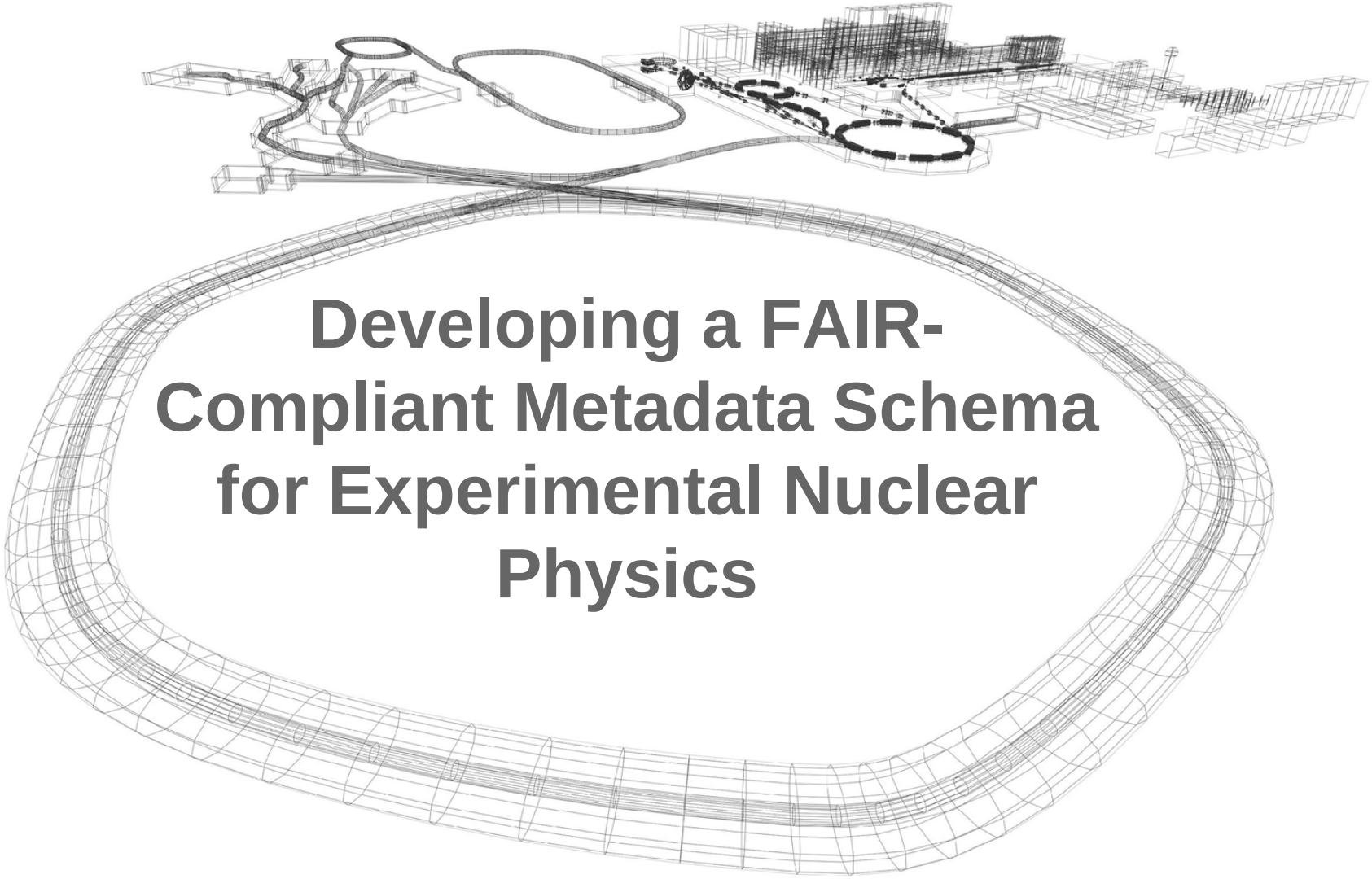
October 19 - 25, 2024
CHEP 2024
Conference on Computing in High Energy and Nuclear Physics



Developing a FAIR-Compliant Metadata Schema for Experimental Nuclear Physics

Ivan Knežević, Dr Andrew K. Mistry GSI Helmholtzzentrum für Schwerionenforschung

GSI Helmholtz Centre for Heavy-ion Research

Heavy-ion accelerator laboratory in Darmstadt, Germany with 1350+ employees & 700 external users

Who we are?

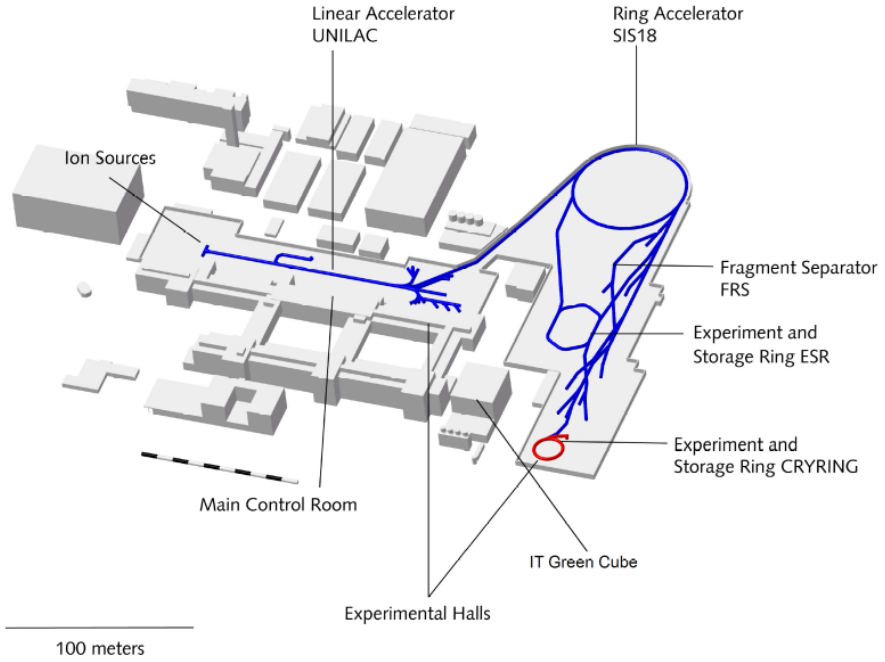
- Founded in 1969
- Elements H to U can be accelerated

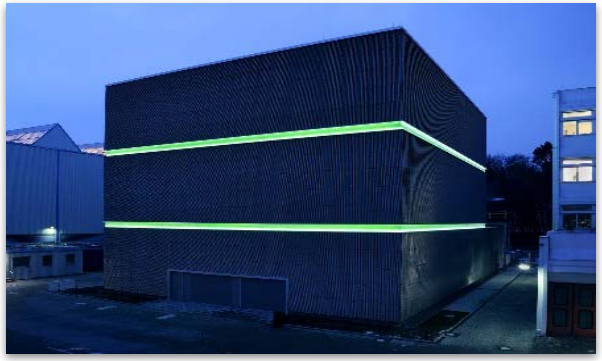
Past Achievements

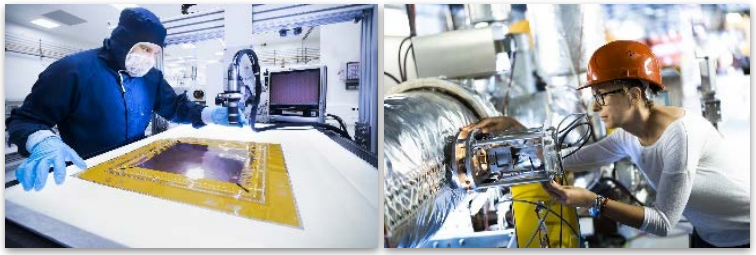
- Discovering **six new elements, many exotic nuclei**
- Developing a **new cancer treatment**

Our future

- First stage injector for the **FAIR facility**
- Experiments continuing (FAIR Phase-0)







