intimately linked to international efforts, and continue to contribute to the evolution of global standards. The inspiration provided by approaches pursued in neighbouring fields has been an essential element in reaching the current status of data management. Conversely, ”exporting” solutions to other fields has strengthened the versatility of own concepts. It verifies the desired reusability in a broad range of applications (e.g. the globally used FITS format enabling a seamless combination of data obtained from many different research infrastructures). **Established links shall be exploited** to support the developments in all task areas in PUNCH4NFDI. A communication platform shall facilitate the information exchange with co-investigators, scientists in research fields of the PUNCH4NFDI initiative that do not actively participate in the programme, and with other NFDI initiatives.

This **exchange of ideas** will be promoted and managed via a marketplace. It will include a series of notice boards which will be proactively curated, an exchange platform, and measures to exploit products and services provided by cooperation partners.

The notice boards will provide links between existing notification platforms operated by subcommunities in physics and astronomy, the exchange platform will link to services and solutions available through the PUNCH4NFDI scientists engaged in international collaborations, and active exchanges will be established to other NFDI consortia.

The marketplace will thus serve as a platform for the exchange of cross-cutting topics, concepts, services, and joint development activities within the NFDI. It will be a point of contact for scientists offering or seeking solutions and will facilitate the exchange with other consortia and communities, international data management projects and research infrastructures.

Status: Many research groups in the PUNCH community established cooperation on data management with other scientists well before the NFDI initiative. Examples are image processing technologies, archiving and statistical methods or the provision of services for computing and data storage. Scientists in many fields wish to associate data sets describing the same object, but obtained by different means and facilities. This is very similar to the needs encountered in ”multi-wavelength” astronomy and exemplifies the need for FAIR principles. During the preparatory stage of PUNCH4NFDI, TA 6 has already established active collaborations on this topic with researchers in other consortia. First contacts with representatives from topical task areas have been initiated so that specific work can be addressed already from the start of the project. Joint work programmes have been agreed upon and are coordinated with the respective consortia.

PUNCH4NFDI has set **collaboration agreements** with the NFDI consortia MaRDI, NFDIxCS, NFDI4Earth, NFDI4Ing, NFDI4Culture, NFDI4Chem, NFDI4Microbiota and further initiatives mentioned in section 3.2.

The PUNCH community enjoys long-established and close international cooperation and carries out most of its research at internationally operated research infrastructures and in international collaborations. The latter are often supported by European initiatives. Solutions in data management developed by these research infrastructures, collaborations and initiatives are generally freely available for non-commercial use and may be shared with other NFDI initiatives. Specific examples are results of ESCAPE, EOSC, ErUM-Data, and ErUM-Pro.

Goals: The marketplace shall provide information about existing solutions and developments as well as services, both for the PUNCH4NFDI consortium and the NFDI. **Synergies** among the task areas and communities of PUNCH4NFDI as well as with other consortia are identified to enable further cooperative work. It will be the main point of interaction, collaboration and common developments between PUNCH4NFDI and other NFDI consortia as well as between the different communities and task areas within PUNCH4NFDI . A well connected and informed network for data management solutions will be set up and operated. In particular, TA 6 plans to make services available that PUNCH4NFDI can provide to a new target community. Additionally, services of external consortia will be adjusted to the needs of PUNCH4NFDI.

Work programme: TA 6 will **set up, operate and curate communication channels and tools** that will allow all researchers within the PUNCH community to promote services, to access external repositories or solutions in terms of data management questions. It will contain a corresponding notice board for all researchers that seek solutions. This exchange platform shall be extended to services and solutions that are available by **international organisations, research infrastructures and IT initiatives** to which members of the PUNCH community contribute as developers, partners, or co-investigators. A staffed forum will manage this exchange. Furthermore the exchange platform will facilitate **interaction with other NFDI initiatives** (see section 3.2) or national, European and international programmes by actively following the work programme of funded initiatives and by establishing contacts between other initiatives and the relevant task areas within the PUNCH4NFDI project. TIB and PTB will support this exchange building on their overarching methodological expertise and their connections to other initiatives and research areas. Similar assistance will be given by the community-targeted expertise of the participants DPG and VdR who reach out to other areas of physics and informatics.

Deliverables:

- **D-TA6-WP1-1 (01 Jan 2022):** A marketplace noticeboard and tools for communication and exchange among communities and consortia will be set up to be operated throughout the duration of PUNCH4NFDI support.
- **D-TA6-WP1-2 (01 Jan 2023):** An exchange platform for information on archives, data management software, services and hubs supporting the communities active in PUNCH4NFDI will be set up.
- **D-TA6-WP1-3 (01 Jan 2023):** Options for synergies (exchange of information and common work) on cross-cutting topics will be discussed with all funded NFDI initiatives.
- **D-TA6-WP1-4:** Services and tools developed by PUNCH4NFDI will be made available to the NFDI.
- **D-TA6-WP1-5:** Active collaborations with other NFDI initiatives (as well as national and international projects on research data management) are pursued; details are given in 3.2 and below.

Common projects have been defined with the following consortia, structures, and initiatives:

- **MaRDI** and PUNCH4NFDI plan to explore viable analysis methods and **statistical procedures**, in particular the analysis in fields where there are no Gaussian errors. Collaboration is also foreseen on the optimisation of neural networks for specific PUNCH4NFDI use cases like particle tracking or identification. Another topic are critical **optimisation problems** in physics where gradient-based algorithms cannot be applied. This can be the case if it is numerically expensive to control the accuracy of a gradient evaluation or if the optimisation surface is too rough. Examples are the quark-mass dependence of hadron masses in QCD [55, 56] or amplitude analyses in hadron physics. While a scalable implementation of an evolutionary algorithm is already available within the Geneva framework [57], it is desirable to develop and benchmark further scalable algorithms for applications on large-scale clusters. This will be done also in collaboration with TA 3.
- **NFDI4HPC/NFDIxCS:** An important topic is the identification and application of standardised interfaces to compute and storage resources of HPC systems. In order to be able to categorise the type of PUNCH jobs running on HPC resources, PUNCH4NFDI will provide the necessary data and metadata to NFDIxCS in order to address the limitations and scalability of computer science approaches.
- **NFDI4Ing:** The specific TAs in NFDI4Ing PUNCH4NFDI plans to interact with are ”Base services”, ”Metadata and terminology services”, ”Repositories and storage”, as well as ”Overall NFDI software architecture – data security and sovereignty”.
- **NFDI4Earth:** Collaboration is foreseen in the publication, exchange and mutual use of data

related to common fields of interest.

- **NFDI4Culture:** Collaboration is planned on legal aspects (data protection and data “property” rights), AAI, data and software publication, education and training, metadata. Both consortia will strive for a single NFDI-wide AAI solution. They are engaged in citizen science and plan to exchange experience in promoting data science in planetaria and museums. NFDI4Culture will explore use cases treatment within the PUNCH-SDP.
- **NFDI4Microbiota:** Collaboration is foreseen in the areas of metadata and metadata standards. Additionally, dynamic caching technologies will be provided to NFDI4Microbiota.
- **NFDI4Chem:** Collaboration is foreseen in the areas of extended metadata and AAI.
- **EOSC, ESCAPE and RDA:** Task area 6 will take care of the harmonisation of service specifications as well as communication interfaces within the international context. Members of PUNCH4NFDI are active in EOSC, ESCAPE, and RDA.
- **Helmholtz Metadata Collaboration (HMC):** Several metadata services are currently shaped and implemented in the HMC platform. PUNCH4NFDI plans to adopt and utilise these services. PUNCH4NFDI supports the goal to establish machine-actionable FAIR Digital Objects as well as human readable information systems.

5.6.2 Work package 6.2: Authorisation and authentication infrastructure

Status: This work package addresses the PUNCH objectives 1, 4, and 6. A transparent and distributed access to resources needs a common authorisation and authentication infrastructure. This has been **realised early on** by the particle physics community, and a security infrastructure based on X509 certificates has been established more than 15 years ago. However, this approach appears less attractive to many other areas of science. In the recent years alternative methods, e.g. using Shibboleth or OpenID Connect, have been developed. These methods use the login credentials of the home institutions of the users for authentication. With the obtained token the user acquires the configured privileges.

Within the Helmholtz association an AAI-related working group was established for the Helmholtz Data Federation (HDF) and a **Helmholtz-wide AAI** has been set-up.

On the other hand horizontal authorisation across different research and e-infrastructures remains difficult. The first EU-funded AARC (Authorisation and Authentication for Research Collaborations) project (2015-2017)[8] gathered requirements from e-infrastructures on federated authorisation and authentication and created a blueprint architecture. The Helmholtz working group developed a prototype for an **AARC blueprint compatible AAI**. Members of this working group are members of the EOSC AAI working group as well. The prototype proves the functionality of the blueprint architecture and makes it possible to evaluate the Unity software as a proxy solution. The legal aspects and the AARC policy starter pack has been adapted to the Helmholtz situation and are operated in the context of the prototype.

Goals: WP 6.2 will provide an interface for user authorisation and authentication to the NFDI, to national, European (e.g. EOSC), and international AAI-solutions as well as those from industry.

[8] `https://aarc-project.eu/`

The work package will contribute to implementing **one NFDI-wide AAI** solution which fulfils the NFDI and PUNCH requirements. A second objective is to offer user and group management that allows for authorisation based on group affiliations.

Work programme: As a prerequisite of a standardised layer computing model, PUNCH4NFDI will work with the Helmholtz HIFIS platform, the DFN and European stakeholders like GEANT, EGI and EUDAT to simplify and consolidate authentication, single sign-on and identity management. The Helmholtz-AAI[9] and the EOSC/EUDAT B2ACCESS[10] are blueprints for a PUNCH-wide AAI. The **software base** for both implementations is Unity. These AAIs serve as proxies for institutional (DFN-AAI/edugain) and ”social” IdPs (Orcid, Github) offering OAuth and SAML end points and token translation for service endpoints. The work package will contribute to the implementation of necessary adjustments/interfaces so that all PUNCH4NFDI use cases can be fulfilled.

