universe

MDPI

Review

Evolution of Data Formats in Very-High-Energy Gamma-Ray Astronomy

Cosimo Nigro [1,*], Tarek Hassan [2] and Laura Olivera-Nieto [3]

[1] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, Bellaterra, 08193 Barcelona, Spain

[2] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), E-28040 Madrid, Spain; tarek.hassan@ciemat.es

[3] Max Planck Institut für Kernphysik, Saupfercheckweg 1, 69117 Heidelberg, Germany; laura.olivera-nieto@mpi-hd.mpg.de

* Correspondence: cosimo.nigro@ifae.es

Abstract: Most major scientific results produced by ground-based gamma-ray telescopes in the last 30 years have been obtained by expert members of the collaborations operating these instruments. This is due to the proprietary data and software policies adopted by these collaborations. However, the advent of the next generation of telescopes and their operation as observatories open to the astronomical community, along with a generally increasing demand for open science, confront gamma-ray astronomers with the challenge of sharing their data and analysis tools. As a consequence, in the last few years, the development of open-source science tools has progressed in parallel with the endeavour to define a standardised data format for astronomical gamma-ray data. The latter constitutes the main topic of this review. Common data specifications provide equally important benefits to the current and future generation of gamma-ray instruments: they allow the data from different instruments, including legacy data from decommissioned telescopes, to be easily combined and analysed within the same software framework. In addition, standardised data accessible to the public, and analysable with open-source software, grant fully-reproducible results. In this article, we provide an overview of the evolution of the data format for gamma-ray astronomical data, focusing on its progression from private and diverse specifications to prototypical open and standardised ones. The latter have already been successfully employed in a number of publications paving the way to the analysis of data from the next generation of gamma-ray instruments, and to an open and reproducible way of conducting gamma-ray astronomy.

Keywords: very-high-energy gamma-ray astronomy; astroparticle physics; open science; data format

check for updates

Citation: Nigro, C.; Hassan, T.; Olivera-Nieto, L. Evolution of Data Formats in Very-High-Energy Gamma-Ray Astronomy. *Universe* **2021**, *7*, 374. https://doi.org/10.3390/universe7100374

Academic Editors: Ulisses Barres de Almeida, Michele Doro and Binbin Zhang

Received: 9 September 2021
Accepted: 30 September 2021
Published: 8 October 2021

1. Introduction

Gamma-ray astronomy, currently observing the non-thermal universe over more than 7 decades in energy, is conducted with different classes of instruments operating in two complementary energy ranges [1]. Space-borne telescopes, sensitive in the so-called high-energy regime (HE, 100 MeV $< E <$ 100 GeV), directly detect the gamma rays through their pair-conversion in an instrumented volume [2]. Ground-based telescopes, sensitive in the so-called very-high-energy regime (VHE, $E >$ 100 GeV), detect the particle cascade (or shower) generated by gamma rays interacting with atmospheric nuclei (via $e^{\pm}$ pair production and Bremsstrahlung) using two different techniques [3]. Imaging atmospheric Cherenkov telescopes (IACTs) use a large reflector ($\sim$10 m) and a photomultiplier camera to image the Cherenkov light emitted by the charged component of the shower. Particle samplers rely on an array of detectors (distributed over a surface up to a $\sim$km^2) to directly sample the charged component using, for example, scintillators or water tanks in which further Cherenkov light is produced and detected (water Cherenkov detectors, WCD). VHE astroparticle physics will be revolutionised in this decade by an upcoming generation of ground-based instruments built with the objective to improve by an order of magnitude the sensitivity of the current ones: the Cherenkov telescope array (CTA) [4] for IACTs; the Large High Altitude Air Shower Observatory (LHAASO) [5] and the Southern Wide-field Gamma-ray Observatory (SWGO) [6] for particle samplers.

