PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

German-Russian Astroparticle Data Life Cycle Initiative to foster Big Data Infrastructure for Multi-Messenger Astronomy

Victoria Tokareva,[a,*] **Igor Bychkov,**[b,c] **Andrey Demichev,**[d] **Julia Dubenskaya,**[d] **Oleg Fedorov,**[e] **Andreas Haungs,**[a] **Donghwa Kang,**[a] **Yulia Kazarina,**[e] **Elena Korosteleva,**[d] **Dmitriy Kostunin,**[g] **Alexander Kryukov,**[d] **Andrey Mikhailov,**[b] **Minh-Duc Nguyen,**[d] **Frank Polgart,**[a] **Stanislav Polyakov,**[d] **Evgeny Postnikov,**[d] **Alexey Shigarov,**[b,c] **Dmitry Shipilov,**[h] **Achim Streit,**[f] **Doris Wochele,**[a] **Jürgen Wochele**[a] **and Dmitry Zhurov**[e]

[a] *Karlsruhe Institute of Technology, IAP, 76021 Karlsruhe, Germany*
[b] *Matrosov Inst. f. System Dynamics and Control Theory, Irkutsk 664033, Russia*
[c] *Irkutsk State University, Irkutsk 664003, Russia*
[d] *Lomonosov Moscow State University, SINP, Moscow 119991, Russia*
[e] *Irkutsk State University, Applied Physics Institute, Irkutsk 664003, Russia*
[f] *Karlsruhe Institute of Technology, SCC, 76021 Karlsruhe, Germany*
[g] *DESY, 15738 Zeuthen, Germany*
[h] *X5 Retail Group, Moscow, 119049 Russia*
E-mail: victoria.tokareva@kit.edu

Challenges faced by researchers in multi-messenger astroparticle physics include: computing-intensive search and preprocessing related to the diversity of content and formats of the data from different observatories as well as to data fragmentation over separate storage locations; inconsistencies in user interfaces for data retrieval; lack of the united infrastructure solutions suitable for both data gathering and online analysis, e.g. analyses employing deep neural networks. In order to address solving these issues, the German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI) was created. In addition, we support activities for communicating our research field to the public. The approaches proposed by the project are based on the concept of data life cycle, which assumes a particular pipeline of data curation used for every unit of the data from the moment of its retrieval or creation through the stages of data preprocessing, analysis, publishing and archival. The movement towards unified data curation schemes is essential to increase the benefits gained in the analysis of geographically distributed or content-diverse data. Within the project, an infrastructure for effective astroparticle data curation and online analysis was developed. Using it, first results on deep-learning based analysis were obtained.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)938

1. Introduction

One of the principal ways to study high-energy processes taking place in the Universe is to detect ultra-high-energy particles and radiation generated by these processes. The crucial benefit of gamma radiation and neutral particles is that they can be directly traced back to their sources. Though the charged particles do not support a simple "point-and-shoot" type of analysis, their integrated flux can nonetheless bring important knowledge about frequency and intensity of high-energy cosmic sources and the properties of interstellar and intergalactic medium. The ultra-high-energy cosmic rays are of particular importance, since their energies are currently unattainable in accelerators and correspond to the most energetic physical phenomena, featuring matter in exotic states and allowing to probe the limits of modern physical theories.

The cosmic-ray flux drops sharply with energy, reaching several particles per year per km^2 for the events with ultra-high energy $> 10^{18}$ eV [1]. Thus, combining data from different experiments acquires particular significance, as it allows to increase statistics and expand the observed area of the sky. Besides, joint analysis of information from various messengers, like charged particles [2], neutrons [3], gamma-rays [4, 5], or neutrinos [6], makes it possible to obtain new fundamental knowledge of the high-energy Universe.

Multi-messenger astronomy [7, 8] is actively developing over the past few years, implying collaboration between individual experiments in astrophysics of high and ultra-high energies. At the same time, conducting this joint research is still complicated although some technical solutions and methodology are yet to be developed [9].

Successful examples of distributed data and/or computing centers have already been established in astronomy [10, 11] and particle physics [12], and it is very important to take into account this experience and at the same time the specificity of the field when working on appropriate solutions for particle astrophysics.

