**PUNCH4NFDI Consortium**

**Consolidated Work Programme**

November 2021

## **Introduction**

This document summarises the work programme of the PUNCH4NFDI consortium at the beginning of its funded work in October 2021. The programme laid down here differs from the one sketched in the proposal, chiefly because of the funding cut of 30% applied by the DFG and the ensuing necessary reductions and time shifts of contributions from the co-applicants.

The NFDI is considered to be a dynamic, evolving infrastructure, and so we certainly envisage changes to the PUNCH4NFDI work plan, be they on the small, WP level or affecting the entire consortium. Furthermore, shifts are to be expected because of unexpected problems in the execution of the programme or because of difficulties in hiring, etc.

This document is therefore only a snapshot taken in late 2021, and updates to it will be made with every reporting step[[1]](#footnote-1) (e.g. probably on a quarterly basis). For this purpose, this document shall be mirrored in the PUNCH4NFDI intranet.

## **Task Area 1**

TA 1 is the management task area, with the work packages “Management setup”, “Financial administration”, “Controlling and reporting” and “Consortium organization”.



Figure 1: Analysis of deliverable trend for TA 1. Note that only non-continuous deliverables are displayed.

### TA 1 Work Package 1

WP1 of TA 1 deals with getting the consortium management structures into place. Most of the deliverables are, therefore, due at the start of the consortium lifetime and have partly already been achieved.

**Deliverables:**

* D-TA1-WP1-1 (30 Sep 2021): Memoranda of understanding and contracts

🡺 achieved 15 October 2021.

* D-TA1-WP1-2 (31 Oct 2021): Kick-off workshop; establishing of Executive Board (EB)
🡺 achieved 15 October 2021.
* D-TA1-WP1-3 (30 Nov 2021): Setup of Management Board (MB); employment of project manager (PM); setup of consortium office at DESY
🡺 achieved 15 October 2021 except for employment of PM.
* D-TA1-WP1-4 (30 Sep 2021): Establishing of User Committee (UC), Scientific Advisory Board (SAB), and Infrastructure & Resource Board (IRB)
🡺 achieved 15 October 2021.

### TA 1 Work Package 2

WP 2 of TA 1 is concerned with the financial administration (request of funding, distribution of funding, transfer of funding, reporting of funding towards the DFG). The WP is assisted by the DESY administration.

**Deliverables:**

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

### TA 1 Work Package 3

WP 3 of TA 1 is concerned with the scientific controlling and reporting, insofar as it has been defined so far. There are various reporting streams to be satisfied, where the financial reporting towards the DFG is organized in WP 2.

**Deliverables**

* D-TA1-WP3-1 (quarterly): Quarterly deliverable monitoring.
* D-TA1-WP3-2 (quarterly): Quarterly monitoring of own and in-kind contributions.
* D-TA1-WP3-3 (quarterly ): Quarterly DFG / NFDI reporting (based, if necessary, on D-TA1-WP3-4).
* D-TA1-WP3-4 (quarterly): Quarterly PUNCH4NFDI reports.
* D-TA1-WP3-5 (30 June 2024): Mid-term review report ready (due by the DFG on 30 September 2024).

### TA 1 Work Package 4

WP 4 of TA 1 is concerned with the internal organisation of the consortium.

**Deliverables:**

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

## **Task Area 2**

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.



Figure 2: Analysis of deliverable trends for TA 2.

### TA 2 Work Package 1

WP 1 of TA 2 is mostly concerned with the data and storage infrastructure of PUNCH. Main activities include the integration of existing storage infrastructure as well as in-kind resources into PUNCH structures. To exercise novel technologies, e.g. token based AAI, prototypes will be set up.

Various groups will contribute effort to make data accessible within PUNCH. This includes the preparation of suited storage technology and exposing relevant metadata such that they can be consumed by other PUNCH services. Where possible interfaces to existing systems are envisaged. When necessary new products are to be developed or adopted, e.g. in the case of Lattice QCD, where a new catalog service for meta data is required. The to be developed service shall be generic and independent of the specific domain.

