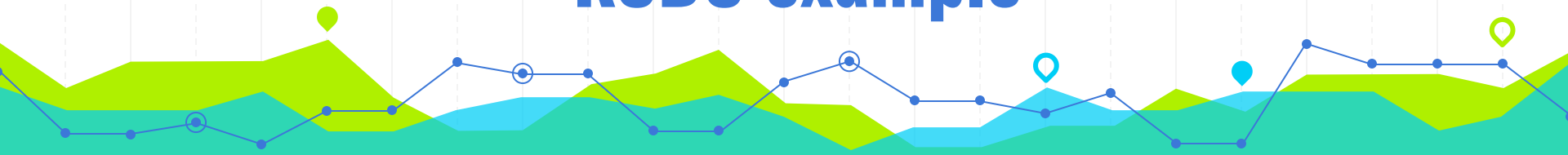
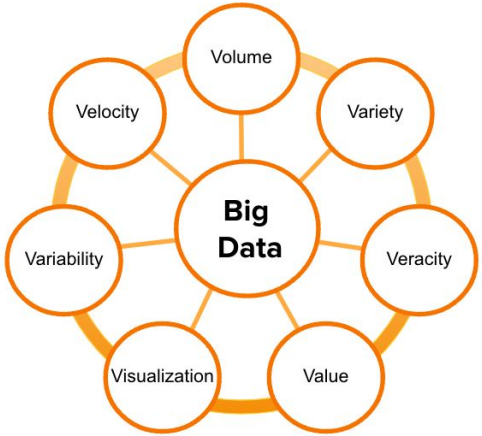
Open data and open-source tools throughout research data life cycle: KCDC example



V. Tokareva for KCDC Team
Karlsruhe Institute of Technology, Institute for Astroparticle Physics
DPG Spring Meeting of the Matter and Cosmos Section (SMuK)
Hacky Hour (joint session AKjDPG/AGI)
Dresden, 23 March 2023

Data-trends in particle, astro-, astroparticle, hadron and nuclear physics

Big data



FAIR data & software

Findable

Accessible

Interoperable

Reusable

Open science

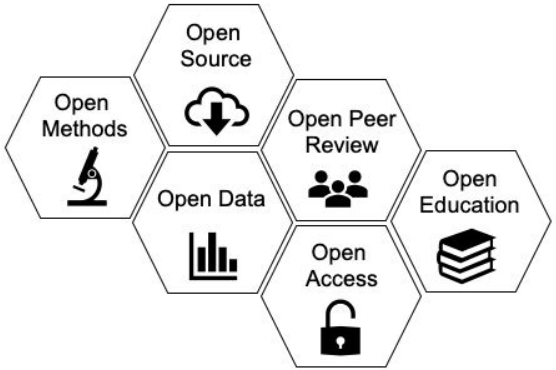
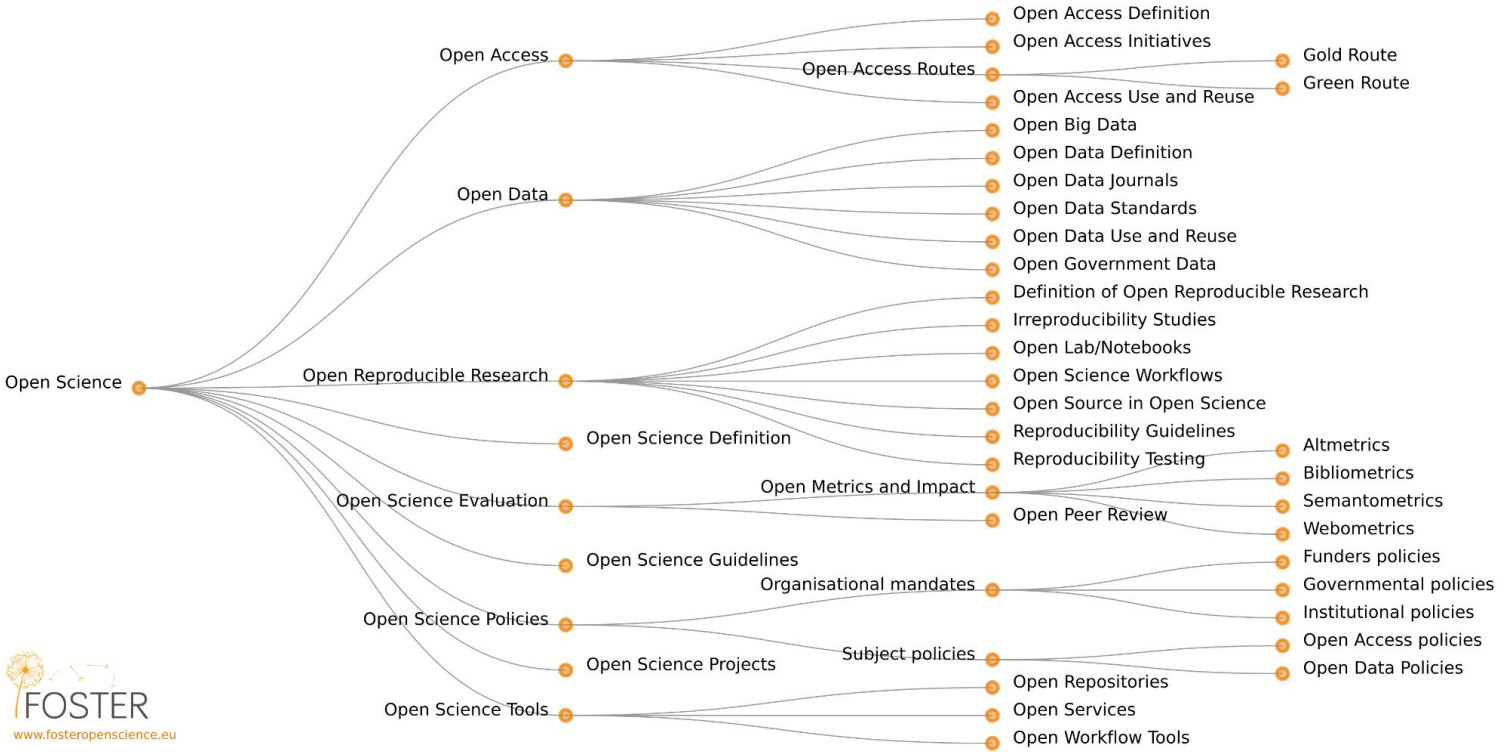