In addition to different detection techniques, the current generation of HE and VHE instruments adopt distinct data and software policies. As typical for space observatories, HE gamma-ray telescopes retained their data proprietary for a limited amount of time (usually one year) before releasing them publicly. This has been the case for both currently operating HE gamma-ray telescopes: the *Fermi* Large Area Telescope (*Fermi*-LAT) [7] and the Astrorivelatore Gamma ad Immagini Leggero (AGILE) [8]. Their data are nowadays made promptly available via web-based platforms, referred to as science data centers, providing astronomers with an interface to retrieve the data and the science tools to perform their analysis [9,10]. VHE telescopes of the current generation, on the other hand, have been operated under more strict data and software policies. Telescopes like the high energy stereoscopic system (H.E.S.S.) or the major atmospheric gamma-ray imaging Cherenkov (MAGIC) have traditionally produced scientific results with proprietary data and closed-source software [11,12]. Few examples of public VHE data or software exist, worth mentioning are: the very energetic radiation imaging telescope array system (VERITAS), that has publicly released under an open-source license one of its analysis chains [13]; the first g-Apd Cherenkov telescope (FACT), that has made public its analysis chain [14], a small sub-sample of its data and quick-look analyses results on all the data collected [15]; and the high-altitude water Cherenkov (HAWC), that has provided some high-level data, mostly meant to reproduce results of major publications [16]. More recent efforts of data sharing in a standardised format will be covered later in this review. Generally speaking, beside sparse endeavours, VHE gamma-ray data largely remain inaccessible to astronomers outside the collaborations gathering them. This situation is bound to change with the forthcoming CTA that will represent the first gamma-ray experiment operated as a proposal-driven open observatory [17]. External scientists will be able to submit observational proposals; data will be proprietary to the principal investigators typically for one year and then released to the public. This implies, as in the case of HE gamma-ray instruments, the necessity to produce accessible data and tools for users external to the collaboration to perform their scientific analyses.

In light of these requirements, VHE gamma-ray astronomers have started developing open-source data-analysis tools (e.g., `ctools` [18] and `Gammapy` [19]) and, in parallel, a standardised format for astronomical gamma-ray data. This review will focus on the latter. The data level expected to be shared by the next generation of VHE observatories with external observers (as already routinely done by *Fermi*-LAT and AGILE) is a *high* data level whose purpose is the production of scientific results (i.e., measurement of the properties of an astrophysical source: flux, morphology, etc.). It contains a reduced amount of information compared to the *low* (or calibrated) data level strictly connected with the particular detection or analysis technique. Specifically it contains lists of detected photons with their estimated physical observables (energy, direction, etc.) and a characterisation of the response of the system. It is abstract enough to represent data from instruments employing diverse detection techniques, such as IACT and WCD. Being difficult to detach the discussion on high-level data format from the software provided to analyse it, we might comment as well upon aspects of software development and policies.

This review is thus structured: the progression of the data format from previous specifications is discussed in Section 2, along with its current status and working principles. In Section 3 we review some projects that have already successfully employed the format, either to validate the capabilities of the science tools, to illustrate the possibility of multi-instrument analysis with current gamma-ray instruments and to extend the format to particle samplers. In Section 4, we gather some ideas for the future of the format and its possible expansion. We provide our conclusions in Section 5.

2. Data Formats for Very-High-Energy Gamma-Ray Astronomy

2.1. Background: Data Model in the Current Generation of VHE Instruments

VHE gamma-ray astronomy inherited, along with the hardware techniques, the software solutions of particle physics. In the late 1990s and early 2000s, C++ and the ROOT [20] framework dominated the field. Hence, software for VHE data reduction and analysis has been mostly built in this environment. As already commented, even if some of these tools are accessible, little documentation is publicly available about the private analysis chains and the data they produce. Nonetheless, from the available material, a common data reduction workflow can be inferred for VHE gamma-ray telescope, sketched in Figure 1.

