



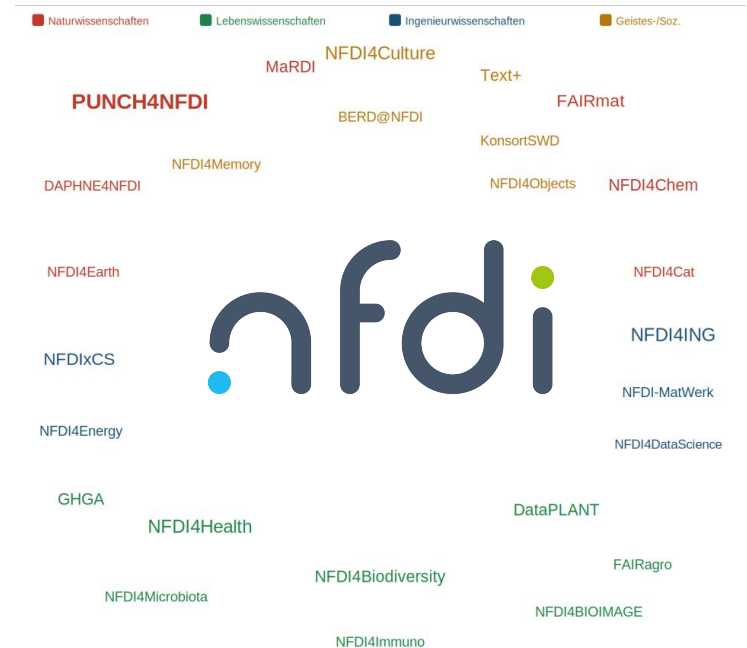
Harmonising heterogeneous metadata ecosystems in PUNCH4NFDI

Victoria Tokareva for PUNCH4NFDI TA4 “Data portal”

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

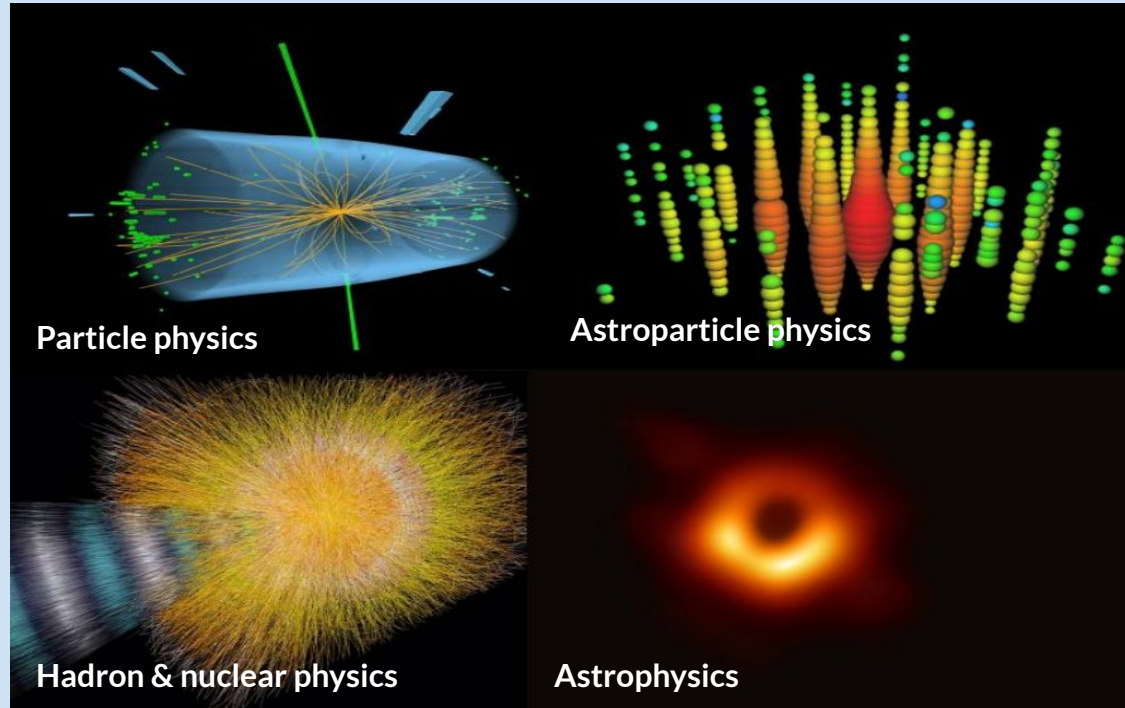
NFDI: Germany's national research data infrastructure