Image sources: https://shorturl.at/nzFN8, https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html, https://opensocialwork.org/research/open-science/

Open Science Taxonomy



FOSTER
www.fosteropenscience.eu

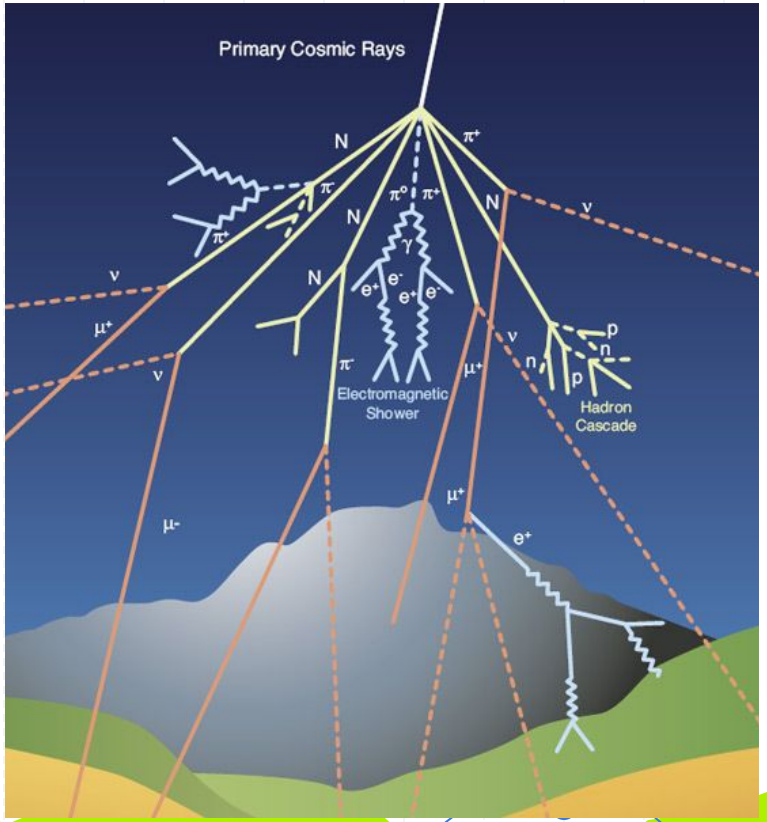
KCDC - KASCADE Cosmic Ray Data Centre

- 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/

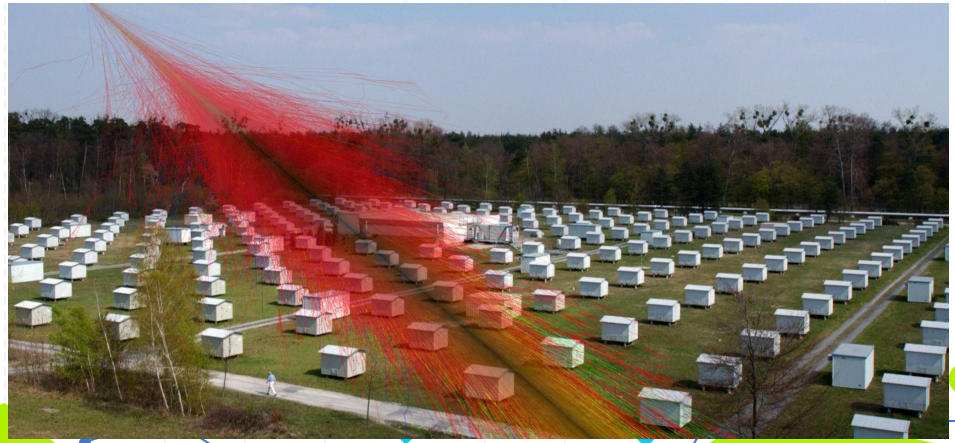
Functionality

- Archive of KASCADE software and data
- Provides free, unlimited, reliable open access to data of the KASCADE cosmic ray experiment
  - Selection of fully calibrated quantities and detector signals
  - Custom user data cuts
  - Allows data selection using both GUI and RESTless API
- Allows interactive analysis with integrated Jupyter Notebooks
- Information platform: physics and experiment backgrounds, tutorials, reference information