In order to provide a sustainable infrastructure, that in the long term has to be fully embedded into the international structures of the experiments and efforts like EOSC, a commitment to support operations and development is required by all involved partners.

**Deliverables**

* D-TA2-WP1-1 (30 Jun 2022): Prototype Data Lake setup
* D-TA2-WP1-2 (30 Jun 2023): Initial integration of data resources: GLOW, MAMI, MESA, ELSA
* D-TA2-WP1-3 (31 Dec 2023): Metadata catalogue, initial reference implementation for LQCD
* D-TA2-WP1-4 (30 Sep 2024): Further integration of Astro and HEP resources, incl. access to external sources
* D-TA2-WP1-5 (30 Jun 2026): Documentation of PUNCH Data Lake technology
* D-TA2-WP1-6 (continuous after 31 dec 2022): Support of PUNCH Data Lake technology

### TA 2 Work Package 2

The WP aims for the integration of compute resources into a common infrastructure and provide suitable access points. The integration is planned to be realised using the COBalD/TARDIS software framework. It allows for a dynamic provisioning of compute resources from var- ious providers using cloud and batch system interfaces and enables a transparent integration of those resources into one common infrastructure. Access to compute resources in this infras- tructure can be enabled using different point of entries utilising well established technologies like JupyterHub, Grid Compute Elements or traditional login nodes to cover the broad needs of the PUNCH community. In addition, well established products like modern containers 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.

### Deliverables (DESY, FZJ, GSI, JGU Mainz, KIT, TLS, U Bonn, and U Göttingen):

* D-TA2-WP2-1 (30 Jun 2022): Demonstrator for Compute4PUNCH (C4P)
* D-TA2-WP2-4 (31 Dec 2024): Integrate a variety of compute resources into C4P
* D-TA2-WP2-5 (30 Jun 2023): Provide entry points to C4P as JupterHub and batch system
* D-TA2-WP2-6 (30 Jun 2024): Integrate opportunistic cache systems into C4P
* D-TA2-WP2-8 (30 Sep 2026): Provide a fully capable container registry

### TA 2 Work Package 3

(GSI, Mainz, DESY, Bielefeld, Münster, KIT)

This work package introduces intelligent workflow techniques. Optimised data placement in and data movements across large-scale computing resoures for huge distributed datasets will be developed. Intelligence can only be added by including monitoring information into the decision-making process of the applied tools. These policies can then enforce the placement of hot and cold data objects. It is also important to use forecasts about future access patterns to proactively migrate or replicate data or jobs. Exabyte datasets also require new algorithms and protocols to scale concerning their management. An alternative to file catalogues is to use distributed hashing which has shown to scale to (nearly) arbitrary sizes in peer-to-peer computing. 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. This work package 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.

## **Deliverables:**

* D-TA2-WP3-1 (31 Dec 2022): Data lake monitoring infrastructure prototype.
* 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-WP3-7 (30 Sep 2025): Data-locality aware scheduling available in the overlay batch system.
* optional: D-TA2-WP3-2 (30 Jun 2023): Middleware and backend storage jointly supporting QoS.

## **Task Area 3**

"Data transformations" to higher levels of abstraction are an essential component of the scientific workflow to gain knowledge from huge amounts of recorded or simulated data. The task of TA3 is to make sure that common data transformation tools and workflows required by large parts of the community will be available on the data portal.



Figure 3: Analysis of deliverable trends for TA 3Analysis of deliverable trends for TA 2..

### TA 3 Work Package 1

WP 3.1 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.

## **Deliverables:**

* D-TA3-WP1-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.

### TA 3 Work Package 2

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.

## **Deliverables:**

* D-TA3-WP2-1 (30 Sep 2026; FZJ, UB, UHH, UR): 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; FZJ, UB, UHH, UB): Provision of data-/compute-heavy algorithms with a focus on algorithmic/technical aspects of scientific reproducibility (resiliency, uncertainties).

### TA 3 Work Package 3

WP 3.3 will design and implementWP3 overview: 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 au- tomated 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.

## **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.

### TA 3 Work Package 4

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 and in continuous exchange with the user community (together with TA 6), 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.