- 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



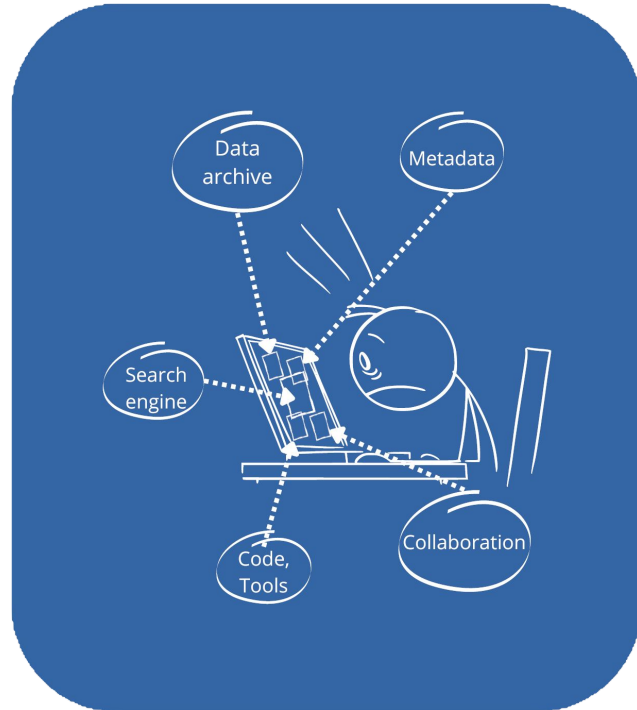
Particles,
Universe,
Nu-
Clei and
Hadrons
4 (for)
Nationale
Forschungs-
Daten
Infrastruktur