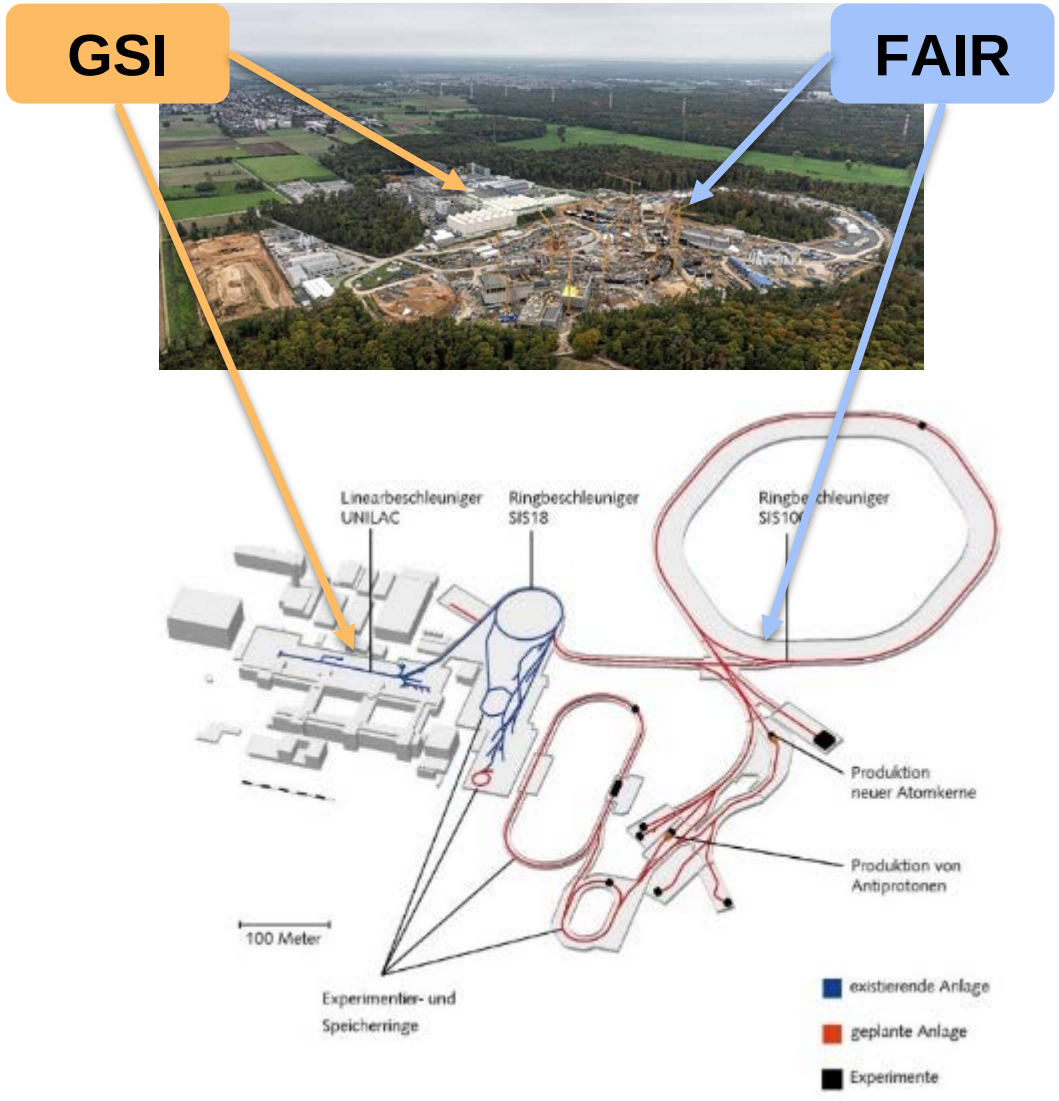
Research Interests

- Accelerator Physics
- Detector Development
- Atomic Physics
- Nuclear and particle Physics
- Plasma Physics
- Biophysics
- Materials Research
- High-Performance Computing
- Theoretical Physics

FAIR (Facility for Antiproton and Ion Research in Europe)

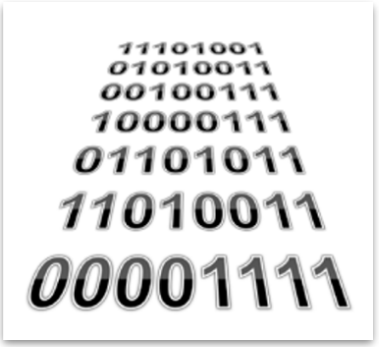
New accelerator facility

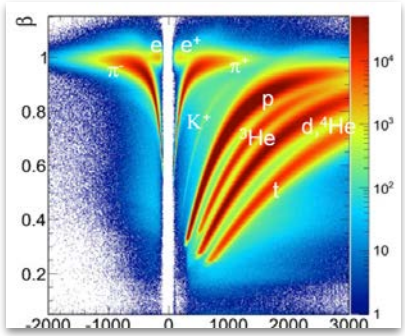
- **Top priority** for European Nuclear Physics Community
- **International**: 50 countries, 3000 researchers
- **Diverse community** from atomic to particle physics
- **'FAIR goes F.A.I.R.': commitment to open science**
- Towards the next generation "data challenge" Volume, Velocity, Veracity, Variety, and Complexity
- **~TB/s data rates**, online processing, ~5x105 cores
- Data stored to **disk 40+ PB/year**
- **Distributed computing** with a large user community
- **Data preservation and accessibility key to success**

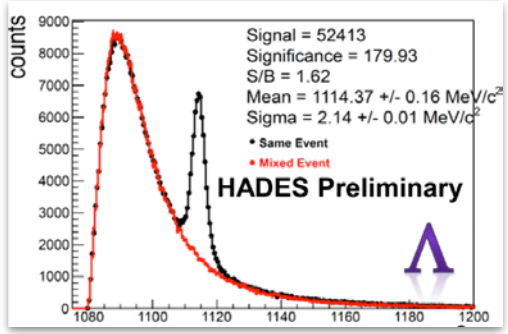


Research Data and Software at GSI: Rich and varied

Raw data → Pre-processed data → Results







Research Area	Exp. run time	Raw data	Simulations	Final datasize
Materials science	~Minutes	~MB-GB	-	10MB
HADES	3 weeks	130TB	150TB	<1TB
FAIR CBM	2 months	18PB	9PB	4PB

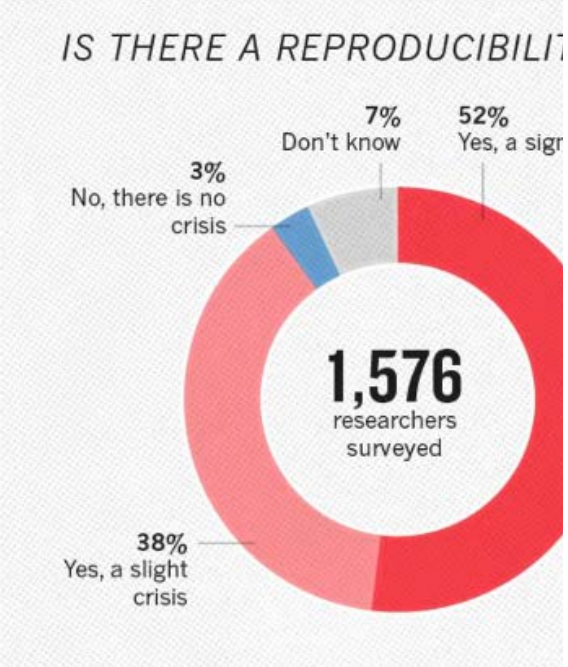
F.A.I.R. data principles and metadata

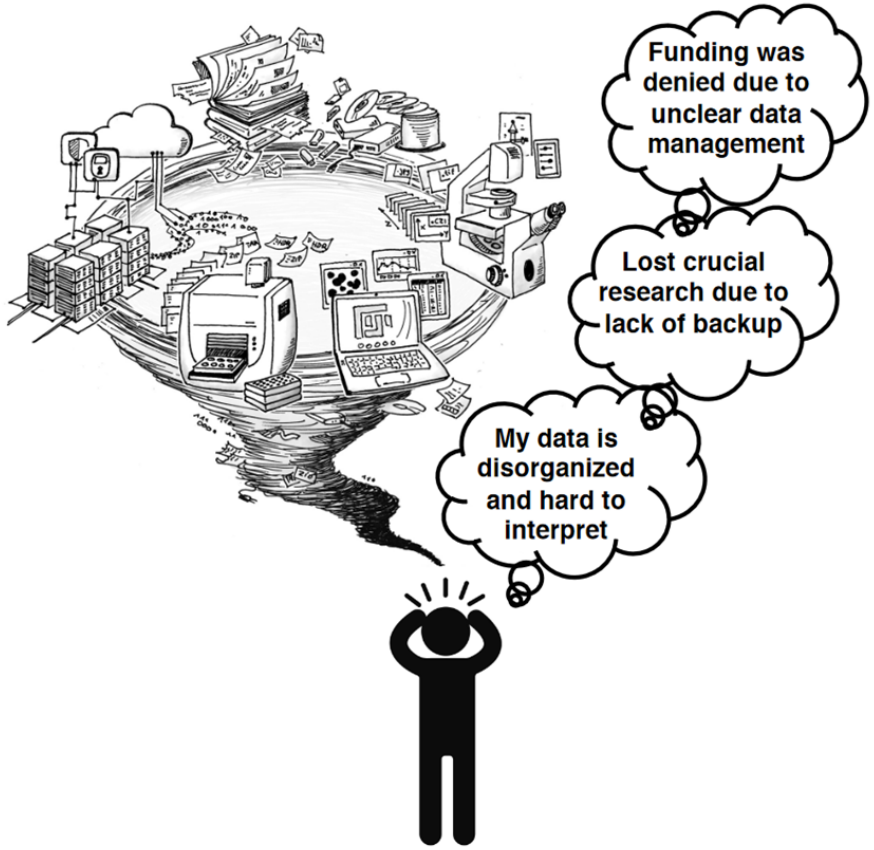
Metadata is essential in Research Data Management and Open Science to enable FAIR data (and code)

