Towards a Global Analysis and Data Centre in Astroparticle Physics [⋆]

Andreas Haungs[0000−0002−9638−7574] [⋆⋆]

Karlsruhe Institute of Technology, KIT-IKP, 76021 Karlsruhe, Germany
andreas.haungs@kit.edu
http://www.kceta.kit.edu

Abstract. Astroparticle Physics is a young and evolving research area, where the amount of data continuously increase due to the modern instruments. Due to the geographical and experimental diversity a dedicated strategy for the digitalisation of the research field is fundamental to astroparticle physics. For an effective implementation with benefits for science and society four strategic points were defined: (i) Establishment of one or more global data centres. (ii) Development of methods for conservation of the measurement data. (iii) Development of applications of modern methods in data analysis. (iv) Expansion of courses for the training of young scientists in modern analytical methods. Aim of the initiative of an analysis and data centre in astroparticle physics is to develop and implement an interdisciplinary concept, which meets the needs of the digitization of the research field and which is also attractive to society. The goal is to enable a more efficient analysis of the data that are recorded in different locations around the world (multi-messenger analyses), as well as modern training for Big Data Scientists in the synergy between basic research and the information society.

Keywords: Analysis and Data Centre · Data Infrastructure · Astroparticle Physics.

1 Introduction: The High-Energy Universe

Understanding the high-energy Universe in the context of Astroparticle Physics means first and in foremost to answer the urgent question of the origin of high-energy cosmic rays. This question could not be answered by roughly 100 years of measurements of cosmic rays since their discovery in 1912 by a series of balloon flights of Victor Hess [1]. Whereas the cosmic rays at lower energies can be directly measured by balloon or satellite based experiments and are interpreted
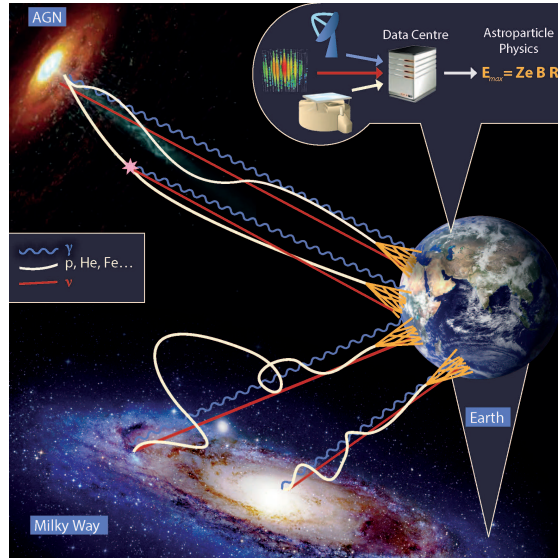
Fig. 1. Motivation for a global data and analysis centre in astroparticle physics. Cosmic rays, neutrinos and gamma rays of galactic (Milky Way) or extra-galactic (e.g. supermassive black holes, named AGN) origin reach our Earth and are observed by different kind of instruments. For a multi-messenger analyses these data have to be combined in a dedicated infrastructure.

as of Galactic origin, i.e. generated and accelerated within the Milky Way, above 100 TeV primary energy the low flux does not allow to measure them directly. Instead, they are studied by the detection of extensive air showers. These are cascades of particles produced when a high-energy cosmic ray enters the Atmosphere and interacts. Experiments measuring these air-showers could prove that the highest-energy cosmic rays (above ca. 8 EeV) are of extra-galactic origin [2]. It is unknown which sources in our deep Universe are responsible for these particles and at which energy exactly the transition from cosmic rays of galactic and extra-galactic origin happens.

Only charged particles can be accelerated in the source regions of the Universe. However, in these acceleration processes also secondary neutral products are generated, like gamma-rays and neutrinos. Once produced, these particles should travel straight on from the source to Earth. Neutrinos are very weakly interacting with material (detectors) and therefore difficult to measure. Gamma-ray measurements are among others motivated by studying them as tracers from charged cosmic-ray sources, but suffer from absorption by the infrared and microwave background in our Universe. Due to all these difficulties it became clear that only a combination of various measurements of the different tracers - not to forget the rare, but meanwhile detected, catastrophic events in our Universe generating Gravitational Waves - will bring us closer to an understanding of

the astrophysics of the High-Energy Universe. This approach is called Multi-Messenger Astroparticle Physics (fig. 1).