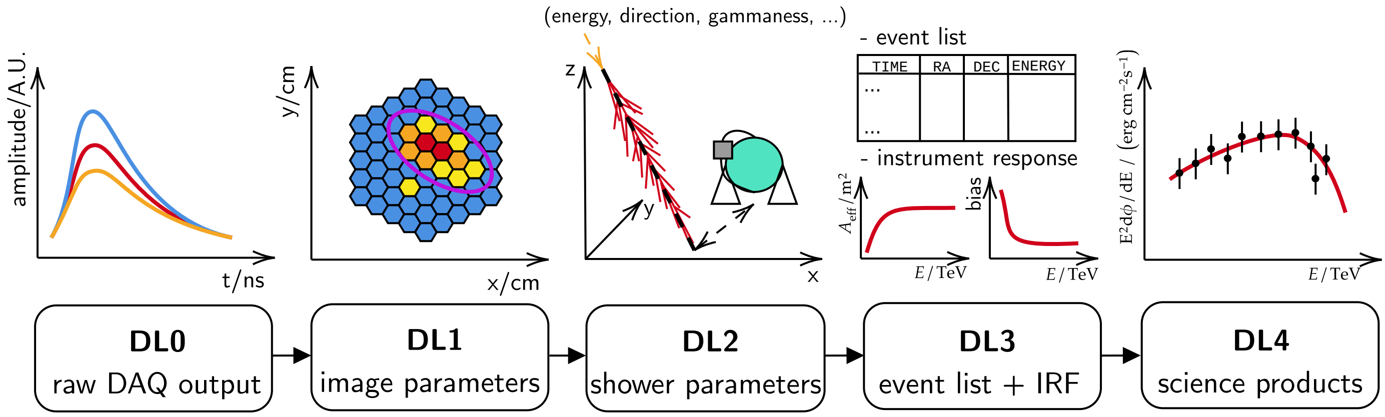


Figure 1. Schematisation of the progressive data reduction and data levels of an IACT. Raw data contain the signal sampled from the photomultipliers at the occurrence of a trigger event (Data Level 0). Calibrated data (Data Level 1) contain the pixelated image of the Cherenkov light of the shower. The latter can be parametrised with few geometrical quantities and used to determine the observables of the original shower, including its probability of being a gamma-ray shower (Data Level 2). The detected events can be gathered in a list of gamma ray candidates, together with the functions representing the response of the system (the so-called instrument response function, IRF), e.g., the collection area of the system as a function of the energy or the bias of its energy reconstruction (Data Level 3). This information can be used to perform a statistical analysis obtaining the so-called science products, in this case the spectrum of the source (Data Level 4).

In the case of an IACT, the raw output of the data acquisition typically consists of binary files containing the waveforms of all the camera pixels, sampled at the occurrence of a trigger event. The raw data are reduced to a list of quantities per pixel (e.g., charge and arrival time) aggregated in the so-called *calibrated* files with size of several GB for each observational run, typically around ∼30 min (in what follows the sizes indicated per each data level are taken from [21], so they refer to VERITAS. One can compare with similar figures reported in [22] for MAGIC). The Cherenkov light of the shower typically illuminates a few pixels in the camera, this pixelated image, representing the distribution of Cherenkov photons, can be parametrised with simple geometrical quantities [23] connected to the shower properties. Image parameters can be fed, at the next data level, to algorithms estimating these properties (e.g., energy and direction of the primary) and classifying the showers initiated by gamma rays against those initiated by cosmic rays, the irreducible background of ground-based gamma-ray telescopes. In the case of particle samplers, such as WCD, the data reduction workflow is similar but instead of camera images, the information is extracted from the pattern in the charge deposited by the shower across the array, as well as from its time evolution. Raw parameters derived from this charge distribution are fed into reconstruction algorithms that, in turn, estimate the relevant shower parameters, such as those mentioned above (see [24] for an overview of the HAWC data reduction pipeline). Having estimated the properties of the shower and of the primary particle generating it, a list of gamma-ray candidates can hence be assembled at the next data level.

