PROCEEDINGS OF SCIENCE

The Very-high-energy Open Data Format: towards a shared, open data format in very-high-energy astronomy

B. Khélifi,[a,1,*] **R. Zanin,**[b] **K. Kosack,**[c,b] **L. Olivera-Nieto**[d] **and J. Schnabel**[e]

[a] *Université Paris Cité, CNRS, Astroparticule et Cosmologie, F-75013 Paris, France*
[b] *Cherenkov Telescope Array Observatory gGmbH, Via Gobetti, Bologna, Italy*
[c] *IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France*
[d] *Max-Planck-Institut für Kernphysik, P.O. Box 103980, D 69029 Heidelberg, Germany*
[e] *ECAP, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nikolaus-Fiebiger-Str. 2 , 91058 Erlangen, Germany*

E-mail: khelifi@in2p3.fr

In very-high-energy (VHE) gamma-ray astronomy, the community is converging towards the use of a common open data format, called "Data formats for Gamma-ray Astronomy", for the high-level data products. This format is in use for ground-based TeV observatories like H.E.S.S, MAGIC or HAWC, some of whom plan to openly release high-level data products. These efforts are parallel to the development and use of open analysis software such as the Gammapy package. This open initiative has shown that it is possible to define common standards even without governance. With the advent of open VHE observatories (e.g. CTAO, KM3NeT) and an increase in both multi-wavelength and multi-messenger studies, such standards should evolve to support all of VHE multi-messenger astrophysics. For these reasons, a new initiative has been created to specify formats of high-level data from very and ultra high energy gamma-ray facilities and from VHE neutrino detectors. It also aims to better respect the FAIR principles and the IVOA recommendations.This communication will present the Very-high-energy Open Data Format (VODF) project that has been established by eleven VHE astroparticle facilities. Its structure, its organisation and its goal will be presented. Anchored in Open Science, our goal is to solicit comments and future contributions from the VHE astrophysics community.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

[1] For the VODF Steering Committee
[*] Speaker

PoS(ICRC2023)1510

1. Landscape of the VHE data analysis

The field of ground-based very-high-energy (VHE) gamma-ray astronomy is rapidly evolving with more than 300 detected sources [1, 2] and with the commissioning and operation of large detector arrays, such as CTAO [3] or LHAASO [4]. Known gamma-ray sources are associated with objects such as supernova remnants, pulsar wind nebulae, binary systems, novae, stellar clusters, star-forming regions, star-forming galaxies, active galactic nuclei, or gamma-ray bursts. Their studies often require the use of data from other wavelengths. In parallel, VHE neutrino astronomy has recently reached a stage that allows the study of neutrino emission from Galactic and Extragalactic sources [5–7], leading to multi-messenger astrophysical analyses of both isolated sources and large-scale emission. The next generation of more sensitive instruments [8, 9] is currently in deployment.

The improvement of sensitivities of VHE astrophysical facilities is accompanied by an increasing amount of public data (e.g. [10, 11]) and open analysis libraries (e.g. [12, 13]). The opening of research products (publications, data, software) is stimulated by national and transnational roadmaps for Open Science (e.g. [14, 15]) that recommend or require free access. Some newly created VHE observatories follow these requirements and will openly grant access to data and analysis software. These policies will stimulate even more the multi-wavelength and multi-messenger analyses. As demonstrated in [16] with VHE gamma-ray data and in [17] using gamma-ray and neutrino data, both with the library Gammapy [18], joint analysis of multi-instrument data can be performed with a coherent statistical treatment while dealing with systematics of/between measurements.

Such multi-instrument analyses are becoming a reality thanks to the use of a common open FITS-based[1] [19] data format for the high level or science-ready data (see section 3) of the VHE gamma-ray instruments, called "Data formats for Gamma-ray Astronomy" (or Gamma-ray Open Data Format, GADF)[2]. The work of data formatting started back in 2010 by TeV experts from H.E.S.S. and CTA following the example of the Fermi-LAT and OGIP formats. And it leads to the first version of GADF in 2016 [20]. Today, the running gamma-ray experiments H.E.S.S., MAGIC, VERITAS and HAWC are using this data format for some of their data sets. The use of this data format for testing purposes and for scientific publications has shown that its design describes efficiently the VHE gamma-ray data and permits accurate scientific analyses. After few years of usage of this format, some main user and core contributors (GADF, CTAO, H.E.S.S., etc) realise the need to evolve because the GADF open initiative started to show some limitations [21], both on the format requirements and its organisation.

