Metadata curation use cases in astroparticle physics*

Victoria Tokareva** for PUNCH4NFDI Consortium

PUNCH4NFDI in-person annual meeting & 4th PUNCH4NFDI General Meeting | Göttingen | 28–29 Sep 2022

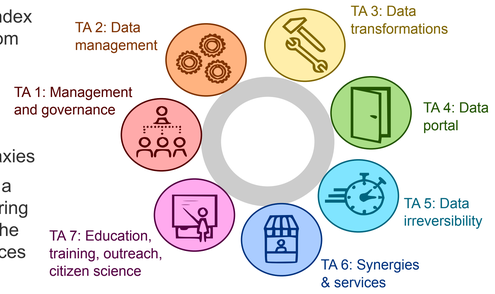
Demanding requirements of fundamental physics at large-scale facilities are forcing researchers to use and further develop sophisticated computer science for high-efficient data processing, analysis, curation and preservation. PUNCH4NFDI (Particles, Universe, NuClei and Hadrons for the NFDI) is a consortium of particle, astroparticle, astro-, hadron, and nuclear physics, looking forward to developing advanced techniques and concepts for scientific big data. An important part of these developments represents in-depth studies of best practices of big data access and transfer, as well as adaptation of effective metadata curation strategies. Prerequisites for development of a user-level metadata schema include a deep knowledge of all the peculiarities of the heterogeneous data supplied to the system from various distributed data sources, as well as a comprehension of the relevant user experiences and the necessary system functionality.

Moreover, there is a significant variety in the practices of working with data and research conduction within the consortium. In this regard, study of user scenarios within individual research groups is of particular importance. In this contribution, a comparative analysis of two metadata curation use cases from the PUNCH4NFDI consortium will be presented. We will consider the experience of two projects in the field of astroparticle physics: KASCADE Cosmic-ray Data Centre (KCDC) and German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI), in the context of the aims and requested functionality, chosen data architectures, technical solutions and, especially, metadata management approaches.

Particles, Universe, NuClei and Hadrons for Nationale Forschungs-Daten Infrastruktur (PUNCH4NFDI)

PUNCH4NFDI is the NFDI consortium of particle, astro-, astroparticle, hadron and nuclear physics.

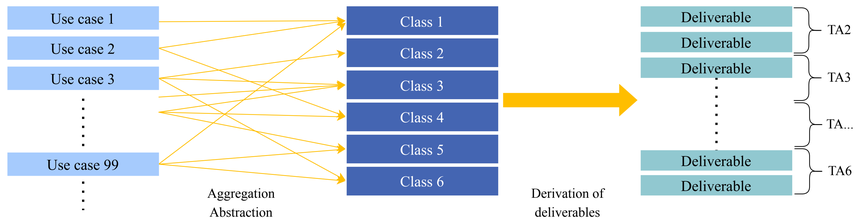
- The **objective of the NFDI** is to systematically index and make available the valuable stock of data from science and research
- **PUNCH physics** addresses the fundamental constituents of matter and their interactions, as well as their role for the development of the largest structures in the universe - stars and galaxies
- The prime **goal of PUNCH4NFDI** is the setup of a federated and "FAIR" science data platform, offering the infrastructures and interfaces necessary for the access to and use of data and computing resources of the involved communities and beyond



Use case studies for NFDI

Currently goes in 6 classes:

- Validating and publishing scientific data collections
- Analysis of local or distributed data sets
- Execution of analysis of numerical simulations
- Community-overarching data challenges
- Real-time challenges & data irreversibility
- Use cases from external partners



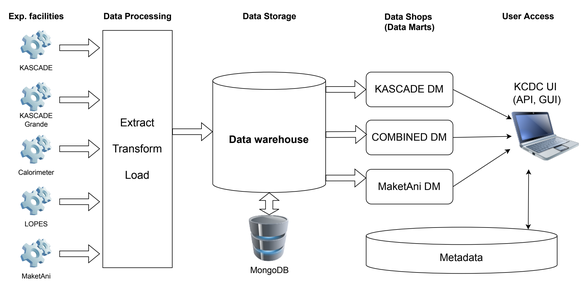
KCDC and GRADLCI data centers

KASCADE Cosmic Ray Data Centre (KCDC) is a public data center for high-energy astroparticle physics based on the data of the KASCADE experiment. Established in 2013, it incorporates KASCADE experiment data archive, information center and outreach platform. Available at: https://kcdc.iap.kit.edu

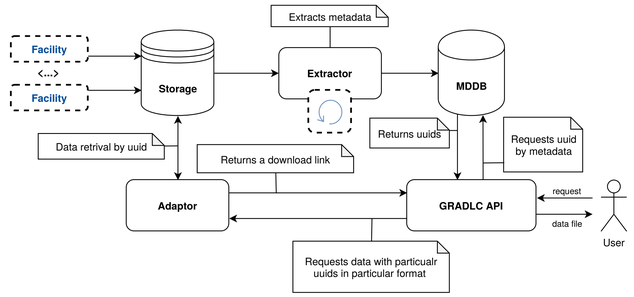
German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI) is an international project aimed into development of big data solutions to support experiments of astroparticle physics at every stage of their data life cycle. The project took place in 2018-2021.