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

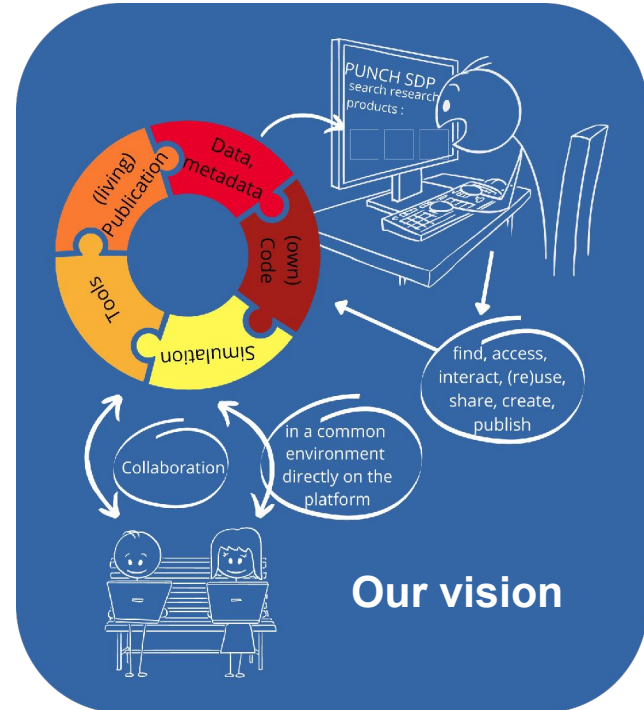


PUNCH4NFDI Science Data Platform

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

Common discovery

Interoperability

Standardisation

Rich metadata

Reproducible & reusable workflows



Challenge

Domain-specific semantics

Community autonomy

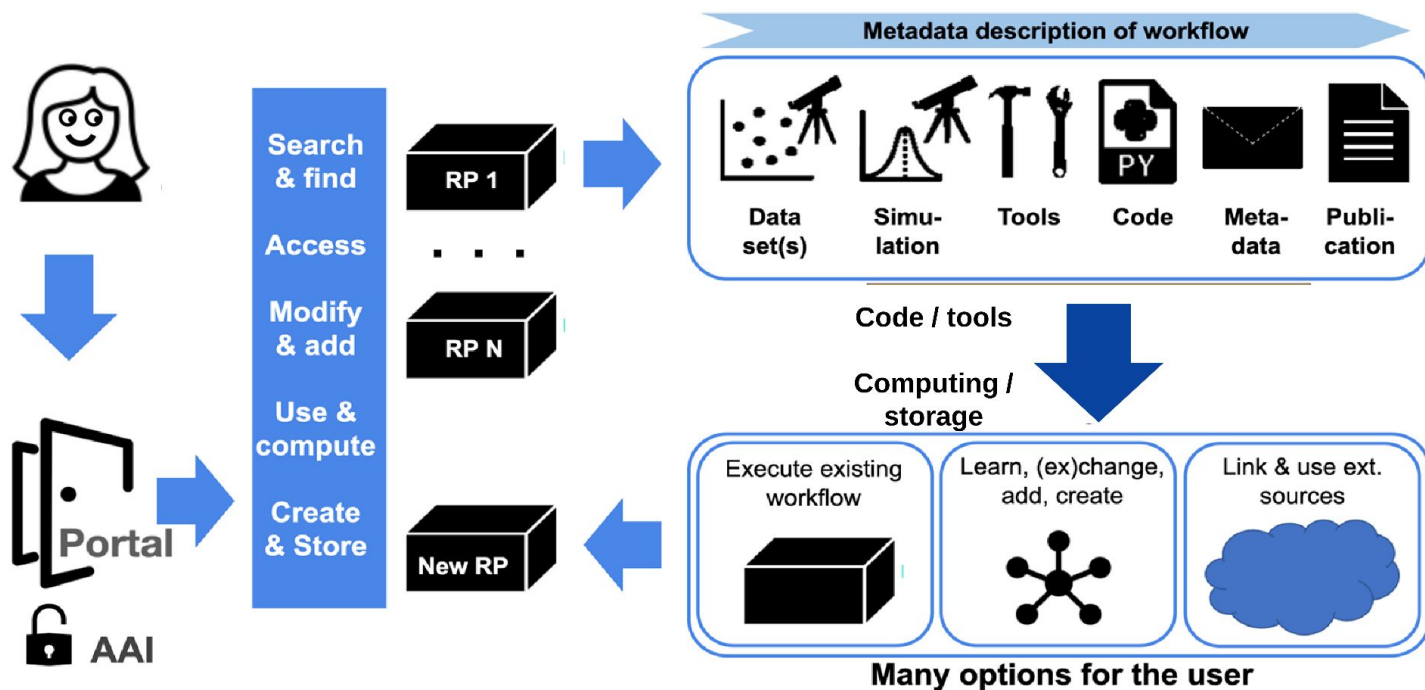
Legacy infrastructures

Expert workload / Harvest complexity

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.

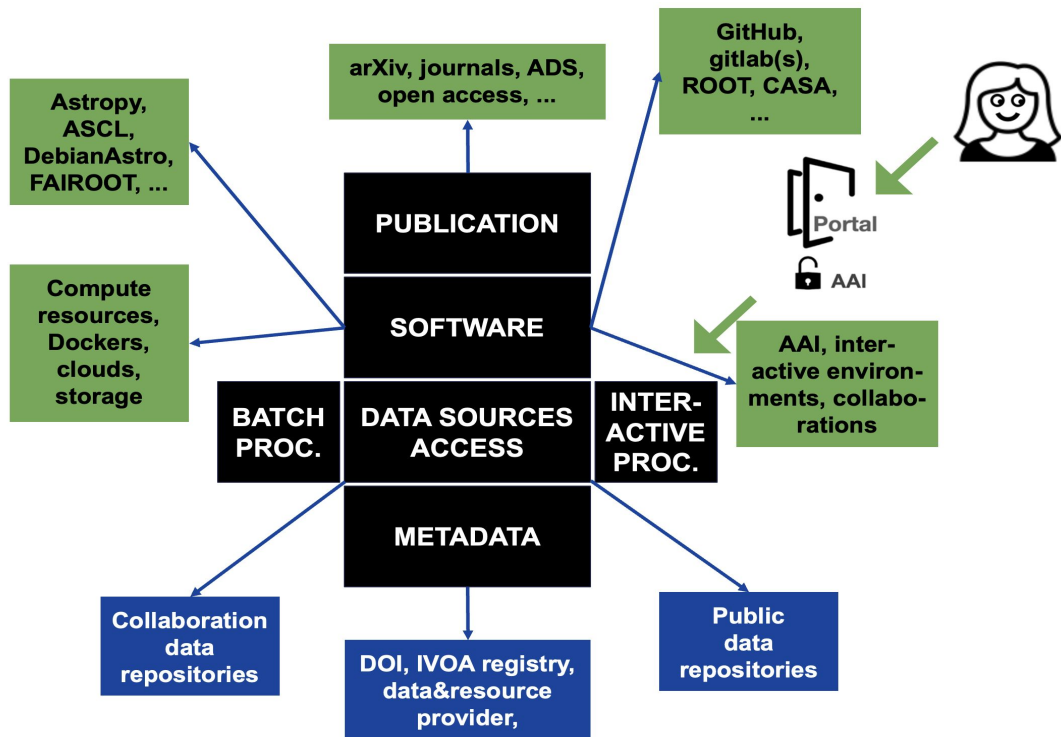
Allows:

- 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

ttbar workflow with atlas open data
by baida.achkar

Description

Tags
Goe_DRP_ATLAS htcondor reana c4p

Technical Requirements
Git Repository S3 Storage
HPC Access Reana Workflow

Repository
https://gitlab-p4n.aip.de/pyatutorials/ttbaranalysis

Clone Information
Times this card was cloned: 1

Created: 2025-11-20 | Last Modified: 2026-02-09 | Total Runs: 17

Bookmark Download Clone Edit Run on Reana

Project: ttbaranalysis

Commit: c7f1c26

Name	Last commit	Last update
< C4P_TTbarAnalysis.C	Edit C4P_TTbarAnalysis.C. Adding explan...	3 months ago
C4P_TTbarAnalysis.jdl	Update comments	3 months ago
C4P_TTbarAnalysis.sh	Edit C4P_TTbarAnalysis.sh. Explanation c...	3 months ago
Samples_list.txt	Update analysis and plotting scripts; excl...	7 months ago
TTbarAnalysis.C	Edit TTbarAnalysis.C. Explanation comm...	3 months ago
TTbarAnalysis.h	Some commented lines of codes that are...	9 months ago
TTbarAnalysisHistograms.h	The TTbarAnalysis is modified to use the...	11 months ago
all_sample_urls.txt	Update analysis and plotting scripts; excl...	7 months ago

Workflow submitted
The workflow has been successfully submitted.

ttbaranalysis #2
Created a few seconds ago

pending step 0/0

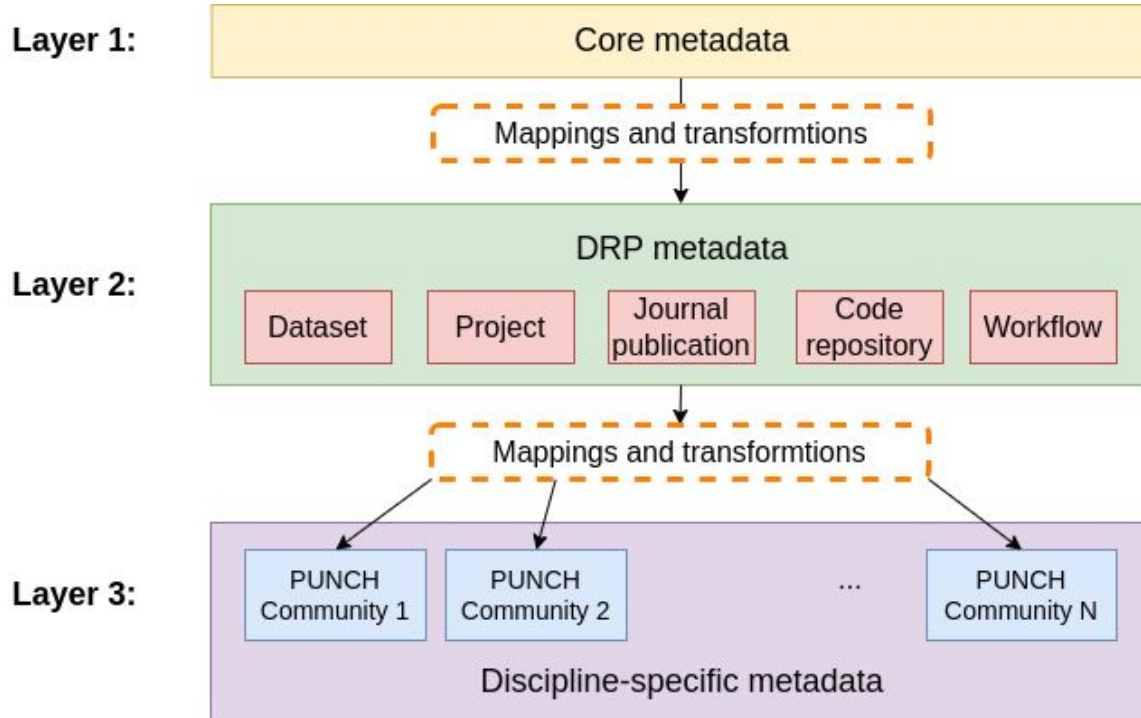
Name	Modified	Size
reana_entry.sh	2026-03-06T15:50:04	1.1 KIB
TTbarAnalysis.h	2026-03-06T15:50:04	10.02 KIB
C4P_TTbarAnalysis.C	2026-03-06T15:50:04	4.24 KIB
TTbarAnalysis.C	2026-03-06T15:50:04	15.82 KIB
TTbarAnalysisHistograms.h	2026-03-06T15:50:04	7.82 KIB
Samples_list.txt	2026-03-06T15:50:04	7.15 KIB