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



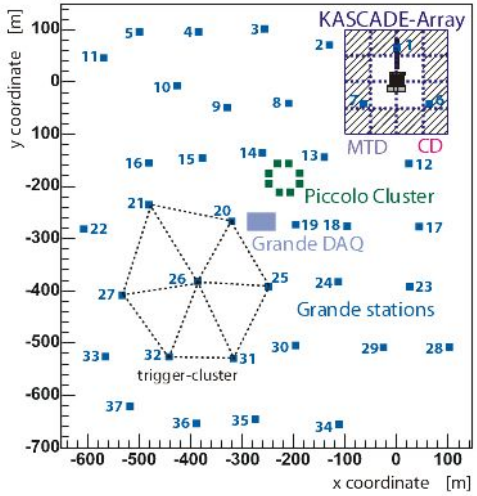
Other open datasets at KCDC

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 measureв 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



Datasets and formats at KCDC

The following **data formats*** are supported:

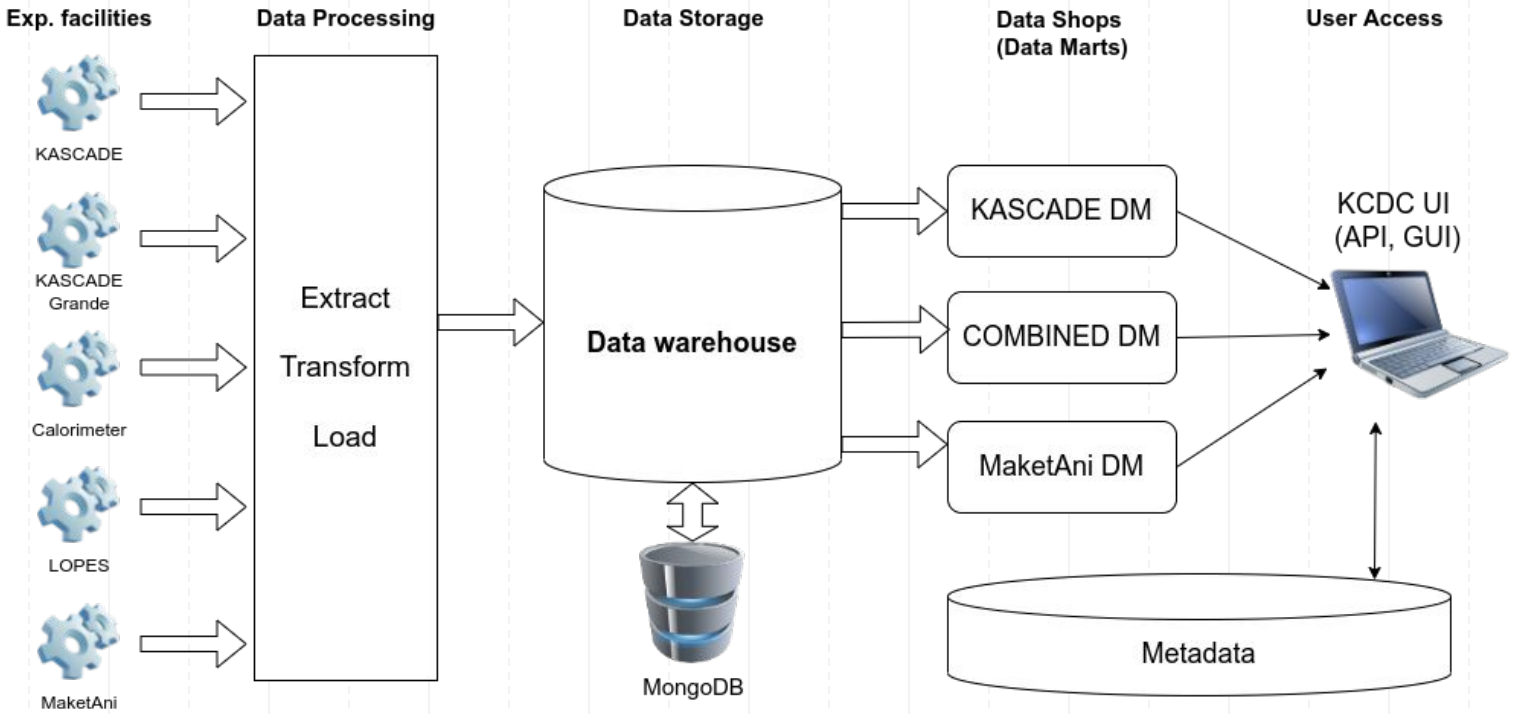
- ASCII - plain text format
- ROOT - object oriented framework developed by CERN
- HDF5 - hierarchical data format

* Selectable by the user and depending on the quantities chosen

Kinds of datasets, available at KCDC:

- Through Web GUI:
  - Preselected datasets
  - Simulations, based of the following models:
    - High energy models from three different model families: QGSjet, EPOS, SIBYLL
    - Low energy model FLUKA
- Through Web GUI + API:
  - Custom data selections