Enriched by the gained experience of the GADF initiative, their main contributors contacted the current and under-deployment VHE gamma-ray and neutrino facilities to establish a new open initiative to build together an extended open data format for the VHE astroparticle experiments, the *Very-high-energy Open Data Format* (VODF). Now officially supported by eleven VHE facilities, ASTRI, CTAO, FACT, Fermi-LAT, HAWC, H.E.S.S., IceCube, KM3NeT, MAGIC, SWGO and VERITAS, this newly born initiative has started its activities. This contribution describes its objectives and some of its requirements (section 2), the targeted level of data that are considered to

[1] https://fits.gsfc.nasa.gov/fits_home.html
[2] https://gamma-astro-data-formats.readthedocs.io/en/v0.3/

be formatted (section 3), the organisation of the initiative (section 4) and the anticipated perspectives (section 5).

2. Guidelines of the open VODF initiative

VODF is an open data model and format for VHE gamma-ray and neutrino astronomy. These experiments strongly differ by their detection techniques, their instrumentation and their type of raw data. Ground-based gamma-ray detectors fall into two categories, Imaging Atmospheric Cherenkov Telescopes (IACTs) and Water Cherenkov Detectors (WCDs), whereas neutrino detectors are using optical modules immersed in a km3-scale volume of ice or water. They all share the following high-level properties, after a low-level data processing consisting of calibration, reconstruction and background reduction:

- they provide a list of gamma-ray or neutrino candidates characterised by their arrival time, their energy and their celestial arrival direction,
- their instrumental response to events can be characterised by the same quantities and have the same type of observational dependencies, such that their associated Instrument Response Functions (IRFs) can be factorised in the same manner,
- as these instruments study the same astrophysical processes, the scientific products are similar: sky maps, spectra, light curves, along with associated statistical information like significance or likelihood profiles.

In this context, the formats can be homogeneous and shared. This approach has been tested and has shown its success to generate scientific results and in addition to offer real interoperability between instruments [16, 22].

The goal of the VODF initiative is to provide a standard set of data models and file formats, starting at the reconstructed event level (*science-ready data*) as well as higher-level products such as N-dimensional binned data cubes (including sky images, light curves, and spectra) and source catalogues. With these standards, common science tools can be used to analyse data from multiple high-energy instruments.

In parallel, several facilities will publicly open their data (e.g. CTAO, KM3NeT) or their archive [23] and the use of certified repositories is strongly recommended. As a consequence, the data should be correctly curated such that they respect as closely as possible the FAIR principles [24] (*Findable, Accessible, Interoperable and Reusable*). Metadata describing the associated data are mandatory. VODF aims to standardise the format of these metadata for each level of data.

In the field of astronomy, the community started to think about common standards in order to share and use more efficiently data. Formed in 2002, the *International Virtual Observatory Alliance* (IVOA)[3] has the mission to create and maintain standards to "facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory". IVOA has established many advanced data models and standards that are an excellent guideline for our VHE domain (e.g. [25]). In addition,

[3] https://www.ivoa.net/

they offer new possibilities to find and access data with the Virtual Observatory (VO) services (such as Aladin or Topcat).

Technical tests have been realised to respect more closely the FAIR principles and to use some of the IVOA standards. The H.E.S.S. test data release [10] or the ANTARES data release [11] illustrate the new possibilities [26, 27] available to the new VHE observatories using standards set by astronomers.

Given this context of state-of-art data formating, the open initiative VODF is built on the following guidelines:

- building a new VHE open data format common to IACT, WCD and neutrino instruments,
- being compliant with the FAIR principles,
- following as much as possible the IVOA standards,
- working by consensus with open contributions under the supervision of supporting experiments.

3. Data Levels

VODF aims to settle standards of different levels of VHE data resulting from the calibration, reconstruction and background reduction of the instrument analysis pipelines. The data types and data levels to format are illusted in Fig. 1 and described below.