Example DRP by B. Achkar

Metadata integration principles

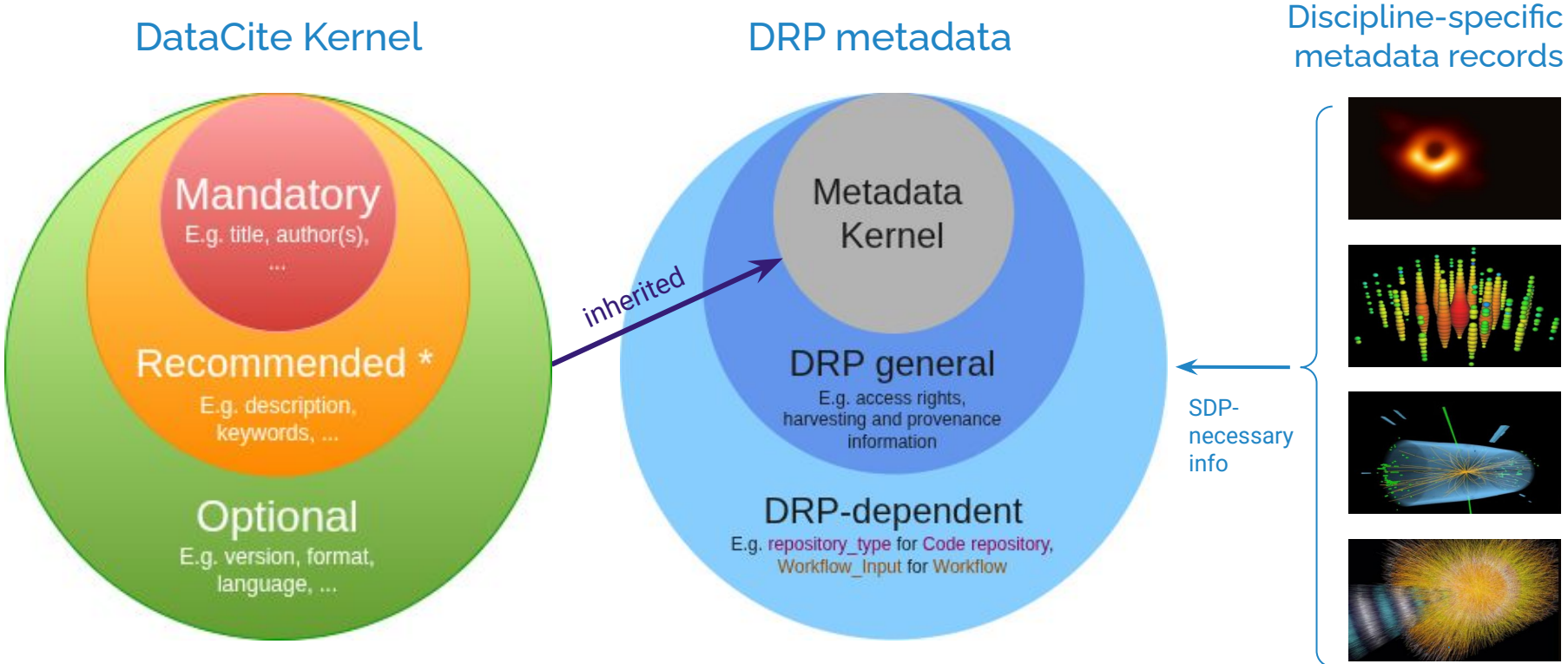
- Maximal possible reuse of already used standards (semantic artifacts: metadata standards, controlled vocabularies, ontologies, taxonomies) and methods
- Based on one of widely used metadata kernels
- Defines minimal and recommended metadata for DRP
- Allows coherent treatment and integration of metadata 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**

