Data Structure Adaption from Large-Scale Experiment for Public Re-Use *

Doris Wochele[1 [0000−0001−6121−0632]], Jürgen Wochele[1 [0000−0003−3854−4890]], Frank Polgart[1 [0000−0002−9324−7146]], Victoria Tokareva[1 [0000−0001−6699−830X]], Donghwa Kang[1 [0000−0002−5149−9767]], and Andreas Haungs[1 [0000−0002−9638−7574]]

[1]Karlsruhe Institute of Technology, Institute for Nuclear Physics, 76021 Karlsruhe, Germany
doris.wochele@kit.edu
http://www.kceta.kit.edu

Abstract. Large-scale experiments in astroparticle physics are usually operated several decades by international collaborations of partly several hundreds of scientists. Experiments launched some decades ago, trying to make their data publicly available, suffer from the fact that their data structures cannot be evaluated using modern information technologies. To overcome this situation and to guarantee a FAIR (findable-accessible-interoperable-reusable) [2] data preservation, the data must be restructured and reformatted. A step in this direction is to provide the data and meta-data as well as the tools to analyse the measured data of the meanwhile dismantled cosmic ray experiment KASCADE, which operated from 1996 to 2013. The project to make the entire scientific data public is called the 'KASCADE Cosmic Ray Data Centre' (KCDC, https://kcdc.ikp.kit.edu). The activities within KCDC are used as blueprint for a sustainable data life cycle including aspects of data curation in astroparticle physics. With this paper we give an overview of the current status of the KASCADE Open Data Publication via the KCDC web portal with focus on the adaption of the initial structure of the KASCADE data for the KCDC data publication.

Keywords: Astroparticle Physics · Data Structure · Data Curation · Public Data Centre

1 Introduction

A major topic of Astroparticle Physics is to investigate the nature of the cosmic radiation manifesting in many ways. The detection of high-energy cosmic rays is performed by the registration of extensive air showers, i.e. the secondary

* Supported by KRAD, the Karlsruhe-Russian Astroparticle Data Life Cycle Initiative (Helmholtz Society Grant HRSF-0027). The authors acknowledge the cooperation with the Russian colleagues (A. Kryukov et al.) in the GRADLC project (RSF Grant No. 18-41-06003) as well as the KASCADE-Grande collaboration for their continuous support of the KCDC project.

particles generated by the primary cosmic ray when entering the atmosphere. In various air-shower experiments located all over the world the cascades of particles are detected, generated in interaction processes of the relativistic cosmic rays (mainly fully ionized atomic nuclei, but also cosmic gamma-rays and neutrinos)with the molecules of the Earth's atmosphere. As these air-showers consist of a huge number of particles which are spread over a vast area, large detectors are required to measure certain characteristic parameters. Dedicated reconstruction algorithms [1] are then used to determine the direction of the incoming particle as well as its energy and mass.

The combined analysis of observations of various components of the cosmic radiation like charged particles, gamma rays and neutrinos is widely known as 'Multi-Messenger Astroparticle Physics'. From this, currently as hot topic classified research field, we hope to gain new and exciting information to extend our knowledge of the origin and transport of what we understand as cosmic radiation. The Multi-Messenger Astroparticle Physics requires access to data (in a reasonable and standardized format) of the diverse experimental installations. Our studies described here are important steps towards establishing a global data and analysis centre [2] for Multi-Messenger Astroparticle Physics.

2 From Proprietary Experimental Data to an Easy Accessible Open Data Format

KASCADE-Grande [3],[4],[5] was an extensive air shower experiment array to study the cosmic ray primary composition and the hadronic interactions in the energy range $E_0 = 10^{14}-10^{18}$ eV. The experiment was situated on site of the KIT, Campus North (49.1 °N,8.4 °E) at 110 m asl, corresponding to an average atmospheric depth of 1022 g/cm^2 [3] and operated between 1996 and 2013. One of the main results obtained by KASCADE is a picture of increasingly heavier composition above the 'knee' caused by a break in the spectrum of the light components. Conventional acceleration models predict a change of the composition towards heavier components.