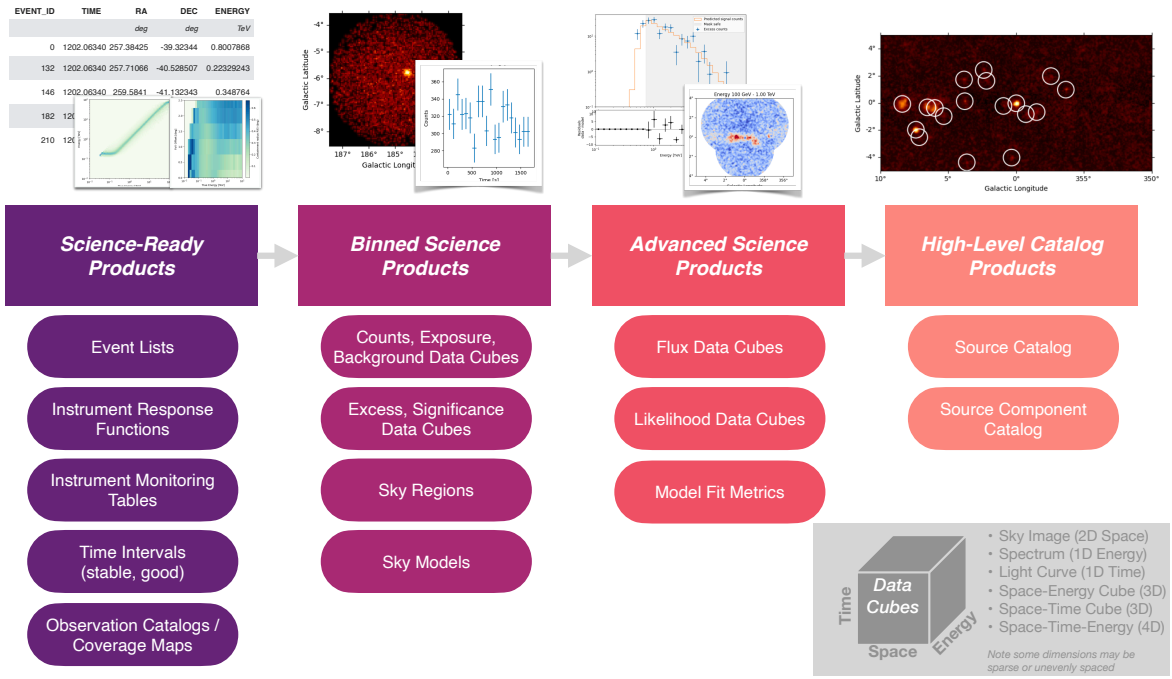

Figure 1: Diagram of the different data levels. Each rounded box stands for data and their associated metadata.

Common Data Structures. The considered data are sharing common elements and structures. One can list in particular:

- the time formats, following the FITS standards [28],
- the coordinate formats, following the IAU resolutions [29],
- the N-dimensional maps, with regular or sparse axis that can contain either bins or points and that handle physical units,
- the metadata, with standard keywords associated to e.g. the instrument, maintainer, data release identifier, data format version, VO standards[4]; each of the following data levels should contain metadata information,
- the provenance information, following the IVOA data model[5]; each of the following data levels should contain provenance information.

Science-Ready Data. They are data ready for delivery to users that contain sets of selected events (typically a mix of signal and irreducible background) with a single final set of reconstructed parameters per event. This data level also contains the IRF of the instrument, currently factored into four independent components: the Effective Area, the Energy Dispersion, the Point Spread Function and the Background model. These contain also auxiliary data describing the astronomical, environmental, and instrumental conditions required for science analysis, such as the stable observation interval, pointing position(s), livetime. And this level should contain index tables such that these data can follow the FAIR standards, in particular their findability and access.

Binned Science Data. These data are produced by binning the spatial, temporal, and/or spectral components of all the Science-Ready Data over a target-specific or a region of interest of the celestial sphere with the user choice of binning. These are in instrumental units, e.g. counts.

Advanced Science Data. They consist on the astrophysical products derived from the Binned-Science Data (combining the binned events with the binned IRFs), such as spectral energy distributions, light curves, sky maps, and phasograms. These are in physical units such as fluxes, energy or angles.

High-Level Catalog Data. They contain a collection of astrophysical objects or VHE sources with a description of their properties and associations, such as source components, upper-limits on size or flux, observation periods or science alerts.

In addition to these formats, the VODF initiative aims to provide tools to check the compliance of a data set with a given format version. These open tools, which use Python and the community-standard `astropy` library, could be used by observatories during their data release process and included in their Verification and Validation (V&V) plan.

The definition of such format will be preceded by a data modeling analysis. The data models of each data level could be considered to be released in addition to the data formats.

The definition and serialisation of sky models are a major bottle neck in the VHE and HE domain. Up to know, there is no convergence between the communities of experiments or analysis software to create standards on models with a their spectral, spatial and/or temporal representation

[4] e.g. https://www.ivoa.net/documents/RM/20070302/index.html
[5] https://www.ivoa.net/documents/ProvenanceDM/

(see for example the difference between `astropy.modeling`, the Sherpa models that include the XSpec ones, or the stand-alone library `astromodels`). Given the complexity of the task, the VODF initiative has postponed to work on this aspect for the moment.

