Material Collection PUNCH-2.0

All PIs further interested in PUNCH4NFDI

Abstract

This document is intended as a living document for material collection; from this raw material, the input for the PUNCH-2.0 proposal should come. In particular, this collection of materials, ideas and interests shall serve as input to the discussion meetings in Kassel and Hamburg on 18 February and 4 March.

In case of interest, please first enter your institute in the institute list, together with the role you envisage (co-applicant / participant). Please mark entries in the document with your or your institute's initials.

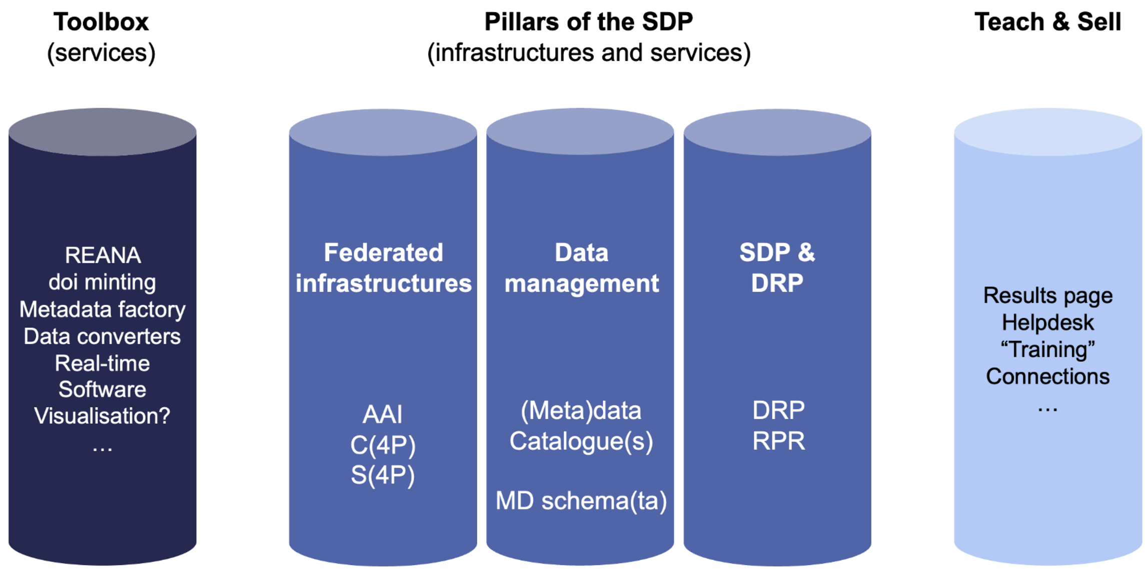
1 List of Contributors

- DESY, applicant institution, PI Thomas Schörner,
- KIT, co-applicant institution; PI Andreas Haungs
- FZJ, co-applicant institution, PI Susanne Pfalzner
- Institute XYZ, participant, PI NN
- FAU Erlangen-Nürnberg/ECAP, co-applicant institution, PI Kay Graf
- AIP Leibniz Institut für Astrophysik Potsdam, co-applicant institution, PI Matthias Steinmetz
- Universität Hamburg, PI Marcus Brüggen
- Leibniz Supercomputing Centre (LRZ), probably participant - to be discussed, PI Stephan Hachinger
- Unversity of Freiburg (UFR), co-applicant, PI Markus Schumacher
- University of Bonn (Ubo), probably co-applicant, PI Philip Bechtle and maybe Sebastian Neubert
- Johannes Gutenberg University (JGU), co-applicant, PI TBD.
- TU Dortmund University (TUDO), participant, PI Kevin Kröninger
- Frankfurt Institute for Advanced Studies (FIAS), co-applicant, PI Andreas Redelbach
- GSI, co-applicant institution, PI Mohammad Al-Turany
- TUM PI Lukas Heinrich
- Universität Bielefeld (UBI), co-applicant institution, PI Dominik Schwarz or Olaf Kaczmarek
- DZA, co-applicant institution, PI Hermann Heßling
- Universität Regensburg (UR), co-applicant institution, PI Sara Collins
- LMU, co-applicant institution, PI Joe Mohr or Thomas Kuhr

2 Overview

The model discussed so far for PUNCH-2.0 is depicted in the following figure. This motivates the choice of task areas (or sections in this document) and should structure our discussions below. Please consider this when inserting your ideas and interests.

Note: In contrast to the first funding period, it seems obvious that in case of diverging ideas for projects, but also for implementations and technologies, potentially harsh decisions will need to be made for the proposal, for the sake of a unified appearance and the chance of increased impact later.