At this stage, the information stored within the data products, generally denoted as *high-level*, is independent of the detection technique, as well as the calibration and analysis methods. High-level data typically consist of a list of gamma-ray events along with a parametrisation of the response of the system, the so-called instrument response function (IRF). The latter provides the information necessary to perform a statistical analysis estimating, for example, the significance of the signal, the flux spectrum, or the light curve of the source, which we refer to as science products.

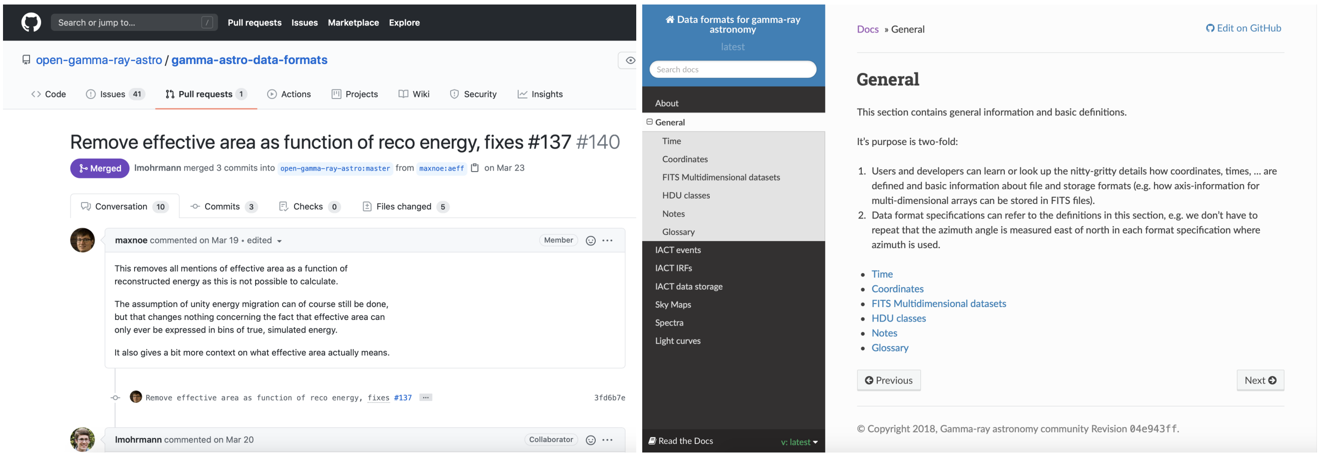
All along the current-generation closed-source analysis chains the data, progressively reduced, are stored in the format associated with the `ROOT` framework, with each collaboration reiterating the effort of defining custom specifications for a data model that shares several commonalities between different experiments. Moreover, even if readable via `ROOT`, the content of these data products cannot be interpreted by a non-expert analyser. There are noticeable efforts to provide analysis tools wrapping these diverse analysis software such as the multi-mission maximum likelihood framework [25]. The ultimate limitation of these tools is though the availability of the experiments to expose their closed-source software and data format and the necessity to implement a new plug-in for each of the instruments considered.

Without a common data model or a general software tool oriented to external users, the current generation of VHE instruments faces different concerns in different time perspectives. At present, multi-instrument analyses simply cannot be performed within a common analysis framework using their proprietary data products. For what concerns the future, as the end of their operation approaches, it is worth to start considering the access to the wealth of data they gathered. If their *legacy* data are to be made public then a release in their original format will make necessary a release of the analysis software as well, which, in turn, has to be maintained. In addition to not being designed for the usage by a large community, this software can rely on libraries that will eventually become deprecated.