2 Multi-Messenger Astroparticle Physics

From the theory point of view multi-messenger astroparticle physics is to connect the physics of the sources with the observations in gamma-rays, neutrinos, and cosmic rays, and to use these observations to learn about the generation, acceleration and propagation properties [3]. From the experimental point of view the multi-messenger idea is mainly pursued by sending real time alerts from one observatory of one tracer to many others. This is partly organized by the individual experiments (e.g., via the Astrophysical Multimessenger Observatory Network (AMON) [4]). Two recent examples have shown first real successes of experimental multi-messenger physics: (i) the detection of a neutron star merger in parallel by a gravitational wave detector and by many telescopes observing the event in the entire electromagnetic spectrum [5]. This event and the corresponding publication is somehow seen as the birth of multi-messenger astronomy (as extension to multi-wavelength astronomy) though no particles (cosmic rays or neutrinos) have been observed from this event. (ii) An alert from the IceCube neutrino observatory for a detection of a high-energy neutrino lead to the corresponding observation of a high-energy gamma ray flare of a distant Blazar [6,7]. This shows that the multi-messenger approach is indeed the most promising idea for gaining new insights in the High-Energy Universe.

More events of parallel observations of the same source region will certainly be provided in the next years, in particular by the new or enhanced large-scale observatories CTA [8], IceCube(-Gen2) [9,10], Pierre Auger Observatory (Prime) [11,12], and LIGO/VIRGO [13,14]. All these observatories are operated independently by large international collaborations with hundreds and partly thousands of physicists. However, the maximum information of the measurements can only be elaborated if a common analysis based on measurements of high accuracy can be achieved. The here presented initiative aims to provide and validate new tools and methods for a sophisticated multi- messenger analysis strategy. An important part of this is to enhance current first steps to a comprehensive analysis and data centre for astroparticle physics [15,16].

3 An Analysis and Data Centre for Astroparticle Physics

With the help of modern Information Technology, Big Data Analytics and Research Data Management, the basic goal is to enter a new era of multi-messenger astroparticle physics. This can only be reached if several aspects of a coherent development are considered, where a dedicated analysis and data centre is a very important ingredient. It will not only provide the tools and environment to take the step into this new physics era, but also allows to take the digitized society along the path a must do in modern Big Data Science. Nowadays, basic
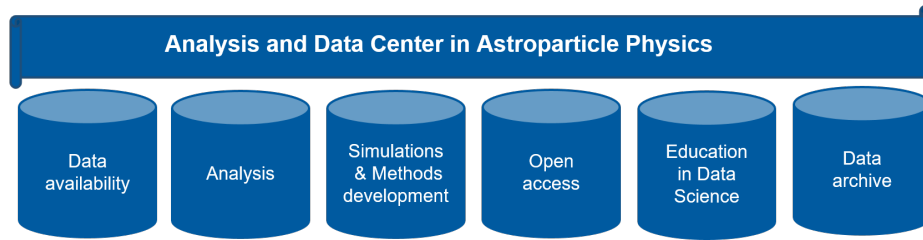
Fig. 2. The main pillars of a possible global Analysis and Data Centre in Astroparticle Physics.

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 large amounts 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 has to provide a vast of functionalities, at least covering the following pillars (fig. 2):

– Data availability: All participating researchers of the individual experiments or facilities need a fast and simple access to the relevant data.
– Analysis: A fast access to the Big Data from measurements and simulations is needed.
– Simulations & 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.
– Education in Data Science: The handling of the centre as well as the processing of the data needs specialized education in Big Data Science.
– Open access: It becomes more and more important to provide the scientific data not only to the internal research community, but also to the interested public: Public Data for Public Money!
– Data archive: The valuable scientific data need to be preserved for a later (re-)use.

