Peer-reviewed version available at *Data* **2018**, *3*, 56; doi:10.3390/data3040056

Article

Russian-German Astroparticle Data Life Cycle Initiative

Igor Bychkov[1,2], Andrey Demichev[3], Julia Dubenskaya[3], Oleg Fedorov[4], Andreas Haungs[5], Andreas Heiss[6], Yulia Kazarina[4], Elena Korosteleva[3], Dmitriy Kostunin[5], Alexander Kryukov[3], Andrey Mikhailov[1], Minh-Duc Nguyen[3], Stanislav Polyakov[3], Evgeny Postnikov[3], Alexey Shigarov[1,2], Dmitry Shipilov[4], Achim Streit[6], Viktoria Tokareva[5], Doris Wochele[5], Jürgen Wochele[5], Dmitry Zhurov[4]

1 Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of Russian Academy of Sciences, Irkutsk, Russia
2 Irkutsk State University, Irkutsk, Russia
3 Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia
4 Applied Physics Institute, Irkutsk State University, Irkutsk, Russia
5 Institute for Nuclear Physics, Karlsruhe Institute of Technology, Karlsruhe, Germany
6 Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Karlsruhe, Germany
* Correspondence: editor@astroparticle.online

Abstract: Modern experimental astroparticle physics features large-scale setups measuring different messengers, namely high-energy particles generated by cosmic accelerators (e.g. supernova remnants, active galactic nuclei, etc): cosmic and gamma rays, neutrinos and recently discovered gravitational waves. Ongoing and future experiments are distributed over the Earth including ground, underground/underwater setups as well as balloon payloads and spacecrafts. The data acquired by these experiments have different formats, storage concepts and publication policies. Such differences are a crucial issue in the era of big data and of multi-messenger analysis strategies in astroparticle physics. We propose a service ASTROPARTICLE.ONLINE in the frame of which we develop an open science system which enables to publish, store, search, select and analyse astroparticle physics data. The cosmic-ray experiments KASCADE-Grande and TAIGA were chosen as pilot experiments to be included in this framework. In the first step of our initiative we will develop and test the following components of the full data life cycle concept: (i) describing, storing and reusing of astroparticle data; (ii) software for performing multi-experiment and multi-messenger analyses like deep-learning methods; (iii) outreach including example applications and tutorial for students and scientists outside the specific research field. In the present paper we describe the concepts of our initiative, and in particular the plans toward a common, federated astroparticle data storage.

Keywords: astroparticle physics, cosmic rays, data life cycle management, data curation, meta data, big data, deep learning, open data

1. Introduction

Research in astroparticle physics addresses some of the most fundamental questions in nature. There is an intimate connection between measurements and theoretical descriptions of astrophysical phenomena to provide the foundation for the sophisticated models of macroscopic astrophysical systems. Scientists between experiments from ground-based and space-based devices have to share their incredibly detailed observations to study processes in astrophysical environments. Moreover, information from various messengers, like charged particles, gamma-rays or neutrinos, measured by different large-scale facilities globally distributed, should be combined to obtain increased knowledge of the high-energy Universe. For that, also named as multi-messenger astroparticle physics, a diverse set of astrophysical data is required to be made available and public.

The current trend, not only in astroparticle physics but also in particle physics is that people from all over the world can use data as soon as they are posted, and scientists can immediately download scientific data. This trend demonstrates from the initial step the power of the Internet and the ability of the scientific community to share data quickly with other colleagues and with the general public. Some experiments in astroparticle physics have already adopted this fascinating idea and they have involved their scientific data in electronic publishing, such as KASCADE Cosmic ray Data Centre (KCDC) [1]. KCDC, presently in its beta-phase, is a web portal where the KASCADE-Grande [2] scientific data is made available for the interested public. However, KCDC is a small project, driven within the KASCADE-Grande experiment, only. In Russia, there is the operating Tunka Advanced Instrument for cosmic rays and Gamma Astronomy (TAIGA) [3] facility, which continuously produces data. There are many scientific reasons to combine the TAIGA and KASCADE-Grande data to perform coherent analyses with sophisticated methods (e.g. deep learning). For such a next step in big data analytics for data from different facilities the information and computing infrastructure is still not available.

This paper presents the current status of the Russian-German astroparticle data life cycle initiative named as ASTROPARTICLE.ONLINE. The initiative strives to develop an open science system to be able to collect, store, and analyze astrophysical data having the TAIGA and KASCADE-Grande experimental facilities as initial data providers. The project, ASTROPARTICLE.ONLINE, aims at a common data portal of two independent observatories and at the same time for a consolidation and maturation of an analysis and data centre of astroparticle physics experiments. There are four main goals of the project:

1. KCDC extension: the already existing data centre released an initial data set of parameters of more than 400 million extensive air showers of the concluded KASCADE-Grande experiment. The initiative extends KCDC by more scientific data from the TAIGA experiment, i.e. current data, so to say up-to-the-date data, allowing on-the-fly multi-messenger-analysis. Our goal is to extend and improve KCDC and make it more attractive to a broader user community.
2. Big Data Science software: such an extension of the data centre allowing not only access to the data but also the possibility of developing specific analysis methods and corresponding simulations in one environment needs a move to most modern computing, storage and data access concepts, which is only possible by a close co-operation between the participating groups from both, physics and information technology. A possible concept to reach this goal is the installation of a dedicated so-called “data life cycle lab”, where this project is aiming for. Dedicated access, storage, and interface software have to be developed.
3. Reliability tests: some specific analyses of the data provided by the new data centre will be performed to test the entire concept. This will give important contributions and confidence to the project as a valuable scientific tool.
4. Go for the public: the full outreach part of the project, including example applications for all level of users, from pupils to the directly involved scientists to theoreticians, with detailed tutorials and documentation is an important goal of the project.