- Allows datasets to be searched for and found
- Enables interoperability between datasets
- Facilitates data reprocessing
- Promotes efficient use of resources

Underlying problems in Data Management

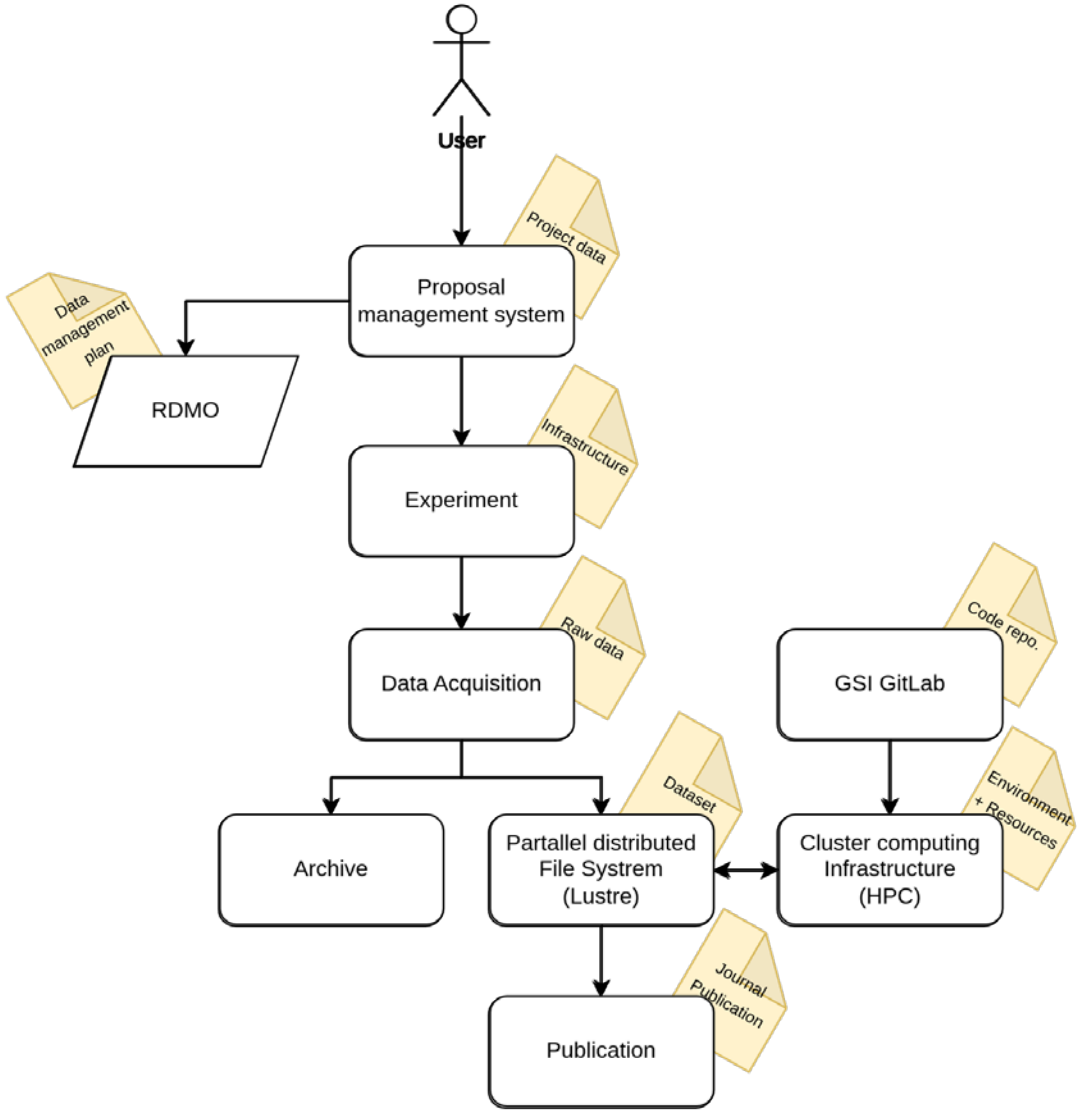
- Growing data volumes
- Unstructured data
- Reproducibility crisis
- Lack of user awareness in RDM practices





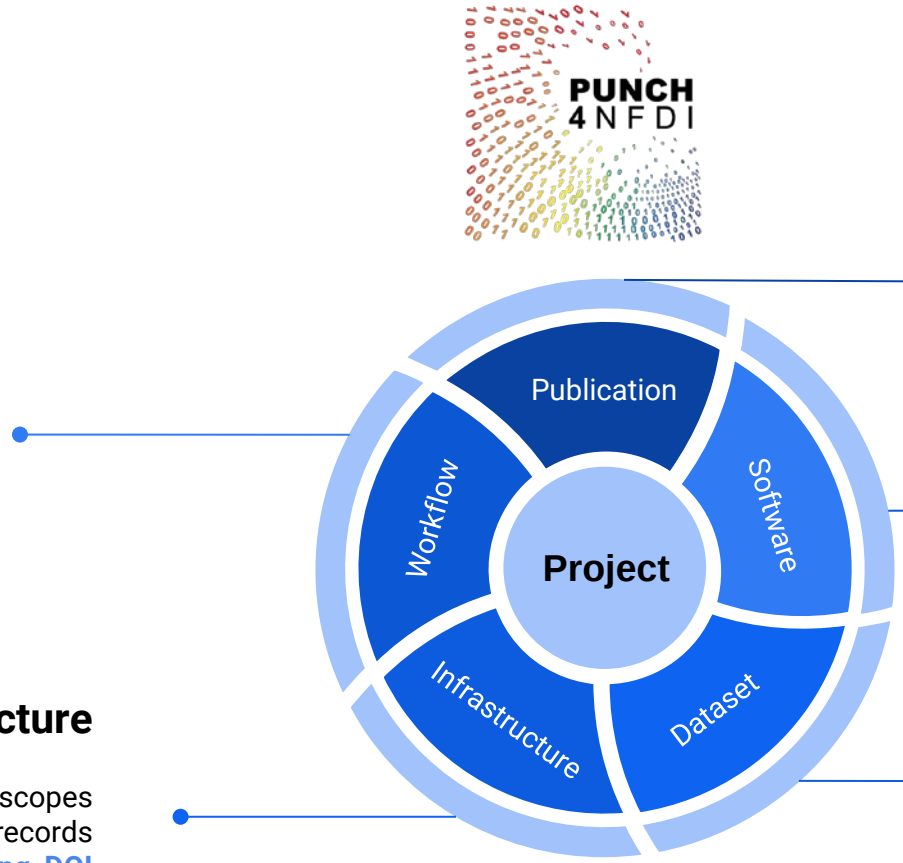
Standard Workflow in Experimental Nuclear Physics

- Workflow is representative of experimental procedures at GSI and similar facilities.
- Each stage generates specific metadata crucial for data management.
- Our metadata stratification and categorization is grounded in this workflow.
- Definition of **D**igital **R**esearch **P**roduct (PUNCH4NFDI) - https://www.punch4nfdi.de/



Digital Research Product

“All digital outputs of the research process have to be referenceable”