## **Deliverables:**

* D-TA3-WP4-1 (30 Sep 2026; FZJ, UB): 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; LMU): Tools to define, test, and execute scalable workflows; library of template workflows.
* D-TA3-WP4-3 (31 Dec 2024; LMU, MPIK): Standard interface for the publication of likelihoods, including a catalogue for the definition of common parameters.

## **Task Area 4**

TA 4 provides 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.



Figure 4: Analysis of deliverable trends for TA 4.

### TA 4 Work Package 1

WP 4.1 will design and implement an open research catalogue to find and access digital Re- search Products (RPs) from across the PUNCH community in a unified way through the Data Science Platform. A central innovation will be the inclusion of Dynamic Research Products (DRPs) in the catalogue. A DRP can be any machine-executable program, such as physics models, functions of data and parameters (e.g., likelihood functions), or even complete data analysis workflows. The catalogue will provide an entry point to search, discover, access, and interlink static and dynamic Research Products using the services developed by TA2, TA3 and TA4 (WP2, WP3 & WP4). Thus, initially unconnected data sources will be made inter-linkable to allow for execution of new innovative use cases. The disposal of research catalogue templates based on the experience gained within PUNCH4NFDI will provide guidance for the implementation of procedures for publication datasets from e.g., smaller experiments and observatories to FAIR standards.

**Deliverables:**

* D-TA4-WP1-1 (31 Dec 2022): Design and implementation of the PUNCH DRP Database on the infrastructure provided by TA2.
* D-TA4-WP1-2 (31 Dec 2023): Dynamic Ingestion and Curation Processes of selected existing research products from different PUNCH subcommunities; demonstration and routine processing.
* D-TA4-WP1-3 (31 Dec 2024): Enrichment of Digital Research Products 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 dynamic DRPs and their interfaces. Prototype demonstrators and consistent integration for running them on the Data Science Platform; Continuously integrate needs of PUNCH community.

### TA 4 Work Package 2

This work package will focus on the coherent treatment of metadata from the different PUNCH science data providers for the entire PUNCH4NFDI data science platform and in particular for the data science portal and its DRPs as provided by TA4. As the various PUNCH4NFDI communities have a different metadata culture and also rely on different metadata structures, ontologies and layers, we will have to provide extensions and standards to integrate all of them into the PUNCH4NFDI science platform.

One of the most important objectives of the Consortium therefore is the development of cross-community metadata schemas in order to provide user-friendly catalogues and portals for data repositories such as the PUNCH4NFDI data science portal and for analysis, simulation, preservation software. For reaching this objective it is of utmost importance to preserve back- ward compatibility with already existing systems, because otherwise the PUNCH4NFDI science platform will not be acceptable for the different communities and the manyfold (international) collaborations.

**Deliverables**

* D-TA4-WP2-1 (31 Dec 2022): Prototype metadata describing the interaction of software and data
* D-TA4-WP2-2 (31 Dec 2023): Define metadata layers and minimal extensions for existing meta- data standards
* D-TA4-WP2-3 (30 Sep 2025): Define cross-layer connections for transformations fo digital objects

### TA 4 Work Package 3

The work package will focus on the implementation of interfaces for the functionality of the PUNCH science data portal respecting the FAIR principles. This includes interfaces for ac- cess management to the central platform, its decentralized branches and external resources linked into the platform; furthermore, interfaces to the content and format of initially heteroge- neous data both inside and outside the platform, enabling 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.

**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

### TA 4 Work Package 4

Data portals exist in many variants in the PUNCH community. They are almost always build for data collections from an instrument or experiment, sometimes they are an assembly of het- erogeneous 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 one of the drivers for forming the PUNCH4NFDI. Some institutes or communities have started to address the issues, e.g., the KCDC, or the SciServer of SDSS and MPE, or AIP using CoCalc based execution environments. In other disciplines, portal solutions built on CKAN and similar software are used for managing in-house data collections.

The goal is the instantiation of the Science 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 taking these out of the context, 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.