Storage architectures

KCDC



GRADLCI

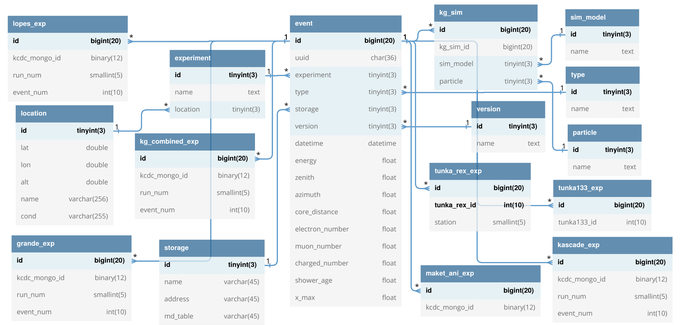


Metadata schemas

JSON metadata schema, example of a record from **KCDC**

```
"model": "kaos_datashop.quantity",
"fields": {
  "quant_type": "num",
  "allow_cuts": true,
  "head_description": "<p class=dcInfoBoxHeaderDS>Zenith Angle Info</p>",
  "descr_type": "HTML",
  "composite_data_handler": "",
  "unit": "\\u00B0",
  "detector": [
    "",
    "grande"
  ],
  "quant_sub_type": "f64",
  "display_format": "default",
  "min_value": "0.0",
  "display_name": "Zenith Angle",
  "description": "<div>\r\n<span class=dcMathFunct>0&deg;</span> to <span class=dcMathFunct>60&deg;</
  span> where <span class=dcMathFunct>0.44deg;</span> corresponds to a vertical shower. The angular resolution is
  between <span class=dcMathFunct>0.44deg;</span> and <span class=dcMathFunct>0.1&deg;</span> depending
  on the energy.\r\n<br>\r\n<b>We recommend to use data only up to 42&deg;</b><br><br></span>\r\n <span
  class=dcInfoBoxReference> [details see  <b>KCDC-Manual</b>]</span> \r\n</div>\r\n",
  "name": "Ze",
  "max_value": "60.0",
  "selection_mode": "D",
  "order": 2,
  "descr_head_html": "<p class=dcInfoBoxHeaderDS>Zenith Angle Info"
}
},
```

Metadata data base schema for **GRADLCI**



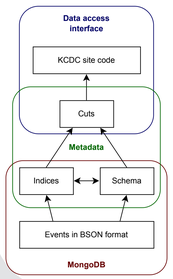
Data overview and acquisition

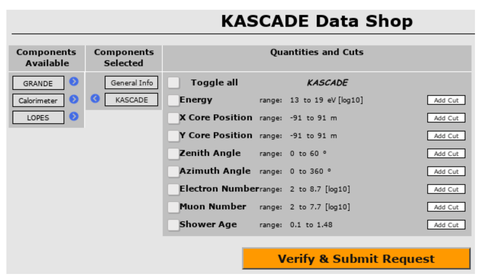
KCDC

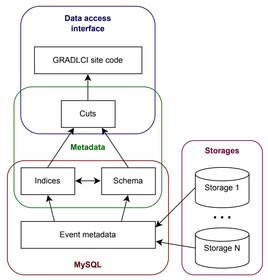
Setup / Detector component	Experimental data		Simulations	
	Events	Size	Events	Size
KASCADE	433 209 340	3 200 GB	22 490 883	26.8 GB
GRANDE	35 310 393	260 GB	4 149 416	4.2 GB
COMBINED	15 635 550	120 GB	2 030 227	2.6 GB
LOPES	3 058	25 MB	—	—
MAKET-ANI	2 682 264	1 GB	—	—

GRADLCI

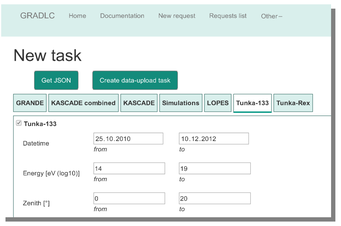
Setup / Detector component	Experimental data		Simulations	
	Events	Size	Events	Size
KCDC datasets	See KCDC table			
Tunka-133	7 421 630	0.5 GB	—	—
Tunka-Rex	107 360 524	3 TB	—	—
TAIGA-IACT	2 700 000 000	605 GB	—	—







Main idea: provide a common data selection interface for various data sources



Comparative analysis of the usecases

Data centers	Characteristics				
	Aim	Task areas / Functions	Datasets	Architectures	Technologies
KCDC	Provision of the free, unlimited, reliable open access to the data of various experiments measuring cosmic radiation by different methods and techniques both for scientists and the broad public	○ Data archive ○ Data analysis platform ○ Information center ○ Outreach platform	KASCADE, KASCADE-GRANDE, COMBINED, Lopes, MaketAni	Data marts	NoSQL (MongoDB) Django Celery, RabbitMQ, Docker/ Syngularity, REST API
GRADLCI	Development of the automatisation the maintenance of astroparticle-physics data throughout their entire life cycle	○ KCDC extension ○ Prototype analysis and data center for multimessenger astronomy ○ Analysis platform for machine learning for astroparticle physics ○ Outreach and education initiative	KASCADE, KASCADE-GRANDE, COMBINED, Lopes, MaketAni, Tunka-133, Tunka-GRANDE, Tunka-IACT (restricted), Tunka-Rex	Data virtualisation platform	File-based, SQL for metadata DB, Flask, Custom task queueing, Docker/ Syngularity, JSON-RPC

*This work was partially supported by DFG fund “NFDI 39/1” for the PUNCH4NFDI consortium.

** Karlsruhe Institute für Technologie, Institut für Astroteilchenphysik (IAP)
mailto: victoria.tokareva@kit.edu

https://www.iap.kit.edu