Harmonising heterogeneous metadata ecosystems in PUNCH4NFDI

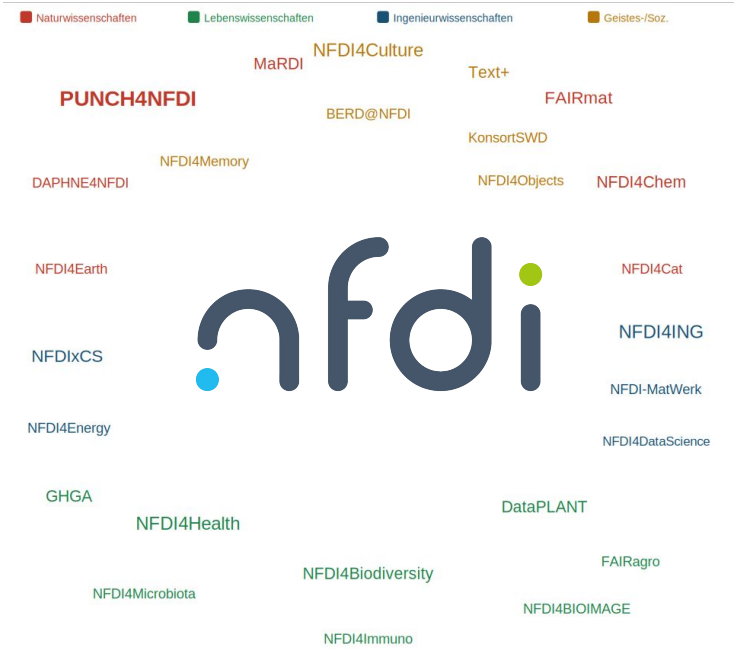
Victoria Tokareva for PUNCH4NFDI TA4 “Data portal”

NAPMIX Training Workshop 2026 – Metadata, RDM & FAIR Integration
Madrid, 18-19 May 2026

NFDI: Germany's national research data infrastructure

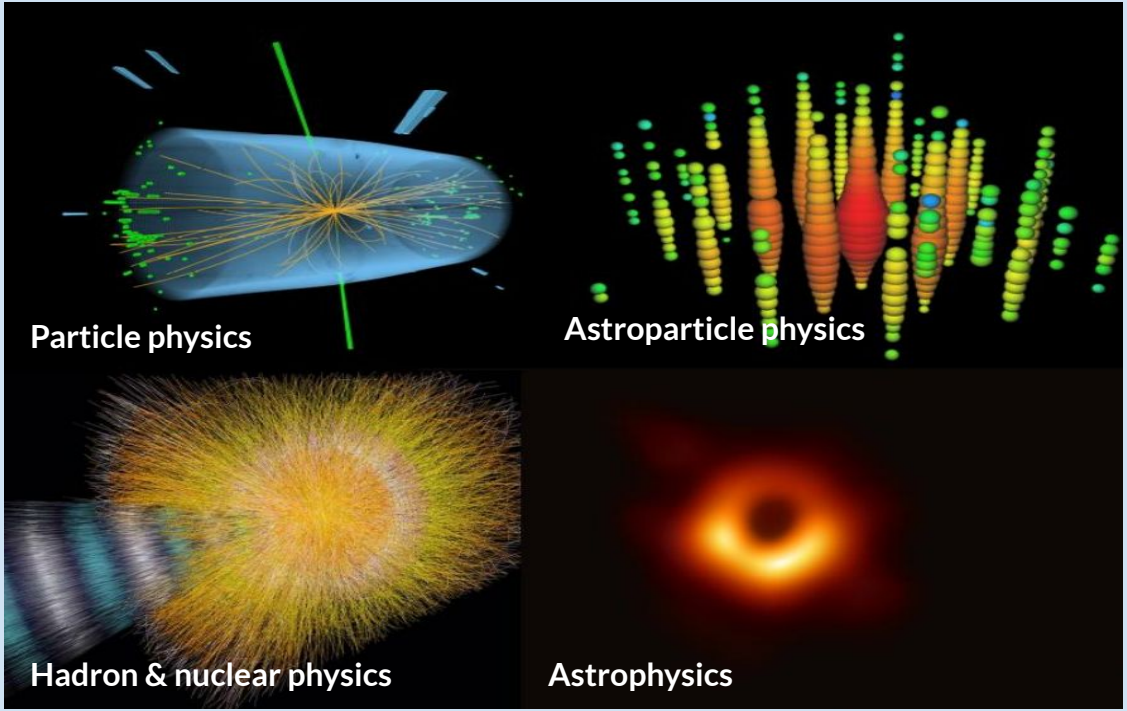
- Federated national initiative for research data management
- Organised through 26 domain-specific consortia for engineering, humanities, life and natural sciences
- **Supports FAIR and interoperable research infrastructures at national scale**
- Bridges research communities, data services and long-term curation