In the framework of a German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI) [13], we approached the task of developing a working prototype of such a data centre via extension of the existing KASCADE Cosmic-ray Data Centre (KCDC) [14], initiated by the KASCADE [15] collaboration, with keeping in mind focus on the data life cycle (i.e. set of steps every item of data goes through from it's creation to is't archival) specific to astroparticle physics (see Fig. 1). Within our project we committed into fulfilling the following aims:

- **Open Science:** We are providing unlimited, barrier free, open access to experimental data, simulations, manuals, and outreach materials, and develop new tools for research and scientific collaboration in order to make scientific knowledge more easily accessible;
- **FAIR data:** Developing principles of data and metadata curation in a way the data which meet principles of findability, accessibility, interoperability, and reusability [16];
- **Flexibility** for small and middle size experiments: we aimed to provide standardized web access to the data of small and middle size experiments without making them changing their specific long-established practices, data processing methods and customized software.

In order to approach these aims, an investigation was carried out in the following directions: KCDC extension with data of more experiments; development of distributed data storage; development of data analysis methods, both employing analytical and deep learning approaches; outreach

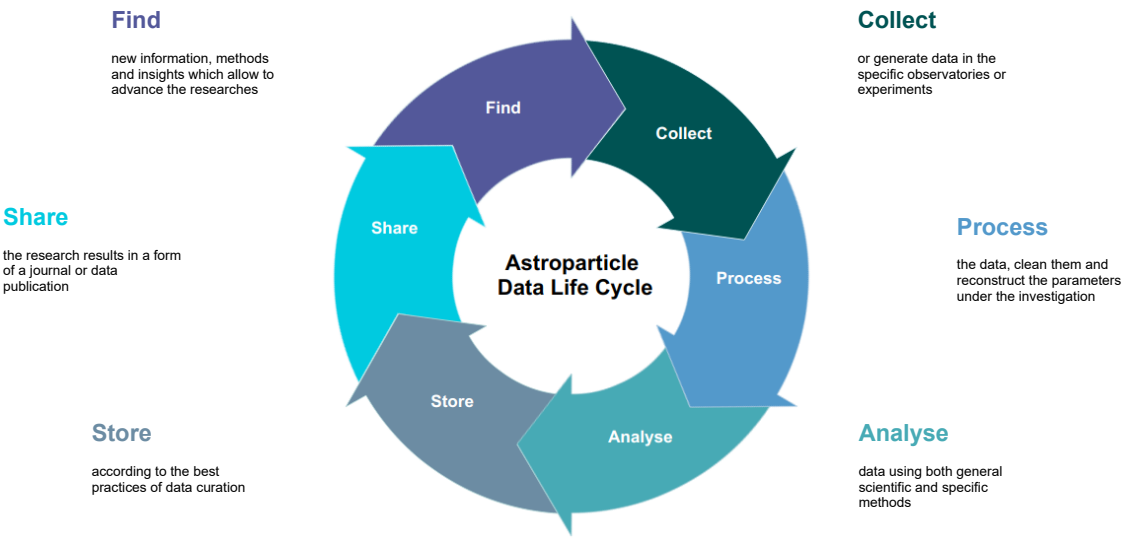


Figure 1: Data Life Cycle in Astroparticle Physics

and education. In the next sections we describe the progress done in the mentioned directions and then summarize the overall project experience.

2. Infrastructure for distributed data curation

The development of the global data center prototype covers unification of metadata, development of APIs for remote storages, scrupulous study of data structure at different reconstruction levels, middleware development and integration of the developed services with the KCDC platform.

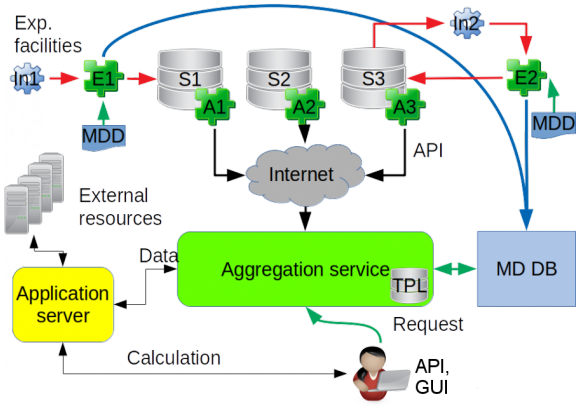


Figure 2: Architecture of GRADLCI distributed platform.

A general (simplified) data center scheme is shown in Fig. 2. Its crucial components are: remote data storages S_i, aggregation server, application server, metadata database (MDDB), data sources In_i, as well as related software - adapters A_i and metadata extractors E_i.