2.2. GADF: A Unifying Effort

In the second half of the 2010s, partly to prototype the high-level data format of the forthcoming CTA and partly to exploit the newly available open-source data-analysis software such as `ctools` and `Gammapy`, VHE astronomers started to explore several software-independent implementations of these high-level data. In 2016, in order to coordinate the parallel efforts and to foster the definition of a common and standardised data model, the *Data Formats for Gamma-ray Astronomy* forum (shortly referred to as the "gamma astro data formats", GADF) [26] was established. A community-driven initiative, the GADF consists of a documentation [27] hosted on `GitHub` [28] (Figure 2), specifying the naming scheme, the content, and the metadata of the files containing high-level gamma-ray observations. Though high-level products are the focus of the initiative, specifications for science products are also under discussion. The documentation, openly provided with a Creative Commons Attribution 4.0 license, evolves with the typical `GitHub` workflow: any interested user can propose changes via *issues* that will be discussed among the active members of the initiative, and implemented via *pull requests* that will be ultimately merged once a consensus is reached. Despite the bias towards IACTs, the flexible development of the format allows to accommodate data from other types of instruments, such as space-borne telescopes or WCD. The format has achieved a stable definition and counts already two minor releases, the present being `0.2` [29].

This section illustrates the guiding principles adopted in the development of the GADF specifications, gives an overview of their actual content and highlights the features that make them generalisable to different gamma-ray instruments. The first version of the GADF was designed for IACT, since the major contributors were VHE astronomers preparing for CTA. The data model and the breakdown of the data levels foreseen for CTA are presented in [30], introducing the following naming convention (see also Figure 1): the raw output of the data acquisition is defined as data level 0 (DL0); calibrated files as data level 1 (DL1); reconstructed shower parameters as data level 2 (DL2); sets of selected gamma-ray events and the instrument response as data level 3 (DL3); science products (spectra, light curves, sky maps) as data level 4 (DL4), and observatory results as catalogues, such as data level 5 (DL5). This nomenclature is used within the GADF and will be also adopted in the following text.



open-gamma-ray-astro / gamma-astro-data-formats

Remove effective area as function of reco energy, fixes #137 #140

lmohrmann merged 3 commits into `open-gamma-ray-astro:master` from `maxnoe:aeff` on Mar 23

maxnoe commented on Mar 19 • edited

This removes all mentions of effective area as a function of reconstructed energy as this is not possible to calculate.

The assumption of unity energy migration can of course still be done, but that changes nothing concerning the fact that effective area can only ever be expressed in bins of true, simulated energy.

It also gives a bit more context on what effective area actually means.

`Remove effective area as function of reco energy, fixes #137` `3fd6b7e`

lmohrmann commented on Mar 20

General

This section contains general information and basic definitions.

It's purpose is two-fold:

1. Users and developers can learn or look up the nitty-gritty details how coordinates, times, ... are defined and basic information about file and storage formats (e.g. how axis-information for multi-dimensional arrays can be stored in FITS files).
2. Data format specifications can refer to the definitions in this section, e.g. we don't have to repeat that the azimuth angle is measured east of north in each format specification where azimuth is used.

- Time
- Coordinates
- FITS Multidimensional datasets
- HDU classes
- Notes
- Glossary

© Copyright 2018, Gamma-ray astronomy community Revision 04e943ff.

Figure 2. (**Left**): `GitHub` repository hosting the development of the *data formats for gamma-ray astronomy* specifications. (**Right**): The repository contains a documentation written in `sphinx` whose `html` version can be explored on `readthedocs`.