★ PUNCH4NFDI is the consortium for data-intensive physics communities



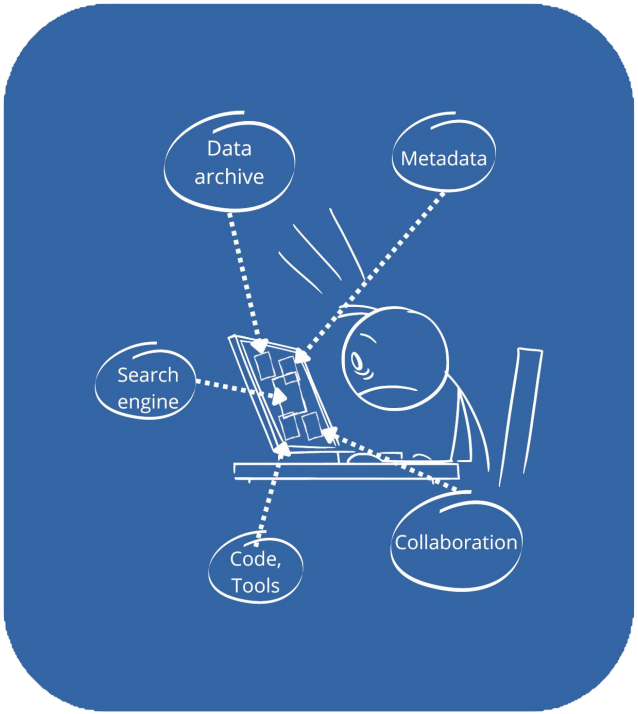
Particles,
Universe,
Nu-
Clei and
Hadrons
4 (for)
Nationale
Forschungs-
Daten
Infrastruktur

is the consortium of particle, astroparticle, hadron & nuclear physics and astrophysics (PUNCH Communities)

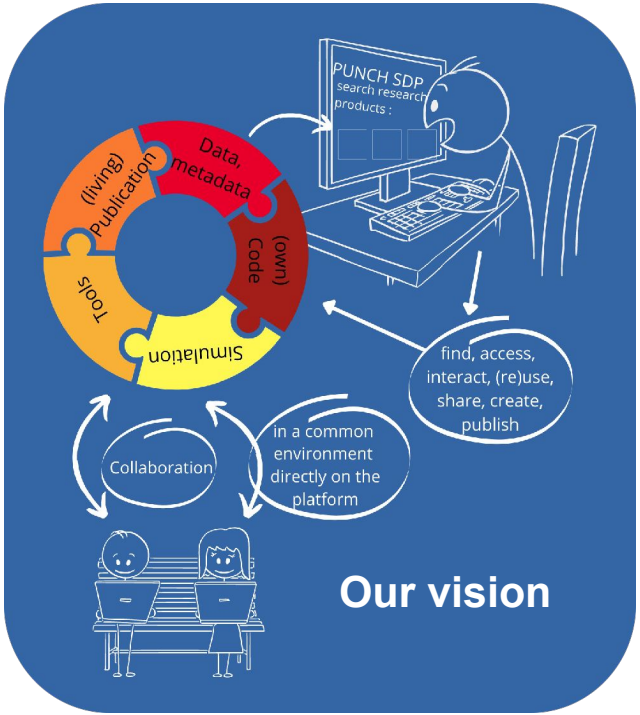


PUNCH4NFDI Science Data Platform

Past



Future



Challenges of metadata integration

- Rapid growth of experimental and simulation data (TB → PB → EB)
- Increasing complexity of analysis pipelines (computationally intensive research)
- Collaboration-internal data practices -> limited transparency and external reuse
- Variety of historically developed data management practices, tools, specific software, data formats -> lack of interoperability
- Intrinsic expert knowledge in research workflows:
  - Expertise-dependent reuse
  - Knowledge loss
- Growing interdisciplinarity & multi-messenger research
- **Need for reproducibility and long-term reuse**

Heterogeneous metadata ecosystems in PUNCH4NFDI

Category	Semantic artefacts used by different PUNCH use cases
Metadata models & catalogues	Astro-WISE · LTA Catalogue · TMSS · ObsCore DM · QCDml · Portal-defined dataset metadata · NAPMIX schema
Formats & serialization	MS · ROOT · FITS · PSRFITS · HDF5 · LIME · VOTable · JSON · XML
Semantic standards	Dublin Core · DataCite · UCD · PROV-O · QUDT · SOSA · openPMD · NeXus
Access & query	VO services · TAP · ADQL · LTA interface · Open Data portal · REST APIs · MDC queries
Protocols	OAI-PMH · HTTP(S) · REST · GridFTP · SRM · SAMP
Identifiers	DOI · ORCID · ROR · IVORN · LFN · SURL · ObsIDs · dataset IDs