The novelty of the proposed approach is reflected in developing integrated solutions including:

- Distributed data storage algorithms and techniques with a common meta-data catalog to provide a common information space of the distributed repository;
- Data transmission algorithms as well as simultaneous data transmission from several data repositories thus significantly reducing load time;
- Deep-learning techniques for identifying mass groups of impinging cosmic particles and their properties in a fully remote access mode;
- KCDC-based prototype system of Big Data analysis and exporting the experimental data from KASCADE-Grande and TAIGA to test the technology of data life cycle management.

Peer-reviewed version available at Data 2018, 3, 56; doi:10.3390/data3040056

- An educational system based on the HUBzero[1] platform dedicated to astroparticle physics.

2. Concept of Astroparticle Data Life Cycle

Nowadays, the exponential growth of the amount of experimental data can be observed. While there was 1-10 Tb of data per year in astrophysics 10-15 years ago, new experimental facilities generate datasets ranging in size from 100's to 1000's of terabytes per year. It can be illustrated by a growth of the amount of data acquired by satellites. While the Integral satellite [4] downloaded to the ground 1.2 Gb of data per day in 2002, now the Gaia spacecraft [5] transfers about 5 Gb of data per day. The other example is the ground-based experiment LSST [6], providing over 3 gigapixels per image with an exposure every 15 seconds. It is expected to produce about 10 petabytes of information per year.

These trends give rise to a number of emerging issues of a big data management. Obviously, various activities should be performed continuously across all stages of the data life cycle to support effective data management: the collection and storage of data, its processing and analysis, refining the physical model, making preparations for publication, and data reprocessing taking refinement into account. An important topic for modern science in general and astroparticle physics, in particular, is open science, the model of free access to data (e.g. [7]): data are accessible not solely to collaboration members but to all levels of an inquiring society, amateur or professional. This approach is especially important in the age of Big Data and Open Science Culture, when complete analyses of the experimental data cannot be performed within one collaboration.

Usually, basic research in the field of particle physics, astroparticle physics, nuclear physics, astrophysics, or astronomy is performed in large international collaborations with partly huge infrastructures producing a big volume of valuable scientific data. To efficiently use all the information to solve the still mysterious question about the origin of matter and the Universe, a broad, simple and sustainable access to the scientific data from these infrastructures has to be provided.

In a general way, such a global data centre should provide a vast of functionality, at least covering the following pillars (see Fig. 1):

1. Data availability: all participating researchers of the individual experiments or facilities needs fast and simple access to the relevant data.
2. Simulations and methods development: to prepare the analyses of the data the researchers need a mighty environment on computing power for the production of relevant simulations and the development of new methods, e.g. by deep machine learning.
3. Analysis: a fast access to the (probably distributed) Big Data from measurements and simulations is needed.
4. Education in Data Science: the handling of the data centres as well as the processing of the data needs specialized education in "Big Data Science".
5. Open access: more and more important is to provide the scientific data not only to the internal research community but also to the interested public.
6. Data archive: the valuable scientific data need to be preserved for a later reuse.

Whereas in astronomy and particle physics data centres are already established, which fulfill a part of these requirements (although, not the same parts), in astroparticle physics only first attempts are presently under development. The reason is the diversity of the experimental facilities in astroparticle physics and their distribution all over the world (partly in really harsh environments), without dedicated research centres like CERN[2] in particle physics, FAIR[3] in nuclear physics, or ESO[4] in astronomy.

---

1 https://hubzero.org
2 https://home.cern
3 https://fair-center.eu
4 https://eso.org

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2018 doi:10.20944/preprints201810.0273.v1