Publication

arXiv, journals, ADS, open access
DOI, ISBN

Workflow

REANA, Compute Resources, Containers
ENV, distors, ContainerID

Software

GitHub, GitLab, ROOT, CASA
Versioning DOI

Infrastructure

Accelerators, Detectors, Radio telescopes instrument records
Versioning, DOI

Dataset

Zenodo, Public and Collaboration data repositories
DOI, PID

Project Overview

Objectives:

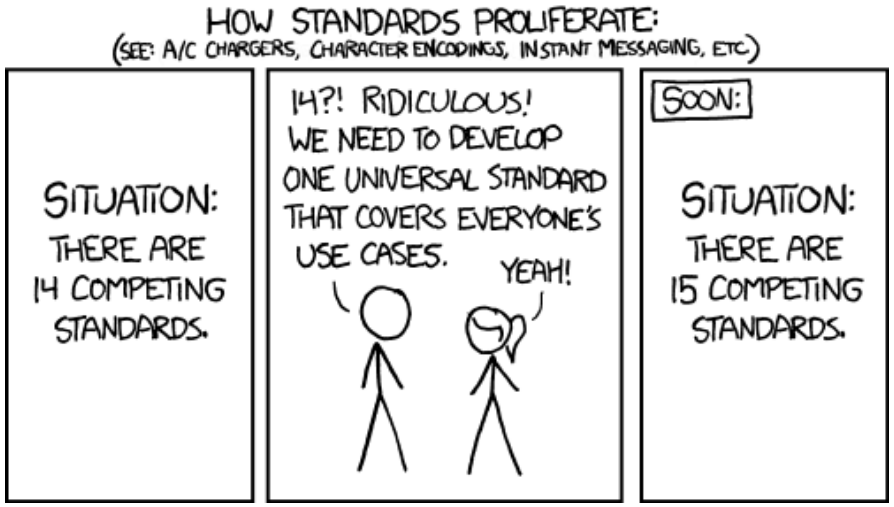
- Develop a standardized, adaptable metadata schema.
- Facilitate data input, referencing, management and publication.

Addressing problems:

- Growing unstructured data,
- Diverse data formats and nomenclatures,
- Difficulty in data sharing across institutions,
- No common schema yet between nuclear physics experiments.

NAPMIX

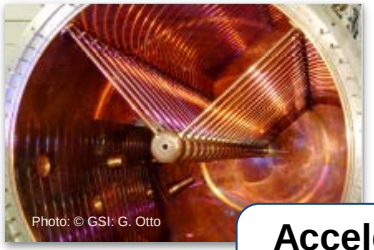
Nuclear, Astro, and Particle Metadata Integration for eXperiments