Shared interoperability layers across communities

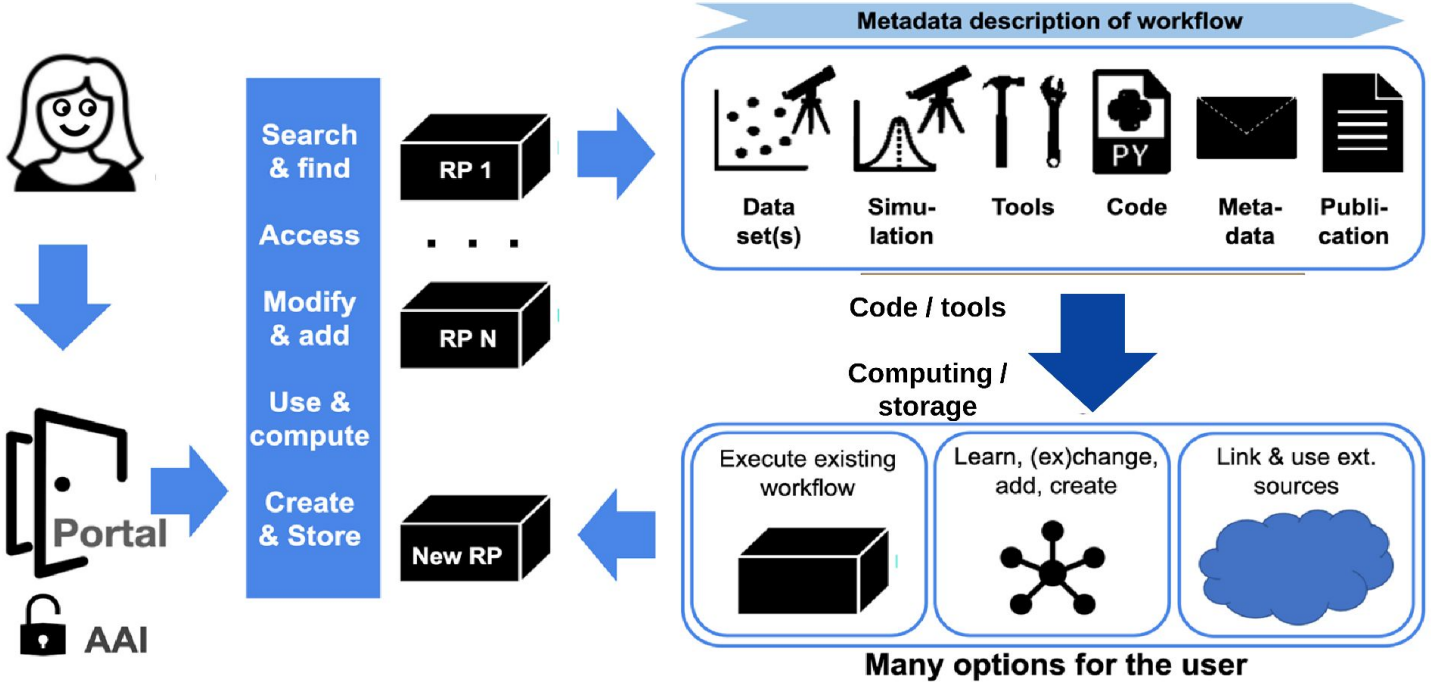
- Identifier layer: DOI · ORCID · ROR · structured dataset identifiers (naming conventions)
- Metadata Core: Dublin Core · DataCite
- Exchange layer: OAI-PMH · Custom REST APIs
- Serialization layer: XML · JSON

Interoperability tensions in federated infrastructures

Goal		Challenge
Common discovery	→	Domain-specific semantics
Interoperability	→	Community autonomy
Standardisation	→	Legacy infrastructures
Rich metadata	→	Expert workload / Harvest complexity
Reproducible & reusable workflows	→	Infrastructure complexity

PUNCH4NFDI Science Data Platform

Data Portal is the central element of the Science Data Platform, which serves as a habitat for the entire lifecycle of PUNCH4NFDI Digital Research Products (DRP)



PUNCH4NFDI Digital Research Product

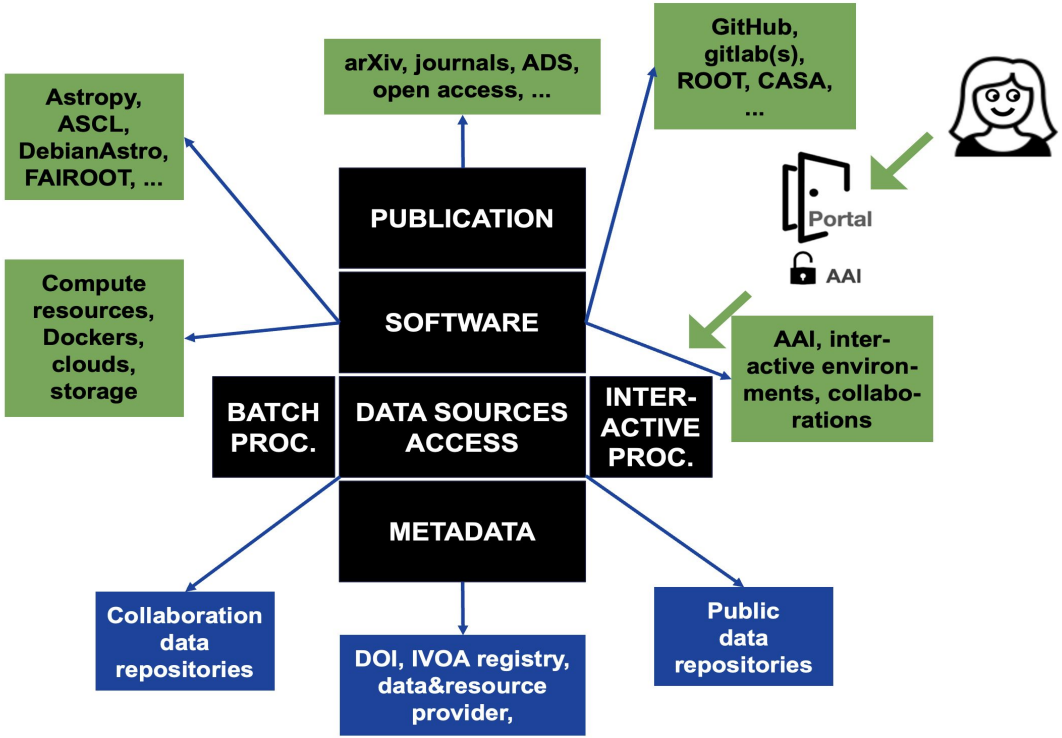
DRP encapsulates digital research outputs of PUNCH4NFDI communities within SDP ecosystem.