2.2.1. Format Specifications

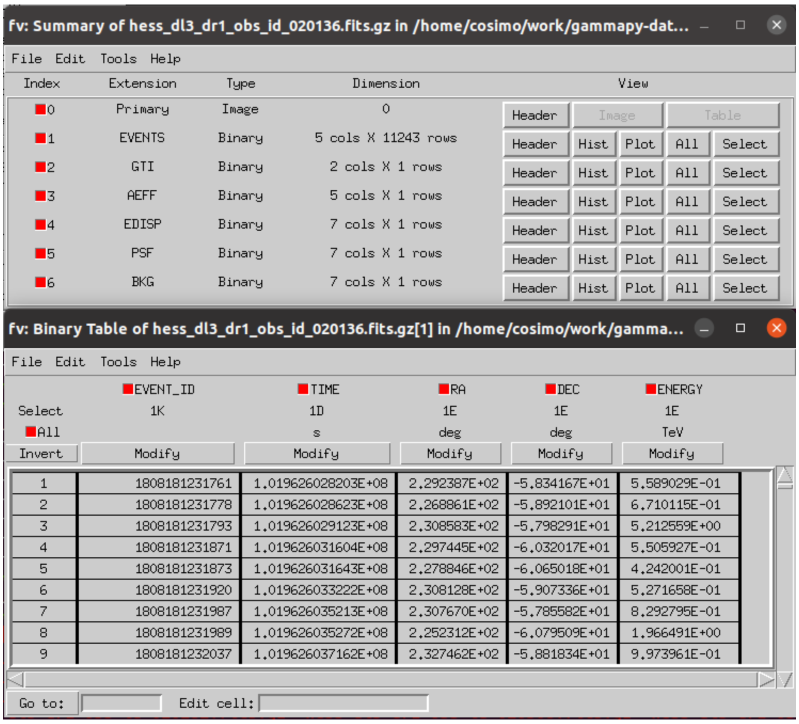
As the GADF is currently the only provider of standardised specifications for high-level VHE gamma-ray data, science tools as `ctools` and `Gammapy` base their data structures on them. Compatibility with open-source data-analysis software is not the only objective of the standardisation effort. One of the guiding principles of the GADF is to produce data whose content is clearly documented and easy to interpret. The file format chosen to host the data is the flexible image transport system (FITS) [31], representing a long-time standard in astronomy at all wavelengths. Another fundamental requirement in the design of the data specifications was to rely as much as possible on already well-established standards used in other FITS files productions, such as those by the missions gathered under NASA's High Energy Astrophysics Science Archive Research Center (HEASARC) [32]. NASA's Office of Guest Investigator Program (OGIP) FITS working group [33] already disseminates to the high-energy astrophysics community recommendations on FITS data productions. These include standards on keyword usage in metadata, on storage of time information, representation of response functions that the GADF extensively follows. The adherence of the GADF to widely used standards ensures additional compatibility with tools already in use by high-energy astrophysicists, such as the `FTOOLS` [34].

As pointed out, the aim of the GADF initiative was to produce specifications for high-level data, therefore, it mostly focuses on the DL3. Nonetheless, the forum discusses data levels higher than the DL3. For example, the OGIP spectral file format [35] is adopted to represent VHE gamma-ray one-dimensional (energy-dependent) spectral data. The compatibility with the OGIP standards ensures that DL3 products can be reduced to spectral data digestible by other established multi-mission analysis tools such as `sherpa` [36,37]. Prototypical specifications for DL4 (such as sky maps, flux points and lightcurves) are under discussion and not yet stable.

2.2.2. GADF DL3 Data

The DL3 is the data level that contains a list of gamma-ray event candidates and the response of the system. All the information in the DL3 files is therefore post-calibration, i.e., already incorporating all the low-level information related to the detector (calibration, gain corrections, digital-count-to-photo-electron conversion) that is hence omitted. A FITS file consists of many extensions, called header data units (HDUs). Each HDU is composed by a header unit, typically containing metadata, and a data unit, containing a n-dimensional array (an image) or a table (in ASCII or binary format). All data units in DL3 files are stored as binary tables.