A user and group management that allows for authorisation based on group affiliations will be set-up by exploiting the **group management** within Unity. Again adaptations and missing features needed by PUNCH4NFDI use cases will be worked on by this work package.

The work package will pay close attention to **licensing** issues of data and software belonging to the ”Accessibility” part of the FAIR principles.

Deliverables:

- **D-TA6-WP2-1 (30 Sep 2026):** Prototype of PUNCH-AAI allowing first use cases to authenticate users (31.03.2022). Requirements from all use cases will lead to the ”Basic PUNCH-AAI” (31.12.2023) and further developed to the final ”Extended PUNCH-AAI”.
- **D-TA6-WP2-2 (31 Dec 2024):** Design of NFDI-AAI: Coordination of AAI deployment between PUNCH4NFDI NFDI, national and international stakeholders; Negotiate a coherent deployment of a distributed AAI with the consortium and beyond aiming for a (preferably common) AAI infrastructure usable by the PUNCH4NFDI consortium and most of the NFDI. A draft design will be ready 31.12.2022.
- **D-TA6-WP2-3 (30 Sep 2026):** PUNCH4NFDI group management. A prototype based on Unity group management enabling services to authorise users based on their group affiliations (31.12.2022) will be revised based on requirements of PUNCH4NFDI use cases (31.12.2024) and aligned with NFDI activities.

5.6.3 Work package 6.3: Cross-community efforts towards FAIR data

This work package addresses the objectives 2 and 6 (section 2.2). Complementary to the ”standard metadata” considered in TA 2–4, particular emphasis will be placed on ”**extended metadata**” needed for accessing cross-community data, and on ”**dynamic metadata**” necessary for coping with emerging demands due to the *dynamic life cycle model*, see TA 5.

Status: The various fields in PUNCH have individually developed advanced data management concepts and solutions. Most of them are based on international efforts, are embedded in Euro-

[9] https://login.helmholtz.de

[10] https://b2access.eudat.eu

pean projects and supported by large international infrastructures. Driven by different research methods, the communities have pioneered different elements of data management (see section 4.2). While many elements of the FAIR data concept have been realised, or are to be implemented, **exchanges between different research fields remain challenging** and hamper an exchange of data management methods and data. In particular truly open access, the necessity to extend metadata schemes to enable open access to data, and the link of data to the final scientific exploitation via publications is pursued to very different degrees and in very different approaches.

Goals: Within the PUNCH4NFDI project research infrastructures shall be supported when making data publicly accessible. This may include a treatment within the workflows provided by TA 2–5 or be limited to supporting access, curation and documentation.

A key element to enable identification, access and reuse of data being made available to the public is an appropriate extension of basic metadata to facilitate **cross-disciplinary data reuse**. Based on existing examples, it is intended to develop a **marker system** for metadata extending the existing frameworks. Indexing will facilitate access for other communities. E.g.: Long-term archives of astronomical data contain extended metadata on meteorological measurements obtained for calibration purposes which are unique archives of primary data in climate science in their own right. The activities in TA 2–5 will be carried out by a joint team for metadata treatment in PUNCH4NFDI. The astronomy, astrophysics, and space community have developed the globally used FITS data format, which is used by thousands of data providers and archives worldwide. The excessive amount of data-handling tools for this format resulted into its propagation in many other and very diverse fields (e.g. the Vatican library). A similarly widely used data format in many fields of physics is ROOT-compatible and used by many of the largest data providers of scientific data in the past decades (e.g. CERN). It is planned to extend early work on converting data between these formats to developing a **flexible format converter** that can be extended to other highly advanced data formats. Reuse shall further be promoted by reciprocal links of data and metadata to resulting publications and **advanced terminology services**.

Work programme: Many datasets stored by data providers within the PUNCH4NFDI community require excessive auxiliary data to be exploited. Parts of these data are treated in different context as metadata and often exceed the data volumes of the primary data. Definitions for these ”**extended metadata**” shall be provided. This will be incorporated into flexible format converters (D-TA6-WP3-3, D-TA6-WP3-5, D-TA6-WP3-6).

The most extensive data sets available to the PUNCH community and not yet processed by other organisations shall be made accessible through the PUNCH-SDP along with the auxiliary data/metadata (D-TA6-WP3-4, D-TA6-WP3-5, D-TA4-WP4).

Existing dynamic metadata frameworks in other communities will be explored. A selected framework will be customised for using in TA 5 (D-TA6-WP3-3).

A reference guide for **publishing data** including guidelines on DOI usage (D-TA6-WP3-1) will be published together with publishers and journals ((e.g. Computing and Software for Big Science (CSBS)). The link of data and corresponding software in publications will be covered in collaboration with NFDI4Culture and MaRDI (D-TA6-WP3-2).

Deliverables:

- **D-TA6-WP3-1 (31 Jul 2022):** Reference guide for scientists and journals on publishing data.
- **D-TA6-WP3-2 (31 Dec 2022):** White paper/reference guide for publication of software.
- **D-TA6-WP3-3 (31 Dec 2023):** Customising extended and dynamic metadata frameworks for integration into the PUNCH-SDP and the *dynamic archiving infrastructure* of TA 5.
- **D-TA6-WP3-4 (31 Mar 2024):** Making Effelsberg data openly available.
- **D-TA6-WP3-5 (31 Dec 2024):** Flexible converter for FITS/ROOT formats; Preparing HESS data for FAIR access.
- **D-TA6-WP3-6 (31 Dec 2025):** Demonstration of formats for metadata-extensions including auxiliary data through pointers and databases.

5.6.4 Work package 6.4: Open-source data analysis tools

This work package addresses the PUNCH objectives 2, 4, and 6. It will evaluate to which degree the requirements of the PUNCH community can be covered by open-source analysis tools and such bring the functionality of already existing open-source libraries into the PUNCH consortium. The use of these tools will be validated using reference data sets of PUNCH4NFDI and promoted by providing detailed analysis examples. In addition contributions by the PUNCH communities to existing open-source software projects are identified and coordinated in collaboration with TA 3 in WP 3.1 and TA 4 in WP4.4 (section 5.3.1). This work package also aims to establish **standard development workflows** and infrastructure inspired from the open-source software community within the PUNCH community.

Status: Traditionally, the communities involved in PUNCH4NFDI used their domain-specific software infrastructure for data analysis based on packages such as CERNLIB [58] or ROOT [59]. However, big data analysis got very popular also outside science and many powerful **general-purpose data analysis tools** have been developed. Those tools often provide alternative functionality to existing domain-specific solutions, but are often not yet validated for use with domain-specific data. The vast majority of those software projects are open source and openly developed, which means that the code base is open to external contributions and not owned by a single community. The step of opening up the development process using hosting services such as GitHub [60] or GitLab [61] not only made the development process more efficient, but also allowed for collaboration between different scientific fields and industry. Finally, it also helped to build an interdisciplinary, more diverse and inclusive online community of software developers with common interests and mutual support. Examples are the Python data-science eco-system with general packages such as SciPy [62], Matplotlib [63] or Pandas [64], more domain specific packages such as Astropy [65], Gammapy [66] and Scikit-HEP [67], front-end solutions such as Jupyter notebooks [68] and Binder [69] hubs and packages such as Dask, Ray and Spark for parallelisation, or Keras and Tensorflow for machine learning and high-performance computing.

Goals: Open-source tools and services get increasingly popular in many scientific fields as well as in the PUNCH community. By providing descriptions and documentation of selected **open-**

source software as well as providing reference analysis examples for most PUNCH4NFDI use-cases, the use of open-source tools will be promoted to science users. The quality of the example analysis results will be assured by using reference datasets provided by the PUNCH community. By opening up ”traditional” PUNCH software development workflows, contributions from external developers and communities shall be motivated. By contributing to open-source software tools, ideas and concepts for data analysis with other NFDI consortia will be exchanged. This includes establishing contacts and cooperation with the MaRDI and HPC communities as well as with international data science projects and organisations such as PYHEP [70] or IRIS-HEP, Python in Astronomy [71] and open-source Linux distributions such as Debian.

Work programme: As a first step, a survey based overview (D-TA6-WP4-1) with respect to use cases will be provided, as well as description and documentation of the open-source data analysis tools currently developed and used within the PUNCH community. A particular focus will be on statistical methods which will be evaluated in close cooperation with WP 3.1. The results will be used to identify which external open-source software needs to be included in the software platform (D-TA6-WP4-5). In addition, **use cases** will be identified where open-source analysis tools can represent an alternative to domain specific solutions. Finally, common approaches to data analysis within the PUNCH community will be identified and used as basis for shared contributions to already existing open-source packages. To open up existing (D-TA6-WP4-2)software development workflows, an example repository will be provided that demonstrates the use of state-of-the-art open development setups. This includes solutions for automated testing, continuous integration, as well as deployment using established packaging managers. The implementation of the solutions will be coordinated with TA 4 WP 4.

Finally detailed analysis examples (D-TA6-WP4-4) of validated reference data sets will be provided based on open-source tools, which are representative for the PUNCH research areas.

Deliverables:

- **D-TA6-WP4-1 (31 Jul 2022):** Survey based overview including documentation of analysis tools currently developed and used within PUNCH4NFDI.
- **D-TA6-WP4-2 (31 Jan 2023) :** Example repository with continuous integration and deployment setup along with general guidelines for open software development.
- **D-TA6-WP4-3 (31 Jan 2022):** Contribute to open-source Linux distributions for software packages relevant for PUNCH4NFDI including guidelines.
- **D-TA6-WP4-4 (30 Jun 2026):** Detailed and non-trivial examples of data analysis based on open-source tools and example data sets representative for PUNCH4NFDI.
- **D-TA6-WP4-5 (30 Jul 2023):** Software platform for PUNCH4NFDI and the NFDI with relevant software tools and frameworks.