Peer-reviewed version available at Data 2018, 3, 56; doi:10.3390/data3040056

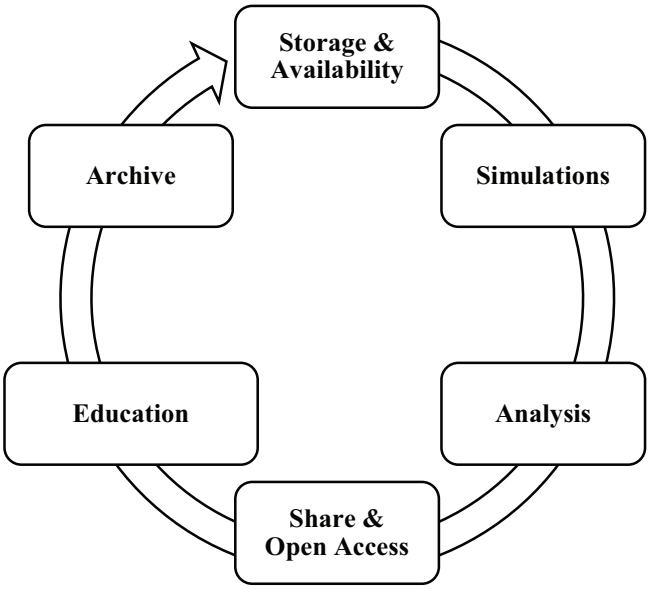

Figure 1. Concept of an Astroparticle Data Life Cycle

Presently, astroparticle physicists do have access or provide to its community several attempts in individual pillars. Namely, a small part of the Tier computing centres is used by astroparticle physics experiments [8], e.g. smaller 5% of GridKa[5] by the Pierre Auger Collaboration [9] for simulations. In addition, first IceCube [10] or Pierre Auger data can be found in the Astronomical Virtual Observatories, like GAVO[6]. The later explained KCDC is an example for a first public release of scientific data. However, these attempts are uncoordinated and mostly specific to individual experiments or collaborations.

It is obvious that astroparticle physics has become a data-intensive science with many terabytes of data and often with tens of measured parameters associated with each observation. Moreover, new highly complex and massively large datasets are expected by novel and more complex scientific instruments as well as simulated data needed for interpretation that will become available in the next decades, probably largely used by the community. Handling and exploring these new data volumes, and actually making unexpected scientific discoveries, poses a considerable technical challenge that requires the adoption of new approaches in using computing and storage resources and in organizing scientific collaborations, scientific education and science communication, where sophisticated public data centres will play the key role.

The methods for a successful performance of the initiative are a mixture of (i) computer scientist's work, i.e. sophisticated programming and usage of modern tools to handle Big Data. This is the major focus of the project and must not be underestimated, as there is (not yet) a standard tool to do it. The project's idea is subject of a relatively new field of research as Big Data handles large or complex data sets where traditional data processing applications are difficult applicable.

Typical challenges include the search, sharing, storage, transfer, or visualization of the data. In information sciences, Big Data is related to the use of predictive analytics or certain other advanced methods to extract scientific value from data. (ii) Astroparticle physics, i.e. understanding the methods of detecting, reconstructing and interpreting particles coming from the deep Universe to Earth as a valuable application of the challenges met in (i). (iii) Outreach tasks and usage of social media, as a public data centre approaching the entire society needs a distinct plan of dissemination.

---

5 http://www.gridka.de
6 http://www.g-vo.org

The FAIR Data Principles are the supreme guideline for all the research data management issues within this project. The FAIR principles [7] intend to provide guidelines for improving usability of digital assets: Findable: The first step in (re-)using data is to find them. Accessible: Once the user has found the required data, he/she must know how to be integrated with them, possibly including authentication and authorization. Interoperable: The data usually has to be integrated with other data. Reusable: The goal of FAIR is to optimize the reuse of data. In order to achieve this, metadata and data should be well described so that they can be used in different ways.

2.1. Storage and Availability

A major goal of the project is to provide scientists with data on requests. The request is a set of conditions and logical operations on them which define what kind of the data the user wants to obtain. All requests proceed by using the metadata information only via special metadata servers. A search within the data will not be available. If one needs to carry out more sophisticated requests, the appropriate information must be extracted from the data and inserted into the meta registry.

The data itself is stored on some local data storage which collecting raw data from the astroparticle facilities such as TAIGA, KASCADE-Grande, etc. Each storage has its own format of data storing, directory structure and policy. We do not touch the internal structure of the storage and traditions of the physics community. Therefore, to provide access to the data storage, it is necessary to deploy a (RESTful) services which will unify the external interface for all the storages. We call such a service an adapter.

One of the candidates for the adapter is the CERNVM-FS[8] service. This service provides export of a local file system over the Internet in read-only mode. For tasks of the data analysis, this mode is sufficient. From the end-user point of view, the set of data storage looks like a local file system of a single virtual storage.

2.2. Simulations

Simulations are one of the important stages in modern experiments. They require a lot of computing resources and produce data volume which is comparable with the volume of raw data. Usually, simulations prove to be "data factory" similar to the experimental facilities. So, we propose to consider the simulations as a specific source of data (like experimental facilities). Thus simulation data should be uploaded to the particular storage by special service.

2.3. Analysis

The next pillar of the data life cycle is the data analysis itself. This task requires the delivery of requested data, access to computing facilities and software for the analysis. There are two main approaches to data analysis in physics: conventional analysis and machine learning. In the first case, a user implements an algorithm which is inspired by the physical model of the phenomena under consideration. In the second case, one uses an artificial neural network technique with supervised or unsupervised learning. For the time being this technique, which has proven its efficiency for image recognition, is actively developed by different experiments equipped with telescopes, particularly by TAIGA for its Imaging Atmospheric Cherenkov Telescopes (IACT) [11].

2.4. Share and Open Access

The open access to the data is provided by the standard way under specially formulated access policy. The policy depends on the local policy of integrated storages which are data owners.

---

[7] https://www.ncbi.nlm.nih.gov/pubmed/26978244
[8] https://cernvm.cern.ch/portal/filesystem

2.5. Education in Data Science

We are going to achieve this target by using special service (web portal) based on the HUBzero platform. The platform supplies users with education courses, documentation and exercises on Monte Carlo simulation, examples of data analysis, introduction to the principle of metadata, and so on.