Allows:

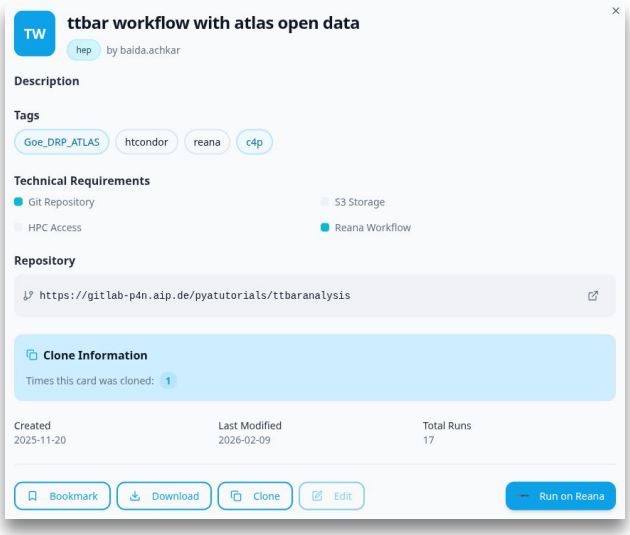
- live analysis and live peer review
- blind discovery

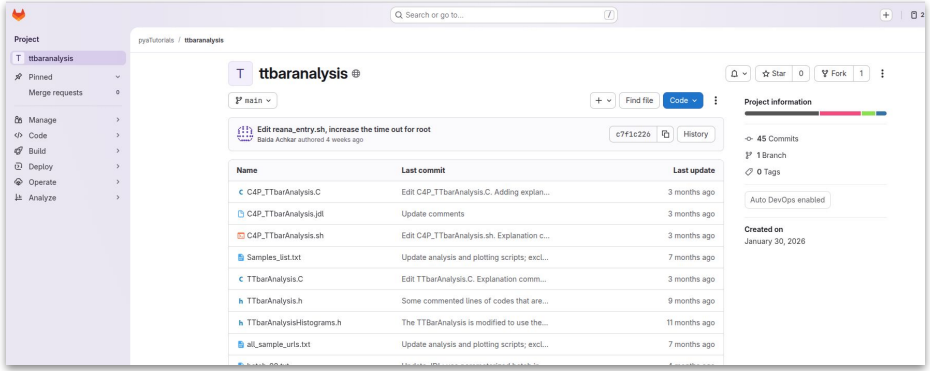
PUNCH DRPHub - a platform to access PUNCH DRPs (beta)