The KASCADE-Grande experiment consisted of four major detector components. An array of 252 detector stations housing separate electron and muon detectors, a central detector for measuring the hadron component with additional muon detection area, a tunnel with streamer tubes to record individual muon tracks and, a second array with 37 detector stations for electron detection which extends the effective KASCADE array area from $200 \times 200\,\mathrm{m}^2$ by a factor of 10 built in 2003. Furthermore, an array of 30 radio antennas is co-located with the KASCADE array (LOPES [10]) which uses the KASCADE trigger and the well-calibrated information of air shower properties to study radio emissions of cosmic rays (see fig. 1).

After the shutdown of the KASCADE experiment the number of scientists who can maintain the databases with old software on old computers degraded fast. The only chance to get the results comprehensible for other experiments is

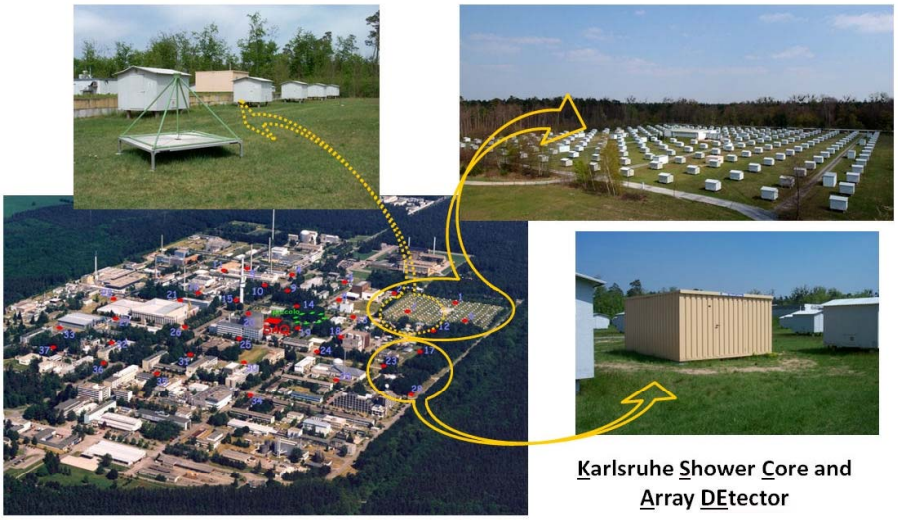

Fig. 1. The Karlsruhe Shower Core and Array Detector (KASCADE)

to start a data curation process and feed the data to a system where they can be kept maintained.

3 KASCADE Data Acquisition and Data Structure

The KASCADE Data Acquisition System was designed and set up in the early nineties of the last century using the most modern types of databases and software available. By that time all off-line programming in high-energy physics was carried out using Fortran 77 programming language. Although this language offered some advantages compared to other common languages at that time, it suffered from a lack of dynamic data structuring facilities. To overcome these disadvantages CERN ZEBRA [11] has been introduced by the particle physicists to allow the programmers to build dynamic data structures even at real (execution) time. Thus, the KASCADE raw data are organized in so-called ZEBRA banks as schematically described in fig. 2.

Below a file header holding information about the current run, properties like start time, detector status, calibration settings etc., the event blocks are organized sequentially, each consisting of an event header and the hierarchically arranged ZEBRA banks for the event data of every detector component being part of the respective event, followed by a direct access table information on number of events and the position of the first data set. Each of the various components of the KASCADE-Grande detector system was designed as an independent experiment with its own data acquisition and control system. Besides the control and monitoring of the hardware, the preparation of the data recorded and the transmission to the Central Event Builder was part of each system. The

Fig. 2. Scheme of KASCADE ZEBRA bank structure.

