Metadata curation efforts at KASCADE Cosmic-Ray Data Centre

V. Tokareva, A. Haungs, D. Wochele, J. Wochele, D. Kang

DPG Spring Meetings 2024

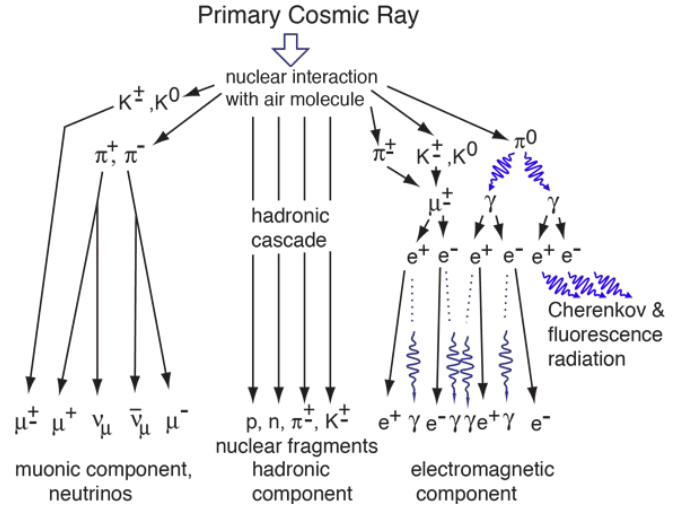
4 March 2024, Karlsruhe

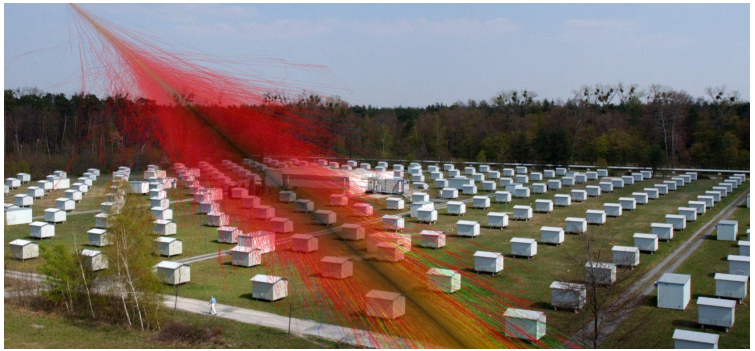
KCDC - KASCADE Cosmic Ray Data Centre

- KCDC is the public data centre for high-energy astroparticle physics
- Based on the data of the KASCADE experiment, contains as well data by KASCADE-Grande, LOPES, Maket-Ani, allows further extensions
- More than 433.000.000 events
- Established in 2013
- https://kcdc.iap.kit.edu/

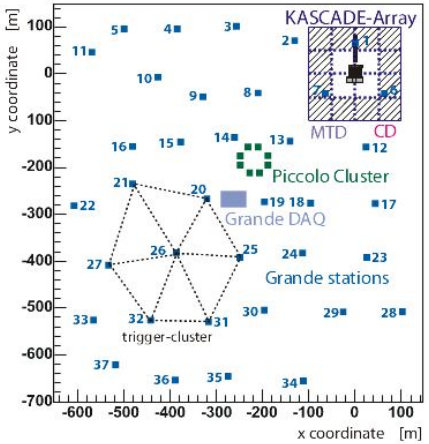
KASCADE - KArlsruhe Shower Core and Array DEtector

- Location: 110 m a.s.l., 49◦ N, 8◦ E, KIT-Campus North, Karlsruhe, Germany
- Operation time: 1996 October – 2010 May ⇒ e/γ detector liquid scintillator effective time ~ 4223.6 days
- Area: 200 × 200 m^2,
- E = 100 TeV – 80 PeV
- 252 scintillator detectors





More open datasets at KCDC



GRANDE (KArlsruhe Shower Core and Array DEtector-Grande) is an extension of the KASCADE experiment. By this the energy range KASCADE was extended to 10^{14}–10^{18} eV. 35 310 393 events are available

LOPES (LOfar PrototypE Station) is an experiment, which measures the radio emission of cosmic ray air showers in the frequency range from 40 to 80 MHz. 3 058 events

COMBINED is a combined dataset, made of data by 'KASCADE' and 'GRANDE' detector components for multi-messenger analysis. 15 635 550 events

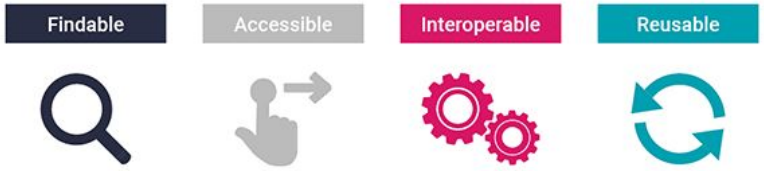
MAKET-ANI is an extensive air shower experiment placed on Mt. Aragats (Aragats Cosmic Ray Observatory, Armenia). 2 682 264 events

Functionality

- Provides free, unlimited, reliable open access to datasets in high-energy astroparticle physics
- Serves as information platform: physics and experiment backgrounds, tutorials, reference information

Towards FAIR data

What is FAIR?