**Deliverables**

* D-TA4-WP4-1 (30 Jun 2023): Working prototype portal (web interface) to access and use the platform
* D-TA4-WP4-2 (30 Jun 2024): Published Research Product examples with stored data and interoperable analysis workflows
* D-TA4-WP4-3 (30 Sep 2026): Feature complete portal service
* D-TA4-WP4-4 (30 Sep 2026): Published and interoperable (D)Research Prod- ucts using the full range of services, including combined analysis of data sets from different sources and experiments

## **Task Area 5**

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. 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.

****

Figure 5: Analysis of deliverable trends for TA 5.

### TA 5 Work Package 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. 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 implications on the potential to make unexpected discoveries.

**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): Concepts towards a general protocol on capturing the decisions made by, and status of, real-time sensors, as a basis for a future demonstrator

### TA 5 Work Package 2

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), TA 5 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 dis- card irrelevant information with both minimal time budget and ability to describe the decision process in metadata that become part of the data itself. Generally, fast real-time clustering and pattern recognition algorithms are also needed to identify and separate background from physically interesting events.

### Deliverables:

* D-TA5-WP2-1 (31 May 2022): Curation & metadata schemes for dynamic filtering.
* D-TA5-WP2-2 (30 Sep 2022): Strategy concept for identifying highly complex (multi-para- metric) signals in huge data streams.
* D-TA5-WP2-3 (31 Dec 2023): Test environment for identifying highly complex (multi-para- metric) 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 (01 Mar 2026): Algorithms for massively parallel real-time sorting, clustering and pattern recognition on specialised hardware.
* D-TA5-WP2-6 (01 Mar 2026): 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.

### TA 5 Work Package 3

The aspects of “Dynamic archiving” are studied in TA 5 WP 3 using astronomical data. This work package 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 opti- mise feedback between online and offline analysis. A test case will be set up to demonstrate a complete dynamic filter and archive feedback loop. A critical task for archives is to continuously monitor the data quality and (re)examine the causal inference of any dynamic filters in use. 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. Indeed, real-time filters are today updated between observations based on updated findings in stored data.

### Deliverables:

* D-TA5-WP3-1 (31 Mar 2023): Specifying the concept of a dynamic archive: Requirements in relationship to other WPs (information loss, dynamic filters, scalable workflows and simulated catalogs) as well as to information present in traditional archives (other TAs).
* D-TA5-WP3-2 (31 Mar 2025): Present a framework in which queries to dynamic archives can be transformed into a dynamic filter (as used by some combination of sensors), and vice versa.
* D-TA5-WP3-3 (30 Sep 2026): Present methods by which queries to dynamic archives also return an estimate on the potential of information loss, i.e. how well the archive response can be assumed to approximate the response of a real-time sensor.

### TA 5 Work Package 4

(FIAS, FZJ, HTW)

This work package of TA 5 aims particularly 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. 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.

**Deliverables:**

* D-TA5-WP4-1 (31 Dec 2024): Porting common off-line packages (e.g. CASA) to a memory-based computing prototype to prepare analysis of “data monster”
* D-TA5-WP4-2 (31 Mar 2025): Standard software (e.g. CASA) compatible with Gen–Z.
* D-TA5-WP4-3 (30 Jun 2025): Caching strategies for processing a set of benchmark files with the evaluated efficiencies and latencies.
* D-TA5-WP4-4 (30 Jun 2026): Definition and initial implementation of an efficient real-time data processing framework
* D-TA5-WP4-5 (30 Sep 2026): Scaled feedback interfaces between off-line software (e.g. CASA) and selected real-time processes using MeerKAT data

### TA 5 Work Package 5

The ”Evaluation and validation of instrument response and characteristics” studied in TA 5 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.

### Deliverables:

* D-TA5-WP5-1 (30 Sep 2024): Development of machine learning prototypes for anomaly detection and predictive maintenance
* D-TA5-WP5-2 (30 Sep 2024): Interference recognition and mitigation schemes for transient discovery leading to a robust triggering system
* D-TA5-WP5-3 (30 Sep2026): Expansion of the concept to a generalized toolkit for predictive maintenance and anomaly detection
* D-TA5-WP5-4 (30 Sep 2026): Evaluation of the machine learning approaches by analyzing false-alarm rates and online feedback

## **Task Area 6**

TA 6 takes care of the “Synergies & Services”.