2.6. Archive

This target will be achieved by provenance tracking of the data. This tracking must store a full history of the data starting from the initial uploading. The history should include who and when processed the data, what kind of software was used and their versions, what kind of calibration was used, etc. It should provide also a fast check of the data consistency. For example, the system have to alarm if two chunks of data are processed under different calibration conditions. For this purpose, the Merkle trees may be used. It is possible also to pack old data and upload to off-line storage like tapes. However, we do not suppose to solve this task in full scope.

2.7. KCDC extension

KCDC is an already existing web portal, where data of the KASCADE-Grande experiment are made available for the interested public, i.e. the methodological concept for this kind of data centers is already developed. The web portal uses modern technologies, including standard internet access and interactive data selections. However, even if the primary target is the user community or the "any interested scientist", both the data and the tools have to be refurbished in order to be usable without the detailed and highly specialized knowledge that is currently only available within the internal collaborations.

The research plan in order to reach the project's goals in terms of providing Big Data Science includes sophisticated methods and tools: the development of adapted distributed system for big data analysis as well as its implementation at the large computing facilities in SINP MSU and SCC KIT. Then the system needs to provide as fast and reliable user access to the full dataset. A fast data exchange is foreseen to be reached via caching filesystems CVMFS and microservice technology using REST architecture. Further, the development of algorithms for the big data analysis of astrophysical experiments, particularly using machine learning, has to be pursued and support of the soft- and hardware for the full data life cycle will be given using, for example, the blockchain technology.

We propose to extend KCDC and even generalize Cosmic Ray Data Centres in order to preserve the intellectual value of the experiments and to further exploit its scientific contents beyond the lifetime of the operation of the instruments. We think that a full return from the collaborations back to the society that has funded this endeavor can best be achieved if the original data and the accompanying software tools are made publicly available in an open-access manner. The data centre(s) will offer great and unique opportunities to people that would not be able to access such data otherwise and will also provide a basis for education and outreach to the general public. This demand is at the heart of Big Data Science. The amount of work needed to install and run such a web portal providing data from two independent experimental facilities with international collaborations should not be underestimated. Constant improvements of the availability and usability are needed. In addition, only a small part of the available data has been made available until now. Adding the remaining detector components will require to process the raw data and to update the documentation to cover the added observables. To enhance the usability of the data, the extensive set of simulations may have to be added, too.

KCDC was implemented as a plugin-based framework to ensure an easy way to adjust, exchange or remove components as needed. Before the software can be released as open source, however, the coverage of the code by functional and unit tests has to be significantly improved. In addition, while there is a lot of documentation on the published data available, there is almost no documentation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2018 doi:10.20944/preprints201810.0273.v1

Peer-reviewed version available at Data 2018, 3, 56; doi:10.3390/data3040056

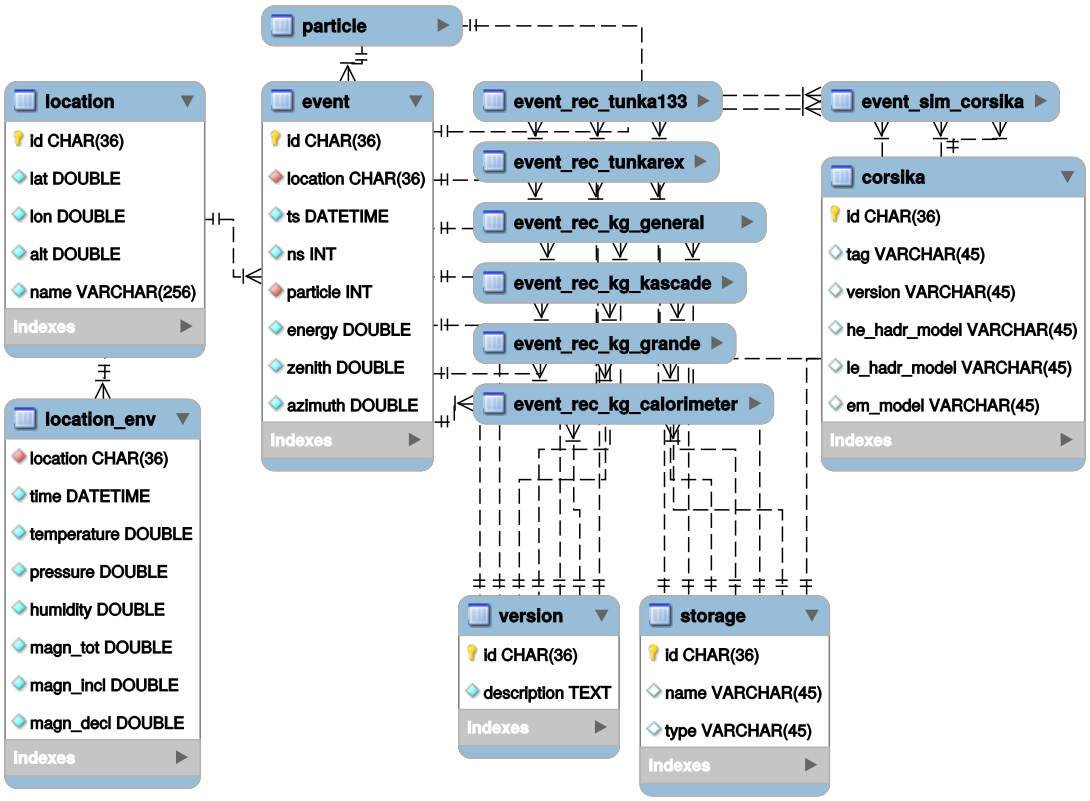

Figure 2. Proposed design of a database containing event metadata. The structure and plot are produced using MySQL Workbench software.