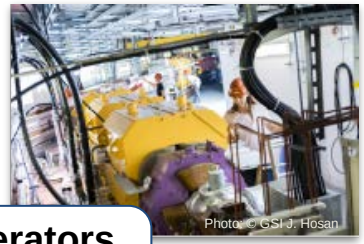


By Randall Munroe: https://xkcd.com/927

Example: GSI accelerator facility

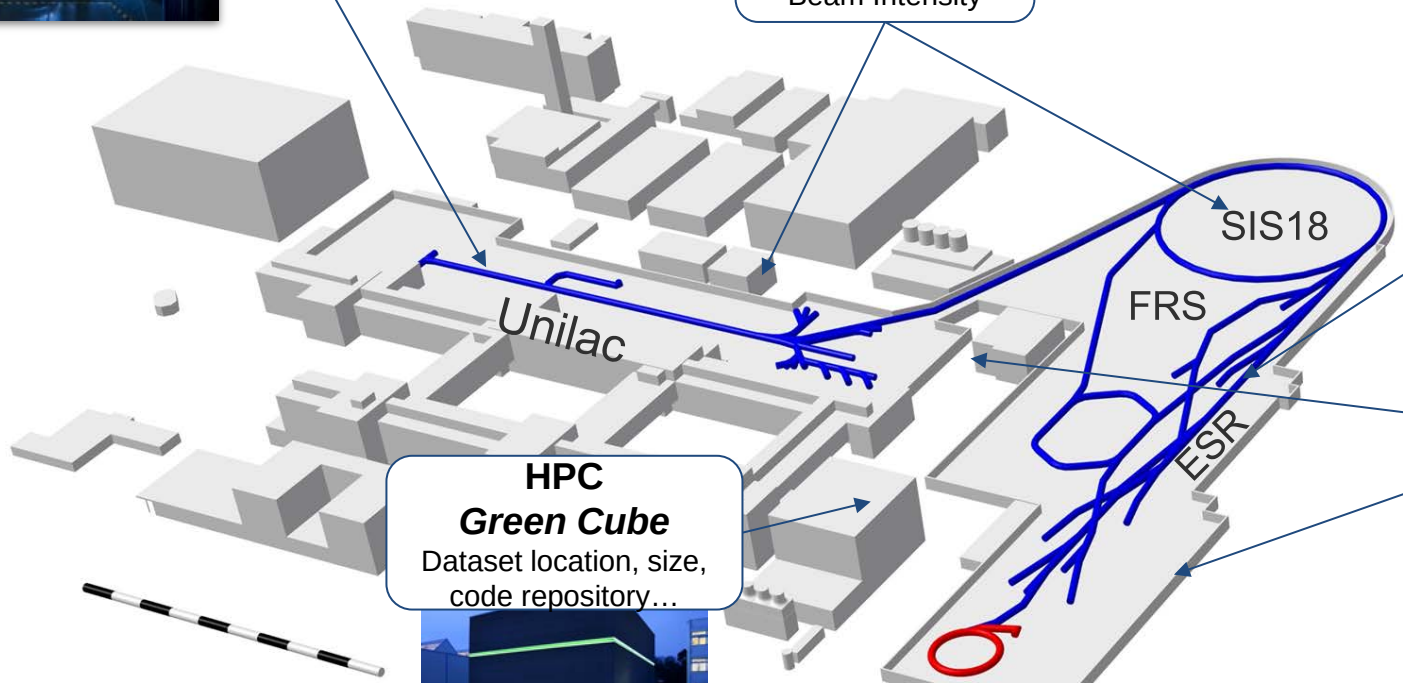
Ion sources
Type,
Projectile Ion…



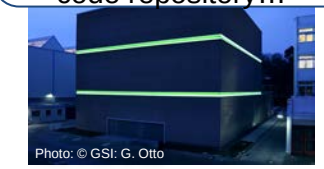


Accelerators
Beam Energy
Beam Intensity

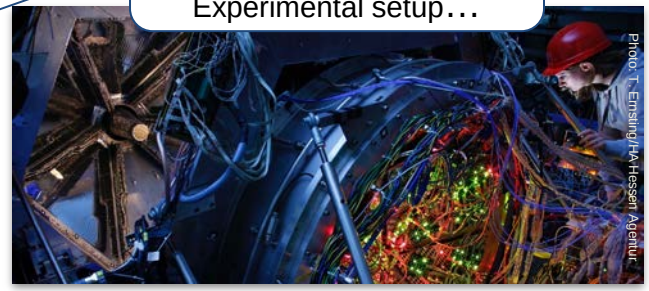
Separators
Type, Reaction…



HPC
Green Cube
Dataset location, size,
code repository…

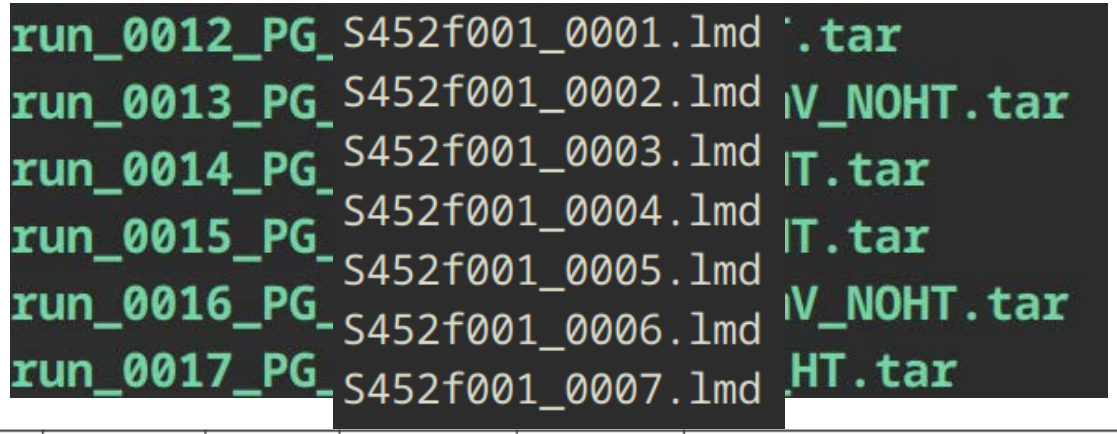


Experimental areas
Produced Isotope
Experimental setup…



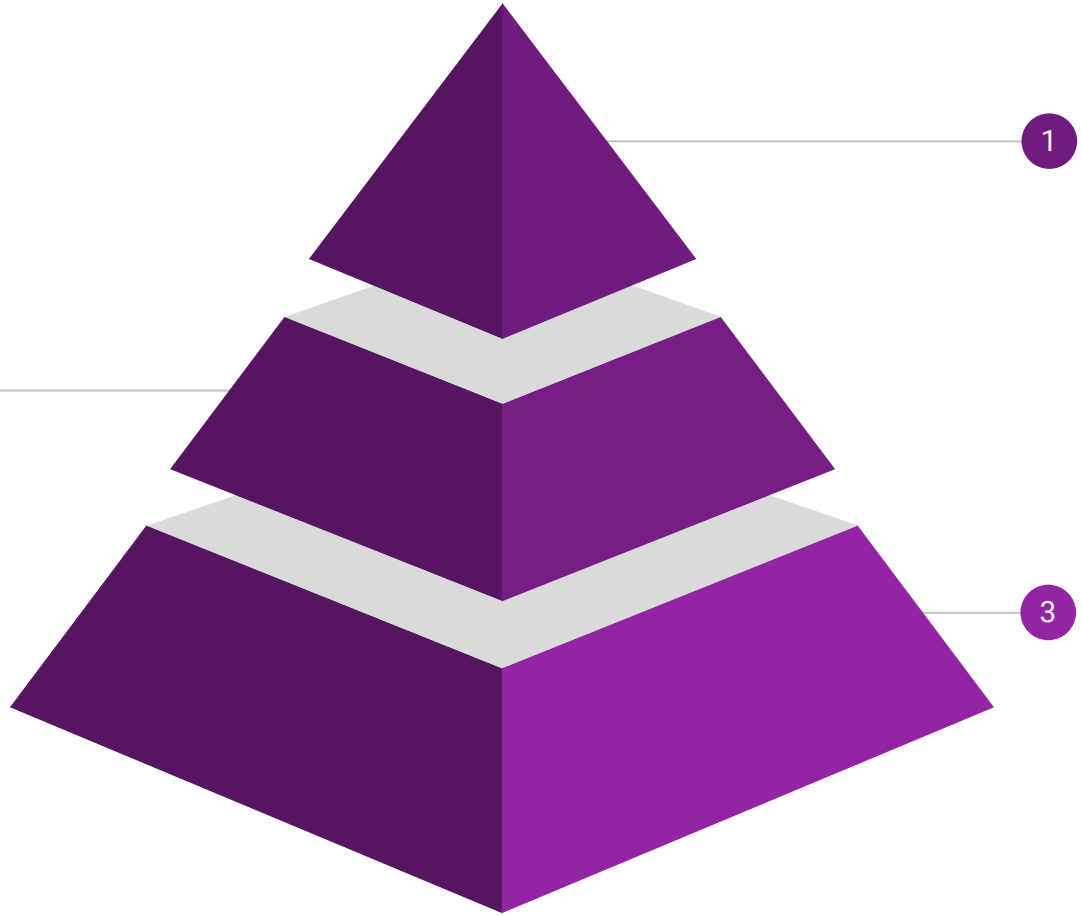
Importance of Standardized Metadata

- Issues with non-standardized metadata,
- Impact on data discovery and utilization,
- Standardization facilitates automation and machine readability,
- Facilitates data sharing and collaboration
- Enhances data quality and integrity



File	Final file	Start	Stop	Time	Slits (mm)	Settings	Comments
S452f001	_0114	03.06 04:24	06:45	3:21	+/- 30 @ s4	206Hg	206Hg fragment in S4 at 04:20
S452f004	_0025	03.06 08:50	09:26	0:36	+/- 30 @ s4	206Hg	Checks on thresholds in Ge
S452f005	_0017	03.06 09:48	10:07	0:19	+/- 30 @ s4	206Hg	Changes to thresholds (200mV→100mV) in Ge
S452f006	_0027	03.06 10:10	12:10	2:00	+/- 30 @ s4	206Hg	The angle of the S2 wedge was adjusted for achromatic beam.

Metadata stratification



Project specific metadata

1

Project name, publication date, collaborations, researchers, institutions, journal references

Discipline specific metadata

2

Projectile data, target data, infrastructure, detector type, datasets, software repositories

Experiment specific metadata

3

Reaction data, gamma-ray energies, time of flight parameters, instrument records

Metadata stratification

Metadata level	Findable	Accessible	Interoperable	Reusable
Project Specific: Standard project information, names, dates, institution…	✓	✓	✗	✗
Discipline Specific: Projectile, target, detector type, datasets…	✓	✓	—	✓
Experiment Specific: Gamma-ray energies, Time of flight parameters…	✗	✓	✓	✓

The Metadata Schema Design

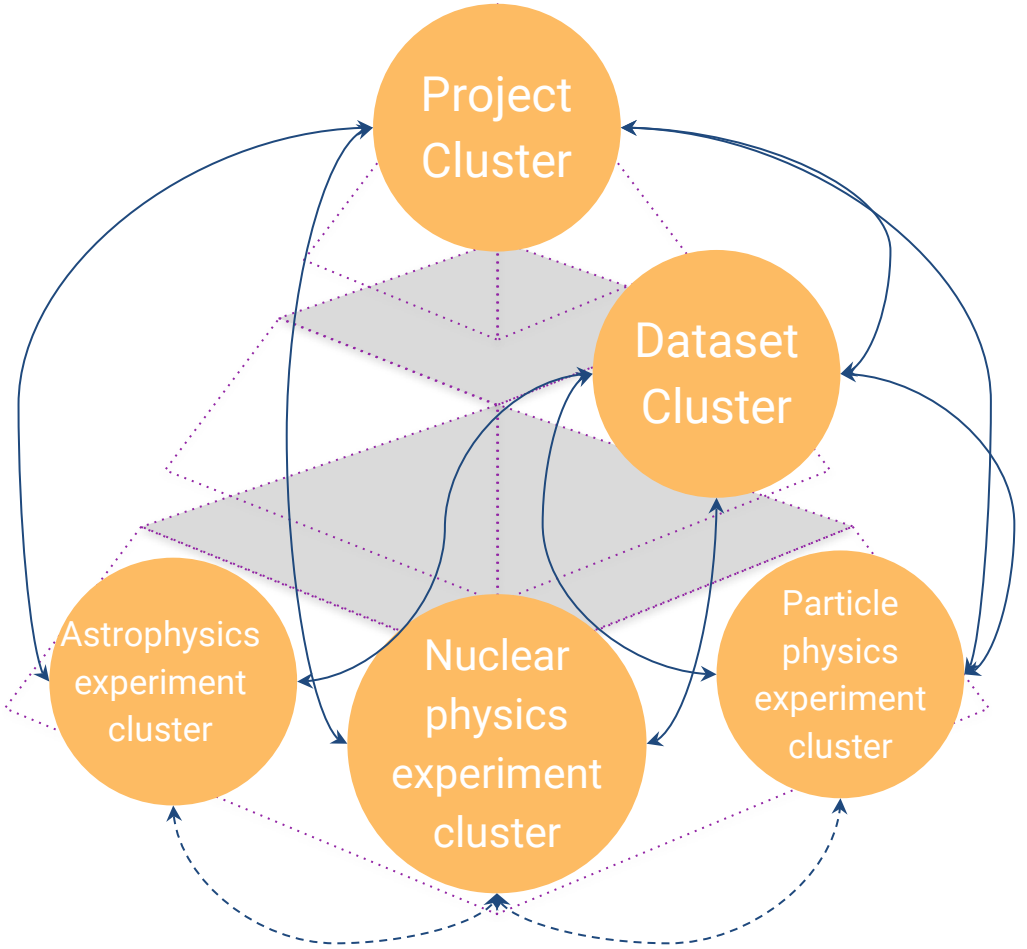
Nodal, Multi-Layered Structure:

- Keeps data in common on top
- Clusters separated but interconnected

Metadata Enrichment:

- New tables or fields can be added to the Experiment cluster to capture additional nuclear physics metadata,