##

Figure 6: Analysis of deliverable trends for TA 6.

### TA 6 Work Package 1

## (UHD, GSI, AIP, DESY, FZJ, UB)

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, via the marketplace 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.

## **Deliverables:**

### D-TA6-WP1-1 (01 Jan 2022): A marketplace noticeboard and tools for communication and exchange among communities and consortia

### D-TA6-WP1-2 (01 Jan 2023): Exchange platform for information on archives, data management software, services, and hubs

### D-TA6-WP1-3 (01 Jan 2023): Options for synergies on cross-cutting topics will be discussed with all funded NFDI initiatives

### D-TA6-WP1-4 (during project runtime): Services and tools developed by PUNCH4NFDI will be made available to the NFDI

### D-TA6-WP1-5 (during project runtime): Active collaborations with other NFDI initiatives (as well as national and international projects on research data management) are pursued (list in the intranet).

### TA 6 Work Package 2

## (GSI, FZJ, KIT)

A transparent and distributed access to resources and services 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. 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. WP6.2 will provide an interface for user authorisation and authentication. 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. The Helmholtz-AAI and the EOSC/EUDAT B2ACCESS are blueprints for a PUNCH-wide AAI. The software base for the implementation is Unity. The work package will contribute to the implementation of necessary adjustments/interfaces so that most PUNCH4NFDI use cases can be fullfilled. A user and group management 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.

## **Deliverables:**

### D-TA6-WP2-1a (31 March 2022): Prototype of PUNCH-AAI allowing first use cases to authenticate users

### D-TA6-WP2-1b (31 December 2023): Requirements from all use cases will lead to the „Basic PUNCH AAI“

### D-TA6-WP2-1c (30 September 2026): Extended PUNCH-AAI“

### D-TA6-WP2-2a (31 December 2022): Draft design of 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.

### D-TA6-WP2-2b (31 December 2024): Complete design.

### D-TA6-WP2-3a (31 December 2022): PUNCH4NFDI group management – a prototype based on Unity group management enabling services to authorise users based on their group affiliations

### D-TA6-WP2-3b (31 December 2024): Revised version based on requirements of PUNCH4NFDI use cases aligned with NFDI activities

### D-TA6-WP2-3c (30 September 2026): Full PUNCH AAI group management

### TA 6 Work Package 3

## (HTW, MPIK, MPIfR, AIP, UHD)

### 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- pean projects and supported by large international infrastructures. Driven by different research methods, the communities have pioneered different elements of data management. 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.

### 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 ob- tained for calibration purposes which are unique archives of primary data in climate science in their own right.

### 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 2025): Exploring existing dynamic metadata frameworks for their integration into the PUNCH-SDP.

### 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.

### TA 6 Work Package 4

## (MPIK, LMU)

The WP is designed to identify the requirements of the PUNCH communities on data analysis tools, evaluate to which degree those 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 those open source tools for data analysis of PUNCH data will be validated using reference data sets and promoted by providing detailed analysis examples. This work package also aims to establish standard development workflows and infrastructure inspired from the open source software community within the PUNCH community.

**Deliverables:**

### D-TA6-WP4-1 (30 Jun 2026): Survey of tools used and/or developed within the PUNCH communities.

### D-TA6-WP4-2 (30 Jan 2023): Reference repository with testing, CI, deployment, guidelines (”code of conduct”) for PUNCH open software

### D-TA6-WP4-4 (30 Jun 2026): Provide detailed and non-trivial examples of data analysis based on open-source analysis tools and example data sets which are representative for the PUNCH research areas.

**Deliverables to be discussed:**

### D-TA6-WP4-3 (01 January 2021?): Contribute to the open-source Linux distributions such as Debian for software packages relevant for the PUNCH community and provide guide- lines for successful contributions.