5.6.5 Work package 6.5: Services in big data management

Status: This work package addresses the PUNCH objectives 3, 4, and 6. In the PUNCH research domain, there exists already a considerable number of efficient **state-of-the-art services** to manage the data challenges from large research infrastructures. Many of these services are also of interest for communities outside PUNCH4NFDI. The number of services is expected to increase considerably in the course of the NFDI as services developed in the task areas reach maturity and transit from the alpha-testing stage to beta-testing (see sections 3.5 and 4.4). Furthermore, as a result of other work packages of this task area, services will be adopted and made available for the use by other communities, within the PUNCH domain and outside.

Goals: The prime objective of this task area is the compilation and provision of a **portfolio of big data services** that are of generic interest for researchers. These services are then adopted to the needs of a broader community and made available for beta-testing. The list of big-data services will continuously be updated according to results from TA 2–5 and reanalysed, focusing on a set of deliverables as outlined below. Means for a sustained operation of these services will be sought for.

Work programme: The goals of this work package will be achieved by **offering large scale infrastructures and services** in a transparent way to the PUNCH community and beyond. This includes large-scale distributed computing and storage facilities, multi-archival cross matches as well as big data management related services. In particular, access will be provided to open data archives such as the open data of CERN's LHC experiments or the International Virtual Observatory (D-TA6-WP5-1). The use of opportunistic storage resources will be enabled and complemented by a service on dynamic disk caches (D-TA6-WP5-2). New scientific challenges include the analysis of"data monster", where a single data set may be as large as 1 PB and more. For that purpose, memory-based computing has been developed and will be pursued and offered to PUNCH4NFDI and beyond (D-TA6-WP5-3). Existing hardware resources such as the supercomputer "HLRN — High Performance Computing in northern Germany", GPU clusters or fractions of the Tier-2 grid computing centres GoeGrid in Göttingen and astronomical database servers at AIP will be made available via interfaces in community-specific applications. In particular the GoeGrid centres can be used as "open analysis resources" for CERN open data by users without explicit CERN or experiment affiliation (D-TA6-WP5-4). Furthermore, small fractions of various decentralised community specific resources will be jointly managed via the COBalD/TARDIS compute resource management software framework and provided to PUNCH4NFDI and beyond (D-TA6-WP5-5). Interactive data analysis for the PUNCH community will be enabled by using multi-cloud resources as compute and storage testbed (D-TA6-WP5-6). The JupyterHub platform developed in TA 2 will be provided for interactive data analysis and visualisation (D-TA6-WP5-7). Jupyter notebooks are an emerging standard for this kind of analysis and are used by researchers and data scientists in many fields. In TA 2, a JupyterHub platform is developed to run Jupyter sessions in the cloud, thereby using cloud resources and enabling access to data in the data lake. The TA 4 data portal will offer a more comprehensive collaborative environment based on Ref. [43]. In TA 6, standard analysis software and newly developed tools are integrated into use case specific Jupyter notebook images. These images will be developed as open source projects and may be of key interest for other communities, potentially leveraging further synergies. A cloud-based environment for PUNCH4NFDI microservices will be offered in deliverable D-TA6-WP5-8. Resources from different providers will be combined and jointly made accessible by clearly defined APIs. A metadata catalogue architecture with user interface and prototype implementation for PUNCH sciences will be implemented and operated (D-TA6-WP5-9). The file transfer service (FTS) allows bulk transfers of files in batch-style manner. The service is widely used in particle physics. Rucio is a data management system designed to cope with large data volumes and many files that are hosted on many geographically distributed storage systems. Its development started in the ATLAS experiment but it is now used by several large particle physics collaborations. FTS and Rucio will be offered to interested partners in PUNCH fields and beyond for evaluation purposes (D-TA6-WP-10).

Deliverables:

- **D-TA6-WP5-1 (30 Jun 2024):** Access to open data archives such as the open data of CERN's LHC experiments and the IVOA.
- **D-TA6-WP5-2 (30 Sep 2022):** Dynamic disk cache for including opportunistic storage resources.
- **D-TA6-WP5-3 (30 Jun 2024):** The Gen-Z open standard for memory-based computing is optimised within the astronomical framework CASA to analyse ”data monsters”. This may also serve as a prototype for genome research.
- **D-TA6-WP5-4 (30 Sep 2024):** Infrastructures and services for community - specific applications as the supercomputer ”HLRN - High Performance Computing in northern Germany” or the GoeGrid grid CPU cluster for education purposes, e.g. using CERN open data.
- **D-TA6-WP5-5 (31 Dec 2024):** Management of decentralised community specific resources via the COBalD/TARDIS compute resource management software framework and application to CERN open data platform for third-party users including the general public.
- **D-TA6-WP5-6 (30 Sep 2022):** Multi-cloud resources as compute and storage testbed.
- **D-TA6-WP5-7 (30 Sep 2026):** Integration of standard analysis software and newly developed tools into the JupyterHub platform of TA 2 via use case specific notebook images.
- **D-TA6-WP5-8 (31 Dec 2023):** Cloud-based environment for PUNCH4NFDI microservices.
- **D-TA6-WP5-9 (31 Dec 2024):** Implementation and operation of a metadata catalog architecture with user interface and prototype implementation based on lattice QCD (LQCD); metadata standards for astrophysical and automated simulation data publication.
- **D-TA6-WP5-10 (30 Jun 2024):** Set-up of FTS and Rucio for evaluation purposes for the upcoming challenges posed by HL-LHC or SKA.

5.6.6 Sustainability, risks and mitigation

Sustainability of the marketplace and of corresponding services is a central element for future interfacing with data management developments beyond the PUNCH4NFDI initiative, including the continued provision of services and access to archives as discussed in section 3.5 and at the beginning of this chapter.

The common risks for all task areas, in particular the recruitment of personnel, and their mitigation are listed in the beginning of this chapter. In TA 6, the availability of qualified personnel is mitigated by a significant involvement of in-kind personnel that will be able to bridge short-term delays. A specific risk results from agreements with other NFDI consortia being based on best-effort only. This is mitigated by the intention to broaden the range of interactions to other consortia after each application cycle.

5.7 Task area 7: Education, training, outreach, and citizen science

Task area 7 addresses objective 3 to train the PUNCH community on data science methods, to provide data access and to educate and engage society at large in data science. It addresses structural challenges related to **gender equality** and regarding access to computing and information resources. Its target groups have a wide range of interests with diverse backgrounds, to which measures are tailored within four main areas. Given the educational nature of the work packages, actions can be easily exported to other NFDI consortia and beyond via the marketplace (section 5.6). Sustainable education in data science aspects and methods lies at the basis of all use cases and aims of PUNCH4NFDIĖducation and public outreach also provide a natural link to other scientific communities and NFDI consortia.

By providing **access** to data, methods, tools, resources for training events and career advancement, work package WP 7.1 supports expert scientists in the framework of a PUNCH *Young Academy* (PYA). Work package WP 7.2 focuses on students and early-stage researchers, for whom **basic training** resources on NFDI-related topics will be developed and supported. In work package WP 7.3, popular interests in astro- and particle physics are addressed by promoting data science methods and infrastructure for **public education and outreach**. By bringing data science to schools and homes, the need to engage children at early ages is emphasised. Active participation in research cultivates an understanding and appreciation of scientific method and reasoning. Through WP 7.4, the diverse public computing and digital communication resources to engage citizens in active science will be utilised, through **citizen science** initiatives such as Einstein@Home or Zooniverse. Regarding data science in physics, PUNCH4NFDI will aim for a lead role in organising professional training, a supporting role in the training of students at universities, and a coordinating and contributing role in the diverse public outreach activities that exist in the German science community.

The institutes involved in this task area have strong competence in education and outreach on different levels and in various scientific disciplines. These range from experimental (astro)particle physics (DESY, KIT, USi, GAU, TUDO) and astrophysics (MPIfR, UB, UoB), to heavy-ion physics (GSI, GU), large-scale computing (FZJ) and data science (HDA, JGU). All university groups involved have a naturally strong background in teaching, training and outreach activities. All research institutes and laboratories involved have been committed to the training of scientists as well, and to public outreach.

5.7.1 Work package 7.1: Training of scientists – **PUNCH** ***Young Academy***

Status: Knowledge about advanced data science and machine learning methods and their application to large and complex data sets are skills in high demand and dynamically growing inside and outside academia. Such skills are essential to exploit the full potential of all fields of science, in particular in the PUNCH community using large scale international research infrastructures in astro- and particle physics. Physicists are typically well-trained in mathematics relating to physical processes, but less so in inductive statistics, data and computer science relating to the extraction of reliable information from complex and often noisy data sets. They are typically conversant with computer programming and processing on a moderate scale, but many are not ready for a world of Big Data with challenges in data storage, access, and efficient analysis on high-performance multi-core computers, or with modern software-development techniques. Proficiency in such skills opens a wide range of **new job opportunities** for young and senior scientists in the dynamically developing data-driven industry as well as in public sectors.

Goals: Consequently, the central goal of this work package is the provision of broad, high-quality training in state-of-the-art **data science techniques** The target groups include Ph.D. students, postdocs, or senior scientists in the spirit of life-long-learning. The training will be online and through in-person workshops, including hands-on experience in tools and applications. Some measures will have broad, general scope, while others will be specialised topical workshops for advanced participants with community-specific needs. In the beginning, a focus will be on training in state-of-the-art methods in the fields, later on, this will be supplemented by courses and online material on the use of the new tools and services developed within PUNCH4NFDI.