KCDC's software architecture

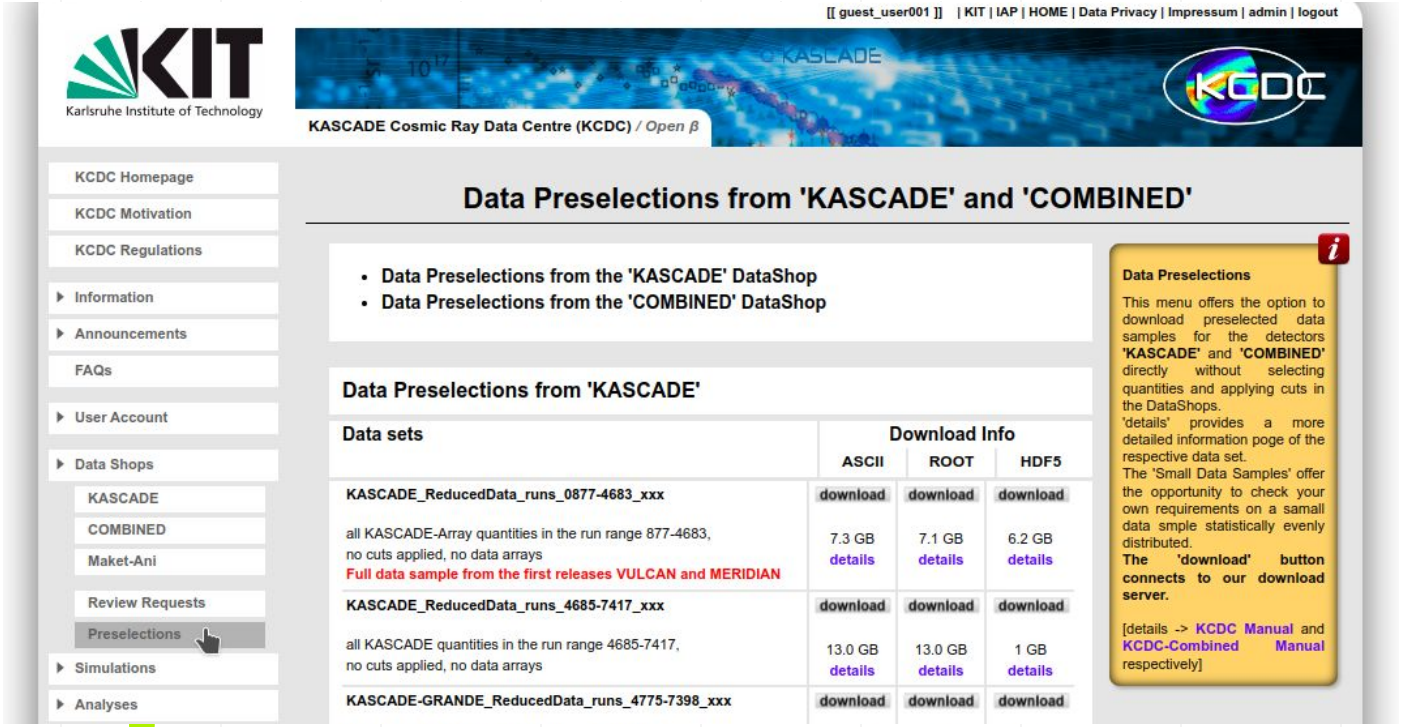


Metadata handling at KCDC

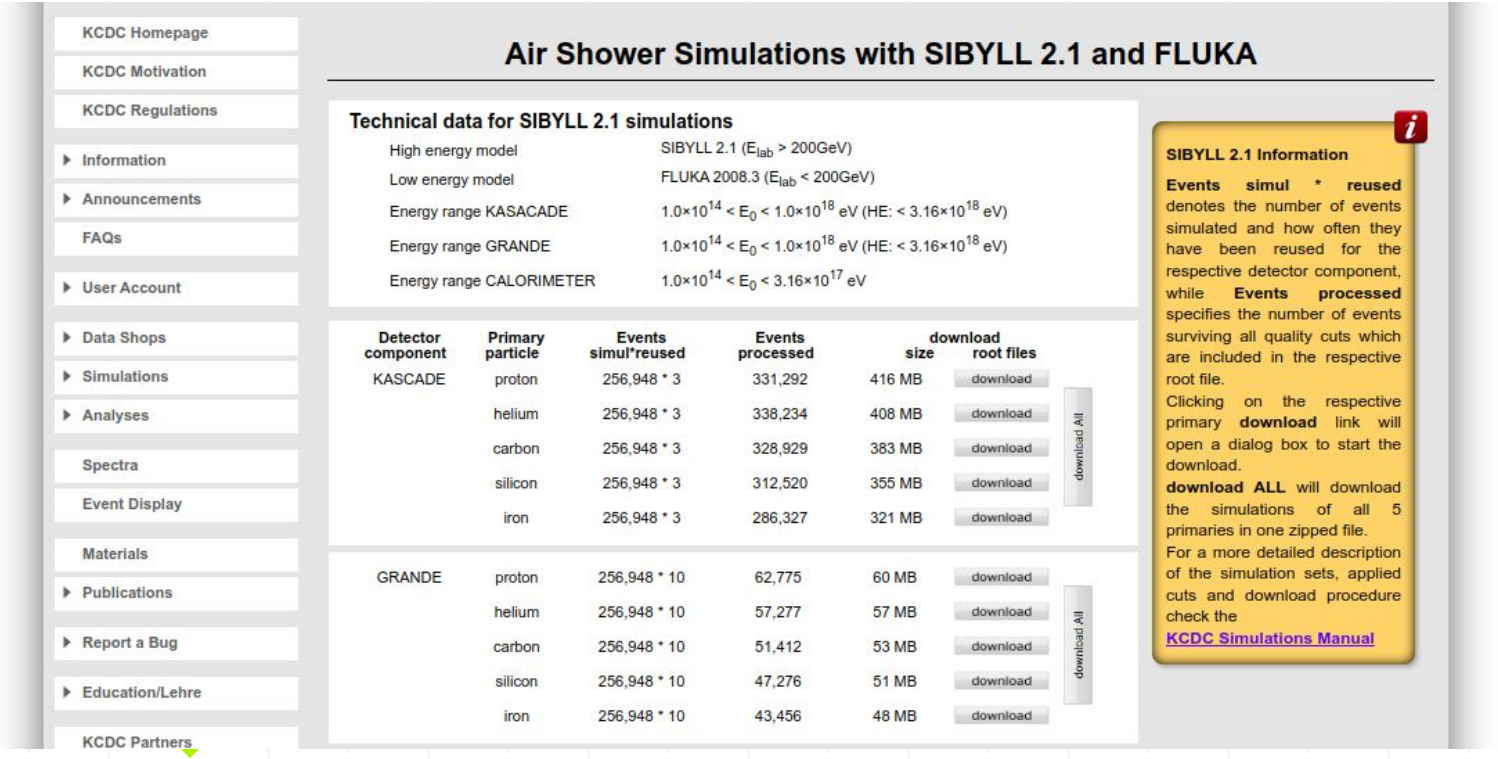
JSON metadata schema, example of a record from KCDC


```
{
  "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=dcInfoBoxDetailsDS>\r\nThe reconstructed Zenith Angle of the
    KASCADE showers is derived from the arrival time distribution of the of the particles at the detector
    stations. The range is from <span class=dcMathFunct>0&deg;</span> to <span class=dcMathFunct>60&deg;</
    span> where <span class=math>0&deg;</span> corresponds to a vertical shower. The angular resolution is
    between <span class=dcMathFunct>0.4&deg;</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"
  }
},
```