### D-TA6-WP4-5 (30 July 2023) Provide a software platform for PUNCH and the NFDI with relevant software tools and frameworks following the example platform as developed by ESCAPE WP3

### TA 6 Work Package 5

(AIP, UG, UM, GSI, ZAH, HTW, DESY)

The prime objective of this task area is to compile a portfolio of big-data services, some existing, some currently under development, that are of generic interest for researchers in (sub)fields other than those were the respective service originated from. These services are then adopted to needs of a broader community and be made available for beta-testing, i.e., testing by other users that are development affine, but were not involved in the design and development of the respective service. The list of big-data services will continuously be updated and reanalysed, focusing on a set of deliverables as outlines below. Means for a sustained operation of these services will be sought for.

In the research domain, there exists already a considerable number of efficient state-of-the-art services to manage the data challenges from large research infrastructures. Topics include data access, visualization, code-to- computer services and so on. These services are not only actively be used by users in the respective sub-community, many of these services are of interest for other communities, within the field but also outside. The number of services is expected to considerably increase in the course of the project as services developed in TA2-5 reach maturity and transit from the alpha-testing to beta- testing, implying the usage by users outside of the project. Furthermore, as a result of other work packages of this task area, services will be adopted and made available for the use by other community, within the larger domain but also outside.

### Deliverables:

### D-TA6-WP5-1 (30 Sep 2022): Dynamic disk cache for including opportunistic storage resources

### D-TA6-WP5-2 (30 Dec 2025): The consortium Gen-Z is developing an open standard for memory-based computing (since 2016). Upcoming extensions are explored concerning their relevance for TA5 where, inter alia, the astronomical framework CASA is optimised for analysing ”data monster” (SKA will provide single images from the cosmos that may be as large as 1 PB). This may also serve as a prototype for genome research.

### D-TA6-WP5-3 (31 Dec 2025): Access interfaces to the GPU cluster in Göttingen will be developed and provided for education and development purposes to the PUNCH4NFDI communities.

### D-TA6-WP5-4 (31 Dec 2025): A fraction of the GoeGrid grid computing cluster in Göttingen will be provided to PUNCH4NFDI and beyond for analysis of the CERN open data to users without explicit CERN or experiment affiliation.

### 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-WP5-7 (30 Sep 2026): Integration of standard analysis software and newly developed tools into the JupyterHub platform via use case specific notebook images.

### D-TA6-WP5-8 (31 Dec 2023): Connect very different domains of IT resource providers like those of PUNCH with microservices as the software architecture within cloud environments to connect services (with limited scope) via clearly defined APIs.

### D-TA6-WP5-10 (30 Jun 2024): Set-up of FTS and Rucio for evaluation purposes inside the PUNCH consortium. Activities beyond need additional resources.

## **Task Area 7**

Task area 7 will train the PUNCH community on data science methods and data access. It will also contribute to educate and engage society at large on data science aspects. It will also address structural challenges related to gender equality and access to computing and information resources. Its target groups have a wide range of interests and diverse backgrounds, to which measures are tailored within four main areas. Given the educational nature of the work packages, actions can be easily linked to other NFDI consortia and different scientific communities.



Figure 7: Analysis of deliverable trends for TA 7.

### TA 7 Work Package 1

The central goal of this work package “Training of scientists – PUNCH Young Academy” is the provision of broad, high-quality training in state-of-the-art data science techniques The target groups include Ph.D. students, postdocs, and senior scientists in the spirit of life-long-learning. Training measures 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. This will be supplemented later on by courses and online material on the use of the new tools and services developed within PUNCH4NFDI.

The PUNCH Young Academy (PYA) 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, e.g., 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 current 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.

**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. Institutes: DESY, Universität Göttingen, FZJ

### D-TA7-WP1-2 (30 Sep 2024): Establishment of the PUNCH4NFDI Young Academy including regular events and courses. Institutes: DESY, ZDV Mainz, Universität Göttingen, FZJ

### D-TA7-WP1-3 (30 Sep 2025): Development of a special programme and specific courses geared towards female scientists. Institutes: ZDV Mainz, Universität Göttingen

### D-TA7-WP1-4 (30 Sep 2024): Critical review of the measures developed and development of a feedback system to guide education and training. Institutes: ZDV Mainz, Universität Göttingen, FZJ

### 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. Institutes: ZDV Mainz, Universität Göttingen

### TA 7 Work Package 2

The “education of students” and early-stage researchers is a central aim of PUNCH4NFDI. A central goal will be to provide and improve proficiency in NFDI-related themes, and thereby widen and 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 recognized 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 an adequate understanding of experimental setups. Universities are the key to transfer knowledge developed within the NFDI to new generations of scientists, who act as knowledge multipliers in academia and society at large by providing best practice examples and becoming the teachers of successive generations.