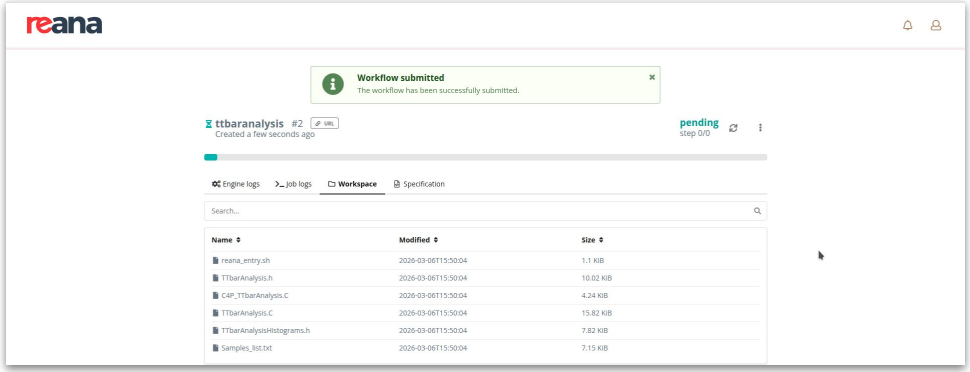
https://drphub-p4n.aip.de/



Example: reproducible analysis of ATLAS Open Data





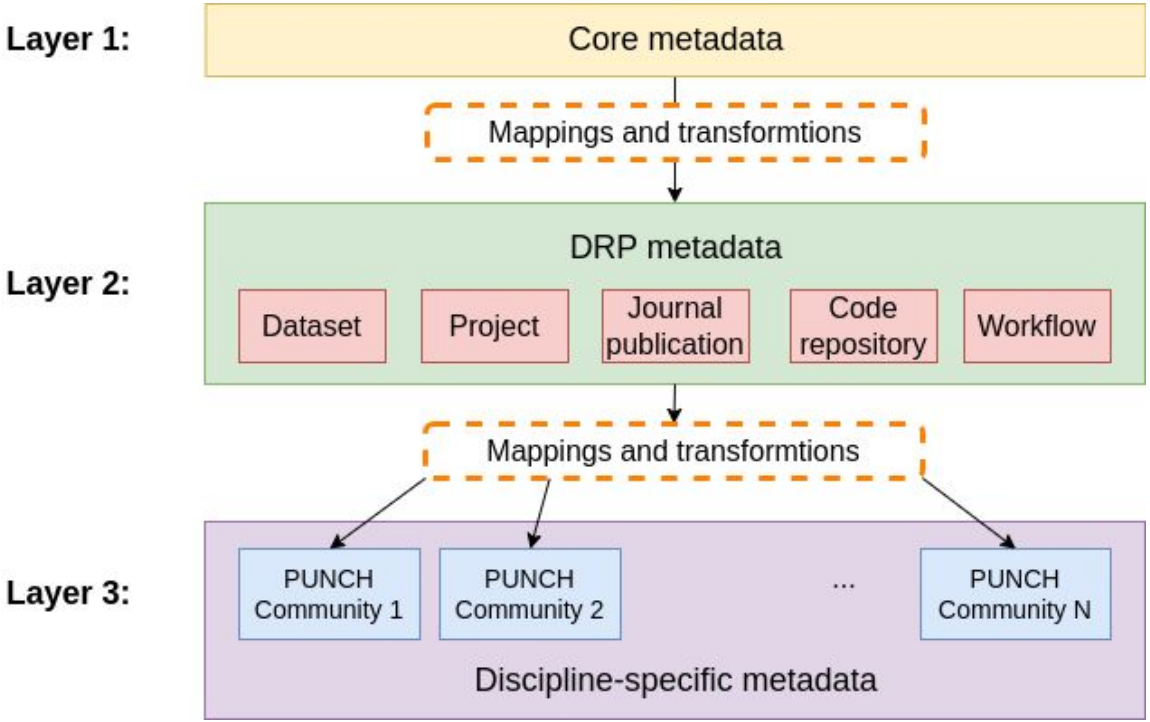


Example DRP by B. Achkar

Metadata integration principles

- **Maximal possible reuse of already used standards (semantic artifacts: metadata standards, controlled vocabularies, ontologies, taxonomies) and methods**
- Based on one of widely used metadata kernels
- Defines minimal and recommended metadata for DRP
- Allows coherent treatment and integration of metadata for different PUNCH Use Cases into the PUNCH SDP
- Compatible with NFDI Core -> Consortia are mapping their metadata to a common ground based on DataCite, DCAT, or schema.org
- Methodology: use-case driven approach, combining hierarchical and recursive structures with cross-community schema alignment to support efficient metadata reuse, traceability, and long-term curation.

PUNCH4NFDI SDP metadata model



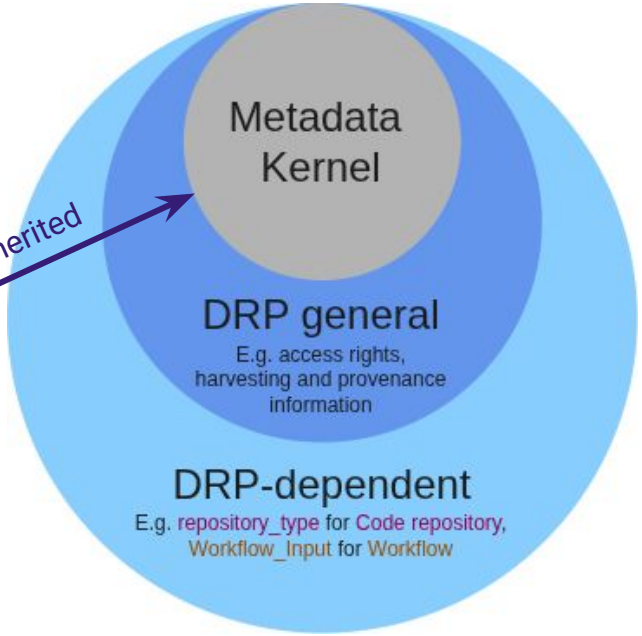
- **Multi-layered model**
- Builds on the widely used **DataCite Kernel** required keywords
- Supports **incorporation of discipline-specific schemas** from PUNCH communities
- Harvesting: preferred **OAI-PMH or REST** based APIs
- Metadata exchange format: **XML, JSON**

Metadata obligation levels explained

DataCite Kernel

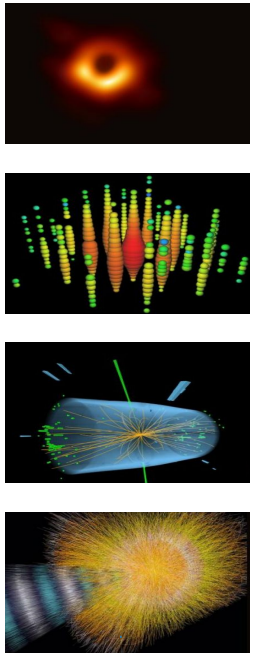
inherited

DRP metadata



SDP-necessary info

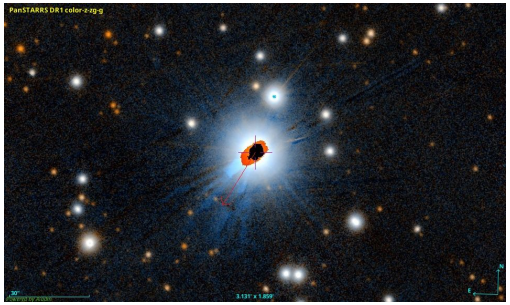
Discipline-specific metadata records



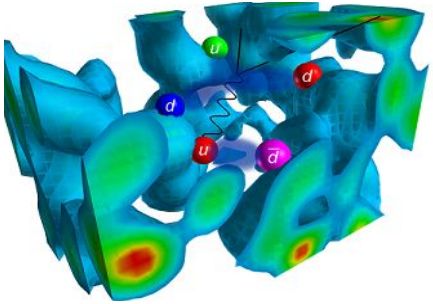
Plenty of discipline-specific information