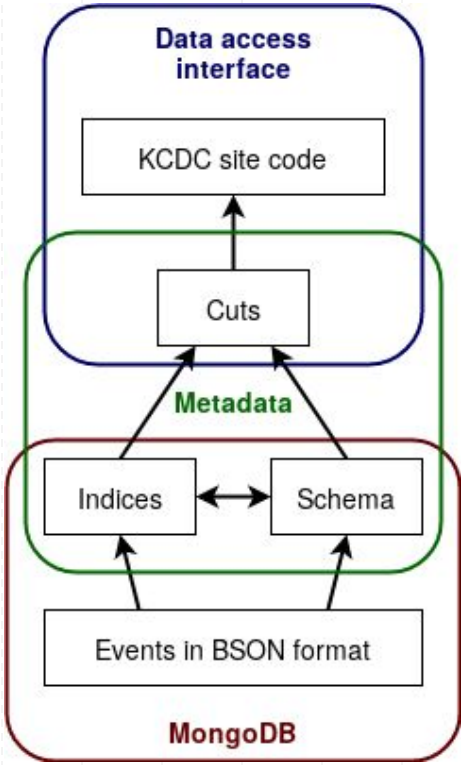
KCDC Preselections



Simulations



KCDC data acquisition



KASCADE Data Shop

Components Available	Components Selected	Quantities and Cuts			
GRANDE	General Info	Toggle all		*KASCADE*	
Calorimeter	KASCADE	Energy	range:	13 to 19 eV [log10]	Add Cut
LOPES		X Core Position	range:	-91 to 91 m	Add Cut
		Y Core Position	range:	-91 to 91 m	Add Cut
		Zenith Angle	range:	0 to 60 °	Add Cut
		Azimuth Angle	range:	0 to 360 °	Add Cut
		Electron Number	range:	2 to 8.7 [log10]	Add Cut
		Muon Number	range:	2 to 7.7 [log10]	Add Cut
		Shower Age	range:	0.1 to 1.48	Add Cut

Verify & Submit Request

KCDC Application Programming Interface (API)

Shell example: Extraction of the all data with an energy range from 17-19eV[log10]

Request:

```
curl --insecure --request POST 'https://kcdc-dev.iap.kit.edu/datashop/api/submit' \
--header 'Authorization: Basic cG92dGVyOmhhcnJ5Kytxb3R0ZXI=' \
--header 'Content-Type: application/json' \
--data-raw '
{
    "reconstruction": "",
    "output_format": "ascii",
    "datasets": [
        {
            "name": "array",
            "quantities": [
                {
                    "name": "E",
                    "cuts": [[17, 19]]
                }
            ]
        }
    ]
}'
```

Response:

job id:

{"id":"dbf1e608b6044223afe472125c020d88"}

or error message:

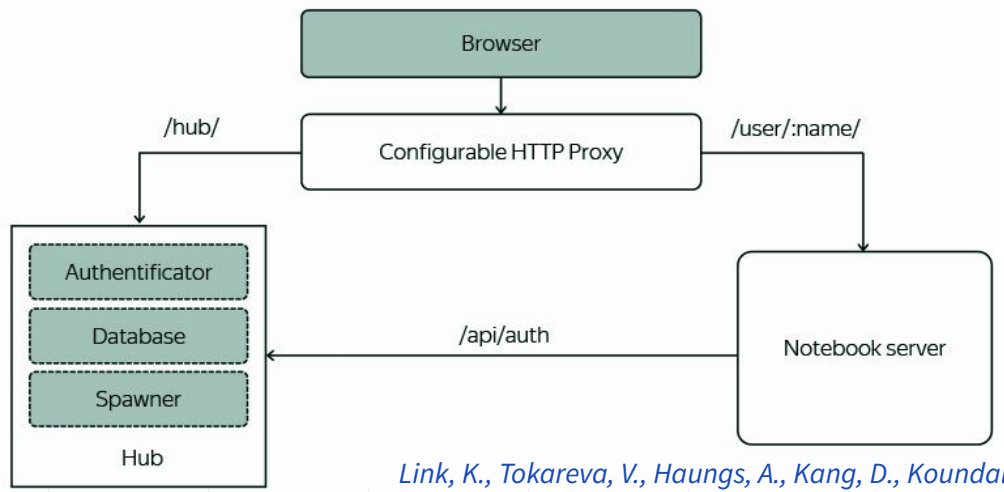
{"detail":"Invalid basic header. Credentials not correctly base64 encoded."}