main task of the Central Event Builder is to merge time information and the recorded data of all running detector components and to transfer it to a mass storage system. The event builder is invoked by a trigger signal from any of the various trigger sources of KASCADE. The conditions to generate a trigger are defined in the local processes of the detector components without any control by the event builder. Data blocks with time labels within a window of 10μs are merged to one event. If a detector component does not supply data for more than 2 minutes, a hardware error is assumed. Then the corresponding component is removed from the data acquisition and a new run is started. Altogether about 1.7 billion events have been recorded during the lifetime of KASCADE from 1996 to 2013 stored in more than 50.000 raw data files, consuming 4 TB of storage space, and analysed with the KASCADE data reconstruction program KRETA. The data have been archived on a tape robot of the central computing department and copied in portions to local disks for fast access during data analysis. Meanwhile all measured and analysed data are stored locally on a Raid system and on the IBM Tivoli Storage Manager of KIT-SCC (Steinbuch Centre for Computing) for long term preservation.

4 KASCADE Data Analysis

The measured air showers in KASCADE are analysed using the reconstruction program KRETA (Kascade Reconstruction for ExTensive Airshowers), which reads the raw data, performs the calibration and reconstructs the basic shower observables, storing all the results in the form of histograms and vectors of parameters (n-tuples).

The calibration parameters and the geometry of the detector layout are stored in a time dependent database called CERN HEPDB [12]. Calibration parameters like energy deposits or correction parameters like time delay offsets are derived from a separate analysis of the data recorded and stored in HEPDB together with a validity time stamp. The validity range of typically several days is mainly caused by the change of the photomultiplier tubes associated with the temperature of the environment, while the geometry database remained mostly constant during the complete measuring period. Other correction data sets like air pressure and temperature were obtained from external measurements. Additionally, a correction table was required to exclude disturbances from a nearby man-made radiation source that occurred periodically for a few minutes per day.

Data sets recorded with irregular hardware conditions (bad runs) have been identified in a preanalysis step and removed for the final reconstruction.

Data quality is commonly described as a state of accuracy for appropriate use of data. The re-usability of an astroparticle experiment dataset relies on a high quality filter of the raw data. Only the researcher can take into account all details of data acquisition and data transformation. The raw KASCADE data are reduced by a factor of four by applying quality cuts within the data analysis procedure, mostly because low energetic events with a too small signal cannot be reconstructed properly. All these cuts were applied during the calibration phase on an event-by-event basis as well as the energy deposits corrections for the detector stations flagged as "not working" or "in saturation". Calibration data, correction data and information on the detector layout are taken from a HEPDB data base where more than 100 different time dependent calibration tables are stored. The reconstruction procedure itself is then done in 3 iteration levels. A rough estimation of shower parameters in level 1 is followed by a more sophisticated analysis in level 2 where the results of level 1 are taken as starting values for level 2. Level 3 aims for an even more tuned accuracy using the results of level 2 and delivers the final reconstruction results. These results for each iteration level are stored in so called ntuples making use of the HBOOK package [14] which is also part of the CERN library.

An ntuple is like a table, where all the variables belonging to a certain event are columns while each event is a row. In this form it is easy to generate one or multi dimensional projections of any of these several hundred variables. Storing requirements become significant for large event samples. The most important output parameters stored in ntuples for KASCADE are:

- the reconstructed position of the shower core,
- the reconstructed shower direction,
- the reconstructed number of electrons and muons,
- most probable reconstructed primary energy E_0 based on a certain variant of the reconstruction process (is always based on a theoretical model of hadronic interactions incorporated in simulations [8]).

In the early years of KASCADE the resulting hbook files have been visualized using the CERN PAW framework (Physics Analysis Workstation [13]) to generate statistical distributions of the measured events. KRETA still produces only hbook files but nowadays they are converted to the CERN ROOT data structure [15] with a disk space requirement of about 1.3 GB and stored on local servers and archived on mass storage robots.

5 KCDC portal as a demonstrator for Big Data Analytics in astroparticle physics

The main goal of KCDC [7,6] is to provide a concept for the open data publication, following the idea of the Berlin Declaration on Open Data and Open Access [9].