on the usage of the software itself as well as on the development of the software. Although it has been kept intuitive, the extensive possibility to configure the web portal via an admin web interface makes such a detailed documentation necessary. Once published, user monitoring and feedback have to be taken into account to further improve the software. The possibility to include plugins and patches implemented by users has to be considered. Legal issues on ownership of the data have to be considered not to hurt the rules of the collaborations.

3. Preliminary results

3.1. Metadata architecture for cosmic-ray experiments

Since the data size is huge and its structure is diverse, a direct search within the data would be extremely slow and resource-consuming, and thus is not going to be implemented. Fortunately, the data have a common metadata format, which includes time, place, atmospheric conditions, etc. A centralized database containing the metadata of all events from both experiments would be used to process data-retrieval requests. The proposed database structure is presented in Fig. 2 [9]. In case any kind of requests use properties not included in this database, the appropriate information must be extracted from the data and inserted into the metadata registry.

[9] https://www.mysql.com/products/workbench/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2018 doi:10.20944/preprints201810.0273.v1

Peer-reviewed version available at Data 2018, 3, 56; doi:10.3390/data3040056

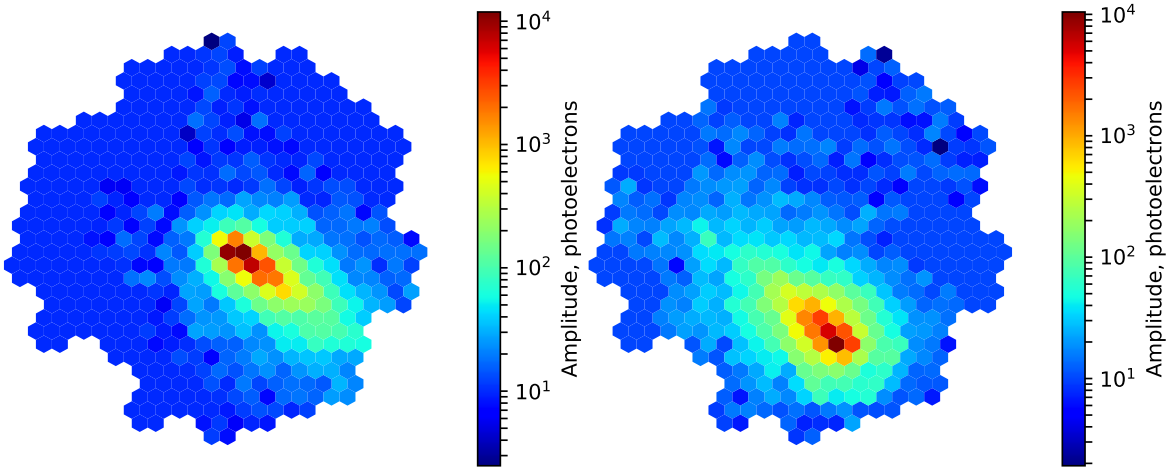

Figure 3. Examples of the TAIGA-IACT simulation images: gamma-ray (left panel) and background particle (proton, right panel). According to conventional techniques of gamma-ray identification, image dimensions and orientation depend on the particle type.

3.2. Data Analysis

We followed machine learning approach to data analysis by the example of Monte Carlo simulation for the TAIGA-IACT imaging air Cherenkov telescope [12]. Our choice was to apply the convolutional neural network (CNN) technique to solving the problem of gamma-ray identification. CNN is a kind of artificial neural network that uses a special architecture which is particularly well-adapted to classify images. Today, CNN is used in most neural networks for image recognition. However, only three CNN-related attempts have been made in IACT data analysis, all of them in the last two years: muon image identification for the VERITAS telescope [13], gamma-ray identification for the Monte Carlo simulation of a standalone telescope in the CTA upcoming project [14], and gamma-ray identification for the stereoscopic mode of the four H.E.S.S. telescopes [15].

In our CNN approach we used Monte Carlo simulation for the TAIGA-IACT telescope. Datasets of gamma-ray images and hadron background (proton) images were simulated for conditions of real observations (Fig. 3). They were split into two parts for learning and testing, and various CNN versions were trained using the PyTorch [16] and TensorFlow [17] software packages. CNN performance estimation was a blinded study: a random proportion of test samples (blind samples) was used to estimate identification quality.

The quality of identification allows suppressing background (proton) events by a factor of 30 while keeping 55% of true gamma-ray events. It's much better than the quality after a simple conventional analysis (a system of 2 consecutive cuts), which allows suppressing proton events only by a factor of 8. After this technique has been improved and verified by experiment, it will become part of the dedicated software for data analysis within the project.