[4] *Online API documentation: https://kcdc.iap.kit.edu/datashop/api/docs/index.html*

[5] *Wochele J. et al. KCDC User Manual: https://kcdc.iap.kit.edu/static/pdf/kcdc_mainpage/kcdc-Manual.pdf*

JupyterHub for data analysis

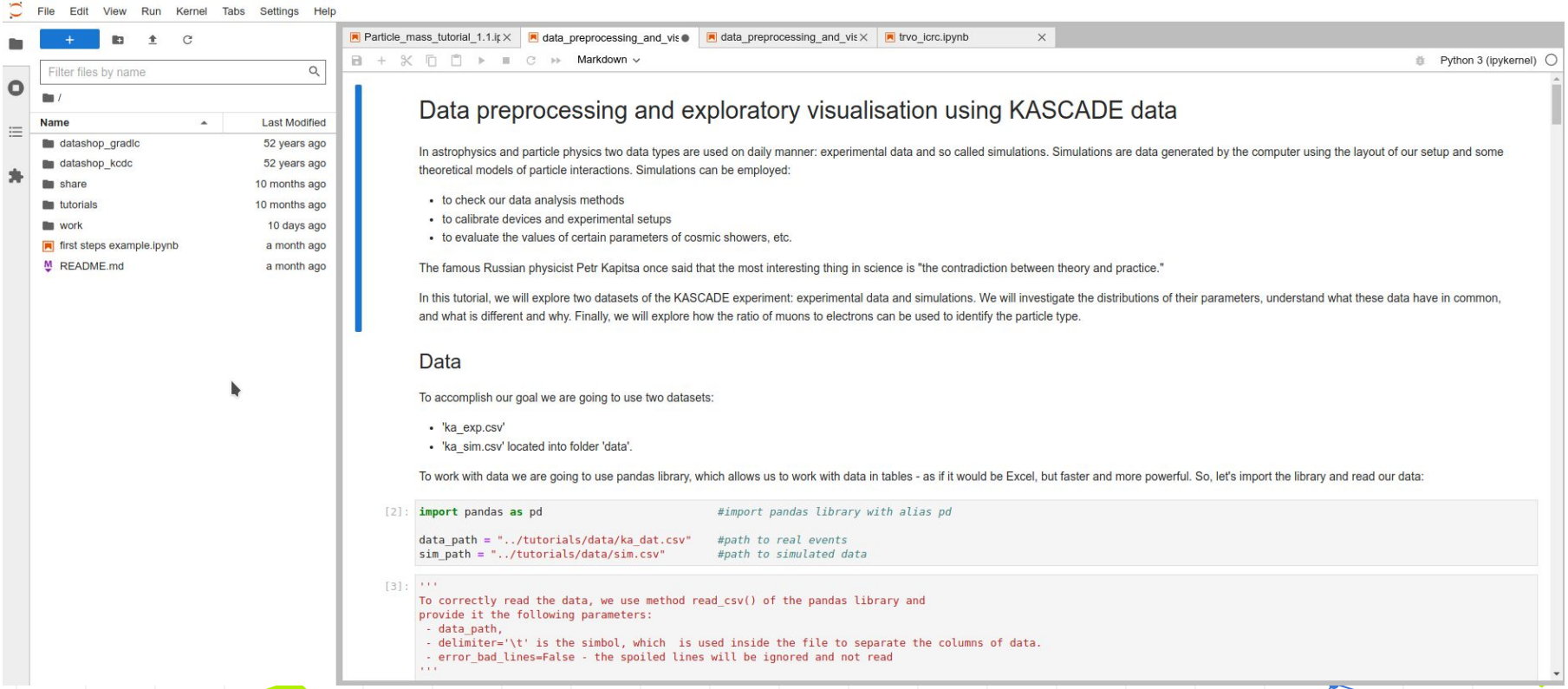
- Login via KCDC credentials
- Administration using Docker Swarm
- Tutorials by: KASCADE, IceCube, TRVO



Authorization	Run individual Jupyter servers
Proxy to Jupyter servers	System and database management

Link, K., Tokareva, V., Haungs, A., Kang, D., Koundal, P., Polgart, F., Tkachenko, O.,Wochele, D., Wochele, J. Online masterclass built on the KASCADE cosmic ray data centre. In 37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07.