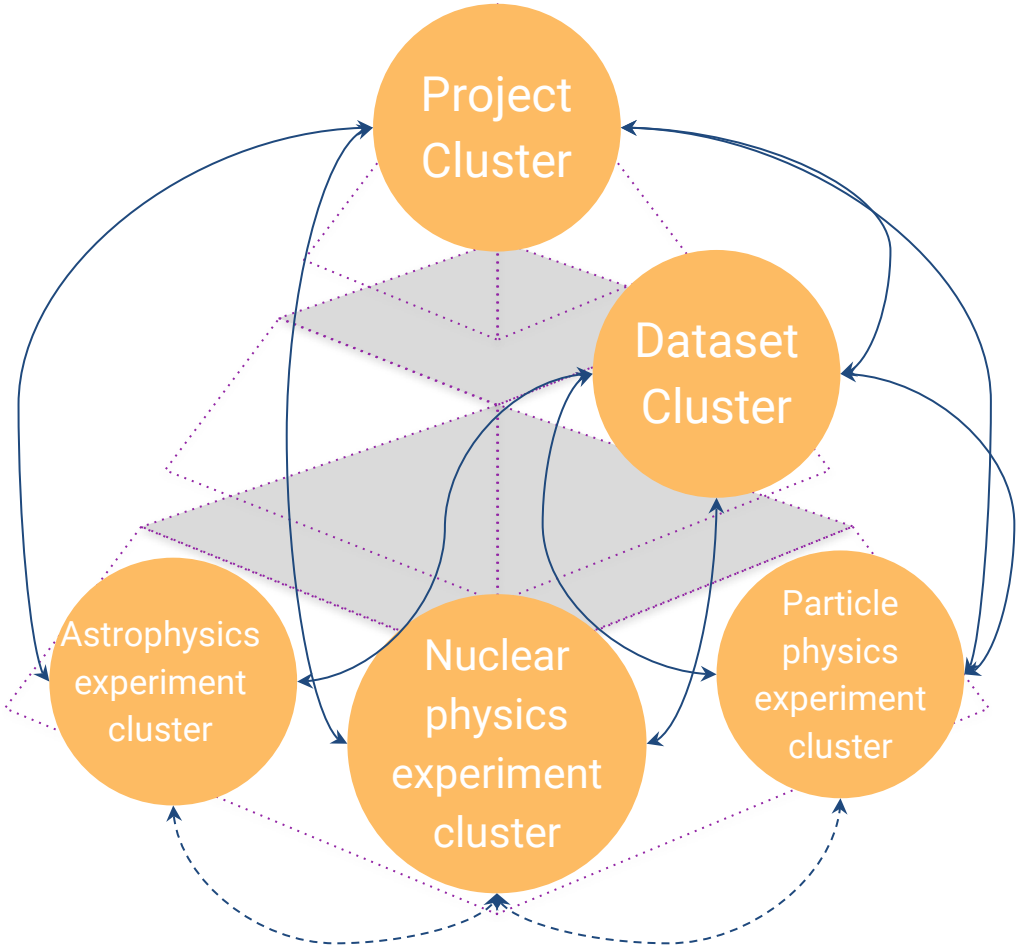
- Other domains can create their own specific clusters while sharing the common Project and Dataset layers.



The Metadata Schema Design

Nodal, Multi-Layered Structure:

- Keeps data in common on top
- Clusters separated but interconnected



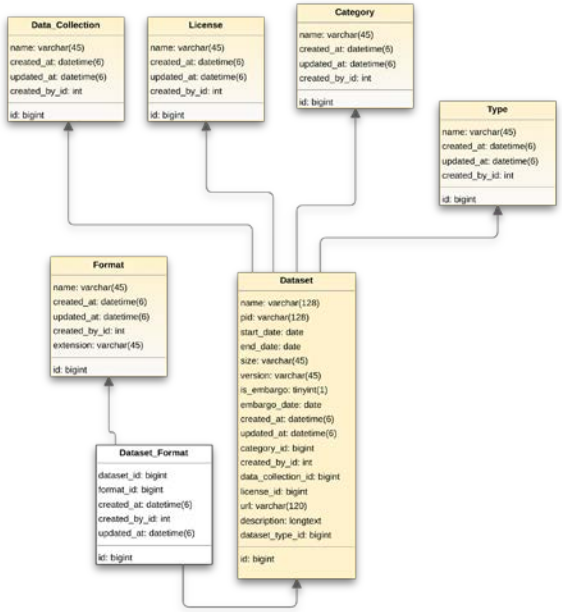
Benefits:

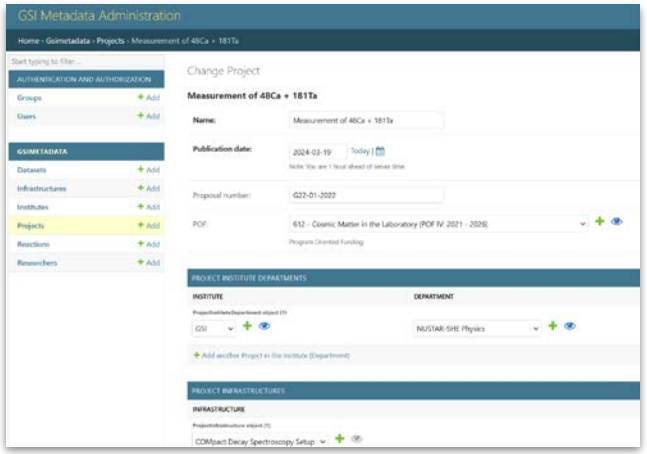
- **Customization**: Researchers can capture all necessary metadata specific to their experiments,
- **Findability**: Clearly defined relations between the clusters makes search easier
- **Interoperability**: The common layers ensure that data remains interoperable with other systems and disciplines,
- **Collaboration**: Facilitates collaboration by allowing different domains to understand and use each other's data.

Implementation process

Concept → Design → Implementation

Model	Cardinality	Metadata	Example Filled Field	Descriptor	Field type
		General Information	Please fill this column for an example dataset		
Project	1	Project Name	Measurement of 48Ca + 181Ta	Name of Project	String
Project	1	Publication Date	08/08/2023	Date of Data Publication	Date
Project	0/1	Proposal Number/Project ID	G22-01-2022	Official experiment number obtained assigned at the proposal s	String
Project	0+	POF Topic	Cosmic Matter in the Laboratory	Helmholtz Program-Oriented Funding	String
Project	0+	Journal references	Journal Ref Name: PID	Links to publications dervied from data	String
Project	0+	Software references	Software Ref Name:PID:Repo URL	Links to Software used with data	String
Project	1+	Facility/Institute of Data Generation	GSI	Of data generation	String
Project	0+	Infrastructure	SHIP, COMPASS	Infrastruture using in data generation	String
Project	1+	Reactions	48 Ca + 181Ta	Links to Reactions	String
Project	0+	Collaboration link to project	SHIP Decay Spectroscopy		String
Project		Created by		Project record creation author	Key
Project		Created at		Project record creation timestamp	Date
Project		Updated at		Project record update timestamp	Date
Department	1	Department/Division Name	NUSTAR-SHE Physics	Define Departments	String
Department	0/1	Department/Division Short name	SHE-Physics	Abbr. Department name	String
Department		Department/Division created by			String
Department		Department/Division created at			Date
Department		Department/Division updated at			Date