3.3. Educational Resources

The HUBzero platform for the educational issues in astroparticle physics have been deployed on the servers of Matorosov Institute for System Dynamics and Control Theory[10]. Currently, the educational resources are under development and it is being filled with the actual documentation, educational courses, data collections, tools for simulation and data analysis. The first experience of the application of this educational resource as a collaboration framework was received at the

[10] http://net.icc.ru

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2018 doi:10.20944/preprints201810.0273.v1

Peer-reviewed version available at Data 2018, 3, 56; doi:10.3390/data3040056

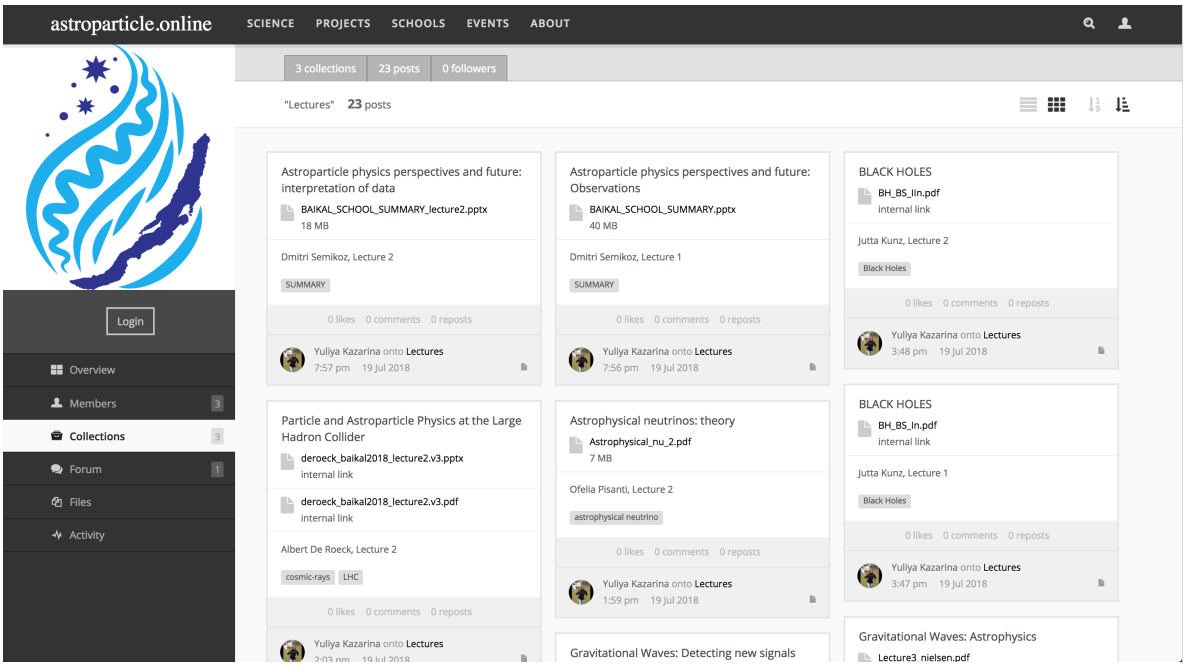

Figure 4. Screen shot of education materials of the ISAPP-Baikal Summer School held in 2018. It can be accessed via https://astroparticle.online/groups/bss.

ISAPP-Baikal Summer School[11] "Exploring the Universe through multiple messengers". Due to lack of the Internet connection at the location of Baikal Summer School, it was proposed to deploy the educational resource locally. This allowed the organizers of the school to spread the conference materials, lectures, student reports on the site, so the participants had the opportunity to access the school materials online. Also, the participants could post their impressions by photos and video comments on the page of the school. After the school, all resources have been synchronized back with the online server. A screenshot of the web page with school materials can be found in Fig. 4.

3.4. Raw Binary Data Sharing and Reuse

One of the important issues is how to archive raw binary data to support their availability and reusing in future [18]. There are five binary file formats used currently in TAIGA projects. They provide a representation of raw data obtained from five TAIGA sub-facilities: gamma-ray setups TAIGA-HiSCORE and TAIGA-IACT [12], and cosmic-ray setups Tunka-133 [19], Tunka-Rex [20] and Tunka-Grande [21]. The long-term preservation of raw binary data as originally generated is essential for rerunning analyses and reproducing research results in future. In this case, the raw data needs to be well documented and accompanied by some readers (i.e. software for parsing these data).