The **PYA - PUNCH Young Academy** will offer career-development support to scientists with non-permanent contracts. PYA will place particular emphasis on the provision of training and career development for **female scientists**, who are still underrepresented in the PUNCH community. Data science-based jobs are especially attractive for women in this field since it is easier to accommodate a **life-family-work balance** here than for example in laboratory work. Many female students have already realised this opportunity, resulting in a relatively high and increasing fraction of female students attending the few already offered ML courses. PUNCH4NFDI will foster this trend by promoting the advantages for women working in this area of the PUNCH community and offering special courses addressed to female researchers.

Work programme: Initially, a **survey** conducted among universities and institutes will provide an overview of the needs and requirements as well as over any existing or newly developed training material. This survey will include available hardware resources for the hands-on sessions. WP 7.1 will identify possible speakers for general and topical workshops and suitable locations for in-person workshops. Regular offers of online and in-person courses and workshops (D-TA7-WP1-1) will be provided, where also **data protection and data property** rights will be covered in close collaboration with NFDI4Culture.

At the start of the PUNCH4NFDI funding period, the PUNCH Young Academy (PYA) will be established. This task includes the identification of potential mentors among the established senior scientists, collection of existing or newly developed material for general career development, collection and continuous update of information for the introduction of PYA fellows to the German academic and funding system, the organisation of best practice and exchange fairs including representatives from German funding agencies or foundations and institutions supporting science, and the establishment and maintenance of an alumni platform as well as the organisation of alumni days where present fellows can meet former members of the PUNCH community who have moved on to industry and now look for new, young talents. PYA will also place a special focus on career development for women and other minorities, their challenges, but also their special opportunities and chances (D-TA7-WP1-2). The promotion of women in PUNCH4NFDI and the NFDI is a special focus. **Roles models** in the national and international community will be invited, special aspects of mentoring will be offered and additional, specialised career development courses will be established (D-TA7-WP1-3).

All measures on professional training within this work package will be continuously evaluated to optimise them, adapt them to the needs of the scientists and the community and take latest developments from the other PUNCH4NFDI task areas or other NFDI consortia into account (D-TA7-WP1-4).

The material of all professional training courses, broad or specialised workshop as well as the PYA-career development, best practice, exchange fares, and alumni day events will be **archived**. They will all be accessible via the PUNCH-SDP. This archive will also include documentation, any further background material or statistical analyses of evaluations or alumni developments, possibly in collaboration with colleagues from social sciences. A specialised course for senior scientist will keep them up with current developments, because their awareness in these developments is crucial, as they have a large influence also on the development of younger staff (D-TA7-WP1-5).

A **Training Manager** will coordinate the professional training effort and the PYA, collect material, propose and discuss topics for schools and workshops with the members of the PUNCH4NFDI consortium as well as other NFDI consortia, take care of the administration and organisation of online or in-person courses and workshops, and conduct and analyse evaluations of all courses and workshops to improve them continuously. Activities also include the organisation of training courses and workshops, their logistics, content, and speakers from within the PUNCH community. The manager will also coordinate joint activities of the PYA and similar efforts by other NFDI consortia.

***Deliverables**:*

- **D-TA7-WP1-1 (30 Sep 2026):** Preparation of online and in-person courses and workshops including hands-on sessions with broad scope or topical workshops geared towards the PUNCH4NFDI specific developments.
- **D-TA7-WP1-2 (30 Sep 2024):** Establishment of the PUNCH4NFDI Young Academy including regular events and courses.
- **D-TA7-WP1-3 (30 Sep 2025):** Development of a special programme and specific courses geared towards female scientists.
- **D-TA7-WP1-4 (30 Sep 2024):** Critical review of the measures developed and development of a feedback system to guide education and training.
- **D-TA7-WP1-5 (30 Sep 2026):** Documentation and long-term archiving of training material, the service documentation and tool descriptions in coordination with the TIB Hannover.

5.7.2 Work package 7.2: Education of students

Status: Modern **university curricula** for astronomy and physics must provide the necessary skills and methods to cope with large and complex data sets. This is mostly not the case, since courses in modern statistics, applied mathematics, or computer science are not mandatory. Many curricula also lack basic training on software design principles, data structures, sampling methodology, statistics, algorithmic design and optimisation, standards and procedures. A systematic training in data science and big data management or modern statistical techniques such as machine learning does not yet exist even on the master or doctoral level. The problem is further aggregated by the fast development of the field and rapid ageing of the course material.

Goals: The education of students and early-stage researchers is a central aim of PUNCH4NFDI. A 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 a commonly recognised **core curriculum** that will also profit other NFDI initiatives.

This work package promotes technological literacy matched with a good skill set in data analysis and the adequate understanding of experimental setups. Universities are the key to transfer knowledge developed within the NFDI to a new generation of scientists, who act as knowledge multipliers in academia and society at large by providing best practice examples and become the teachers of future generations.

Work programme: A **survey** among universities will be conducted to obtain an overview over the conceptual integration of NFDI-related topics into the curricula, on the teaching methods and materials used, as well as on the extra-curricula activities in the NFDI realm (D-TA7-WP2-1). The survey will also include interviews with experts students in the field.

PUNCH4NFDI strives for a deeper and **standardised data-science education** at the university level given the necessity to harmonise physics and data science curricula with respect to statistics, computation, and data science (D-TA7-WP2-2). Based on the results of the survey and expert opinions from inside and outside the PUNCH4NFDI consortium, several levels of recommendations for such an NFDI schedule will be prepared. This includes proposals for minimal schedules to be incorporated in programs with limited flexibility, suggestions for extra-curricula courses, as well as the design of a fully-fledged course program on NFDI topics. The different recommendations will be tailored to the needs of bachelor, master and doctoral students. As a minimum, help will be given to provide students and early-stage researchers with basic training in the NFDI-related topics and tools, in order to guarantee the necessary expertise for scientific research and competitiveness on the labour market.

In addition, PUNCH4NFDI will provide recommendations and support for the development and organisation of **learning materials and tools**, to complement university-based training through extra-curricular training events (D-TA7-WP2-3). This includes the development of teaching material that can be used in the class room but also for independent learning. The material will be aggregated from the experts in the PUNCH4NFDI consortium and revised to match the recommendations for a harmonised curriculum. The development of events for scientists at all levels, from lecture-style workshops, hands-on tutorials, to developers' meetings is foreseen. The training and education programs will be designed to be complete and consistent, and focus on the most relevant state of the are technologies in computing and data management.

University-based education builds a natural link to other consortia with similar needs. The introduction and harmonisation of these topics into the university curricula will be achieved in collaboration with and support by the KFP and DPG, who already provide harmonisation platforms. A particular focus will be placed on the digitisation of teaching material.

While some universities host large computing infrastructures, others rely on the usage of external resources. The use of **computing infrastructure** for educational purposes, e.g. at large computing facilities and research labs, will be enabled (D-TA7-WP2-4).

In addition to the design and preparation of teaching material and schedules, **educational events** in the realm of the NFDI-topics will be offered (D-TA7-WP2-5). Examples are online courses on data management, programming skills and the usage of modern development tools. The events will be done in collaboration by specific hosts, e.g. university groups and lecturers, and carried out as online events or face-to-face meetings. The PUNCH community has a long-standing history in education and training, exploiting the power and diversity of our national research infrastructure. The PUNCH4NFDI efforts will be coordinated with other community-wide educational programs and resources. A bundling and streamlining of existing and newly developed training and education programs is foreseen. The programs developed inside the PUNCH4NFDI consortium will be opened for other consortia. Examples are the CERN School of Computing, the GridKa School, the education programs of the Helmholtz Alliance *Physics at the Terascale*, the *International Virtual Observatory Alliance*, or various data science summer schools and courses offered at our universities.

An **Education Manager** will coordinate and administer the education efforts, perform a **market survey of available concepts and materials** in the PUNCH community, collect material, propose and discuss topics for schools and workshops, and organise educational events. Example course content and material developed by PUNCH4NFDI will be provided to all universities and institutions. Long-term **archiving** of the learning material is foreseen in coordination with scientific libraries, e.g. TIB Hannover.

Deliverables:

- **D-TA7-WP2-1 (30 Sep 2022):** Market survey of available teaching concepts and material.
- **D-TA7-WP2-2 (30 Sep 2022):** Development of standardised curriculum.
- **D-TA7-WP2-3 (30 Sep 2026):** Compilation and development of teaching material for courses and independent learning.
- **D-TA7-WP2-4 (30 Sep 2026):** Aggregation of data resources and access to computing infrastructure.
- **D-TA7-WP2-5 (30 Sep 2026):** Coordination and initiation of educational events, e.g. visiting seminars.

5.7.3 Work package 7.3: Public outreach

Status: Natural science derives from an archaic curiosity about human existence within a vast cosmic space and history. To popular culture, astronomy and physics not only contribute wondrous images and tales of mysterious objects, but also a sense of orientation in space and time, of a certain place, a home in cosmic history. Modern physics became a science with sophisticated methods, big data volumes, and a language that requires professional training. Its fundamental questions and objects of interest enjoy wide public interest, but public perception of contemporary physics ranges from intellectual appreciation, over mystic awe to relativism.

An important goal of science education is that members of modern society are able to locate themselves, their fellows, and their environment as part of a world that can be understood through **scientific enquiry and objective knowledge**. This not only empowers them to participate in rational social and political debates, but it opens career options in fields with high demand and great social and economic relevance.

Modern science relies increasingly on big data, artificial intelligence, and machine learning techniques, which are disruptive technological advancements that meet fear, but also attract young people, and inevitably affects the future of academic and professional life.

Although the importance of **communicating science to a wide public** is evident, outreach has not been near the centre of scientific practice, but is commonly added on without dedicated resources. This can be a dangerous deficit, as fundamental science depends not only on public financial support, but on an educated and well-informed public that understands and appreciates the scientific approach and methods.

With a flood of information through rapidly evolving modes of communication, science communication is increasingly reduced to discovery news items of short-term relevance that often lacks depth, context, and a critical emphasis on the scientific method. This entails a danger that the distinction between fact and belief blurs, while trust in scientific truthfulness and open discourse gets lost and suspicion grows towards a science and technology that appears to follow obscure algorithms and ulterior interests.