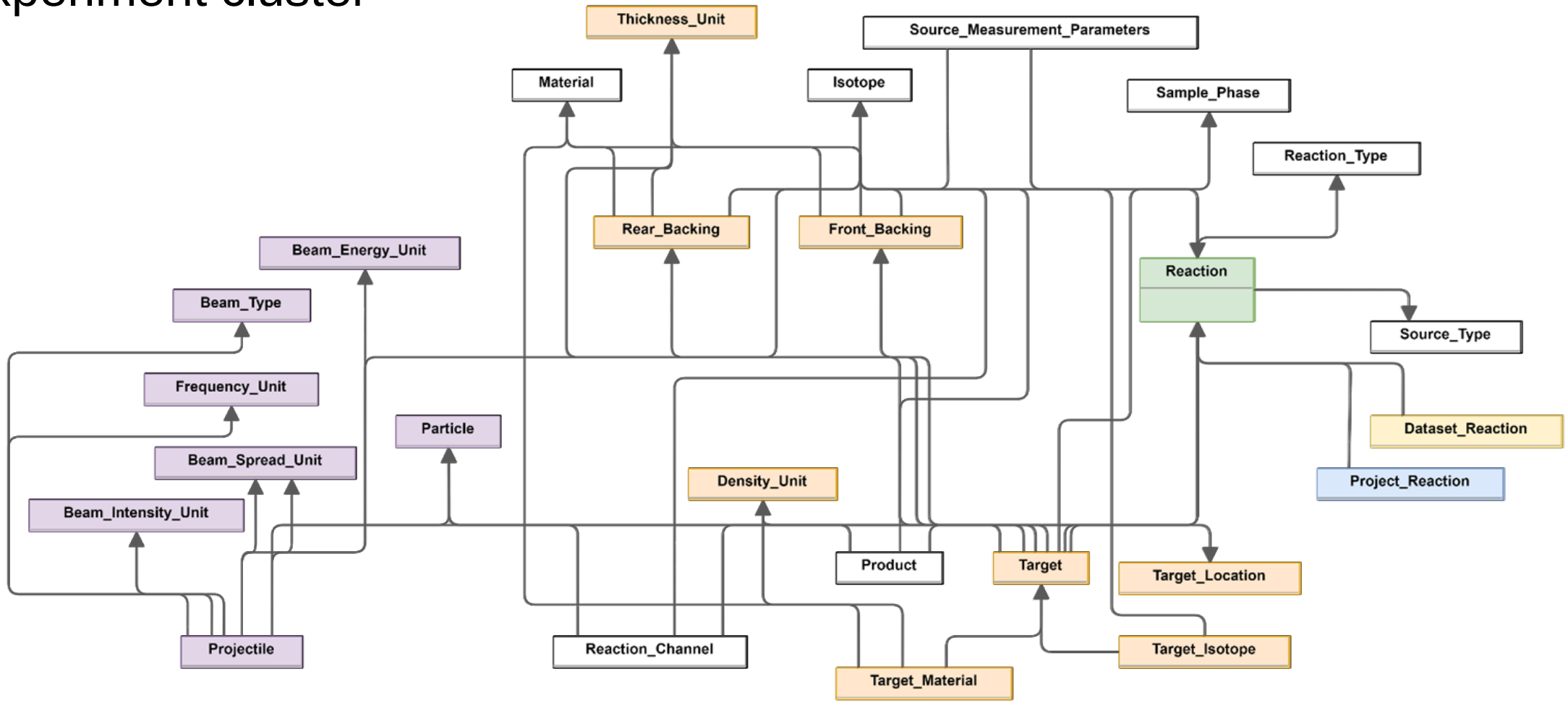


django

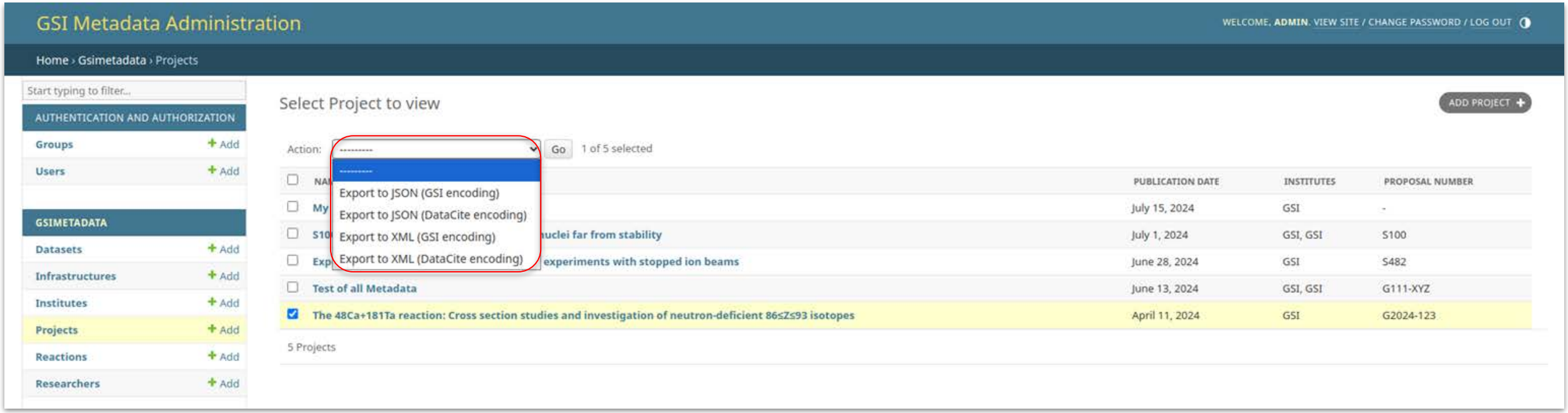
MySQL

The Database Schema Design

Experiment cluster



Metadata Schema Generator



- A tool that allows users to export comprehensive metadata for a selected project.
- Supports multiple formats and encoding standards.

Metadata Schema Generator

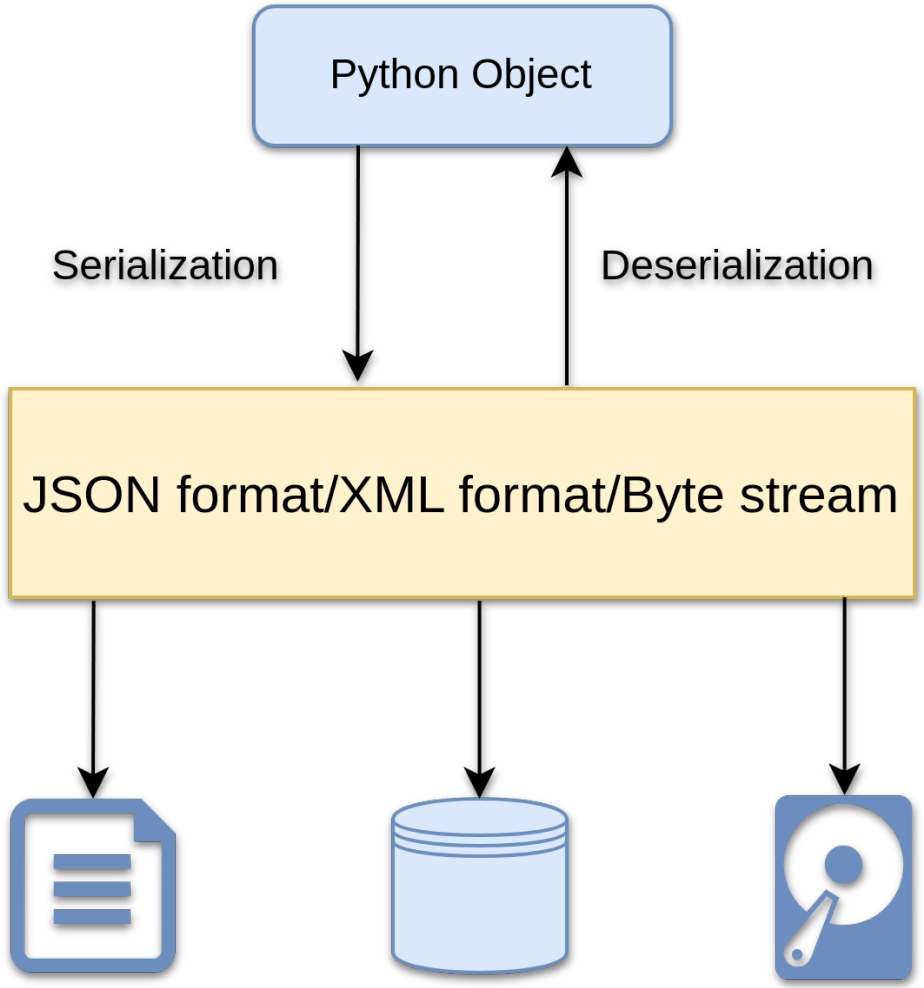
XML schema

```xml
<Reactions>
  <Reaction>
    <name>208Pb + 9Be</name>
    <reaction_type>Fragmentation</reaction_type>
    <source_type>Accelerator</source_type>
    <Projectiles>
      <Projectile>
        <Isotope>Lead (Pb-208)</Isotope>
        <Percentage>100.0</Percentage>
        <Beam_energy>1050.0</Beam_energy>
        <Beam_energy_unit>Megaelectronvolt/nucleon (MeV/u)</Beam_energy_unit>
        <Beam_operation>Pulsed</Beam_operation>
        <Average_beam_intensity>1500000000.0</Average_beam_intensity>
        <Beam_intensity_unit>Particles per spill (pps)</Beam_intensity_unit>
        <Beam_destination>HFS</Beam_destination>
        <Beam_type>Stable</Beam_type>
      </Projectile>
    </Projectiles>
```