4. Organisation of the VODF initiative

THe VODF initiative was formed at the beginning of 2023 and is officially supported by eleven VHE experiments: ASTRI, CTAO, FACT, Fermi-LAT, HAWC, H.E.S.S., IceCube, KM3NeT, MAGIC, SWGO and VERITAS (in alphabetical order). The data format used by the Fermi-LAT experiment was used as a base for the original GADF and has been settled since a long time. Thus, the participation of Fermi-LAT in this initiative will bring in valuable expertise and skills, as well as ensure that the VODF format can serve both ground-based and satellite experiments.

With the official delegates of each experiment, a governance document has been written and it describes the functionning of the initiative. The latter possesses a simple and clear organisation:

- a Steering Committee composed by one official delegate per experiment,
- three Lead Editors, one per experimental technique (IACT, WCD, Neutrino),
- and two Conveners.

The initiative is fully open and a community-supported project. Any contribution will be considered and evaluated. The initiative is committed to fostering an inclusive community with an established Code of Conduct.

The setting and the improvement of the formats will be driven by open contributions from the community (users and experiments). Experts of the VHE domain and of other astronomical fields are welcome to participate, as well as users. All inputs on features, documentation or tools will be discussed. The lead editors will drive the discussions and validate the proposals accepted by consensus. Ultimately, the steering commitee might arbitrate some discussion. The lead editors are responsible for the maintenance of the supporting tools and report the advance to the steering committee. Major orientation plans or roadmaps and major improvements will be prepared by the lead editors and evaluated by the steering committee.

The initiative are using several tools to communicate, exchange and develop the standards. A central web site[6] hosted by `readthedocs.org` has been settled and is still under construction. A dedicated exchange workspace[7] is created on the communication platform `Slack` to exchange between contributors. And a VODF `GitHub` project[8] hosts the standards and the format documentation, the interface to report changes (`Issues`) and to propose improvements (`Pull requests`), the V&V tools and the web site files.

5. Conclusion and perspectives

VODF is a newly-born open initiative aiming to create standards for VHE data produced by ground-based gamma-ray and neutrino instruments. Supported officially by eleven astroparticle

[6] https://vodf.readthedocs.io/en/latest/index.html
[7] vodf-workspace.slack.com
[8] https://github.com/vodf

projects, the VODF initiative aims to settle a new data format for the high-level products of VHE instruments that respects the FAIR principles and follows as closely as possible the IVOA standards. It will propose VHE standards for Common Data Structures, Science-Ready Data, Binned Science Data, Advanced Science Data and High-Level Catalog Data.

The initiative has just been created and standards have not yet been settled. Some data modelling is under way and evaluations of some technical choices are currently made to propose machine-readable formats and to serialise data under this format (e.g. within FITS, YAML, ASDF). The first version of the VODF format will probably be an extension of the GADF format including the experience from CTA and HAWC members. The in-depth data models produced by the Computing Department of CTAO are key elements to take into account. The next versions will respect the FAIR principles as soon as possible, such that the next open data will be easily findable and accessible. The specifities of the neutrino detectors will be included following the experiences gathered from IRF generation for KM3NeT in a common CTA and KM3NeT analysis using Gammapy. And the IVOA recommandations will be added as much as possible, with the willing of providing feedback to the IVOA. The exact roadmap will be soon settled by the lead editors.

The data formatting is a hidden pillar of Open Science. It permits that open data can follow the FAIR principles and that open software can follow the FAIR4RS principles [30]. The VODF initiative is thus an important activity that supports the current and under-construction VHE experiments, and the VHE Science Analysis Tools. It will allow a real interoperability between the different VHE data in order to realise joint multi-wavelength and multi-messenger analyses.

References

[1] S. P. Wakely and D. Horan, *TeVCat: An online catalog for Very High Energy Gamma-Ray Astronomy*, in proceedings of the *30th International Cosmic Ray Conference* (2007) [ads:2008ICRC....3.1341W], TeVCaT catalog.

[2] Z. Cao, et al., *The First LHAASO Catalog of Gamma-Ray Sources*, arXiv:2305.17030.

[3] W. Hofmann, R. Zanin, *The Cherenkov Telescope Array*, *Handbook of X-ray and Gamma-ray Astrophysics* (2023) [arXiv:2305.12888].