Definitions

- Data is *potential* information
- Metadata record is a container for data about an object

Towards FAIR data

What is FAIR?

Why FAIR?

- Enable the discovery and reuse of information by humans and machines
- It allows knowledge to be derived from this information and applied across domains
- Collaboration opportunities

Definitions

- Data is *potential* information
- Metadata record is a container for data about an object

Towards FAIR data

What is FAIR?

Why FAIR?

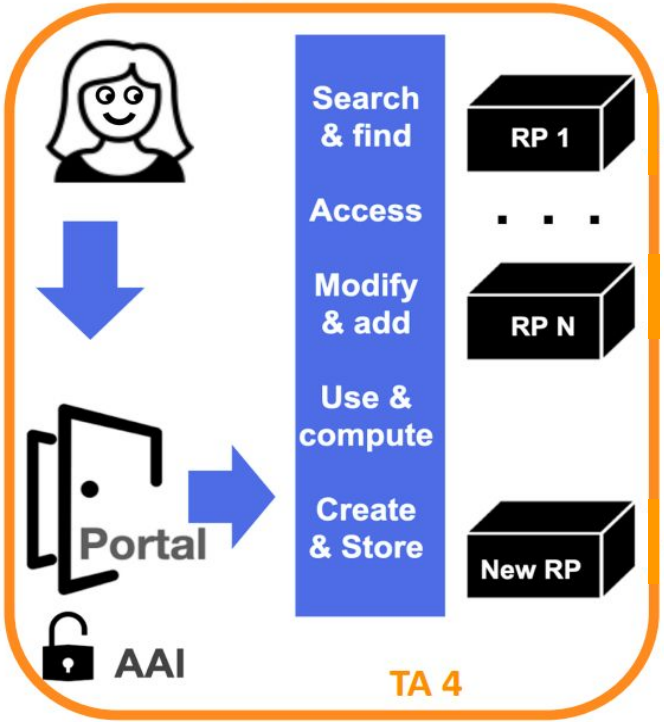
- Enable the discovery and reuse of information by humans and machines
- It allows knowledge to be derived from this information and applied across domains
- Collaboration opportunities

Definitions

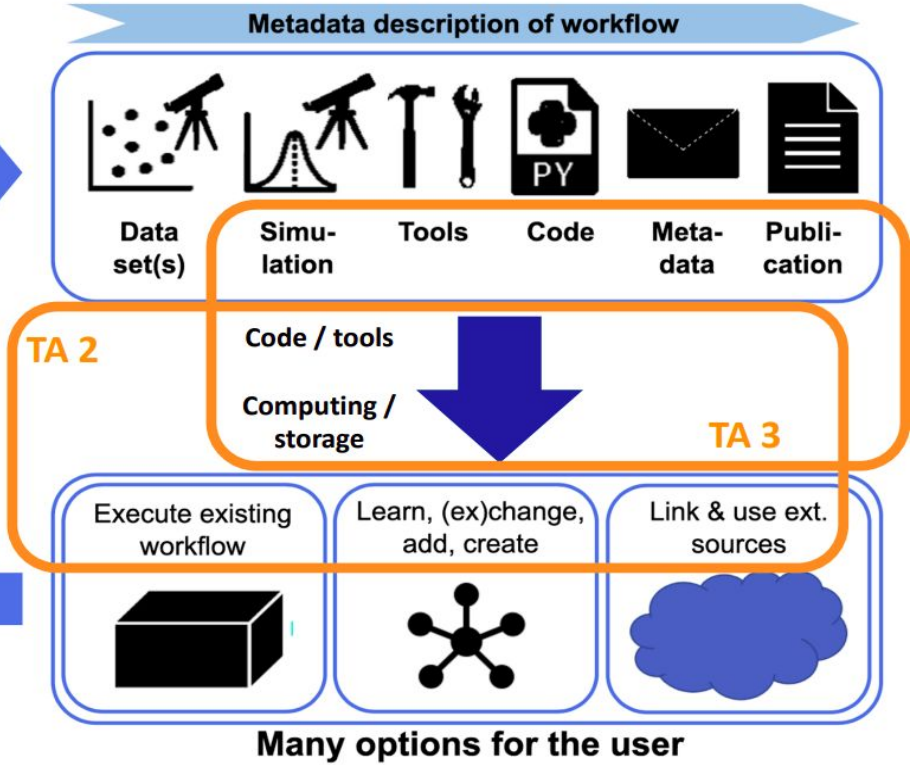
- Data is *potential* information
- Metadata record is a container for data about an object
- Digital Research Products (DRP) include: experiment and simulations data, code snippets, research papers, plots, workflows and the other products of research data life cycle, existing in digital form