**Deliverables**

### D-TA7-WP2-1 (30 Sep 2022): Market survey of available teaching concepts and material. Institutes: TU Dortmund

### D-TA7-WP2-2 (30 Sep 2022): Development of standardized curriculum. Institutes: TU Dortmund, Universität Frankfurt (Participant), AlfA – Universität Bonn, Hochschule Darmstadt (Participant), Universität Bielefeld, Universität Göttingen

### D-TA7-WP2-3 (30 Sep 2026): Compilation and development of teaching material for courses and independent learning. Institutes: TU Dortmund, Universität Frankfurt (Participant), Universität Siegen (Participant), AlfA – Universität Bonn, Hochschule Darmstadt (Participant), Universität Göttingen

### D-TA7-WP2-4 (30 Sep 2026): Aggregation of data resources and access to computing infrastructure. Institutes: Universität Göttingen, FZJ

### D-TA7-WP2-5 (30 Sep 2026): Coordination and initiation of educational events, e.g. visiting seminars. Institutes: Universität Frankfurt (Participant), Hochschule Darmstadt (Participant), Universität Göttingen, GSI Darmstadt

### TA 7 Work Package 3

The PUNCH4NFDI “public outreach” activity aims at developing an adequate infrastructure to foster public interests in science and an understanding of how data provides the foundation of, and data science methods enable discovery in modern natural sciences. Specific goals will be to help educate the public on how to read, cope, and interpret data, and thereby to foster an adequate understanding of issues relating to big data, artificial intelligence, and machine learning.

Physics addresses fundamental and universal questions of humankind. It does so in transcending national, cultural, or gender borders. Promoting gender balance and cultural diversity is therefore a natural goal of PUNCH4NFDI. The planned outreach activity will therefore implement proactive, affirmative measures to counter historically grown or socially ingrained discrimination.

**Deliverables**

### D-TA7-WP3-1 (30 Sep 2026): Networking, outreach training, study on effectiveness of outreach, development of evaluation criteria. Institutes: AlfA – Universität Bonn, Universität Göttingen

### D-TA7-WP3-2 (30 Sep 2026): Access to data and software for education and entertainment. Institutes: AlfA – Universität Bonn, FZJ

### D-TA7-WP3-3 (30 Sep 2026): Tutorials and resources for teachers and students, material for masterclasses. Institutes: AlfA – Universität Bonn, FZJ

### D-TA7-WP3-4 (30 Sep 2026): Study and test pedagogical approaches to promote data science to children. Institutes: AlfA – Universität Bonn

### D-TA7-WP3-5:(30 Sep 2024) Pilot extracurricular activities in schools, promote changes in school curricula. Institutes: Universität Göttingen

### D-TA7-WP3-6 (30 Sep 2026): Support hackathons, rent a scientist, masterclasses, VO days. Institutes: FZJ, Universität Göttingen

### TA 7 Work Package 4

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.

### Deliverables

### D-TA7-WP4-1 (30 Sep 2022): Map out potential research applications in citizen science. Institutes: AlfA – Universität Bonn, MPIfR Bonn

### D-TA7-WP4-2 (30 Sep 2023): Prepare data sets and providing soft- and hardware infrastructure. Institutes: MPIfR Bonn

### D-TA7-WP4-3 (30 Sep 2024): Pilot 6–12 months projects, engage schools and universities, evaluate results. Institutes: MPIfR Bonn, Universität Bielefeld

### D-TA7-WP4-4 (30 Sep 2026): Launch further projects jointly with the physics community. Institutes: MPIfR Bonn, KIT

1. Note that a separate document describing the PUNCH4NFDI reporting scheme is currently being developed. [↑](#footnote-ref-1)