JSON schema

```json
"reactions": [
    {
        "name": "208Pb + 9Be",
        "reaction_type": "Fragmentation",
        "source_type": "Accelerator",
        "projectiles": [
            {
                "isotope": "Lead (Pb-208)",
                "percentage": 100.0,
                "beamEnergy": 1050.0,
                "beamEnergyUnit": "Megaelectronvolt/nucleon (MeV/u)",
                "frequency_type": "Pulsed",
                "beamIntensity": 1500000000.0,
                "beamIntensityUnit": "Particles per spill (pps)",
                "beam_destination": "HFS",
                "beam_type": "Stable"
            }
```

Metadata Schema Generator - data serialization




```python
def serialize_projectile(projectile, exclude_fields=None, field_name_mapping=None):  1 usage  i.knezevic
    exclude_fields = ['reaction'] if exclude_fields is None else exclude_fields + ['reaction']
    projectile_data = {}

    # Prepare reverse mapping for special fields
    reverse_field_name_mapping = {v: k for k, v in (field_name_mapping or {}).items()}
    special_fields = {
        'isotope': reverse_field_name_mapping.get('isotope', 'isotope'),
        'beam_energy_unit': reverse_field_name_mapping.get('beam_energy_unit', 'beam_energy_unit'),
        'beam_intensity_unit': reverse_field_name_mapping.get('beam_intensity_unit', 'beam_intensity_unit'),
    }

    field_order = ['isotope', 'percentage', 'beam_energy', 'beam_energy_unit', 'beam_operation', 'frequency',
        'average_beam_intensity', 'beam_intensity_unit', 'charge_state', 'beam_destination', 'energy_spread',
        'beam_spread_unit', 'beam_spot_size', 'beam_spot_size_unit', 'beam_type']
```

- Serializing each smaller cluster separately,
- Increased schema flexibility,
- Relationships maintained

Metadata Schema Generator - Custom attribute names

Original attribute name

Desired attribute name

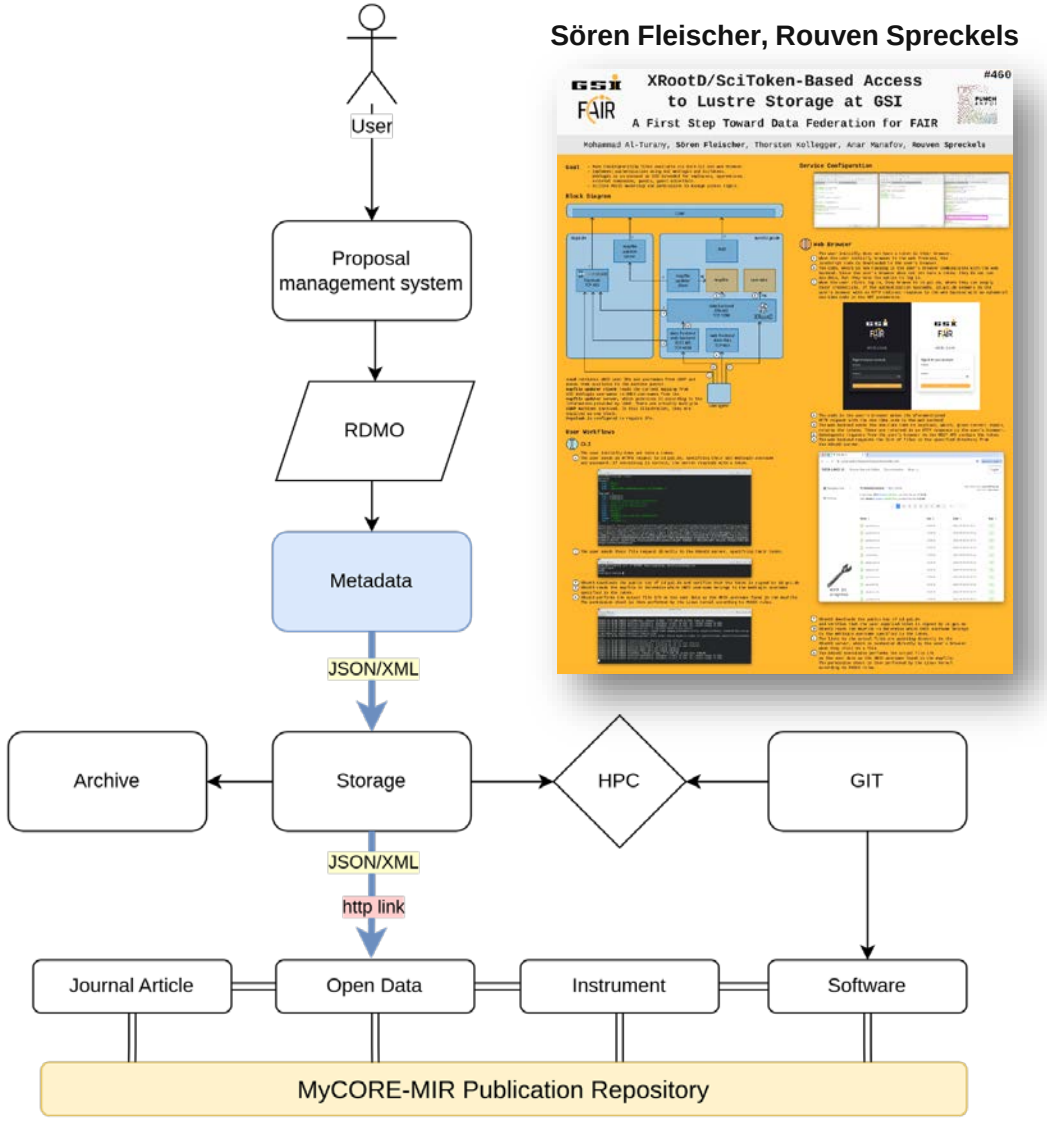
```
def export_to_json_datacite(self, request, queryset):  1 usage  i.knezevic
    if queryset.count() == 1:
        project = queryset.first()
        # Define your custom field mappings here
        custom_field_name_mapping = {
            'journal_article_pid': 'articlePID',
            'is_prime_investigator': 'PrimeInvestigator',
            'energy_spread': 'beamEnergySpread',
            'energy_spread_unit': 'beamEnergySpreadUnit',
            'beam_spot_size': 'beamSpotSize',
            'beam_spot_size_unit': 'beamSpotSizeUnit',
            'beam_energy': 'beamEnergy',
            'beam_energy_unit': 'beamEnergyUnit',
            'average_beam_intensity': 'beamIntensity',
```

- Attribute names can be customized via code (array mapping),
- Supports interoperability across platforms and institutions,
- Provides flexibility to meet specific metadata standards,
- Plans for making attribute customization more accessible through the UI.

Interoperability with Other Systems

- Designed to integrate with research data management systems,
- Outputs fully compliant with international metadata standards like DataCite,
- Can be used when depositing data in repositories like Zenodo,
- Planned support for standard data exchange protocols like OAI-PMH*,
  - Easier harvesting and synchronizing metadata between the systems.

*Open Archives Initiative Protocol for Metadata Harvesting



Sören Fleischer, Rouven Spreckels

Collaboration and Synergies

Involved Institutions:

TÉCNICO LISBOA
CLEAR
CMAM
GSI
GANIL
DESY.
TECHNISCHE UNIVERSITÄT DARMSTADT
FAIR
HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF
CNA Centro Nacional de Aceleradores
<HMC>
HUN REN
atomki DEBRECEN
KIT Karlsruher Institut für Technologie
NAPMIX

Open Science Initiatives:

INSPIRE HEP
ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures
EURO-LABS EUROPEAN LABORATORIES FOR ACCELERATOR BASED SCIENCES
PUNCH4NFDI
eosc

Funded by:

OSCARS Open Science Clusters' Action for Research & Society

Future Outlook

- Frontend development
  - User Interface,
  - Advanced search filters,
- Additional features
  - Automatic metadata import,
  - Attribute customization,
  - Collaboration mechanism,
- Integration with other systems
  - Integration within the PUNCH4NFDI ecosystem,
  - Support for protocol OAI-PMH*.
- Community expansion
  - The service will also be onboarded to the ESCAPE Open Source Software Repository (OSSR),
  - Inclusion to the in the EOSC marketplace,
  - Full accessibility to the wider scientific community.

*Open Archives Initiative Protocol for Metadata Harvesting

Questions?

Contact us

- open-science@gsi.de