PUNCH-SDP

The science data platform for RPs



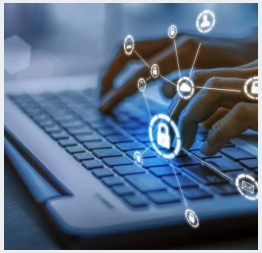
Research product contains executable workflow



KCDC Digital resources

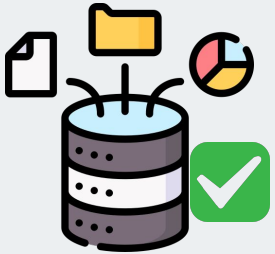
Karlsruher Institut für Technologie

User-selected data



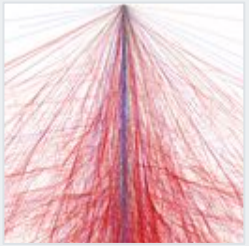
Quantity: ∞
API access: yes

Preselected datasets



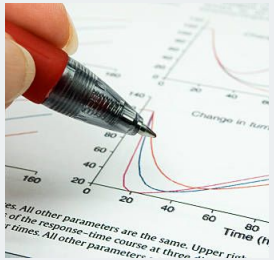
Quantity: 39
API access: no

Simulated data



Quantity: 280
API access: no

Publications

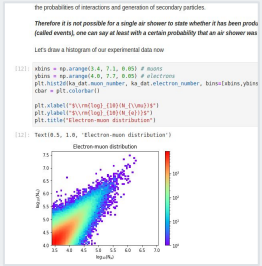


Quantity: 34
API access: no

Software

Quantity: 12
API access: no

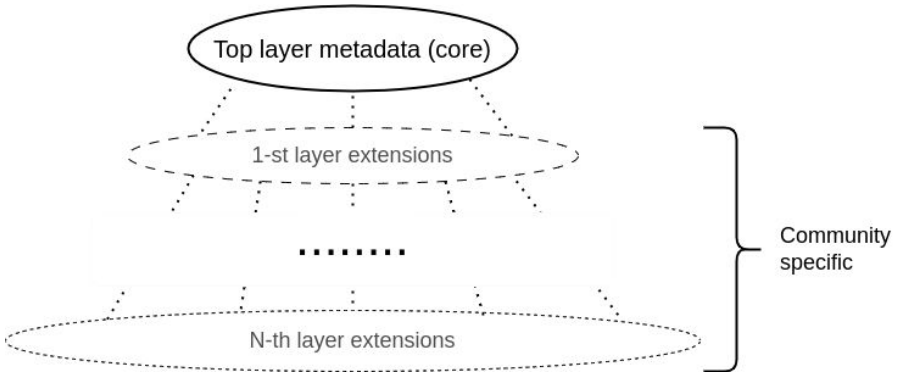
Tutorials



Quantity: 5
API access: no

Metadata standards

- Metadata scope: descriptive and some necessary administrative and legal metadata
- A **metadata schema** establishes and defines data elements and the rules governing the use of data elements to describe a resource
- Discovery, delivery and preservation needs of the project require usage of different metadata structure, content (syntax and vocabulary encoding schemes) and wrapper standards in order to present metadata information in layer scheme
- Observed general purpose schemas:
  Dublin Core, **DataCite**, MODS
- Advantages of DataCite:
  - well-supported,
  - provides interoperability,
  - widely used within data intensive physics community,
  - can be mapped to DublinCore
  - Uses DOIs as permanent identifiers
    => DOI minting



Towards DOI minting for KCDC digital resources

Common approaches:

- DataCite Fabrica
- RADAR KIT
- Side repositories: Zenodo, Arxiv, etc.

Alternative proposals:

- Partial solutions: ISBN existing for some publications, PIDs for software, etc.
- URLs
- UUID-5-based approaches