The remote data storages connected to the system provide access to data of such astroparticle-physics experiments as: KASCADE, KASCADE-Grande, LOPES, Tunka-Rex, Tunka-133, Maket-Ani. One gets access to the data by requesting them from the aggregation server via its API. Then the aggregation server requests MDDB, where metadata of all the events are collected. Every event in the system is identified by its UUID and can be directly acquired from the remote storage by this parameter. Introduction of metadata search helped to speed up data retrieval up to 35-40% for particular requests. The metadata catalog supports file and event-level data search, which allows to handle data of different reconstruction level. The MDDB is kept up-to date using metadata extractors retrieving new event metadata from remote storages. Interaction with various remote storages is unified using adapters. Users can interact with the system using either Web GUI or API [17].

The combination of the services developed resulted in a working prototype of a cloud storage system, showing efficiency for small and medium-sized astroparticle-physics experiments [18].

3. KCDC extension

In order to adapt KCDC to become a significant part of the prototype of the envisaged Global Analysis and Data Center in Astroparticle Physics and improve user's experience, several big changes has been made. We included an update of the data center with a JupyterHub analysis environment [19], development of the RESTless API for access to the KASCADE datashop as well as extending available datashops, simulations, spectra, educational materials as well as corresponding manuals.

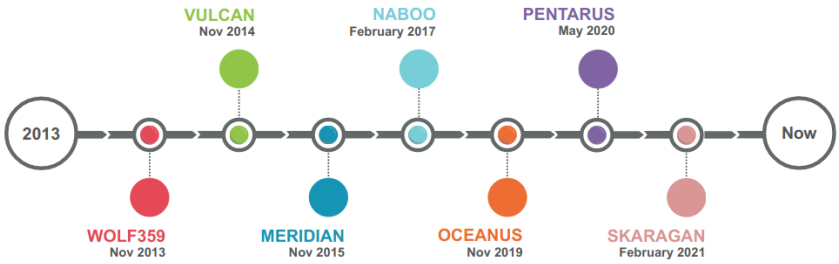


Figure 3: Timeline of KCDC releases.

The newly introduced datashops include a COMBINED datashop, representing calibrated data from detectors KASCADE and GRANDE, allowing more thorough investigation of elemental composition of high-energy cosmic rays, and a Maket-Ani datashop, containing data from Maket-Ani [20], an extensive air shower experiment, which was located on Mt. Aragats (Aragats Cosmic Ray Observatory, Armenia).

The mentioned improvements were introduced in the OCEANUS, PENTARUS, SKARAGAN KCDC releases (Fig. 3) and were followed by many significant internal updates and modifications, which one can inquire in [14] as well as in KCDC Changelogs [21].

4. Analysis of data from multiple sources

Various machine learning techniques were adapted for analysis of extensive air showers (EAS) detected by the TAIGA and KASCADE experiments. They were employed for such tasks as identification of primary particle type, reconstruction of spectrum mass composition, evaluation of EAS energy, and fast simulation generation. Training datasets were obtained using the CORSIKA program [22]. Software implementation was developed using the frameworks TensorFlow [23–25], PyTorch [23, 26], and sklearn [26].

In the study [23], the primary particle for TAIGA IACT data was identified using convolutional neural networks (CNN). The recognition quality parameter of the developed CNN equals Q = 2.7–3.0, which is significantly higher than Q = 1.7 provided by the classification methods based on analysis of Hillas parameters.

The work [24] was dedicated to the estimation of the EAS energy. The developed technique was shown to have an error of 20–25 %, compared to approximately 50 % one for traditional methods. The method was generalized for the simultaneous analysis of data from several Cherenkov telescopes (stereo mode), further reducing the error to only 13–15%, which is a significant improvement in analysis techniques for astroparticle physics.

In the work on analyzing KASCADE data [26], the issues of determining the primary particle and reconstructing the mass composition of the spectrum were investigated using machine learning methods, like DecisionTree and RandomForest. The accuracy of gamma-hadron separation up to 95 % was achieved on simulation data.

In the research [25], a generative adversarial network (GAN) was trained on a sample for protons and for gammas, each one comprising about 25.000 events, which took approximately 12 hours on a Tesla P100 GPU. After that, it takes about 1 second to generate 400 events, that is over 1000 times faster than generation using CORSIKA. The quality estimation using third-party software has shown that 85.7 % of GAN-generated gamma events were identified correctly, while 4.4 % of generated protons were recognized as gammas.