3 Fundamental Questions

There are a few fundamental questions we better have answers to before submitting a new proposal. Feel free to add both answers to questions already written up and new questions.

- How does PUNCH-2.0 tie in with other funding schemes and initiatives — ErUM-Data, institutional funding for compute, ...?
  in first order PUNCH serves as contact point for the national NFDI from the four communities of fundamental physics, reaching and seeking for mutual benefits in tools and services...
- What can PUNCH-2.0 realistically achieve for the KAT, KET, KHuK, RDS communities — what do they want from us, and how can we ensure their support for our plans?
- ...

4 Potential Task Areas

4.1 Federated Infrastructures

- DESY will contribute to Storage4PUNCH implementations and offer own storage resources for that. Also full integration of DESY into Compute4PUNCH?
- DESY will take a role in running the AAI system, for example best practices to integrate production services/applications into an AAI (HelmholtzID, grid, EOSC).
- FZJ will contribute to Compute4PUNCH and Storage4PUNCH implementations and offer own storage and compute resources for that. We are n particular interested in developing tools a next generation archives (intelligent storage, knowledge trees (?)).
- FZJ would like to contribute with a connected code and data repository for the astro and lattice simulation communities full (communication between HPC and storage) etc. LRZ can contribute its experience here for selecting codes etc.
- FZJ will take a role in developing the AAI system connecting it to the BASE4NFDI efforts in this context (transfer the existing mode to state-of-the-art authentification, improvement of user experience with authentification, increase security during by preparing MFA.

– KIT will contribute to Compute4PUNCH and Storage4PUNCH implementations and offer own storage resources for that. In particular in developing tools like COBALD/Tardis etc.

– AIP will contribute to Storage4PUNCH implementations with S3 implementation and own storage resources for that. In particular in developing tools for transparent moving of data between POSIX and non-POSIX file systems

– AIP will contribute to Compute4PUNCH with extending REANA based workflows for computational tasks (e.g. inclusion of SLURM clusters) and working on a 'project based' management of resource allocations and an underlying monitoring system

– TUM is speaker on in-review FAIRUM proposal for federated analysis facilities and can contribute organizing a coherent integration of PUNCH and ErUM Data services and infrastructures

– AIP will run REANA, gitlab and other service components required for deployment, management and execution of workflows on federated PUNCH infrastructures.

– UFR will provide its flexible and extensible accounting ecosystem AUDITOR (Accounting data handling toolbox for opportunistic resources). It will provide support for implementation and operation. It is foreseen to extent or adapt it to further use cases according to the application specific requirements.

– JGU will provide tools to analyze the storages in PUNCH to identify possibilities for data reduction through deduplication/compression

– JGU will provide a service to pool local storages in nodes and provide a unified, large, and fast temporary storage to run jobs with high local storage demands.
(*MS: shall a short description of AUDITOR be provided (see below)?*
AUDITOR is able to cover a wide range of use cases and infrastructures. AUDITOR gathers accounting data via so-called collectors which are designed to monitor batch systems, COBalD/TARDIS, cloud schedulers, or other sources of information. The data is stored in a database and provided to so-called plugins, which act based on accounting records. An action could for instance be creating a bill of utilised resources, computing the CO2 footprint, adjusting parameters of a service, or forwarding accounting information to other accounting systems. Depending on the use case, a suitable collector and plugin can be chosen from a growing ecosystem of collectors and plugins. Libraries for interacting with AUDITOR are provided to facilitate the development of collectors and plugins by the community. Its REST interface and client libraries in Rust and Python allow a broad community of users to contribute to the expansion of the ecosystem and to implement their own collectors and plugins quickly and easily. The existing collectors and plugins can be combined in any way to implement various use cases. If a particular use case cannot be implemented with the existing components, it can easily be extended to cover any other use case in the context of job accounting in distributed computing systems.)

– UBo will continue to contribute to Storage4PUNCH and Compute4PUNCH and to its usage from the analysis side

– UBo will continue to provide storage and compute resources to PUNCH

– UBo will adapt `scitokens-cpp` to support the PUNCH AAI functionality which is being implemented (granular permissions), which is required for authorization handling in Storage4PUNCH (with XRootD) and Compute4PUNCH (HTCondor)

– FIAS will investigate sustainability of workflows on C4P/S4P and integrate a service to evaluate corresponding energy consumption/carbon footprint.

– GSI will support the implementation of Storage4PUNCH by contributing an XRootD-based system with a Lustre backend. Additionally, we will provide our own storage resources and ensure their seamless integration into Compute4PUNCH.