* Recommended for interoperability - but not obligatory

PUNCH Data Providers and Use Cases

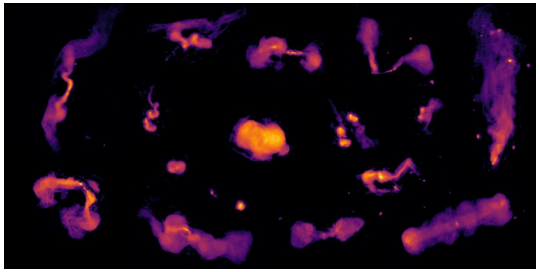


Leibniz-Institut für Astrophysik Potsdam, Astronomisches Rechen-Institut Heidelberg



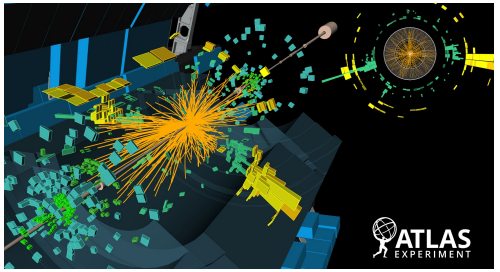
International Lattice Data Grid

NAPMIX



LOFAR

KASCADE Cosmic Ray Data Centre



ATLAS Open Data

Use-case driven analysis of data providers

Collected information:

- metadata harvesting access methods
- service endpoints and APIs
- supported metadata formats
- persistent identifier systems (DOI and others)

Key observations:

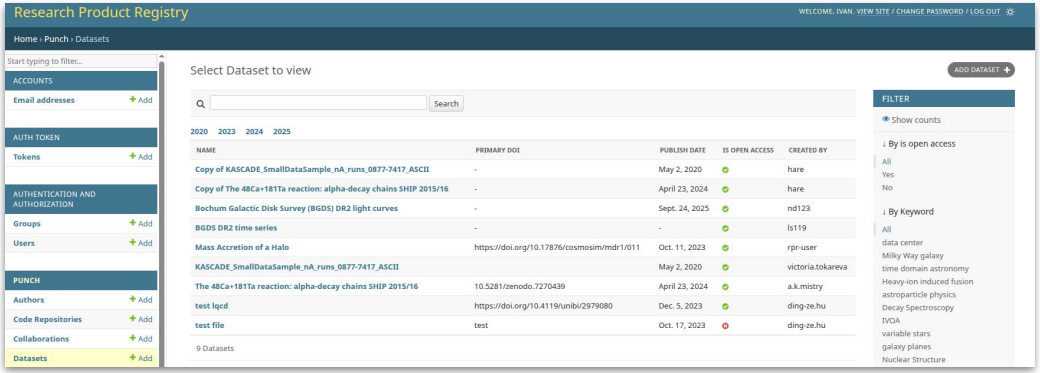
- Dublin Core and DataCite are most common harvesting formats
- Limited DOI adoption by PUNCH communities/research groups
- OAI-PMH widely adopted (5 of 6 analysed providers)
- Several providers additionally expose REST APIs

Implications for PUNCH metadata integration:

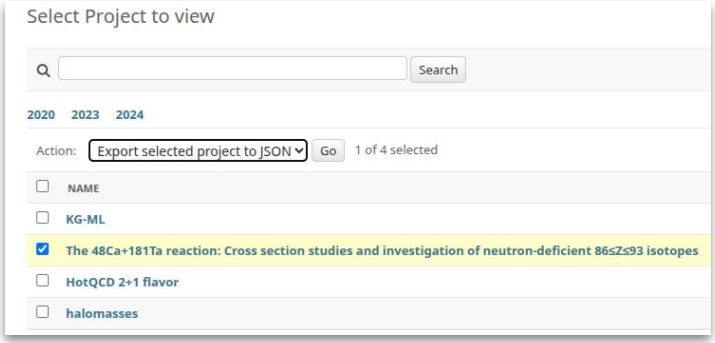
- Shallow metadata can be harvested from most providers
- DataCite to Dublin Core mappings enable transformation of harvested metadata
- Need for crosswalks for domain-specific schemas - work in progress