Given the public interest in its fundamental questions, physics is an ideal science to address these concerns and help to develop a civil understanding and empowerment of the wide public that in turn supports the scientific endeavour. Physics is an natural subject to develop the skills necessary to cope with vast information, not for passive consumers, but active players. Physics and astronomy data are generally public, and its protagonists are mostly free of commercial interest and are eager to share their joy, awe, and reverence.

An important target group of the PUNCH4NFDI education and outreach activity are children and adolescents, teachers, and decision makers. **Children and adolescents** relate to science mostly through what they learn in school. Because astrophysics and data science are usually not part of the school curriculum, most children have little relation to the rapid developments in these areas, which is compounded by the decreasing popularity of physics in schools, and by the dramatic shortage of qualified physics teachers.

However, most children are fascinated by astronomical phenomena and objects. When providing them with tools, data, and an attractive context, their interest can easily be directed toward an **active engagement** in astronomy, physics, and the related data science. This can be aided by a growing interest among children in computer games, robotics, and in informatics as a school subject, as well as by the availability of dedicated platforms like *Scratch* or *App Inventor* that allow them to playfully engage in computing and discovery. A promotion of data science through astronomy and physics is therefore a promising way to strengthen both natural and data science in schools and in extracurricular activity.

Goals: The PUNCH4NFDI outreach activity aims at developing an **adequate infrastructure** to foster public interests in science and an understanding of how data is the foundation of and data science methods enable discovery in modern physics and natural science. It is an important goal of PUNCH4NFDI to educate the public on how to read, cope, and interpret data, to provide an adequate understanding on topics such as big data, artificial intelligence, or machine learning.

Physics addresses fundamental questions of humankind, without national or other borders. Promoting **gender balance and cultural diversity** is therefore also an important, natural goal of PUNCH4NFDI, which outreach activity implements proactive, affirmative measures to counter historically grown or socially ingrained discrimination.

Work programme: PUNCH4NFDI will use the popularity of astro- and particle physics with the media and the general public to communicate the role of data science methods for the advance of science. Specific topics to be communicated are based on input and needs from the NFDI community and will be continuously surveyed and reviewed.

PUNCH4NFDI ensures that **public access points** exist that allow to engage in the exploration of science data, providing tools and interfaces that enable the general public to access, explore, and analyse authentic physics data. Such interfaces will be tailored to different levels of expertise, from school students and teachers, to amateur scientists.

PUNCH4NFDI will be represented to the public through the classical channels such as websites, blogs, and twitter (D-TA7-WP3-1). To help the PUNCH community in their science outreach activity, a common **repository for outreach resources** will be provided that focuses on data science in physics (D-TA7-WP3-3). PUNCH4NFDI will initiate **science communication training** for scientists, including introductions to social media platforms (D-TA7-WP3-2). Public discourse about science will be analysed on how it relates to communication by scientist, professional science writers, and the media. Metrics will be established to evaluate and report on the impact of outreach activities, learning tools, access paths, and impact.

PUNCH4NFDI will initiate and support efforts to **stimulate the interest** of children to engage with natural and data sciences, and consider an academic education in these areas. Based on PUNCH4NFDI results and expertise, written and digital material, hands-on school courses and an app or equivalent media will be provided that can be distributed to teachers, high-school students and the general public. Material will also be presented and distributed via YouTube or the TIB library Hannover AV-Portal, as well as at science and education fares such as didacta, IdeenExpo, Highlights der Physik (D-TA7-WP3-4). This material will help teachers to introduce the topic of data science, data literacy, machine learning algorithms. **Hackathon school com-**

petitions will be offered to get pupils and teachers actively involved in PUNCH topics, and **masterclasses** with NFDI-related topics will be developed. Based on the positive experience collected with the concept of masterclasses over many years, a **rent a scientist** program will be established, where consortium scientists visit schools to promote data science to pupils and teachers (D-TA7-WP3-7).

Leaning on the successful work of various IVOA projects in bringing astronomy to schools, in cooperation with international colleagues PUNCH4NFDI will promote the development of and improvement of tutorials, including their test in classroom situations (D-TA7-WP3-7).

Programs to develop **innovative pedagogical approaches** will be supported or initiated to promote children's active engagement in scientific discovery. PUNCH universities already support the introduction of modern scientific topics and experiments in schools through, e.g. the *Physikwerkstatt Rheinland*, *Schülerlabor* and the *Netzwerk Teilchenwelt* (D-TA7-WP3-5).

Extracurricular *Arbeitsgemeinschaften* (AG, Sek I) and *Projektkurse* (PK, Sek II) will be conceived that combine astronomy, physics and data science, thus taking advantage of both fields' popularity among children. They open a direct way to bring data science into schools and optimise pedagogical approaches without curricular restrictions (D-TA7-WP3-6).

***Deliverables**:*

- **D-TA7-WP3-1 (30 Sep 2026):** Website, blogs, Twitter, communication of PUNCH4NFDI products.
- **D-TA7-WP3-2 (30 Sep 2026):** Networking, outreach training, study on effectiveness of outreach, development of evaluation criteria.
- **D-TA7-WP3-3 (30 Sep 2026):** Access to data and software for education and entertainment.
- **D-TA7-WP3-4 (30 Sep 2026):** Tutorials and resources for teachers and students, material for masterclasses.
- **D-TA7-WP3-5 (30 Sep 2026):** Study and test pedagogical approaches to promote data science to children.
- **D-TA7-WP3-6:(30 Sep 2024)** Pilot extracurricular activities in schools, promote changes in school curricula.
- **D-TA7-WP3-7 (30 Sep 2026):** Support hackathons, rent a scientist, masterclasses, VO days.

5.7.4 Work package 7.4: Support for citizen science

Status: There is broad public interest in Germany in **new discoveries** in astronomy and particle physics, e.g. the discovery of the Higgs particle, the first detection of gravitational waves, or the first image of a black hole. Germany also has a dedicated **amateur astronomer** community that delivers a broad range of educational events and engagement opportunities, much orientated towards sky observations and telescope technology. It nevertheless remains a challenge to actively involve non-traditional audience groups, including seniors, children and females.

The younger generation's increasing **digital literacy**, and the ever more diverse communication technology and social interactions, provide opportunities to engage citizens in novel ways, to serve their interest in astronomy wherever they direct their minds. Today, interested citizens can be easily invited to work on state-of-the-art research data, allowing them to share the research and discovery experience, and receive recognition for valuable contributions to science. In the astro- and astroparticle physics community, educational initiatives such as *Zooniverse* (incl. the *Radio Galaxy Zoo*), *Muon Hunter*, or CREDO increasingly **engage the public** in a more active role. The Albert-Einstein-Institute and the MPIfR continue to use *Einstein@home* to successfully engage the public in the search for new radio pulsars in radio astronomical data and from gamma-ray space observatories. MPIfR Effelsberg 100-m telescope data were also used in the BBC *Stargazing Live* program to engage school children in pulsar astronomy.

Goals: Active participation in research cultivates the understanding of and identification with the scientific method and reasoning. Being an active part of an international scientific mission also helps to **bridge differences** in geography, culture, religion, ethnicity, and gender. We take advantage of the younger generation's increasing digital literacy and diversifying communication to actively engage the public in citizen science projects such as *Einstein@Home*, *Zooniverse*, or *Muon Hunter*. We provide incentives and access to data infrastructure and methods to **involve the public in ongoing research**.

Work programme: While the number of astronomical citizen science projects in Germany has been limited, their success has shown its enormous potential. Whether it was *SETI@home*, the *Radio Galaxy Zoo*, the popular *Einstein@Home* scheme, or **Folding@Home**, the public is very interested to actively participate in research. Building on this experience, PUNCH4NFDI will give the public not only access to data archives, but make infrastructure (e.g. based on BOINC) available to involve the public in ongoing research. Most successful have been projects where the citizen scientists can make actual discoveries, like in pulsar searches. In the first phase, more data sets will be made available, while the PUNCH4NFDI partners develop new ideas. In a second phase, schools and students will be engaged, public and school events will be organised to raise awareness and participation, while delivering science and public education at the same time. PUNCH4NFDI and NFDI4Culture agree to share experience with citizen science projects and develop mutual initiatives to engage a wider audience and utilise complementary infrastructure.

Deliverables:

- **D-TA7-WP4-1 (30 Sep 2022):** Map out potential research applications in citizen science.
- **D-TA7-WP4-2 (30 Sep 2023):** Prepare data sets and providing soft- and hardware infrastructure.
- **D-TA7-WP4-3 (30 Sep 2024):** Pilot 6–12 month projects, engage schools and universities, evaluate results.
- **D-TA7-WP4-4 (30 Sep 2026):** Launch further projects jointly with the physics community.

5.7.5 Sustainability, risks, and mitigation

The training and outreach programme developed will be carried by the PUNCH4NFDI consortium and their partners. Given that these are mostly universities and research laboratories, the programme is expected to be sustainable.

The programme carries an inherently low risk since the initiatives are relatively straight forward to implement and much rely on existing infrastructur, e.g. the *Terascale Alliance* (D-TA7-WP2) or the *Netzwerk Teilchenwelt* (WP 3). The necessary complementary infrastructure can be built with NFDI and in-kind resources.

For the initiatives to develop their full potentials, an active participation of the PUNCH community is needed, which naturally carries uncertainties. Given the broad range of experience and ongoing activity in training and outreach of the consortium members, in particular in deliverables D-TA7-WP1–D-TA7-WP3, the discontinuation of individual partners can be averted by substitutions.

A central risk for the success of the consortium lies in the staffing of the positions defined in the work packages, and in the expenditure of the centrally managed global funds. However, there are already candidates that could be hired, and the PUNCH4NFDI partners envision many concrete networking events and projects to be pursued with additional personnel. Therefore, the related risks are likely small.