– UBI will continue to provide compute and storage resources for the astro and lattice communities, the integration of our resources in C4P and S4P was less trivial than expected and we still need to discuss what we can and should promise for PUNCH2.0. This is on-going.

– UBI is a co-applicant of the BASE4NFDI MultiCloud proposal and would like to follow up with activities along that line.
– DZA will build up the largest data collection in Germany for astrophysics data, including advanced multi-parameter and time-domain resolved cross-matches, compute and storage services and interfaces to major international data bases.
– DESY, FZJ and UBi will contribute the deployment and operation of modular distributed (meta-) data management services for the hep-lat (lattice) and further communities.
– UR will enable fine-grained access controls for modular distributed data and metadata management services and connection BASE4NFDI.
– DESY and FZJ will provide connection/integration to S4P and FTS with the modular distributed (meta-)data framework.
– UBi will provide integration into MD Harvesting and/or EOSC services.

4.2 Data Management

– DESY will contribute to metadata schemata development and metadata catalogue solutions — e.g. SciCat and / or ILDG for all DESY activities and beyond (nuclear, small accelerators, ...).
– DESY will develop a defined procedure for a scientific community to arrive at a metadata schema (best practices, tools, ...)
– DESY will develop metadata schemata & cross-walk registry for conversion between different metadata schemata for use with various catalogues → would allow for higher degree of interoperability between different schemata
– DESY will contribute to DOI minting and publication processes.
– FZJ and LRZ will take a role in the next step towards FAIR simulation data by implementing the strategies developed in PUNCH 1.0 into practice. This includes developing metadata schemes, easy storage tools, data accessibility tools, connective DOIs, training, etc.)
– FZJ and LRZ will develop workflows and tools for publishing data automatically with metadata, but with the possibility to augment this data by the user, if required. Also, the integration of these published datasets in DRPs is envisaged.
– FZJ will take a role in developing strategies, implementing and teaching energy-efficient research data handling.
– FZJ and LRZ will contribute to connect PUNCH to HPC data handing strategies. JGU is also willing to contribute to this.
– AIP will contribute to DOI minting and publication processes using DRP
– KIT will contribute to to metadata schemata development and metadata catalogue solutions, in particular taking into account the demands and needs for Astroparticle Physics.
– FAU will contribute to metadata for software (link to CodeMeta)
– AIP will contribute to improve findability of data across PUNCH communities and accessibility using community specific protocols (astropy/TAP, etc)
– GSI will work on enhancing metadata schema for Nuclear physics communities, with improved end-user interfaces and catalogues.
– GSI will contribute to API developments to enable transfer of metadata across interfaces: development of knowledge graphs to link schema with bibliographic systems, data repositories, data management planning tools, code repositories, etc.
– GSI will contribute to establishing metadata cross-links with broader communities like NAPMIX to ensure interoperability.

- UBI will continue to work on metadata catalogue improvement both in lattice and astro
- DZA will contribute to R&D on metadata in the petabyte range (relevant for SKAO) and provide access to this data.
- DESY, FZJ, UBi and UR will contribute the development of modular distributed (meta-)data management services for the hep-lat (lattice) and further communities.
- FZJ and DESY will contribute the minting of persistent identifiers ("DOI minting") and data publishing workflow as part of a modular distributed (meta-)data management.
- UR will contribute the development and extension of the data provenance information for use in the modular distributed data and metadata management services.

4.3 SDP and DRP

- AIP will work on making the DRP a cornerstone of reproducible workflows and data management in PUNCH
- AIP will work on the main cornerstones of the SDP: DRP, Resource Allocation and Management, Workflow-Management, Authentication and Autorisation, DataSearch Engine, and access to all internal and public services
- DESY will check applicability of HEP workflow ideas for photon science (i.e. REANA)
- DESY will add dCache integration with REANA (EOS integration is already supported).
- FZJ will contribute to the SDP in general, in particular to open data from Astrophysics simulations and Lattice simulations.
- KIT will contribute to the SDP in general, in particular to open data from Astroparticle Physics
- FAU will contribute primarily on software repositories and software in workflows in general
- UBo would be interested in Future Collider related workflows as DRPs and to contribute towards using PUNCH technologies in the physics preparation towards Future Colliders
- FIAS is interested in CBM and ALICE related workflows and their coupling to DRPs and REANA.
- GSI will work on introducing REANA workflows to the nuclear physics community
- GSI will work on design and integrate a workflow catalogue linked to, or within REANA to enhance metadata schema, enabling F.A.I.R. compliant pipelines.
- DZA will perform R&D on workflows for massive cross-matches and comprehensive statistical analyses.
- LMU will exploit the experience with analysis reinterpretation gained during PUNCH 1.0 to conceptually advance and contribute to the implementation of the DRP.