[4] X.-H. Ma, et al., *LHAASO Instruments and Detector technology*, *Chinese Phys. C* 46 *030001 (2022)* [doi:10.1088/1674-1137/ac3fa6].

[5] A. Albert, et al., *Hint for a TeV neutrino emission from the Galactic Ridge with ANTARES*, Physics Letters B *841* (2023) 137951 [arXiv:2212.11876].

[6] R. Abbasi, et al., *Observation of high-energy neutrinos from the Galactic plane*, Science *380* 6652 (2023) [arXiv:2307.04427].

[7] F. Halzen, *IceCube: Neutrinos from Active Galaxies*, in proceedings of *the 2023 Electroweak session of the 57th Rencontres de Moriond*, 57th Rencontres de Moriond [arXiv:2305.07086].

[8] M. G. Aartsen, et al., *IceCube-Gen2: the window to the extreme Universe*, J. Phys. G: Nucl. Part. Phys. *48* (2021) 060501 [arXiv:2306.05900].

[9] S. Adrián-Martínez, et al., *Letter of intent for KM3NeT 2.0*, Journal of Physics G Nuclear Physics (2016) *43* 084001 [arXiv:astro-ph.IM/1601.07459].

[10] H.E.S.S. Collaboration, *H.E.S.S. first public test data release* (2018), [arXiv:1810.04516].

[11] ANTARES Collaboration, *Data sets for searches for cosmic neutrino point sources with ANTARES*, open data release.
[12] *Gammapy, a Python package for gamma-ray astronomy*, Gammapy web site [arXiv:1709.01751].
[13] *3ML, Multi-Mission Maximum Likelihood framework*, 3ML web site [arXiv:1507.08343].
[14] EU Open Science.
[15] NASA Open-Source Science Initiative.
[16] A. Albert, et al., *Validation of standardized data formats and tools for ground-level particle-based gamma-ray observatories*, A&A 667, A36 (2022) [arXiv:2203.05937].
[17] T. Unbehaum, *Joint-instrument analyses with Gammapy*, Master's Thesis in Physics, ECAP Friedrich-Alexander-Universität Erlanger-Nürnberg (2020), link.
[18] *Gammapy: Python toolbox for gamma-ray astronomy, LTS v1.0*, [10.5281/zenodo.7311399].
[19] D. C. Wells, E. W. Greisen, and R. H. Harten, *FITS - a Flexible Image Transport System*, Astronomy and Astrophysics Supplement *44*, 363 (1981).
[20] C. Deil, et al., *Open high-level data formats and software for gamma-ray astronomy*, in proceedings of *the Gamma 2016 conference*, arXiv:1610.01884 [doi:10.5281/zenodo.1409830].
[21] M. Linhoff, et al., *Data Formats for Gamma-Ray Astronomy (GADF)*, oral contribution to *the AstroParticle Symposium 2022*.
[22] C. Nigro, et al., *Towards open and reproducible multi-instrument analysis in gamma-ray astronomy*, *A&A* **625** (2019) A10 [arXiv:1903.06621].
[23] C. Nigro, *Establishing the MAGIC data legacy: adopting standardised data formats and open-source analysis tools*, in proceedings of the *7th Heidelberg International Symposium on High-Energy Gamma-Ray Astronomy*, *Gamma 2022* [arXiv:2302.13615].
[24] M. D. Wilkinson, et al., *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific Data *3*, 160018 (2016).
[25] M. Servillat, *Provenance of astronomical data*, in proceedings of the *Annual meeting of the French Society of Astronomy and Astrophysics*, *SF2A (2021)* [arXiv:2204.11486].
[26] M. Servillat, et al., *FAIR high level data for Cherenkov astronomy*, in proceedings of *ADASS XXXI*, edited by ASP Conf. Ser., *ADASS XXXI (2021)* [arXiv:2201.03247].
[27] J. Schnabel, *Approach to Virtual Observatory link in KM3NeT*, *ESCAPE WP4 meeting (2019)*.
[28] A. H. Rots et al., *Representations of time coordinates in FITS. Time and relative dimension in space*, A&A *574* A36 (2015) [arXiv:1409.7583].
[29] G. H. Kaplan, *The IAU Resolutions on Astronomical Reference Systems, Time Scales, and Earth Rotation Models*, U.S. Naval Observatory Circular No. 179 (2005) [arXiv:astro-ph/0602086]
[30] M. Barker et al., *Introducing the FAIR Principles for research software*, Scientific Data *9*, 622 (2022)

PoS(ICRC2023)1510