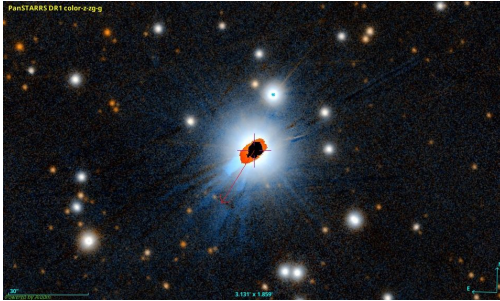
Metadata obligation levels explained



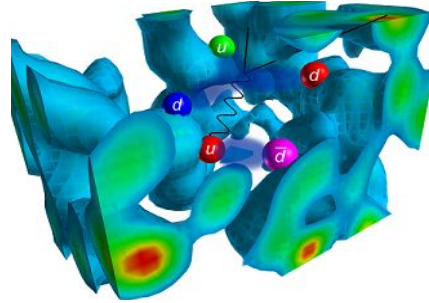
* Recommended for interoperability - but not obligatory

□ Plenty of discipline-specific information

PUNCH Data Providers and Use Cases



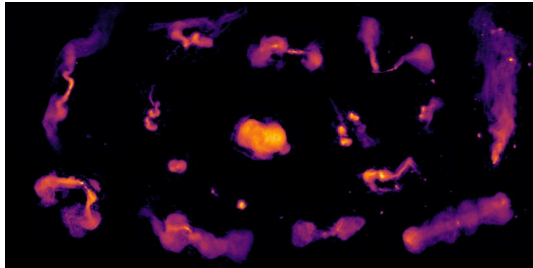
Leibniz-Institut für Astrophysik Potsdam,
Astronomisches Rechen-Institut Heidelberg



International Lattice Data Grid



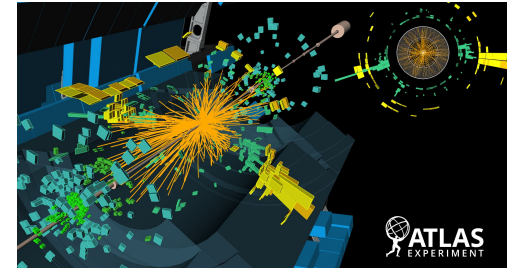
NAPMIX



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

Metadata schema export

Interactive data ingestion platform

Research Product Registry

Home » PUNCH » Datasets

Start typing to filter...

ACCOUNTS

- Email addresses + Add

AUTH TOKEN

- Tokens + Add

AUTHENTICATION AND AUTHORIZATION

- Groups + Add
- Users + Add

PUNCH

- Authors + Add
- Code Repositories + Add
- Collaborations + Add
- Datasets + Add

Select Dataset to view

Q [] Search

2020	2023	2024	2025						
NAME	PRIMARY DOI	PUBLISH DATE	IS OPEN ACCESS	CREATED BY					
Copy of KASCADE_SmallDataSample_nA_runs_0877-7417_ASCII	-	May 2, 2020	🟢	hare					
Copy of The 48Ca+181Ta reaction: alpha-decay chains SHIP 2015/16	-	April 23, 2024	🟢	hare					
Bochum Galactic Disk Survey (BGDS) DR2 light curves	-	Sept. 24, 2025	🟢	nd123					
BGDS DR2 time series	-	-	🟢	ls119					
Mass Accretion of a Halo	https://doi.org/10.17876/cosmosim/mdr1/011	Oct. 11, 2023	🟢	rpr-user					
KASCADE_SmallDataSample_nA_runs_0877-7417_ASCII	-	May 2, 2020	🟢	victoria.tokareva					
The 48Ca+181Ta reaction: alpha-decay chains SHIP 2015/16	10.5281/zenodo.7270439	April 23, 2024	🟢	a.k.mistry					
test_lqcd	https://doi.org/10.4119/unibi/2979080	Dec. 5, 2023	🟢	ding-ze.hu					
test_file	test	Oct. 17, 2023	🔴	ding-ze.hu					

9 Datasets

FILTER

- Show counts
- By is open access
 - All
 - Yes
 - No
- By Keyword
 - All
 - data center
 - Milky Way galaxy
 - time domain astronomy
 - Heavy-ion induced fusion
 - astroparticle physics
 - Decay Spectroscopy
 - IYGA
 - variable stars
 - galaxy planes
 - Nuclear Structure