5. Outreach and education

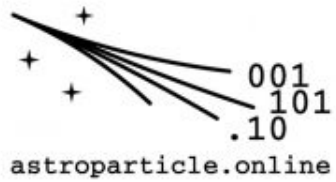
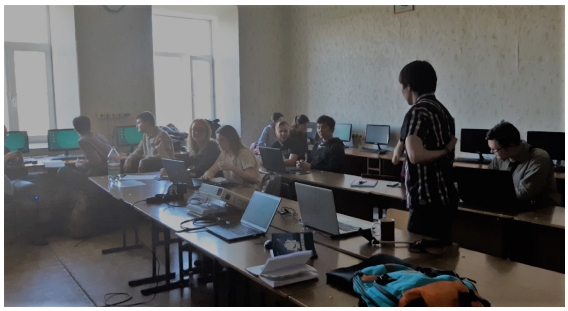


Figure 4: Logo of astroparticle.online.

Outreach activities were organized within two web sites: KCDC and astroparticle.online. The second one was established specifically to advance outreach initiatives in multi messenger astronomy (in Russia), inform publicity about current developments in the field and implicate young students into advanced research work.

The publications, lectures, talks and laboratory works developed for the project have not only become part of the web portal, but have also been used in the IGU course "Introduction to Astroparticle Physics" as well as in tutorials at workshops and conferences for young scientists, such as the ISSAP Baikal Summer School and Data life cycle in physics (DLC) (Fig. 5(b)). In total, more than 300 students took part in educational events organized by the projects [27].

(a)



(b)

Figure 5: a) ICD-18 participants are working on data analysis; b) ISU students are analyzing open data as part of an event organized by the astroparticle.online portal.

Interactive materials were also developed for the project, such as laboratory works involving data processing via programming in Python or C++ languages and a CNN-based particle identification application (see Section 4 for details). Collaboration with the KCDC team played an important role in organizing both the technical support for the site and the content creation.

The development of educational materials for the KCDC portal was largely associated with practical exercises on data analysis organized at KIT—in particular, with the participation of IAP KIT in such annual events as the International Cosmic-ray Day (see Fig. 5) and Woche der Teilchenwelt. New approaches to organize specific masterclasses were also applied in 2020 to conduct an online masterclass [28] using Zoom, KASCADE data from the cloud environment developed within the framework of the project, and JupyterHub as a platform for the online data analysis.

6. Summary

Multi-messenger astronomy is one of scientific fields that already operates petabytes of data. Providing the correct and, most importantly, user-friendly access to such amounts of data is one of the key tasks in the field of big data analysis.

In this project, we aimed at managing big data in multi-messenger astronomy, developing tools for analyzing these data (in particular, using machine learning), and informing the public about the achievements of multi-messenger astrophysics with scientific and popular-science publications.

The results obtained can be taken into account and expanded in further work on creating data centers that support the concepts of open science and FAIR data, and with reasonable amendments can also be extended for use in other fields of science.

Acknowledgments

This work was financially supported by Russian Science Foundation Grant 18-41-06003 and the Helmholtz Society Grant HRSF-0027.

References

[1] W. Apel et al., *The spectrum of high-energy cosmic rays measured with KASCADE-Grande*, *Astroparticle Physics* **36** (2012) 183.

[2] A.V. Olinto, *Cosmic rays: The highest-energy messengers*, *Science* **315** (2007) 68 [https://science.sciencemag.org/content/315/5808/68.full.pdf].

[3] A. Aab et al., *A targeted search for point sources of EeV neutrons*, *The Astrophysical Journal* **789** (2014) L34.

[4] D. Horns, *Gamma-ray astronomy from the ground*, *Journal of Physics: Conference Series* **718** (2016) 022010.

[5] J. Knödlseder, *The future of gamma-ray astronomy*, *Comptes Rendus Physique* **17** (2016) 663.

[6] M. Tluczykont et al., *Connecting neutrino astrophysics to multi-TeV to PeV gamma-ray astronomy with TAIGA*, in *Proceedings, Magellan Workshop: Connecting Neutrino Physics and Astronomy*, S.K. Dahmke, M. Meyer and L. Vanhoefer, eds., DESY-PROC, (Hamburg), pp. 135–142, Verlag Deutsches Elektronen-Synchrotron, 2016, DOI.

[7] I. Bartos and M. Kowalski, *Multimessenger astronomy*, in *Gamma-Ray Bursts*, IOP Publishing (2017).

[8] E. Burns, A. Tohuvavohu, J. Bellovary, E. Blaufuss, T. Brandt, S. Buson et al., *Opportunities for multimessenger astronomy in the 2020s*, *arXiv preprint arXiv:1903.04461* (2019) .

[9] F. Mazzocchi, *Scientific research across and beyond disciplines: Challenges and opportunities of interdisciplinarity*, *EMBO reports* **20** (2019) e47682.