Another, and rather current, risk is the effect of the COVID-19 pandemic on classroom teaching and training events. Mitigation is expected from offering online courses (D-TA7-WP1-1) and preparing material for independent learning (D-TA7-WP2-3) until the situation normalises.

Appendix 1. Bibliography and list of references

In the following, we list data sources and data repositories, information infrastructures and software that PUNCH4NFDI members contributed to at a significant level and which are required for the description of the status quo of our consortium's efforts in research data management.

Data sources and repositories

Table 3: Collection of data sources with relevance for PUNCH4NFDI

Data Source	Owner	Status	Relevance for TA
CERN open data [11] [27]	CERN experiments	public	2,4
Dark Energy Survey (DES)[8]	DES Collaboration		2
Effelsberg RT data	MPIfR Bonn	partially public	4, 5, 6
Experiments at ELSA (Bonn)	Affected experiment	to be published partially	2, 4
GLOW-LOFAR pulsar data [9]	Bielefeld-Bonn pulsar group		2
H.E.S.S. data	H.E.S.S. Collaboration	to be published	2, 6
LOFAR Two-metre Sky Survey (LoTSS) [10]	LOFAR Surveys Key Science Project		2
MAMI/MESA Data	Mainz Prisma$^+$ Excellence Cluster	to be partially published	2
AIP Data Center	AIP	public	4, 6
GAVO Data Center	UHD	public	4, 6
KASCADE Cosmic Ray Data Center (KCDC)	KIT	public	2, 4, 6
Solar Data Collections	KIS	to be published	4
TLS Data collection	TLS	to be published	2, 4
Research Data from S-DALINAC and SFB 1245	TUDa	to be published	4, 6
Research data from 10+ years of experiments from small and medium scale experiments	UzK	to be published	4, 6
ALICE data	ALICE collaboration	mostly internal	
ATLAS data	ATLAS collaboration	mostly internal	
CMS data	CMS collaboration	mostly internal	
LHCb data	LHCb collaboration	mostly internal	
Belle II data	Belle II collaboration	mostly internal	
IceCUBE data	IceCUBE collaboration	internal & public	
Lattice QCD data	corresponding collaboration	mostly public	2, 3, 4

Software and middleware

- Lattice QCD configurations and software
  - openQCD software: `http://luscher.web.cern.ch/luscher/openQCD`.
  - Twisted mass lattice QCD software `https://github.com/etmc/tmLQCD`.
  - Lattice Data Grid tools developed by and for lattice QCD `https://hpc.desy.de/ldg/`.

[11] The ALICE, ATLAS, CMS and LHCb collaborations released reconstructed data from old runs. Raw data and recent data are ”internal” and only accessible within the collaboration.

  - CLS configurations: arXiv:1411.3982.
  - Access to CLS configurations https://www-zeuthen.desy.de/alpha/public-cls-nf21/.
  - Grid lattice QCD software: https://github.com/paboyle/Grid.
- Experiment and Observatory software
  - ALICE Online-Offline Computing Framework.
  - Belle II core software [72].
  - FairRoot.
  - Signal processing software for ATLAS, Belle II, LHCb.
  - Track fitting software for ALICE, ATLAS, Belle II and LHCb.
  - CB-ELSA/TAPS, BGOOD analysis framework ExPlORA.
  - Signal processing software the gamma-ray experiments HESS and VERITAS
  - Control software for LOFAR https://git.astron.nl/ro/lofar/
  - Control software for GLOW-LOFAR stand-alone-mode operations
  - **MUSE pipline** https://data.aip.de/projects/musepipeline.html
- Software for statistical applications
  - **BAT.jl** https://github.com/bat/BAT.jl.
  - **TMVA** https://sourceforge.net/projects/tmva/.
- Middleware
  - dCache https://www.dcache.org https://github.com/dCache/dcache.
  - XRootD and various plug-ins developed by GSI.
  - Rucio https://rucio.cern.ch https://github.com/rucio.
  - COBalD https://github.com/MatterMiners/cobald.
  - Tardis https://github.com/MatterMiners/tardis.
- Other software
  - Optimisation library Geneva.
  - xfitter [73] package for PDF fits https://www.xfitter.org https://gitlab.cern.ch/fitters/xfitter.
  - Alpaka [74] https://github.com/ComputationalRadiationPhysics/alpaka.
  - CORSIKA simulation tool http://www.ikp.kit.edu/corsika/index.php.
  - PCB design software.
  - Design software for FPGA.
  - ACTS http://acts.web.cern.ch/ACTS/.
  - astropy https://www.astropy.org/
  - gammapy https://gammapy.org/
  - AMPEL https://github.com/AmpelProject

  - Debian Astro Pure Blend https://blends.debian.org/astro
  - NIFTy http://ift.pages.mpcdf.de/nifty/
- Data publication software
  - **Django-Daiquiri** (data publication framework) https://github.com/django-daiquiri/daiquiri
  - **RDMO** https://rdmorganiser.github.io/
  - **DaCHS** https://docs.g-vo.org/DaCHS/
  - KCDC open data software http://kcdc.ikp.kit.edu.

Infrastructures

- Computing and storage contributions by PUNCH4NFDI members via formal agreements (contracts, MoU or similar)
  - AIP and UHD operate a mirror for the Gaia satellite mission
  - DESY supports ATLAS, CMS, LHCb as Tier-2 and is regional centre for Belle II and a major site for ILC activities.
  - GSI is a Tier-2 for the ALICE experiment.
  - KIT supports all LHC experiments as Tier-1[12], In addition, Belle II, Pierre Auger Observatory and IceCube is supported.
  - The LMU Munich provides a Tier-2 for ATLAS together with the MPI for Physics, the Leibniz Computer Centre (LRZ) and the Garching computer centre.
  - The Bergische Universität Wuppertal provides a Tier-2 for ATLAS.
  - FZJ operates one of three LOFAR LTAs.
  - RWTH Aachen supports the CMS experiment as Tier-2[13].
  - Universität Freiburg provides a Tier-2 for ATLAS.
  - Universität Göttingen provides a Tier-2 for ATLAS.
- Other major computing and storage resources
  - AIP operates an astronomical data base server
  - AIP and UHD host a VO data base server
  - DESY hosts a National Analysis Facility (NAF) for all German HEP users.
  - FZJ provides access to HPC architectures, prototype systems, and storage.
  - GSI hosts a National Analysis Facility (NAF) for ALICE.
  - The FAIR facility Tier-0[12] is located in the Green IT Cube on the GSI campus.
  - Universität Bonn hosts a Tier-3 for ATLAS.
  - Universität Göttingen provides access to HPC resources and collaboration with HLRN.
  - Universität Bielefeld operates the GLOW computing and storage cluster, hosted by FZJ, for LOFAR, MeerKAT, and the SKA-MPG telescope.

[12] Large international centre with archiving responsibility and 24x7 coverage for critical services.

[13] Typically medium-size centre with 8x5 coverage for their services.

Glossary

4MOST 4-metre Multi-Object Spectroscopic Telescope 54

AAI authentication and authorisation infrastructure 14, 19, 25, 55, 97

AARC authentication and authorisation for research collaborations 97

AG Astronomische Gesellschaft 8, 15

AIP Leibniz-Institut für Astrophysik Potsdam 3, 18, 74, 80, 94, 114

Aladin Aladin Sky Atlas 23

ALICE ALICE 7, 16, 20, 91, 114

ALMA Atacama Large Millimeter Array 23

ALU Albert-Ludwigs-Universität Freiburg 4, 15, 54

API application programming interfaces 55, 76

ASDF advanced scientific data format 39

ATLAS ATLAS 7, 18–23, 69, 114

AutoML automated machine learning 45, 68, 69

BAT Bayesian Analysis Toolkit 22, 64

Belle II Belle II 7, 18, 20, 21, 23, 28, 88, 114

BOINC Berkeley Open Infrastructure for Network Computing 112

CBM Compressed Baryonic Matter experiment 7

CDS Centre de Données astronomiques de Strasbourg 54, 56

CERN Europäisches Kernfornschungszentrum 4, 18, 54

CMS CMS 7, 18, 20, 69, 114

CREDO Cosmic-Ray Extremely Distributed Observatory 112

CSBS computing and software for big science 99

CTA Cherenkov Telescope Array 6, 8, 18, 21, 23, 34

CVMFS CERN virtual machine file system 57

dCache www.dcache.org 18, 60

DES Dark Energy Survey 21, 55, 114

DESY Deutsches Elektronen-Synchtrotron 15, 18, 49, 53, 74, 84, 94

DFG Deutsche Forschungsgemeinschaft 51

DFN Deutsches Forschungsnetz 98

DLR-DW Deutsches Luft- und Raumfahrtzentrum, Institut für Datenwissenschaften 4, 5, 15, 54

DOI Digital Object Identifier 6, 36

DOMA Data Organisation Management and Access 55, 61

DPG Deutsche Physikalische Gesellschaft 4, 5, 15, 19, 24, 26, 94, 108

EB Executive Board 27, 28, 50

EGI advanced computing services for research 98

ELSA electron accelerator ELSA 23, 55, 114

EOSC European Open Science Cloud 13, 18, 25, 27, 28, 30, 72, 76, 95, 97

ESA European Space Agency 8, 18

ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures 13, 18, 21, 27, 30, 55, 72, 95, 97

ESFRI European Strategy Forum on Research Infrastructures 13

ESO European Southern Observatory 8, 18

Euclid euclid-ec.org 6, 21

EUDAT collaborative data infrastructure 76, 98

FAIR facility Facility for Antiproton and Ion Research in Europe 16, 20, 28, 53, 57, 93, 116