Metadata records creation

- Metadata records were created for all preselected datasets
- Sample metadata records were created for: simulations, publications, software snippets


```
<resource xsi:schemaLocation="http://datacite.org/schema/kernel-4 https://schema.datacite.org/meta/kernel-4.4/metadata.xsd">
  <identifier identifierType="DOI">10.17616/R3TS4P</identifier>
  <creators>
    <creator>
      +<creatorName nameType="Organizational"></creatorName>
    </creator>
  </creators>
  <titles>
    <title xml:lang="en-US">KASCADE_SmallDataSample_nA_runs_0877-7417_ROOT</title>
    <title xml:lang="en-US" titleType="Subtitle">
      KASCADE_SmallDataSample_nA_runs_0877-7417_ROOT XML metadata
    </title>
  </titles>
  <publisher xml:lang="en">KASCADE Cosmic Ray Data Centre (KCDC)</publisher>
  <publicationYear>2020</publicationYear>
  <subjects>
    <subject xml:lang="en-US" schemeURI="https://physh.org/browse" valueURI="https://physh.org/concepts/678197d1-8bf1-4c16-86cc-8133fba03b86" subjectScheme="PhySH (Physics Subject Headings)"
    classificationCode="678197d1-8bf1-4c16-86cc-8133fba03b86">Cosmic rays & astroparticles</subject>
  </subjects>
  <contributors>
    <contributor contributorType="ResearchGroup">
      <contributorName nameType="Organizational">KASCADE-Grande Collaboration</contributorName>
      <affiliation affiliationIdentifier="https://ror.org/04t3en479" affiliationIdentifierScheme="ROR" SchemeURI="https://ror.org">Karlsruhe Institute of Technology</affiliation>
    </contributor>
  </contributors>
  <dates>
    <date dateType="Created" dateInformation="Data selection OCEANUS_1s">2020-05-02</date>
  </dates>
  <language>en-US</language>
  <resourceType resourceTypeGeneral="Dataset">Reconstructed events</resourceType>
  <sizes>
    <size>90 MB</size>
  </sizes>
  <formats>
    <format>ROOT</format>
  </formats>
  <version>OCEANUS_1s</version>
  <rightsList>
    <rights xml:lang="en-US" rightsURI="https://kcdc.iap.kit.edu/static/pdf/kcdc_mainpage/EULA.pdf">
      End User Licence Agreement for using the KCDC IAP webportal and the KCDC data (EULA)
    </rights>
  </rightsList>
  <descriptions>
```

Metadata records integration

- Several metadata records for preselected datasets were integrated into PUNCH Metadata Catalog
- DOI search available, search by object's title is in work

```
victoria@victoria-ThinkPad-E490:~/Metadata_catalog_access_try-client$ ./try-mdc -punch -list
-----------------------------------------------------------------------------------------
[ws_show] 'HASH'
   #API   = 'REST'
   #Check = 'OK'
   #Code  = '200 OK'
   #Op    = 'POST'
   #URL   = 'https://euldg.physik.uni-bielefeld.de/ildg/mdc/datacite/query'
   result -> H[4]:
         elements -> A[82]:
                  10.1103/PhysRevLett.113.072001
                  10.22323/1.214.0186
                  10.1088/1742-6596/509/1/012098
                  10.1016/j.nuclphysa.2006.11.159
                  10.1103/PhysRevD.95.074505
                  10.1016/S0370-2693(01)01114-5
                  10.1103/PhysRevD.106.014510
                  10.1088/0954-3899/35/10/104093
                  10.22323/1.020.0163
                  10.1103/PhysRevD.103.094505
                  10.22323/1.105.0214
                  10.1016/j.nuclphysa.2014.08.073
                  10.1016/j.physletb.2019.05.013
                  10.1016/j.nuclphysa.2016.01.008
                  10.1103/PhysRevD.95.054504
                  10.4119/unibi/2985954
                  10.1143/PTPS.186.563
                  10.1103/PhysRevD.88.094021
                  10.1103/PhysRevD.80.014504
                  10.1103/PhysRevD.90.094503
                  10.22323/1.032.0021
                  10.5506/APhysPolBSupp.14.241
                  10.1016/S0010-4655(02)00327-2
                  10.1103/PhysRevD.68.014507
                  10.22323/1.363.0223
```

Conclusion

- KCDC's digital resources were revised
- Metadata strategy was mapped out
- DataCite schema was selected as metadata standard for shallow core metadata
- DataCite Fabrica was chosen for DOI minting
- Sample metadata records were prepared
- Shallow metadata, complying with the DataCite schema, were created for all preselected datasets
- Selected metadata records were integrated into PUNCH Metadata Catalog

Future steps

- Selection of suitable metadata standards for further layers of metadata
- Implementation and testing of DOI minting
- Organisation of metadata harvesting employing OAI-PMH
- Preparation metadata usage guidelines
- Further population of PUNCH Metadata Catalog

Conclusion

- KCDC's digital resources were revised
- Metadata strategy was mapped out
- DataCite schema was selected as metadata standard for shallow core metadata
- DataCite Fabrica was chosen for DOI minting
- Sample metadata records were prepared
- Shallow metadata, complying with the DataCite schema, were created for all preselected datasets
- Selected metadata records were integrated into PUNCH Metadata Catalog

Future steps

- Selection of suitable metadata standards for further layers of metadata
- Implementation and testing of DOI minting
- Organisation of metadata harvesting employing OAI-PMH
- Preparation metadata usage guidelines
- Further population of PUNCH Metadata Catalog

Thank you for your attention! Questions?

Contact me: victoria.tokareva@kit.edu