All the present and future large-scale observatories mentioned above will provide their scientific data via sophisticated infrastructures and data centres for internal and also external use. However, information from various experiments and various messengers like charged particles, gamma-rays or neutrinos, measured by different globally distributed large-scale facilities, have to be combined. For that a diverse set of astrophysical data is required to be made available and public as well as a framework for developing tools and methods to handle the (open and collaborative) data.
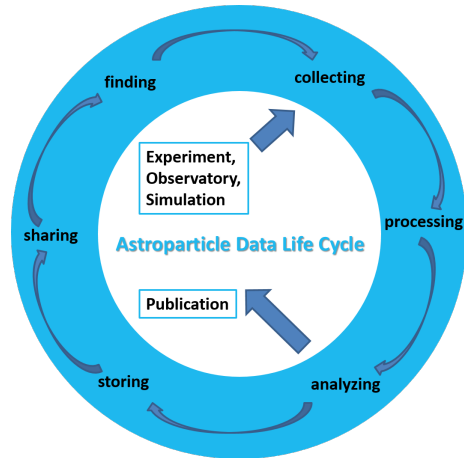
Fig. 3. Data life cycle in Astroparticle Physics. Data are generated in the specific observatories or experiments and the final goal of each analysis is a (journal) publication of the analysis resutls. A global analysis and data centre has to provide the tools and infrastructure to perform each individual part of the cycle.

We aim to extend the current activities in the individual observatories on an experiment-overarching, global and international level. This includes the validation of both, providing public access to (an initial part of) the scientific data and using the computing environment by the involved researchers to perform multi-messenger analyses. A further goal is to standardize the data, to make the publication FAIR [17], and by that to make it more attractive for a broader user community. The FAIR principles require that all parts of an data life cycle (fig. 3) are considered equally important in the realisation of an open data centre. The move to most modern computing, storage and data access concepts will also open the possibility of developing specific analysis methods (e.g. deep learning) and corresponding simulations in one environment opening a new technological opportunity for the entire research field.

4   Steps towards the Analysis and Data Centre

Whereas in Astronomy and Particle Physics data centres are already established, which fulfill a part of the above mentioned requirements (although, not the same parts), in Astroparticle Physics only first attempts are presently under development. For example, at the KASCADE experiment [18] have initiated KCDC for a first public release of scientific data. In addition, some public IceCube or Auger data can be found already now in the Astronomical Virtual Observatories, like in GAVO [19].

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 to 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 far beyond the next decade. 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 heterogeneous computing and storage resources and in organizing scientific collaborations, scientific education and science communication, where sophisticated public data centres will play the key role.

Based on the experience gained with KCDC [20] and at the GriDKA Tier-1 environment at KIT [21], on a long term, a global Astroparticle Physics Data and Analysis Centre (presently under construction and based on KCDC and GridKa) as a large-scale infrastructure is being established.

5 KCDC - The KASCADE Cosmic-ray Data Centre

Here, a brief summary of the realisation of KCDC is given: When publishing data it is not enough to put some ASCII files on a plain webpage. To ensure that potential users can actually use the data, an extensive documentation on how the data has been obtained is needed. Depending on the kind of data, this meta-data is at least a description of the detector and the reconstruction procedures employed, but can also consist of parameters determined by the raw data as well as event identifiers like UUIDs. Since this information, in addition to a license agreement, is not expected to change often, creating some static pages is a viable option. Another important aspect is the user and access management. For KCDC it was the concept that only registered users get access to the data and to the detailed data shop. Therefore, it is sufficient to check if an account is active and the user is authenticated. While there is already a basic implementation of a permission based access limitation, a useful categorization of the users into - possibly hierarchical - groups is needed (no administrator should manually manage privileges of single users) to effectively use it. The heart of the data centre is the data shop. The design goals have to ensure an easy access to an user defined subset of the whole dataset, a natural way to configure additional detector components or observables without the need to change the code-base and a clear overview of previous requests with the possibility to use these selections as a template for future requests. For KCDC, the data are stored in a MongoDB. Its scheme-less design allows us to collect all available information of an event, although the available detector components may vary for each event. In principle KCDC can use any kind of input format, as these are implemented as plugins. Currently three output formats are supported, HDF5, ROOT and ASCII. These can be extended by adding additional plugins, too. The requests are processed using a Celery based task queue. The simultaneous processing of requests can be achieved by adding more worker processes, which can be distributed among several machines. Together with the realisation to run the MongoDB on a sharded cluster, a scaling of the needed processing power with the demand is achieved.