FIAS Frankfurt Institute for Advanced Studies 3, 19, 84

FITS flexible image transport system 33, 39

FYST Fred Young Submillimeter Telescope/CCAT 23

FZJ Forschungszentrum Jülich 3, 15, 17, 19, 30, 53, 63, 84, 94

Gammapy gammapy.org 34, 65

GAU Georg-August-Universität Göttingen 3, 15, 19, 53, 63, 94

GAVO German Astrophysical Virtual Observatory 18, 23, 72

GEANT Europäisches Forschungsnetz 98

GLOW German Long Wavelength Consortium 20, 22

GridKa Grid Computing Centre Karlsruhe 8, 20

GSI Helmholtzzentrum für Schwerionenforschung 3, 20, 49, 53, 74, 94

GU Johann Wolfgang Goethe-Universität Frankfurt 4, 5, 15, 63

H.E.S.S. High Energy Stereoscopic System 17, 21, 23, 114

HDA Hochschule Darmstadt 4, 5

HDF Helmholtz Data Federation 97

HIFIS Helmholtz federated IT services 98

HL-LHC High-Luminosity LHC 53, 57, 68, 88

HLRN high-performance computing in northern Germany 45

HMC Helmholtz Metadata Collaboration 42, 97

HPC high-performance computing 4, 19, 20, 23, 25, 36, 53, 66, 67, 87, 90

HTC high-throughput computing 25, 36

HTW Hochschule für Technik und Wirtschaft Berlin 3, 20, 84, 90, 94

HUB Humboldt-Universität zu Berlin 4, 5

HZDR Helmholtz-Zentrum Dresden Rossendorf 4, 5, 15, 22, 63

ILDG International Lattice Data Grid International Lattice Data Grid 27, 72

ILT International LOFAR Telescope 8, 21

IRB Infrastructure & Resource Board 28, 50, 51, 55

IVOA International Virtual Observatory Alliance 27, 93

JGU Johannes-Gutenberg-Universität Mainz 3, 15, 20, 53, 63, 84

KAT Committee for Astroparticle Physics 8, 15

KCDC KASCADE Cosmic-Ray Data Centre 20, 33, 72, 76, 80, 114

KET Committee for Elementary Particle Physics 8, 15

KFP Konferenz der Fachbereiche Physik 108

KHuK Committee for Hadron and Nuclear Physics 8, 15

KIS Leibniz-Institut für Sonnenphysik 4, 5, 54, 63, 75, 114

KIT Karlsruher Institut für Technologie 3, 15, 20, 49, 53, 74, 94, 114

LHC Large Hadron Collider 9, 16, 19–22, 88, 91, 93

LHCb Large Hadron Collider beauty 8, 21–23, 114

LMU Ludwig-Maximilian-Universität München 3, 15, 21, 63, 94

LOFAR Low Frequency Array 6, 17, 19, 22, 53, 54, 57, 69

LoTSS LOFAR Two-metre Sky Survey 22, 55, 114

LQCD lattice QCD 22, 56, 67, 71, 76

LRZ Leibniz-Rechenzentrum 4, 6, 15, 30, 54, 63, 75, 94

MAMI Mainz Microtron 20, 55, 114

MB Management Board 27–29, 50, 51

MCMC Markov-Chain Monte Carlo 65, 71

MeerKAT MeerKAT 6, 19, 21, 23, 53, 87, 88

MESA Mainz Energy-Recovering Superconducting Accelerator 20, 55, 114

ML machine learning 5, 6, 9, 11, 63, 68

MPA Max-Planck-Institut für Astrophysik 8

MPCDF Max Planck Computing and Data Facility 4, 6, 30

MPE Max-Planck-Institut für extraterrestrische Physik 8, 80

MPIfR Max-Planck-Institut für Radioastronomie 3, 21, 84, 94, 114

MPIK Max-Planck-Institut für Kernphysik 3, 21, 63, 74, 94

MPP Max-Planck-Institut für Physik 8

NFDI Nationale Forschungsdaten-Infrastruktur

10, 16, 36, 49

PANDA panda.gsi.de 6, 8

PM project manager 27, 28, 50, 51

PTB Physikalisch-Technische Bundesanstalt 4, 6, 15, 75, 94

PUNCH-SDP PUNCH science data platform 12

PYA PUNCH Young Academy 104

QoS quality of service 59

RDM research data management 6

RdS Rat deutscher Sternwarten 8, 15

ROOT root.cern 39, 64, 99

RUB Ruhr-Universität Bochum 4, 6, 54

Rubin Observatory Rubin Observatory 6, 8, 53, 54, 57, 69

Rucio rucio.cern.ch 4, 60, 76

RWTH RWTH Aachen University 4, 6, 54, 75

SAB Scientific Advisory Board 28, 29, 50, 51

SDSS Sloan digital sky survey 9, 53, 80, 89

SKA Square Kilometre Array 8, 17, 19, 21, 53, 54, 57, 69, 87, 90

TA task area 1, 12

TIB Technische Informationsbibliothek 4, 6, 15, 94

TLS Thüringer Landessternwarte 3, 21, 53, 114

TOPCAT Tool for Operations on Catalogues and Tables 23

TUDa Technische Universität Darmstadt 4, 7, 15, 75, 114

TUDD Technische Universität Dresden 3, 15, 22, 84

TUDO Technische Universität Dortmund 3, 15, 22, 53, 63, 84

TUM Technische Universität München 4, 7, 15, 63

UB Universität Bielefeld 3, 15, 22, 53, 63, 74, 84, 94

UC User Committee 28, 50, 51, 55

UHD Ruprecht-Karls-Universität Heidelberg 3, 15, 22, 74, 84, 94, 114

UHH Universität Hamburg 3, 15, 23, 63

Unity www.unity-idm.eu 97

UoB Universität Bonn 3, 15, 23, 53, 74

UP Universität Potsdam 4, 7, 75

UR Universität Regensburg 3, 15, 23, 53, 63, 74

USi Universität Siegen 4, 7, 64

UzK Universität zu Köln 4, 7, 15, 75, 114

VdR Verein für datenintensive Radioastronomie e.V. 4, 7, 20, 94

VO Virtual Observatory 9, 10, 23, 33, 56, 72, 76

WebDAV web distributed authoring and versioning 55

WLCG Worldwide LHC Computing Grid 4, 9, 10, 16, 18, 19, 23, 27, 28, 52, 93

WP work package 4, 12

WWU Westfälische Wilhelms-Universität Münster 4, 7, 15, 54, 94

XRootD xrootd.slac.stanford.edu 20, 55, 60, 91

ZAH Zentrum für Astronomie 22, 23

zenodo zenodo.org 76

Bibliography

[1] Mark D. Wilkinson et al.
"The FAIR Guiding Principles for scientific data management and stewardship". en.
In: *Scientific Data* 3.1 (2016). Number: 1 Publisher: Nature Publishing Group, p. 160018.
issn: 2052-4463. doi: 10.1038/sdata.2016.18.
url: https://www.nature.com/articles/sdata201618 (visited on 06/05/2020).

[2] Bird, Ian et al. "Architecture and prototype of a WLCG data lake for HL-LHC".
In: *EPJ Web Conf.* 214 (2019), p. 04024. doi: 10.1051/epjconf/201921404024.
url: https://doi.org/10.1051/epjconf/201921404024.

[3] M Asch et al.
"Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry".
In: *The International Journal of High Performance Computing Applications* 32.4 (2018), pp. 435–479. doi: 10.1177/1094342018778123.

[4] N. Jones. "How to stop data centres from gobbling up the world's electricity". en.
In: *Nature 561, 163* (2018). doi: 10.1038/d41586-018-06610-y.

[5] A. Treloar, D. Groenewegen, and C. Harboe-Ree.
"The Data Curation Continuum - Managing Data Objects in Institutional Repositories". en.
In: *D-Lib Magazine, 13(9/10)* (2007). doi: 10.1045/september2007-treloar.

[6] DataCite Metadata Working Group. *DataCite Metadata Schema 4.3*.
https://doi.org/10.14454/7xq3-zf69. 2020.

[7] Provenance Working Group. *IVOA Provenance Data Model Version 1.0*.
http://ivoa.net/documents/ProvenanceDM/20200411/index.html. 2020.

[8] T. M. C. Abbott et al. "The Dark Energy Survey: Data Release 1".
In: *The Astrophysical Journal Supplement Series* 239.2 (Nov. 2018), p. 18.
issn: 1538-4365. doi: 10.3847/1538-4365/aae9f0.
url: http://dx.doi.org/10.3847/1538-4365/aae9f0.

[9] J. Y. Donner et al.
"First detection of frequency-dependent, time-variable dispersion measures".
In: *AaP* 624, A22 (Apr. 2019), A22. doi: 10.1051/0004-6361/201834059.
arXiv: 1902.03814 [astro-ph.GA].

[10] T. W. Shimwell et al. "The LOFAR Two-metre Sky Survey. II. First data release".
In: *AaP* 622, A1 (Feb. 2019), A1. doi: 10.1051/0004-6361/201833559.
arXiv: 1811.07926 [astro-ph.GA].

[11] S. Alef et al. "The BGOOD experimental setup at ELSA".
In: *Eur. Phys. J. A* 56.4 (2020), p. 104. doi: 10.1140/epja/s10050-020-00107-x.
arXiv: 1910.11939 [physics.ins-det].

[12] Martin Barisits et al. “Rucio: Scientific Data Management”. In: *Computing and Software for Big Science* 3.1 (Aug. 2019), p. 11. issn: 2510-2044. doi: 10.1007/s41781-019-0026-3. url: https://doi.org/10.1007/s41781-019-0026-3.

[13] see https://lcgdm.web.cern.ch/dynafeds-text-documentation-white-paper.

[14] R. Brun and F. Rademakers. “ROOT: An object oriented data analysis framework”. In: *Nucl. Instrum. Meth. A* 389 (1997). Ed. by M. Werlen and D. Perret-Gallix, pp. 81–86. doi: 10.1016/S0168-9002(97)00048-X.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. url: https://www.R-project.org/.