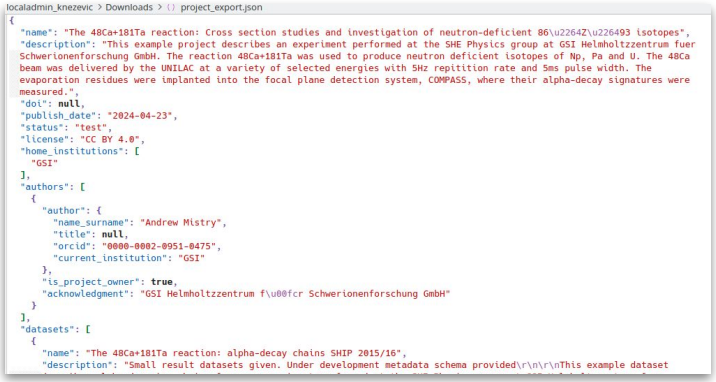
Research Data Repository of PUNCH4NFDI

Interactive data ingestion platform



Metadata schema export





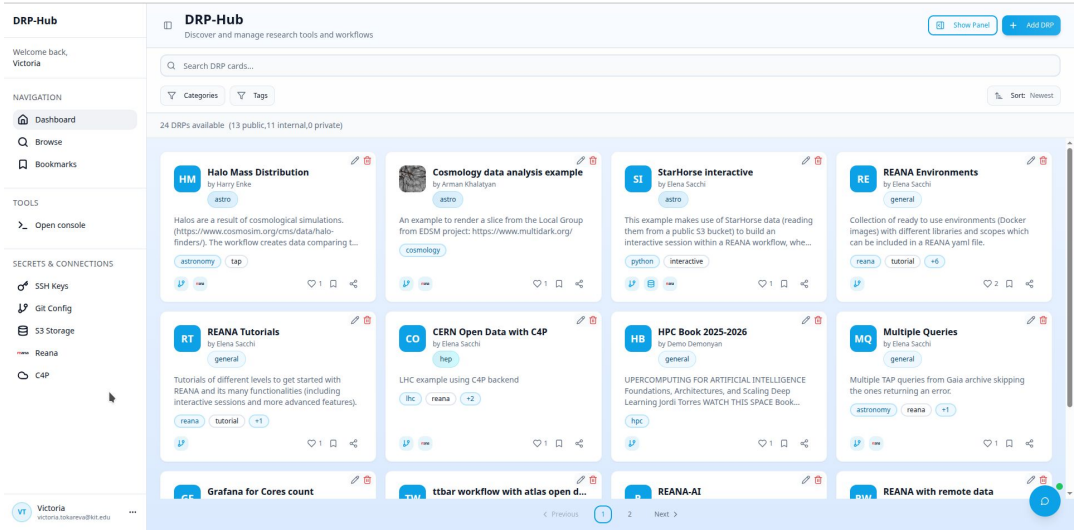
Author: Ivan Knezevic

Digital Research Product Hub

enabling reproducible science

A federated infrastructure provided by the **PUNCH4NFDI consortium**.

FAIR Principles | Reproducible Workflows | PARTNERS | Federated Infrastructure



Features:

- Access management via Helmholtz AAI, PUNCH and Keycloak
- Support for private, internal and public DRPs
- Bookmarking, cloning and sharing of DRPs
- REANA-based multi-backend workflow execution
- Integrated Jupyter environments
- AI-assisted interaction (PhysicsLLM / ERUM-DataHub)

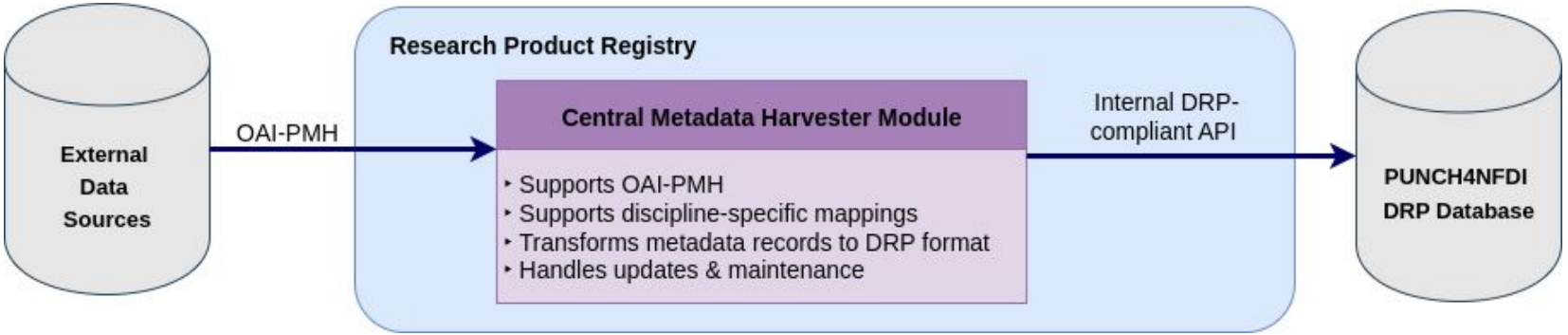
Authors:

Elena Sacchi (AIP),
Arman Khalatyan (AIP),
Olaf Michaelis (AIP),
Harry Enke (AIP)

https://drphub-p4n.aip.de/

Metadata harvesting and transformation

- Metadata records for the digital research products are being harvested in the Research Product Registry (RPR)
- RPR stores DRP metadata and preserves pointers to the [meta]data repositories, software environments (containers), and execution parameters
- Preferred harvesting method: OAI-PMH, with additional other supported endpoints (preferred REST based APIs)
- Support integration of external data sources