6 The German-Russian Astroparticle Data Life Cycle Initiative

KCDC is a web portal where the KASCADE scientific data is made available for the interested public. In Russia, there is the operating TAIGA and Tunka-Rex facilities [22,23,24], where by many reasons combined Tunka-Rex and KASCADE data analyses with sophisticated Big Data Science analysis methods (e.g. deep learning) are of advantage for solving physics questions. These high-statistics experiments can be used as testbed for future multi-messenger astroparticle physics analyses based on data of the big Observatories coming into operation in next years. The project aims, for the first time, for a common data portal of two independent observatories and at the same time for a consolidation and maturation of an astroparticle data centre.

The German part of the GRADCLI project [25] focuses on four items all of them are initial ingredients of the envisaged global data and analysis centre:

Extension of KCDC: The existing data centre KCDC will be extended by scientific data from TAIGA allowing on-the-fly multi-messenger-analysis.

The experimental setups in Russia and Germany generate or have generated large amounts of primary data. This have to be specified and a common language in data description have to be defined. This needs a careful preparation of the data and extension of the KCDC software and web interface. Within this project, we will adapt the Tunka-Rex data to the KCDC concept and provide an extended public data centre [26]. This central and distinct work will apply the concept and software of KCDC to follow the request of the funding agencies to make (at least the high level) scientific data public in a *FAIR* way. In addition, specific data will be included for the experts at both sides to combine the data in common analysis methods. The extended data centre and the experience gained within this project will serve for data releases of present and forthcoming large-scale experiments in astroparticle physics.

Big Data Science Software: Advancement of Big Data Science: The data centre shall allow not only access to the data, but also provide the possibility of developing specific analysis methods and perform corresponding simulations. The concept to reach this goal is the installation of a dedicated Data Life Cycle Lab.

To reach this goal first, the basic software of the KCDC data centre needs to be improved. Then the data centre has to be moved to the Big Data environment provided by KIT-SCC. The software is being built as a modular, flexible framework with a good scalability (e.g. to large computing centres). The configuration is hold to be simple and doable also via a web interface. The entire software will be based solely on Open Source Software (Python, Django, mongoDB, HDF5, ROOT, etc.). Having done this step, the result enables to install and publish a dedicated Data Life Cycle Lab for the astroparticle physics community. Dedicated access, storage, and interface software has to be developed. In addition, the appropriate hardware has to be installed and commissioned. The described concept and working plan is valid also for the next steps, i.e. to generalize and consolidate the software package of a global astroparticle data centre for public

and scientific use.

Multi-Messenger Data Analysis: Specific analyses of the data provided by the new data centre will be performed to test the entire concept giving important contributions and confidence to the centre as a valuable scientific tool.

For a detailed reliability test we aim for common data analyses using the complementary information of independent experiments, in particular to apply deep learning (machine learning) methods to multi-messenger analyses [27]. This will give important contributions and confidence to the project as a valuable scientific tool. The data centre will be also very useful for theoreticians to interpret experimental results.

Go for the public: A coherent outreach of the project, including example applications for all level of users - from pupils to the directly involved scientists - with detailed tutorials and documentation is an important ingredient of any activity in publishing scientific data.

In particular the documentation, i.e. the metadata, of any released dataset has to be prepared with reasonable care. The goal of having detailed tutorials, i.e. an education portal, is to provide the data also to a general public in the sense of a visible outreach of astroparticle physics. The target groups for the tutorials are teachers and pupils in high schools. A tutorial has to provide at least: i) a basic knowledge on the experiment, astrophysics and related topics; ii) the required software and KCDC data (preferably as a pre-selection); iii) a step by step explanation of a simple data analysis; iv) a modern programming language code example; v) the interpretation and discussion of the outcome. With increasing number of users also the user management system (e.g. Q&A sections, discussion blogs, etc.) have to be extended and maintained. This part includes also creating and designing outreach materials, conducting routine website maintenance, preparing text or images. In addition, web pages, press releases, distribution of the progress and news in social networks like Facebook, Twitter, etc. needs to be managed.