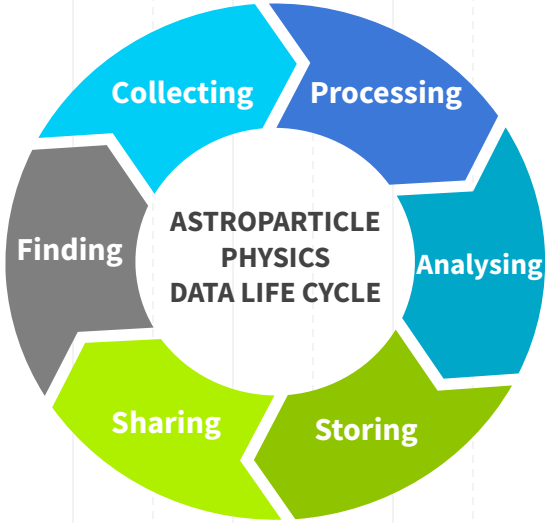
Usage of KCDC's JupyterHub



Open technologies in use:

- Django Web Framework
- Messaging RabbitMQ
- Celery Task Queue
- NoSQL (MongoDB) database
- JupyterHub for Jupyter Notebooks
- Docker/Singularity
- RESTful API
- Python, bash, JSON, HTML

Next steps: data life cycle and good research practice



Good Research Practice

According to DFG Code of conduct:

- Guideline 7: Quality Assurance
- Guideline 10: Usage Rights
- Guideline 12: Documentation
- Guideline 13: Public Access
- Guideline 17: Archiving

https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf

Data Life Cycle - the sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion

Takeaways

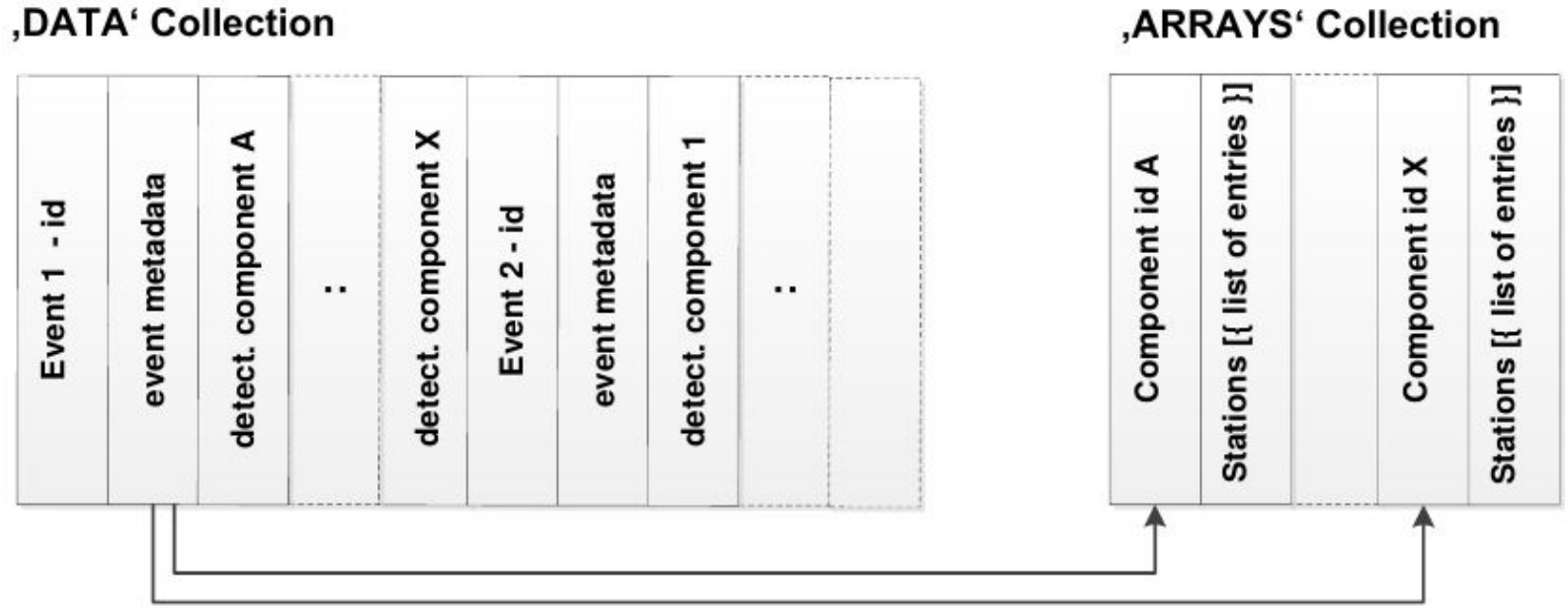
- KCDC is the public data centre for high-energy astroparticle physics
- All-in-one: data center, archive, information and educational platform
- Uses only open-source technologies
- Allows custom data selection using both GUI and RESTless API and interactive analysis with integrated Jupyter Notebooks
- http://kcdc.ikp.kit.edu
- Provides a good set of tools to be used on most of steps of research life cycle
- Promotes open science and good research practices

THANKS!