Towards PUNCH4NFDI-2.0

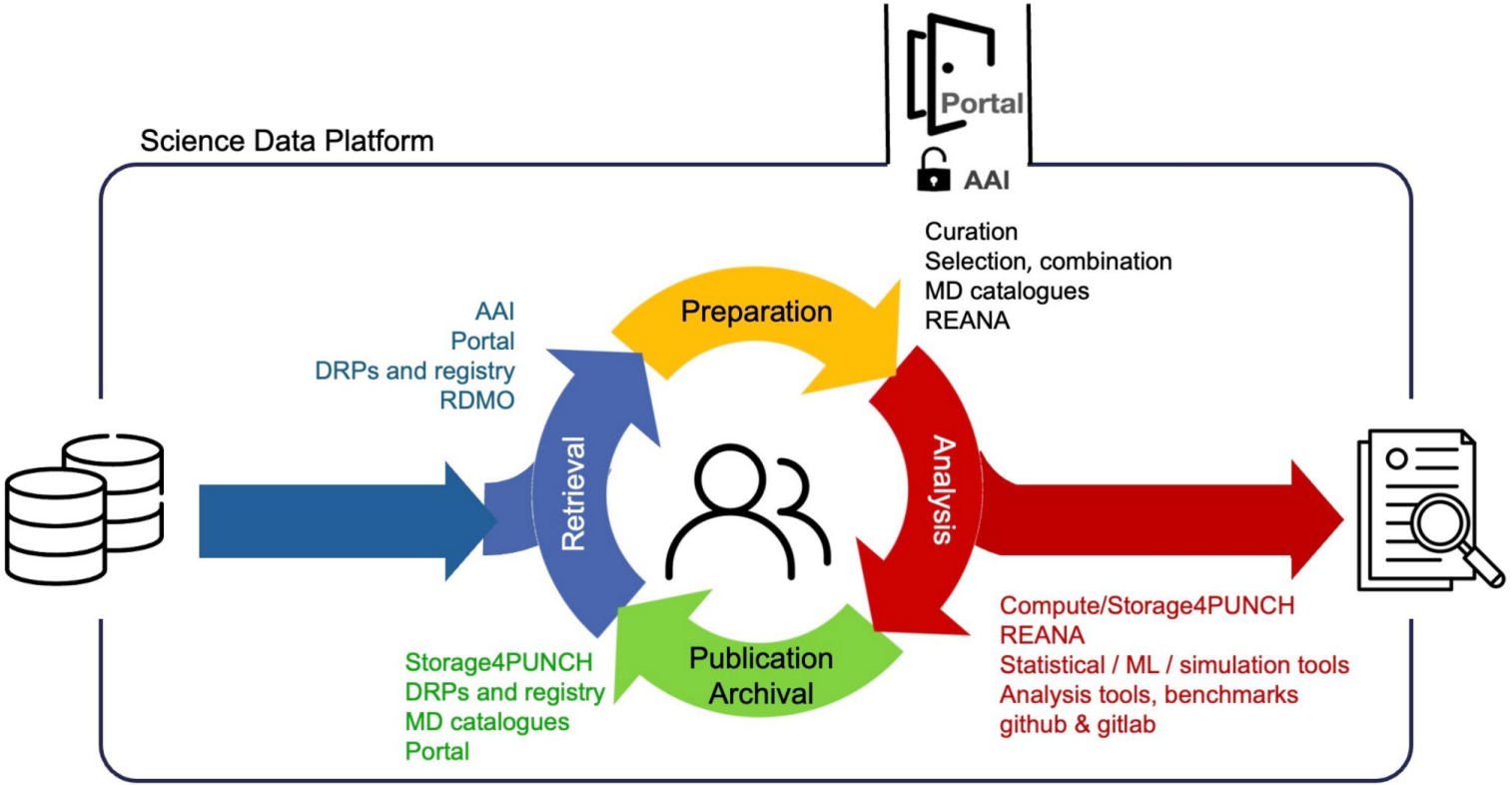


Diagram by T. Schöner

Outlook - Ongoing developments

- Integration of the Research Product Registry (RPR) and DRPHub
- Further development of executable and reproducible DRP workflows
- Extension of metadata mappings and schema transformations
- Support for integration of external metadata providers

Future directions

- Development of formal DRP specifications
- Refinement of the multi-layer metadata architecture
- Cross-community harmonisation of metadata exchange approaches
- Alignment with European and international interoperability initiatives

Thank you for your attention!

Learn more about PUNCH4NFDI: https://www.punch4nfdi.de/
PUNCH4NFDI Results page: https://results.punch4nfdi.de/
Contact me: victoria.tokareva@kit.edu

Example DRP metadata record

This record includes:

- Information about KG-ML (Machine Learning on KASCADE-Grande data)
- Code repository information
- Related publication

https://shorturl.at/094Ja

Backup: DRP vs Fair Digital Objects (FDO)

DRP :

- deliver products which are FAIR (as possible)
- are immediately useable
- combine heterogeneous elements:
  - data and metadata
  - software and workflows
  - exec environments
- resolve interoperability by using
  - conversion of data (on the fly)
  - provided by workflow/software
  - or/and metadata
- works with the communities' different approaches
- syncretic approach

FDO:

- uniform 'view' of digital objects (DO)
- provides a data type registry (DTR) for the DO
- build on a resolving mechanism (implementation) of PID (DOIP)
- abstract from the communities developments etc.
- implement findability
- provide pointers to metadata etc.
- require each FDO to adhere to certain standards, regardless of its state or relations

Source: Enke, H. (2025) "Science Data Portal and Digital Research Product of PUNCH4NFDI". Zenodo, 31 July. doi:10.5281/zenodo.16637860.

Mandatory properties of the metadata kernel

NFDI level discussion on properties obligation priorities: https://doi.org/10.5281/ZENODO.15227235