[10] G.B. Berriman, T.I. Committee, T.I. Group and T.I. Community, *The International Virtual Observatory Alliance (IVOA) in 2020*, *arXiv preprint arXiv:2012.05988* (2020) .

[11] P. Tallada, J. Carretero, J. Casals, C. Acosta-Silva, S. Serrano, M. Caubet et al., *Cosmohub: Interactive exploration and distribution of astronomical data on Hadoop*, *Astronomy and Computing* **32** (2020) 100391.

[12] J. Shiers, *The worldwide LHC computing grid (worldwide LCG)*, *Computer physics communications* **177** (2007) 219.

[13] I. Bychkov, A. Demichev, J. Dubenskaya, O. Fedorov, A. Haungs, A. Heiss et al., *Russian–German astroparticle data life cycle initiative*, *Data* **3** (2018) 56.

[14] A. Haungs, D. Kang, K. Link, F. Polgart, V. Tokareva, D. Wochele et al., *Status and future prospects of the KASCADE cosmic-ray data centre KCDC*, in *37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07. 2021*, 2021.

[15] T. Antoni, W. Apel, F. Badea, K. Bekk, A. Bercuci, H. Blümer et al., *The cosmic-ray experiment KASCADE*, *Nuclear Instruments and Methods in Physics Research Section A: accelerators, spectrometers, detectors and associated equipment* **513** (2003) 490.

[16] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak et al., *The FAIR guiding principles for scientific data management and stewardship*, *Scientific data* **3** (2016) 1.

[17] V. Tokareva, A. Haungs, D. Kang, F. Polgart, D. Wochele and J. Wochele, *Data aggregation platform for experiments of astroparticle physics*, in *DLC 2020: Proceedings of the 4th International Workshop on Data Life Cycle in Physics. Ed.: A. Kryukov*, p. 134, 2020.

[18] A. Kryukov, I. Bychkov, E. Korosteleva, A. Mikhailov and M.-D. Nguyen, *AstroDS–A distributed storage for astrophysics of cosmic rays. Current status*, *arXiv preprint arXiv:2010.04938* (2020) .

[19] F. Polgart, A. Haungs, D. Kang, D. Wochele, J. Wochele and V. Tokareva, *An analysis framework for KCDC*, in *DLC 2020: Data Life Cycle in Physics: Proceedings of the 4th International Workshop on Data Life Cycle in Physics, Moscow, Russia, June 8-10, 2020*, p. 111, 2020.

[20] A. Chilingarian, G. Gharagyozyan, S. Ghazaryan, G. Hovsepyan, E. Mamidjanyan, L. Melkumyan et al., *Study of extensive air showers and primary energy spectra by MAKET-ANI detector on mountain Aragats*, *Astroparticle Physics* **28** (2007) 58.

[21] "KCDC Announcements - ChangeLogs." https://kcdc.iap.kit.edu/announcements/changeLogs/.

[22] D. Heck, J. Knapp, J. Capdevielle, G. Schatz, T. Thouw et al., *CORSIKA: A Monte Carlo code to simulate extensive air showers*, *Report fzka* **6019** (1998) .

[23] E. Postnikov, A. Kryukov, S. Polyakov, D. Shipilov and D. Zhurov, *Gamma/hadron separation in imaging air cherenkov telescopes using deep learning libraries tensorflow and pytorch*, *Journal of Physics: Conference Series* **1181** (2019) 012048.

[24] E. Postnikov, A. Kryukov, S. Polyakov and D. Zhurov, *Deep learning for energy estimation and particle identification in gamma-ray astronomy*, in *CEUR Workshop Proceedings*, pp. 90–99, 2019.

[25] J. Dubenskaya, *Modeling images of proton events for the TAIGA project using a generative adversarial network: features of the network architecture and the learning process*, in *The 5th International Workshop on Deep Learning in Computational Physics (DLCP-2021)*, 2021.

[26] D. Kostunin et al., *New insights from old cosmic rays: A novel analysis of archival KASCADE data*, in *37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07. 2021*, 2021.

[27] V. Tokareva, Y. Kazarina, V. Samoliga, A. Haungs, A. Kryukov, E. Postnikov et al., *Multi-messenger astroparticle physics for the public via the astroparticle.online project*, in *37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07. 2021*, 2021.

[28] K. Link, V. Tokareva, A. Haungs, D. Kang, P. Koundal, F. Polgart et al., *Online masterclass built on the KASCADE cosmic ray data centre*, in *37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07. 2021*, 2021.