[16] Bob Carpenter et al. “Stan: A Probabilistic Programming Language”. In: *Journal of Statistical Software, Articles* 76.1 (2017), pp. 1–32. issn: 1548-7660. doi: 10.18637/jss.v076.i01. url: https://www.jstatsoft.org/v076/i01.

[17] Allen Caldwell, Daniel Kollar, and Kevin Kroninger. “BAT: The Bayesian Analysis Toolkit”. In: *Comput. Phys. Commun.* 180 (2009), pp. 2197–2209. doi: 10.1016/j.cpc.2009.06.026. arXiv: 0808.2552 [physics.data-an].

[18] Oliver Schulz et al. “BAT.jl – A Julia-based tool for Bayesian inference”. In: (Aug. 2020). arXiv: 2008.03132 [stat.CO].

[19] Jeff Bezanson et al. “Julia: A Fresh Approach to Numerical Computing”. In: *CoRR* abs/1411.1607 (2014). arXiv: 1411.1607. url: http://arxiv.org/abs/1411.1607.

[20] Allen Caldwell et al. “Integration with an Adaptive Harmonic Mean Algorithm”. In: *Int. J. Mod. Phys. A* 35.24 (2020), p. 1950142. doi: 10.1142/S0217751X20501420. arXiv: 1808.08051 [physics.data-an].

[21] Vasyl Hafych et al. “Parallelizing MCMC Sampling via Space Partitioning”. In: (). eprint: 2008.03098.

[22] Julien Lesgourgues. “The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview”. In: *arXiv e-prints*, arXiv:1104.2932 (Apr. 2011), arXiv:1104.2932. arXiv: 1104.2932 [astro-ph.IM].

[23] Diego Peteiro-Barral et al. “Toward the scalability of neural networks through feature selection”. In: *Expert Systems with Applications* 40 (June 2013), pp. 2807–2816. doi: 10.1016/j.eswa.2012.11.016.

[24] N. Aghanim et al. “Planck 2018 results: VI. Cosmology parameters”. In: *Astronomy and Astrophysics* 641 (Sept. 2020), A6. issn: 1432-0746. doi: 10.1051/0004-6361/201833910. url: http://dx.doi.org/10.1051/0004-6361/201833910.

[25] Yasmine Sara Amhis et al.
"Averages of b-hadron, c-hadron, and τ-lepton properties as of 2018". In: (Sept. 2019). arXiv: 1909.12524 [hep-ex].

[26] S. Bocquet et al.
"Cluster Cosmology Constraints from the 2500 deg^2 SPT-SZ Survey: Inclusion of Weak Gravitational Lensing Data from Magellan and the Hubble Space Telescope".
In: *ApJ* 878.1, 55 (June 2019), p. 55. doi: 10.3847/1538-4357/ab1f10.
arXiv: 1812.01679 [astro-ph.CO].

[27] *CERN Open Data Portal*. http://opendata.cern.ch. 2020.

[28] see https://www.hepdata.net.

[29] R. J. Hanisch et al.
"The Virtual Astronomical Observatory: Re-engineering access to astronomical data".
In: *Astronomy and Computing* 11 (June 2015), pp. 190–209.
doi: 10.1016/j.ascom.2015.03.007. arXiv: 1504.02133 [astro-ph.IM].

[30] Peter Athron et al.
"GAMBIT: The Global and Modular Beyond-the-Standard-Model Inference Tool".
In: *Eur. Phys. J. C* 77.11 (2017). [Addendum: Eur.Phys.J.C 78, 98 (2018)], p. 784.
doi: 10.1140/epjc/s10052-017-5321-8. arXiv: 1705.07908 [hep-ph].

[31] *MasterCode*. https://mastercode.web.cern.ch. 2020.

[32] J. De Blas et al. "HEPfit: a code for the combination of indirect and direct constraints on high energy physics models". In: *Eur. Phys. J. C* 80.5 (2020), p. 456.
doi: 10.1140/epjc/s10052-020-7904-z. arXiv: 1910.14012 [hep-ph].

[33] Philip Bechtle et al. "HiggsBounds: Confronting Arbitrary Higgs Sectors with Exclusion Bounds from LEP and the Tevatron".
In: *Comput. Phys. Commun.* 181 (2010), pp. 138–167.
doi: 10.1016/j.cpc.2009.09.003. arXiv: 0811.4169 [hep-ph].

[34] see http://reanahub.io.

[35] see http://analysispreservation.cern.ch.

[36] *KCDC KASCADE Cosmic Ray Data Centre*. https://kcdc.ikp.kit.edu/. 2020.

[37] *ESCAPE - European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures*. http://projectescape.eu. 2020.

[38] *ARCHIVER - Archiving and preservation for research environments*.
http://archiver-project.eu. 2020.

[39] Mark G. Beckett et al. "Building the International Lattice Data Grid".
In: *Comput. Phys. Commun.* 182 (2011), pp. 1208–1214.
doi: 10.1016/j.cpc.2011.01.027. arXiv: 0910.1692 [hep-lat].

[40] see https://zenodo.org.

[41] see https://invenio-software.org/.

[42] see https://ui.adsabs.harvard.edu/.

[43] see https://github.com/sagemathinc/cocalc.

[44] R. P. Eatough et al. “Selection of radio pulsar candidates using artificial neural networks”. In: *MNRAS* 407.4 (Oct. 2010), pp. 2443–2450. doi: 10.1111/j.1365-2966.2010.17082.x. arXiv: 1005.5068 [astro-ph.IM].

[45] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Phys. Lett. B* 716 (2012), pp. 1–29. doi: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[46] S. Chatrchyan et al. “Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC”. In: *Phys. Lett. B* 716 (2012), pp. 30–61. doi: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[47] B.P. Abbott et al. “Observation of Gravitational Waves from a Binary Black Hole Merger”. In: *Phys. Rev. Lett.* 116.6 (2016), p. 061102. doi: 10.1103/PhysRevLett.116.061102. arXiv: 1602.03837 [gr-qc].

[48] D.R. Lorimer et al. “A bright millisecond radio burst of extragalactic origin”. In: *Science* 318 (2007), p. 777. doi: 10.1126/science.1147532. arXiv: 0709.4301 [astro-ph].

[49] Johannes Albrecht and et al. “A Roadmap for HEP Software and Computing R&D for the 2020s”. In: *Computing and Software for Big Science* 3.1 (Mar. 2019). issn: 2510-2044. doi: 10.1007/s41781-018-0018-8. url: http://dx.doi.org/10.1007/s41781-018-0018-8.

[50] E. Petroff et al. “A real-time fast radio burst: polarization detection and multiwavelength follow-up”. In: *MNRAS* 447.1 (Feb. 2015), pp. 246–255. doi: 10.1093/mnras/stu2419. arXiv: 1412.0342 [astro-ph.HE].

[51] Emily Petroff et al. “VOEvent Standard for Fast Radio Bursts”. In: *arXiv e-prints*, arXiv:1710.08155 (Oct. 2017), arXiv:1710.08155. arXiv: 1710.08155 [astro-ph.IM].

[52] R. Aaij et al. “Allen: A High-Level Trigger on GPUs for LHCb”. In: *Computing and Software for Big Science* 4.1 (Apr. 2020). issn: 2510-2044. doi: 10.1007/s41781-020-00039-7. url: http://dx.doi.org/10.1007/s41781-020-00039-7.

[53] David Rohr, Sergey Gorbunov, and Volker Lindenstruth. “GPU-accelerated track reconstruction in the ALICE High Level Trigger”. In: *Journal of Physics: Conference Series* 898 (Oct. 2017), p. 032030. issn: 1742-6596. doi: 10.1088/1742-6596/898/3/032030. url: http://dx.doi.org/10.1088/1742-6596/898/3/032030.

[54] Shreyasi Acharya et al.
"Real-time data processing in the ALICE High Level Trigger at the LHC".
In: *Comput. Phys. Commun.* 242 (2019), pp. 25–48. doi: 10.1016/j.cpc.2019.04.011.
arXiv: 1812.08036 [physics.ins-det].

[55] M.F.M. Lutz et. al.
"On the convergence of the chiral expansion for the baryon ground-state masses".
In: *Nucl. Phys A* 977 (2018), pp. 146–207.

[56] M.F.M. Lutz et. al. "A generalized Higgs potential with two degenerate minima for a dark QCD matter scenario". In: (June 2019). arXiv: 1907.00237 [hep-th].

[57] R. Berlich et. al. "Distributed Parametric Optimization with the Geneva Library".
In: *Data Driven e-Science, Conference proceedings of ISGC* (2010), p. 303.

[58] see https://cernlib.web.cern.ch/cernlib/.

[59] see https://root.cern.ch.

[60] see https://github.com.

[61] see https://gitlab.com.

[62] see https://scipy.org.

[63] see https://matplotlib.org.

[64] see https://pandas.pydata.org.

[65] see https://astropy.org.

[66] see https://gammapy.org.

[67] see https://scikit-hep.org.

[68] see https://jupyter.org.

[69] see https://mybinder.org.

[70] see https://hepsoftwarefoundation.org.

[71] see https://openastronomy.org.

[72] Belle II Framework Software Group. "The Belle II Core Software".
In: *Comput. Softw. Big Sci.* 3.1 (2019), p. 1. doi: 10.1007/s41781-018-0017-9.
arXiv: 1809.04299 [physics.comp-ph].

[73] S. Alekhin et al. "HERAFitter, Open Source QCD Fit Project". In: (2014).
arXiv: 1410.4412 [hep-ph].

[74] Erik Zenker et al. "Alpaka – An Abstraction Library for Parallel Kernel Acceleration".
In: *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (May 2016). doi: 10.1109/ipdpsw.2016.50.
url: http://dx.doi.org/10.1109/IPDPSW.2016.50.