The large amount of data stored in ROOT files makes it impossible to offer the entire data directly for download. Thus we choose to fill the reconstructed data and meta data into a database and install a web portal, named the KCDC data shop [6], in a way that a registered user can select parameters and apply specific cuts on most of the quantities to limit the download volume for his own analysis. For event object storage a NoSQL database was chosen because of the unfixed scheme per event. For multi-messenger analyses in cosmic ray physics the experimental setup usually varies within time. This makes it necessary to add components and new joins which is still a 'show-stopper' in big relational database management systems (RDBMS). NoSQL databases can be more easily adapted to new setups. In KCDC MongoDB was chosen as an easy-to-use and scalable document-oriented storage well proofed in Big Data environments.

The KCDC web portal is publicly accessible, no special software installation is required on the user side. All data are offered for ftp download in ROOT, which is standard data format in High Energy Physics, or in HDF5, mostly used to store big amounts of data, or in ASCII for simple education examples.

Most of the parameters stored in the ROOT files gained from the KASCADE data reconstruction with KRETA can only be used by experts with detailed background knowledge of the detector geometry and the detector properties. In order to give a wider audience the opportunity to use these data, they must be prepared in such a way that they can be handled without special knowledge.

In modern terms this procedure is known as data curation. Curation includes a range of activities and processes done to create, manage, maintain, and validate data elements. Specifically, data curation is the attempt to determine what information is worth saving and for how long. The data curation workflow is determined from data quality management, data protection, life cycle management and data movement. Therefore the routine filling KASCADE data in the MongoDB of KCDC determines which data should be published in that specific version of KCDC. The KCDC Major Version Number is an indicator for a complete dataset. Any changes within the database will be indicated by a new version number and we can only guarantee a valid representation of KASCADE using data from the same published versions.

Another important point for this procedure is that in KRETA each detector component generates its own ROOT files because no data merging has been applied. This disadvantage we overcome by publishing the combined data analysis in a new KCDC data shop presently under construction. For the time being, we merge the data of the different components when filling the MongoDB.

As the KASCADE Events can be uniquely identified by their Run- and Event numbers, the filling routine uses these two parameters to merge the data sets from the different components of KASCADE and some additional information from separate tables like exclude lists for special events. Most of the space required in the MongoDB is occupied by the data arrays holding information on e/γ and μ-energy deposits of each of the 252 KASCADE detector stations and

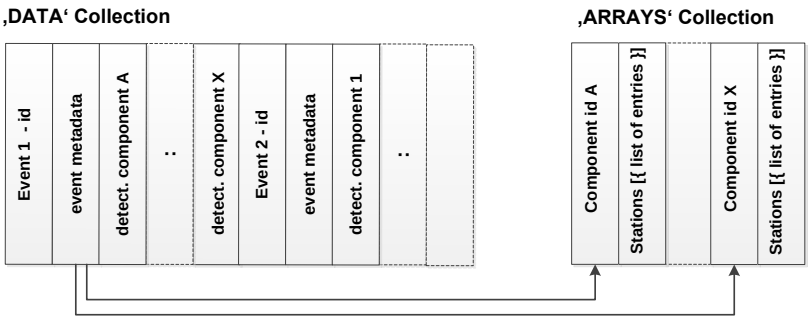

Fig. 3. Scheme of the used MongoDB data storage structure.

the arrival time of the first particle hitting the respective station and producing a valid signal in the detector electronic. The structure scheme of the data stored in the MongoDB is shown in fig. 3. In MongoDB we have two collections named ‘DATA‘ and ‘ARRAYS‘, where the first is filled with (meta)data from the reconstruction process with information on parameters like event number, event time etc. and shower parameters like core position, angle of incidence and number of particles separately for each detector component. In the second, the ‘ARRAYS‘, relevant data from every detector station are kept.

6 Outlook

Publishing data for other scientist enhances the sensibility for correct documentation and cross-checking the results. Using only one source of data for researchers from the KASCADE collaboration and for external researchers led automatically to a high integrity of the published data sets and increased our confidence in the data sets published. Published data sets in experimental physics underlie changes whenever analytical methods are improved or errors are discovered. A change of the version number indicates a non-semantic change of the data sets whereas the elements e.g. a specific event, is still part of the data set. A unique UUID, implemented in the next release, represents such an object independent from versions and the reference is immutable. Records that were once published in KCDC are frozen by versions, which means that even if the data are extended or changed the reproducibility is maintained.