Some of the state-of-the-art tools for formal describing binary data formats can provide a sufficient solution for the issues of raw astroparticle physics data documenting and parsing. We use two of them, Kaitai Struct[12] and FlexT[13] for describing TAIGA binary data formats formally, documenting, and parsing library generation. For example, Fig. 5 demonstrates a diagram for Tunka-133 file format specification presented in Kaitai Struct. As a result, we generated reader libraries on target languages (including C/C++, Java, Go, JS, and Python) for each format. The

11 https://astronu.jinr.ru/school/current
12 http://kaitai.io
13 http://hmelnov.icc.ru/FlexT

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2018 doi:10.20944/preprints201810.0273.v1

Peer-reviewed version available at Data 2018, 3, 56; doi:10.3390/data3040056

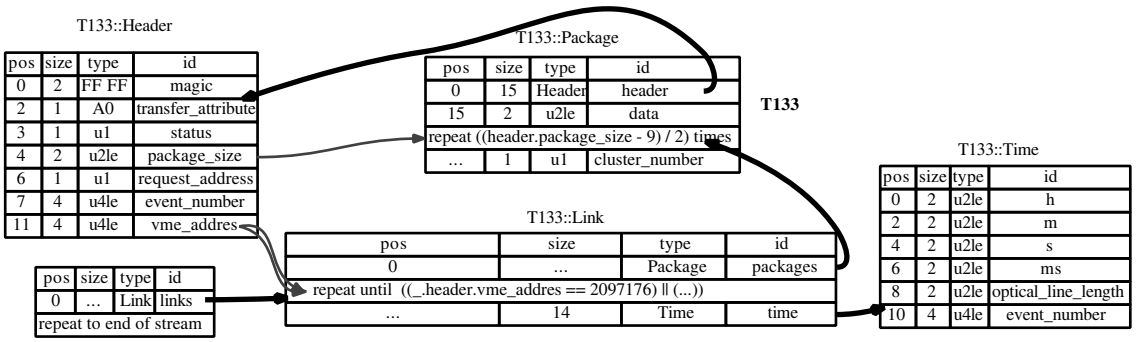

Figure 5. Tunka-133 file format specification presented in Kaitai Struct.

libraries were successfully tested on real TAIGA data, what also helped us to indicate the small fraction of corrupted data and fix it.

This result demonstrates the ways for describing binary file formats for astroparticle raw binary data sharing and reuse. It can be useful in other experiments too, where raw binary data formats remain weakly documented or some parsing libraries for contemporary programming languages are required. We suppose to use them for supporting the transmission of the raw data between web services we develop. They also can simplify the software development for data aggregation from various sources in the case of multi-messenger analysis.

4. Conclusions

The described initiative as a pilot project should have a significant long-term impact on the publication and release policies of future facilities in nuclear, particle and astroparticle physics. The resulting data centre and the experiences gained within this project will serve as a proof-of-principle that a public data centre opens the door to new methods of data analysis as well as to a new strategy of open science. In addition, it provides a concept for the required data release of forthcoming large-scale experiments in astroparticle physics, in particular as a dedicated facility in spanning over many different experiments.

This new and strategic approach for astroparticle physics is possible as KCDC is already now accepted by the community as a forerunner, but needs to be consolidated and matured in its scientific and technological performance to be ready for the global use. The present initiative is a necessary step in this direction. In this sense, the results of this project will validate the concept of a widely usable and public data centre in astroparticle physics.

We believe that our innovative approach will be used in astroparticle physics also beyond the present project. Plans are underway to expand the number of experiments by exporting data from other scientific collaborations. It will rapidly advance the research of fundamental properties of matter and the Universe. It is noteworthy that the suggested approach can be used not only in the specified field of science but also adapted to other scientific disciplines.

Acknowledgments: This work was financially supported by Russian Science Foundation Grant No. 18-41-06003 (Sections 2 and 3) and Helmholtz Society Grant HRSF-0027. The developed educational resources were freely deployed on the cloud infrastructure of the Shared Equipment Center of Integrated Information and Computing Network for Irkutsk Research and Educational Complex (http://net.icc.ru).

Author Contributions: Conceptualization, Andreas Haungs, Dmitriy Kostunin and Alexander Kryukov; Investigation, Oleg Fedorov, Elena Korosteleva, Andrey Mikhailov, Stanislav Polyakov, Evgeny Postnikov, Dmitry Shipilov and Dmitry Zhurov; Methodology, Igor Bychkov, Andrey Demichev, Julia Dubenskaya, Oleg Fedorov, Andreas Haungs, Andreas Heiss, Yulia Kazarina, Elena Korosteleva, Dmitriy Kostunin, Alexander Kryukov, Andrey Mikhailov, Minh-Duc Nguyen, Stanislav Polyakov, Evgeny Postnikov, Alexey Shigarov, Dmitry Shipilov, Achim Streit, Viktoria Tokareva, Doris Wochele, Jürgen Wochele and Dmitry Zhurov; Project

administration, Andreas Haungs and Alexander Kryukov; Resources, Igor Bychkov, Andreas Haungs, Yulia Kazarina, Andrey Mikhailov, Minh-Duc Nguyen, Stanislav Polyakov, Alexey Shigarov and Dmitry Shipilov; Software, Andrey Demichev, Julia Dubenskaya, Andreas Heiss, Dmitriy Kostunin, Alexander Kryukov, Andrey Mikhailov, Minh-Duc Nguyen, Stanislav Polyakov, Dmitry Shipilov, Achim Streit, Viktoria Tokareva, Doris Wochele, Jürgen Wochele and Dmitry Zhurov; Supervision, Andreas Haungs, Dmitriy Kostunin and Alexander Kryukov; Validation, Oleg Fedorov, Elena Korosteleva, Andrey Mikhailov, Stanislav Polyakov, Evgeny Postnikov, Dmitry Shipilov and Dmitry Zhurov; Writing – original draft, Andreas Haungs, Yulia Kazarina, Dmitriy Kostunin, Alexander Kryukov, Evgeny Postnikov and Alexey Shigarov; Writing – review editing, Andreas Haungs, Dmitriy Kostunin, Alexander Kryukov, Minh-Duc Nguyen and Evgeny Postnikov.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KASCADE: KArlsruhe Shower Core and Array DEtector
TAIGA: Tunka Advanced Instrument for cosmic rays and Gamma Astronomy
IACT: Imaging Atmospheric Cherenkov Telescope
HiSCORE: High-Sensitivity Cosmic ORigin Explorer
KCDC: KASCADE cosmic ray data centre
SCC KIT: Steinbuch Centre for Computing Karlsruhe Institute of Technology
SINP MSU: Skobeltsyn Institute of Nuclear Physics Lomonosov Moscow State University