7   Conclusion

Several initiatives have been started towards a dedicated and global Analysis and Data Centre in Astroparticle Physics. Aim of these initiatives is to develop and implement an interdisciplinary framework, which meets the needs of the digitization of the research field and which is also attractive to society. One of the immediate goals is to enable a more efficient analysis of the data that is recorded in different locations around the world for coherent multi-messenger studies in order to better understand the high-energy processes in our Universe.

References

1. Victor F. Hess; Über Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten; Phys.Z. **13** (1912) 1084-1091

2. R. Aloisio, V. Berezinsky, A. Gazizov; Transition from galactic to extragalactic cosmic rays; Astropart.Phys. **39-40** (2012) 129-143.

3. A. Palladino, W. Winter; A Multi-Component Model for the Observed Astrophysical Neutrinos; Astron.Astrophys. **615** (2018) A168

4. M.W.E. Smith et al.; The Astrophysical Multimessenger Observatory Network (AMON); Astropart.Phys. **45** (2013) 56-70.

5. B.P. Abbott et al.; Multi-messenger Observations of a Binary Neutron Star Merger; Astrophys.J. **848** (2017) no.2, L12.

6. M.G. Aartsen et al.; Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A; Science **361** (2018) 6398, 1378.

7. M.G. Aartsen et al.; Neutrino emission from the direction of the blazar TXS 0506+056 prior to the IceCube-170922A alert; Science **361** (2018) 6398, 147.

8. Edited by J. Hinton, S. Sarkar, D. Torres, J. Knapp; Seeing the High-Energy Universe with the Cherenkov Telescope Array - The Science Explored with the CTA; Astropart.Phys. **43** (2013) 1-356.

9. M.G. Aartsen et al.; Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector; Science **342** (2013) 1242856.

10. J. van Santen et al.; IceCube-Gen2: the next-generation neutrino observatory for the South Pole; PoS ICRC2017 (2018) 991.

11. A. Aab et al.; The Pierre Auger Cosmic Ray Observatory; Nucl.Instrum.Meth. A **798** (2015) 172-213.

12. R. Engel et al.; Upgrade of the Pierre Auger Observatory (AugerPrime) By Pierre Auger Collaboration (). PoS ICRC2015 (2016) 686.

13. B. P. Abbott et al.; Exploring the Sensitivity of Next Generation Gravitational Wave Detectors; Class.Quant.Grav. **34** (2017) no.4, 044001.

14. F. Acernese et al.; Advanced Virgo: a second-generation interferometric gravitational wave detector; Class.Quant.Grav. **32** (2015) no.2, 024001.

15. M. Spiro; Open data policy and data sharing in Astroparticle Physics: the case for high-energy multi-messenger astronomy; J.Phys.Conf.Ser. **718** (2016) no.2, 022016.

16. A.Haungs et al.; The KASCADE Cosmic-ray Data Centre KCDC: Granting Open Access to Astroparticle Physics Research Data; Eur.Phys.J.C. **78** (2018) 741

17. M. D. Wilkinson et al.; The FAIR Guiding Principles for scientific data management and stewardship Sci Data. **3** (2016) 160018.

18. T. Antoni et al; The Cosmic-Ray Experiment KASCADE; Nucl.Instr. and Meth **A513** (2003) 490-510

19. GAVO: see `https://www.g-vo.org/`

20. KCDC: see `https://kcdc.ikp.kit.edu/`

21. GridKa: see `https://www.gridka.de/`

22. 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

23. F.G. Schröder et al.; Tunka-Rex: Status, Plans, and Recent Results; EPJ Web Conf. **135** (2017) 01003

24. P.A. Bezyazeekov et al.; Measurement of cosmic-ray air showers with the Tunka Radio Extension (Tunka-Rex); Nucl.Instrum.Meth. **A802** (2015) 89-96

25. I. Bychkov et al.; RussianGerman Astroparticle Data Life Cycle Initiative; Data **3(4)** (2018) 56

26. D. Wochele et al.; Data Structure Adaption from Large-Scale Experiment for Public Re-Use; (2019) these proceedings.

27. V. Tokareva; Data infrastructure development for a global data analysis centre; (2019) these proceedings.