In a next step we will adapt the data of a totally independent experiment (TAIGA [16] or TUNKA [17]) to the scheme described above and include them into KCDC. With a first multi-messenger like analysis applied to the data of both experiments a proof-of-principle of the demonstrator will be given [18] (see fig. 4).

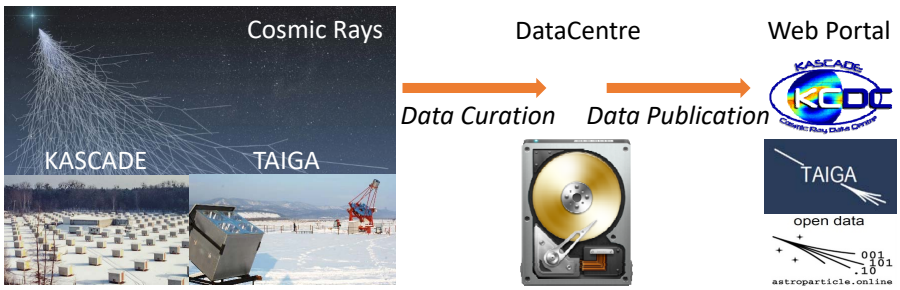

Fig. 4. Scheme of the data flow of the exemplary data life cycle initiative: Cosmic ray data from two different experiments are stored with a standardized data structure and comparable meta data before the made available for a re-use.

References

1. A. Haungs, H. Rebel, M. Roth; 'Energy spectrum and mass composition of high-energy cosmic rays'; Rept.Prog.Phys. **66** (2003) 1145-1206
2. A. Haungs; 'Towards a global analysis and data centre in Astroparticle Physics'; (2019) these proceedings.
3. T. Antoni et al; The Cosmic-Ray Experiment KASCADE; Nucl.Instr. and Meth **A513** (2003) 490-510
4. W.-D. Apel et al; The KASCADE-Grande Experiment; Nucl.Instr. and Meth. **A620** (2010) 202
5. KASCADE homepage, `https://www.ikp.kit.edu/KASCADE/`
6. KCDC homepage, `https://kcdc.ikp.kit.edu`
7. A. Haungs et al; The KASCADE Cosmic-ray Data Centre KCDC: Granting Open Access to Astroparticle Physics Research Data; Eur. Phys. J. C (2018) **78**:741 ; https://doi.org/10.1140/epjc/s10052-018-6221-2
8. CORSIKA homepage, `https://www.ikp.kit.edu/corsika`
9. Berlin Declaration, `https://openaccess.mpg.de/Berlin-Declaration` accessed Jan 2015
10. H. Falcke et al; Detection and imaging of atmospheric radio flashes from cosmic ray air showers; Nature **435**:313 (2005)
11. ZEBRA reference Manual, `https://cdsweb.cern.ch/record/2296399/files/zebra.pdf`
12. HEDB reference manual, `http://cds.cern.ch/record/2296379/files/hepdb.pdf`
13. PAW reference Manual, `http://www2.pv.infn.it/~sc/cern/paw.pdf`
14. HBOOK reference manual p.e., `http://osksn2.hep.sci.osaka-u.ac.jp/~taku/doc/hbook.pdf`
15. ROOT users guide, `https://root.cern.ch/guides/users-guide`
16. N. Budnev et al; The TAIGA experiment: From cosmic-ray to gamma-ray astronomy in the Tunka valley; Nucl.Instrum.Meth. **A845** (2017) 330-333
17. F.G. Schröder et al.; Tunka-Rex: Status, Plans, and Recent Results; EPJ Web Conf. **135** (2017) 01003
18. I. Bychkov et al.; 'RussianGerman Astroparticle Data Life Cycle Initiative'; Data **3(4)** (2018) 56