Select Project to view

Q [] Search

2020 2023 2024

Action: Export selected project to JSON Go 1 of 4 selected

- NAME
- KG-ML
- The 48Ca+181Ta reaction: Cross section studies and investigation of neutron-deficient 86Zr93 isotopes
- HotQCD 2-1 flavor
- halomasses

```
localadmin_knezevic > Downloads > ( ) project_export.json
{
  "name": "The 48Ca+181Ta reaction: Cross section studies and investigation of neutron-deficient 86Zr93 isotopes",
  "description": "This example project describes an experiment performed at the SHE Physics group at GSI Helmholtzzentrum fuer Schwerionenforschung GmbH. The reaction 48Ca+181Ta was used to produce neutron deficient isotopes of Nb, Pa and U. The 48Ca beam was delivered by the UNILAC at a variety of selected energies with SHz repetition rate and Sms pulse width. The evaporation residues were implanted into the focal plane detection system, COMPASS, where their alpha-decay signatures were measured.",
  "doi": null,
  "publish_date": "2024-04-23",
  "status": "test",
  "license": "CC BY 4.0",
  "home_institutions": [
    "GSI"
  ],
  "authors": [
    {
      "author": {
        "name_surname": "Andrew Mistry",
        "title": null,
        "orcid": "0000-0002-0951-8475",
        "current_institution": "GSI"
      },
      "is_project_owner": true,
      "acknowledgment": "GSI Helmholtzzentrum fuer Schwerionenforschung GmbH"
    }
  ],
  "datasets": [
    {
      "name": "The 48Ca+181Ta reaction: alpha-decay chains SHIP 2015/16",
      "description": "Small result datasets given. Under development metadata schema provided\r\n\r\nThis example dataset"
    }
  ]
}
```

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

The screenshot displays the Digital Research Product Hub interface. On the left is a navigation sidebar with sections: DRP-Hub, WELCOME BACK, VICTORIA, NAVIGATION (Dashboard, Browse, Bookmarks), TOOLS (Open console), and SECRETS & CONNECTIONS (SSH Keys, Git Config, S3 Storage, Reana, C4P). The main content area is titled 'DRP-Hub' and contains a search bar, filters for 'Categories' and 'Tags', and a 'Sort: Newest' option. Below this, there are 24 DRPs available (13 public, 11 internal, 0 private). The visible cards include: 'Halo Mass Distribution' by Harry Enke (astro), 'Cosmology data analysis example' by Arman Khalatyan (astro), 'StarHorse interactive' by Elena Sacchi (astro), 'REANA Environments' by Elena Sacchi (general), 'REANA Tutorials' by Elena Sacchi (general), 'CERN Open Data with C4P' by Elena Sacchi (hep), 'HPC Book 2025-2026' by Demo Demiryanyan (general), 'Multiple Queries' by Elena Sacchi (general), 'Grafana for Cores count', 'ttbar workflow with atlas open d...', 'REANA-AI', and 'REANA with remote data'. Each card shows a title, author, category, and a brief description.

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)

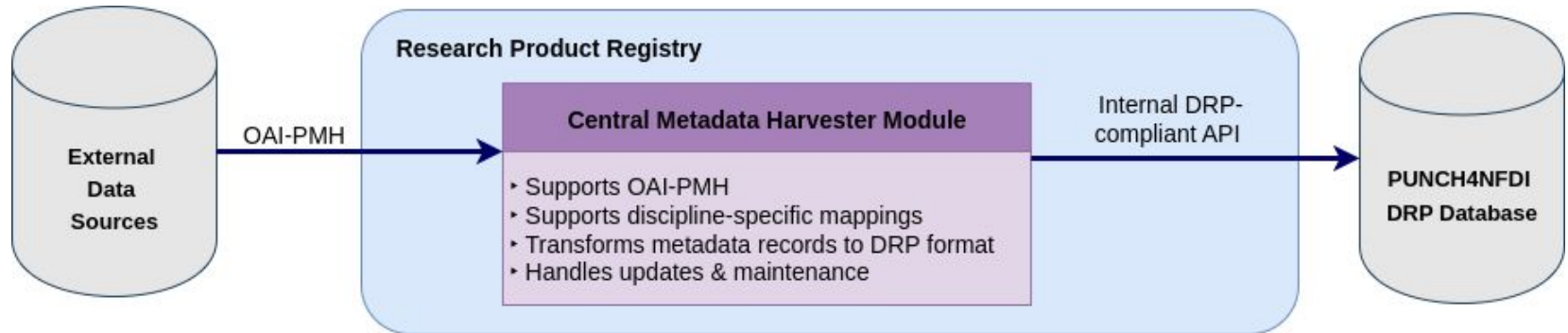
Authors:

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

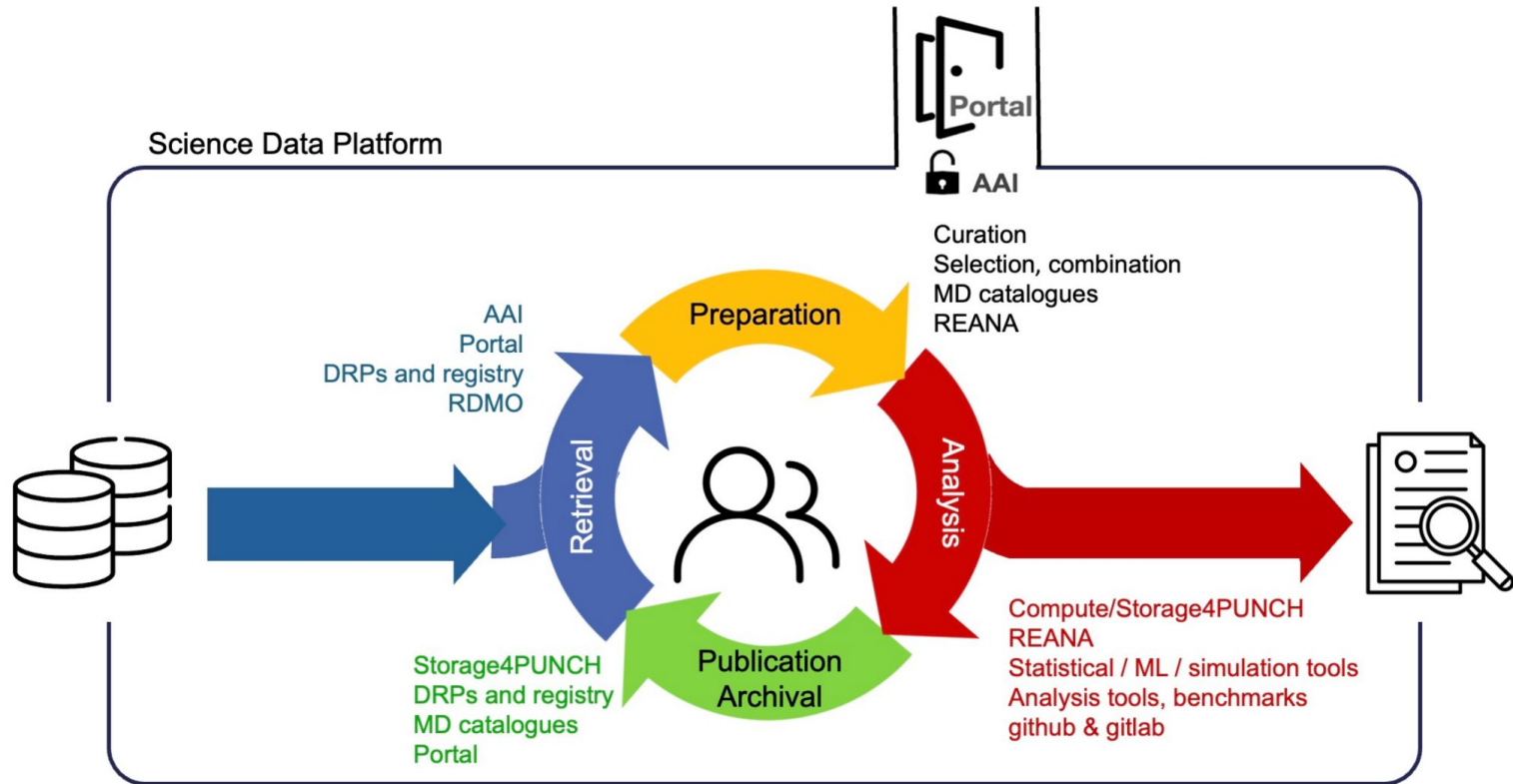


Diagram by T. Schöner

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

Mandatory properties of the metadata kernel



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