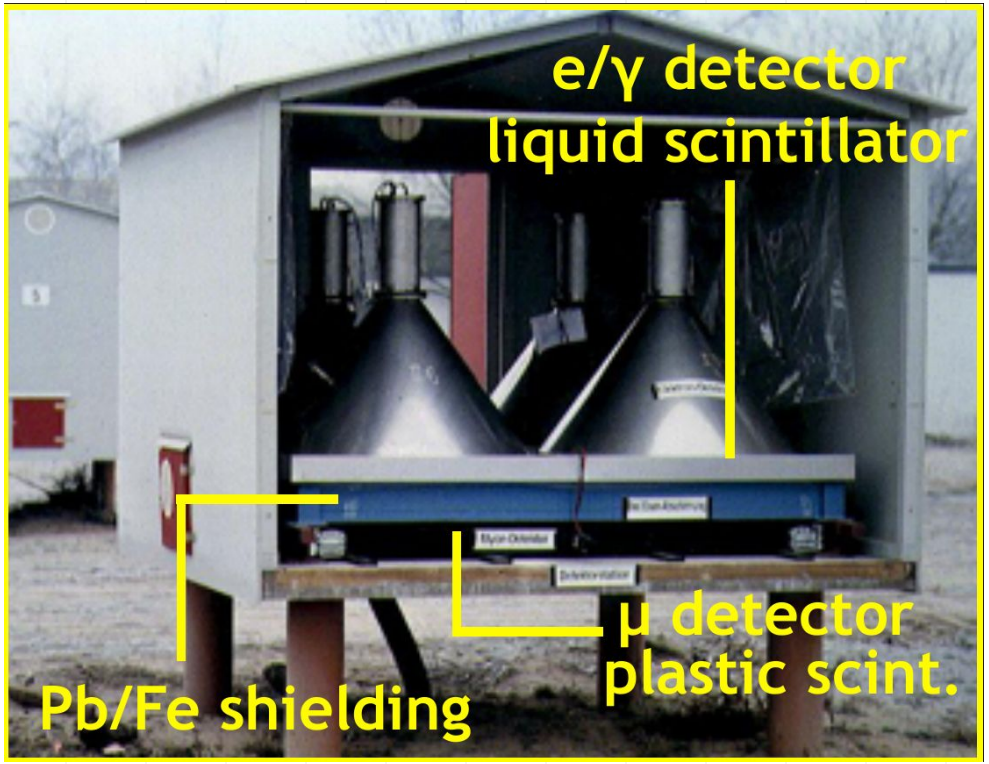
Hands-on time!
The materials are at https://uni-muenster.sciebo.de/s/VNenNLMwxopn1kR

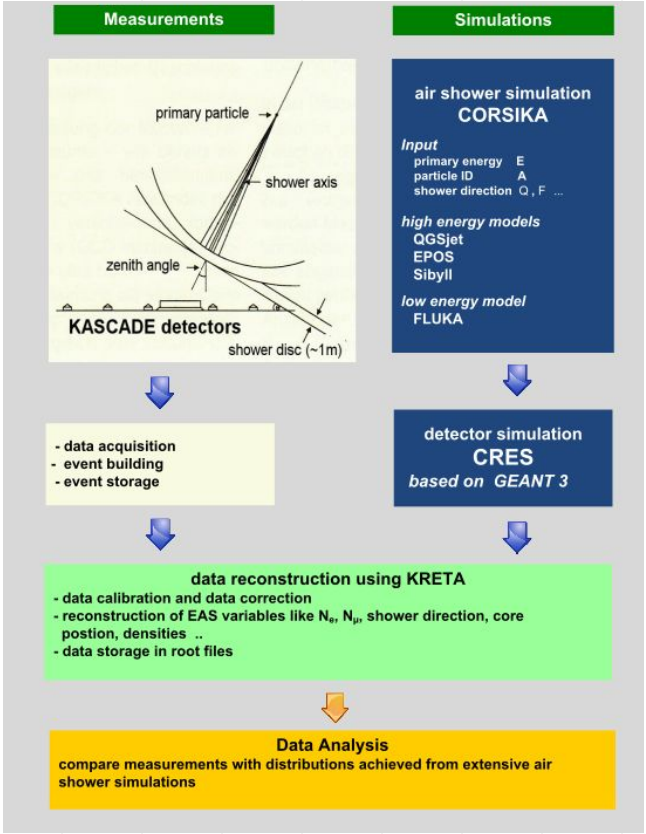
Contact me: victoria.tokareva@kit.edu

MongoDB data storage structure



Wochele, D., Wochele, J., Polgart, F., Tokareva, V., Kang, D., & Haungs, A. Data Structure Adaption from Large-Scale Experiment for Public Re-Use. CEUR-WS (2019) 2406, 114


e/γ detector
liquid scintillator
μ detector
plastic scint.
Pb/Fe shielding


Measurements
Simulations
primary particle
shower axis
zenith angle
KASCADE detectors
shower disc (~1m)
air shower simulation
CORSIKA
Input
primary energy E
particle ID A
shower direction Q , F ...
high energy models
QGSjet
EPOS
Sibyll
low energy model
FLUKA
- data acquisition
- event building
- event storage
detector simulation
CRES
based on GEANT 3
data reconstruction using KRETA
- data calibration and data correction
- reconstruction of EAS variables like N_e, N_μ, shower direction, core postion, densities ..
- data storage in root files
Data Analysis
compare measurements with distributions achieved from extensive air shower simulations

Data quantities example

CALORIMETER Quantities

Var	Name	Available Data Range	Unit	Representation
Nhad	Nr of Hadrons	0. - 511.		
Ehad	Hadron Energy Sum	0.; 1.e10 - 1.e16	eV	log10 -> 10.0 - 16.0

GRANDE Quantities

Var	Name	Available Data Range	Unit	Representation
Xc	X-Core Position	-500.0 - +100.0	m	
Yc	Y-Core Position	-600.0 - +100.0	m	
Ze	Zenith Angle	0.0 - 40.0	°	
Az	Azimuth Angle G	0.0 - 360.0	°	
Nch	Number of charged part	11111. - 1,000,000,000.		log10 -> 4.0 - 9.0
Nmu	Number of Muons	1500. - 100,000,000.		log10 -> 3.2 - 8.0
Age	Shower Age G	-0.385 - +1.485		
GDeposit	Energy Deposit charged	0.0 - 100,000.0	MeV	/station
GArrival	Arrival Time	1000. - 10,000.0	ns	/station