4.4 Toolbox

Here, individual services provided by and for the consortium can be listed that do not fit under the other headings.

- FZJ will contribute to ML methods in general, in particular in optimizing the use of HPC computing (GPU-computing and energy considerations).
- Uni Hamburg would like to provide a service for researchers unfamiliar with machine learning which offers services in ML data analysis. Building on experience gained within our cluster of excellence, we would provide a specialised consulting service to facilitate the application of ML in astrophysics/particle physics research. Building on our previous work and backed by three machine learning professorship that we have established in recent years, we would offer expertise in important areas such as data pre-processing, model selection and evaluation, algorithm implementation, and the interpretation of results. the service will operate as a collaborative hub. Workshops

and training sessions will be conducted to enhance the computational skills of researchers, fostering a self-sustaining, knowledgeable community.
In addition we will offer support for the use of large language models (LLMs) (we organised an international conference on LLMs in physics in 2024 in Hamburg) This means that we will test/tune/share models and act as consultants for the use of LLMs in teaching (e.g. specialised chatbots) and research (e.g. interface for data inference, statistics and plotting).

- similar to above TUM has been organizing AI block courses and knowledge transfer events as part of our EXC ORIGINS. TUM has experience in large-scale data processing with e.g. dataframe based / pyhton tools, where we can put in more development work
- DZA will provide R&D on algorithms for selecting rare events in real-time out of huge astronomical data streams, for analysing huge data objects, for data visualization, and for Smart Green Computing.
- FZJ, DESY and UR will contribute the development of the user and administration toolbox for a modular distributed (meta-)data management, including GUI.

4.5 Management Tasks

- DESY takes over the financial management of the consortium.
- KIT to be discussed....
- FZJ to be discussed, ... possibly connection to other European efforts in astrophysics (SPECTRUM, Space-CoE, etc.)
- Development of a business model for PUNCH2.0 and beyond (Andreas)
- The DZA will support the management of storage and computing, in cooperation with its international connections in astrophysics.

4.6 Training, Communication, Outreach Tasks

- DESY can further host the INDICO and web pages of the consortium.
- KIT could be interested in specific outreach tasks like citizen science... (comment HE: do we really need this to do?)
- FZJ can expand its existing formats for high performance computing to new formats like youtube videos and podcasts on PUNCH topics. A strong emphasis would be on new teaching resources on efficient use of resources.
- TUM can contribute to training on AI, Open Science, REANA
- UFR may offer tutorials on how to install and use the accounting ecosystem AUDITOR and how to use it to create awareness of the produced $(CO_2)_{eq}$ footprint by performing scientific computations.
- TUDO would like to maintain the link between science/RDM and university education, e.g. by collecting and developing courses for MSc. students ("Basics of scientific data processing") and by continuing to help integrate these topics into physics curricula.
- FIAS wants to offer teaching material/tutorials for high performance computing, resource-efficiency (also relevant for NFDI)
- LRZ can use its participation in various NFDI consortia and within GCS for contacts and to harmonise PUNCH methods with others in the NFDI and GCS frameworks (cf. goals of InHPC-DE project).
- DESY, FZJ, UBi and UR will provide user support for the modular distributed (meta-)data management services.

5 Other Ideas, Issues, Questions, Worries, ...

- FZJ would like to contribute to help desks for the PUNCH community. In particular, we would like to help users with their simulation data management on HPC machines.
- FZJ would like to offer tech sprints and hackathons to solve data mangement problems.
- DESY will involve in open data policy discussions with CERN.
- Hamburg would be interested in running public outreach campaigns and develop material for educators.
- KIT should be included in discussions around EOSC, EURO-HPC, NHR-centers, etc...
- FZJ should also be included in discussions around ESOC, NHR.centres, VO, CDS etc. ...
- We should spend reasonable time and efforts on sustainability (Andreas)
- FAU could be linking to ESCAPE (EOSC)
- UFR shall participate in discussions with NHR alliance and can serve as link to ErUM-Data funded research networks in the topical area of federated infrastructures (running until 9/2024 FIDIUM, proposed new networks FAIRUM, FUSE, SUSFECIT)
- what is included in the central virtual help-desk / ticket system? is it used ? how is the feedback? how to include new tools and their user support? (MS from UFR)
- JGU can provide links to NHR-centers for integration of C4P and S4P.
- DZA will coordinate its data policy and data access with international data providers.
- DZA will contribute to future initiatives in EU computing.