References

1. Haungs, A.; others. The KASCADE Cosmic-ray Data Centre KCDC: Granting Open Access to Astroparticle Physics Research Data. *Submitted to: Eur. Phys. J. C* **2018**, [arXiv:astro-ph.IM/1806.05493].
2. Apel, W.D.; others. The KASCADE-Grande experiment. *Nucl. Instrum. Meth.* **2010**, *A620*, 202–216.
3. Budnev, N.; others. The TAIGA experiment: from cosmic ray to gamma-ray astronomy in the Tunka valley. *J. Phys. Conf. Ser.* **2016**, *718*, 052006.
4. Krivonos, R.; Revnivtsev, M.; Lutovinov, A.; Sazonov, S.; Churazov, E.; Sunyaev, R. INTEGRAL/IBIS all-sky survey in hard X-rays. *Astron. Astrophys.* **2007**, [arXiv:ASTRO-PH/astro-ph/0701836]. [Astron. Astrophys.475,775(2007)].
5. de Bruijne, J.H.J. Science performance of Gaia, ESA's space-astrometry mission. *Astrophys. Space Sci.* **2012**, *341*, 31–41, [arXiv:astro-ph.IM/1201.3238].
6. Abell, P.A.; others. LSST Science Book, Version 2.0 **2009**. [arXiv:astro-ph.IM/0912.0201].
7. David, P.A. Understanding the emergence of 'open science' institutions: functionalist economics in historical context. *Industrial and Corporate Change* **2004**, *13*, 571–589.
8. Berghöfer, T.; others. Towards a Model for Computing in European Astroparticle Physics **2015**. [arXiv:astro-ph.IM/1512.00988].
9. Aab, A.; others. The Pierre Auger Cosmic Ray Observatory. *Nucl. Instrum. Meth.* **2015**, *A798*, 172–213, [arXiv:astro-ph.IM/1502.01323].
10. Ahrens, J.; others. Icecube - the next generation neutrino telescope at the south pole. *Nucl. Phys. Proc. Suppl.* **2003**, *118*, 388–395, [arXiv:astro-ph/astro-ph/0209556]. [,388(2002)].
11. Postnikov, E.; others. Commissioning the joint operation of the wide angle timing HiSCORE Cherenkov array with the first IACT of the TAIGA experimen. *PoS* **2018**, *ICRC2017*, 756.
12. Kuzmichev, L.A.; others. TAIGA Gamma Observatory: Status and Prospects. *Physics of Atomic Nuclei* **2018**, *81*, 497–507.
13. Feng, Q.; Lin, T.T.Y. The analysis of VERITAS muon images using convolutional neural networks. *Proc. the International Astronomical Union Symposium S325: Sorrento, Italy, October 19-25, 2016* **2016**, *12*, 173–179, [arXiv:astro-ph.IM/1611.09832].
14. Nieto, D.; Brill, A.; Kim, B.; Humensky, T. Exploring deep learning as an event classification method for the Cherenkov Telescope Array. *Proceedings of Science* **2017**, *301*, 809, [arXiv:astro-ph.IM/1709.03483].

15. Shilon, I.; others. Application of Deep Learning methods to analysis of Imaging Atmospheric Cherenkov Telescopes data **2018**. [arXiv:astro-ph.IM/1803.10698].
16. Ketkar, N. Deep Learning with Python; Apress: Berkeley, CA, 2017; pp. 195–208.
17. Martín, A.; others. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*; USENIX Association: Savannah, GA, 2016; pp. 265–283.
18. Kryukov A., Korosteleva E., B.I.K.A.M.A.S.A. Specifying Binary File Formats for TAIGA Data Sharing and Reuse. Book of abstracts of 26th Extended European Cosmic Ray Symposium / 35th Russian Cosmic Ray Conference, 2018, pp. 171–172.
19. Prosin, V.V.; others. Results from Tunka-133 (5 years observation) and from the Tunka-HiSCORE prototype. *EPJ Web Conf.* **2016**, *121*, 03004.
20. Bezyazeekov, P.A.; others. Measurement of cosmic-ray air showers with the Tunka Radio Extension (Tunka-Rex). *Nucl. Instrum. Meth.* **2015**, *A802*, 89–96, [arXiv:astro-ph.IM/1509.08624].
21. Monkhoev, R.D.; others. The Tunka-Grande experiment: Status and prospects. *Bull. Russ. Acad. Sci.* **2017**, *81*, 468–470.