One of the file extensions contains the event list and, in the associated data unit, a flat table with a column for each event property (see Figure 3). In the current specifications columns listing the events identification number (in the DAQ system), energy, sky coordinates (right ascension and declination) and timestamp are mandatory. Optional columns might include results of the classification algorithms (e.g., a *gammanness* score) and quantities related to the reconstruction (e.g., image or shower parameters). Each file corresponds to a single observing run, therefore the events header unit contains the identification number of the data acquisition run, the type and number of telescopes used in the observation, information about the location of the instrument and its observation mode along with time and duration of the observation. Another HDU is dedicated to a list of good time intervals (GTI), specifying the time periods within the event lists with adequate scientific quality.



fv: Summary of hess_dl3_dr1_obs_id_020136.fits.gz in /home/cosimo/work/gammapy-dat...

Index	Extension	Type	Dimension
🟥0	Primary	Image	0
🟥1	EVENTS	Binary	5 cols X 11243 rows
🟥2	GTI	Binary	2 cols X 1 rows
🟥3	AEFF	Binary	5 cols X 1 rows
🟥4	EDISP	Binary	7 cols X 1 rows
🟥5	PSF	Binary	7 cols X 1 rows
🟥6	BKG	Binary	7 cols X 1 rows

fv: Binary Table of hess_dl3_dr1_obs_id_020136.fits.gz[1] in /home/cosimo/work/gamma...

	🟥EVENT_ID	🟥TIME	🟥RA	🟥DEC	🟥ENERGY
Select	1K	1D	1E	1E	1E
🟥All		s	deg	deg	TeV
1	1808181231761	1.019626028203E+08	2.292387E+02	-5.834167E+01	5.589029E-01
2	1808181231778	1.019626028623E+08	2.268861E+02	-5.892101E+01	6.710115E-01
3	1808181231793	1.019626029123E+08	2.308583E+02	-5.798291E+01	5.212559E+00
4	1808181231871	1.019626031604E+08	2.297445E+02	-6.032017E+01	5.505927E-01
5	1808181231873	1.019626031643E+08	2.278846E+02	-6.065018E+01	4.242001E-01
6	1808181231920	1.019626033222E+08	2.308128E+02	-5.907336E+01	5.271658E-01
7	1808181231987	1.019626035213E+08	2.307670E+02	-5.785582E+01	8.292795E-01
8	1808181231989	1.019626035272E+08	2.252312E+02	-6.079509E+01	1.966491E+00
9	1808181232037	1.019626037162E+08	2.327462E+02	-5.881834E+01	9.973961E-01

Figure 3. Example of a DL3 file compliant with the GADF specifications. Top: the header data units (or extensions) of the file contain the event lists, under (`EVENTS`), followed by those representing the good time intervals (`GTI`) and the instrument response components: effective area (`AEFF`), energy dispersion (`EDISP`), point spread function (`PSF`) and background (`BKG`). Bottom: event list table and its content.

The response of the system is needed to properly relate the reconstructed events with astrophysical source properties. It is assumed that this response can be factorised in different components. The components considered are: the effective area, describing the acceptance of the system to gamma-ray events; the energy dispersion (or migration matrix), describing the probability distribution of the energy estimator and the point spread function (PSF), describing the probability distribution of the direction estimator. The background rate (measuring the rate of cosmic ray events misclassified as gamma rays) might be included among the IRF components, however it is not mandatory. The IRF components depend on observational (e.g., atmospheric conditions, zenith and azimuth angle of the pointing) and physical quantities (e.g., the energy or direction of the showers). The IRF components considered in the format are valid for a single exposure, which is typically defined by constant observational conditions (e.g., zenith range, atmospheric quality, etc.), hence considering any such dependency of the IRF averaged out. In the current specifications, the dependencies on physical quantities considered are the photon energy and the offset of its position from the centre of the instrument field of view (a response symmetric with the offset coordinate is assumed). As an example, Figure 4 illustrates the energy and offset dependency of the effective area component for a H.E.S.S. observation stored in the GADF DL3 format. IRF components are not stored in flat tables: energy and offset bin edges are stored in separate columns, and a last column contains a multi-dimensional array corresponding to the response in each bin. OGIP specifications are followed in storing both events and IRF components.

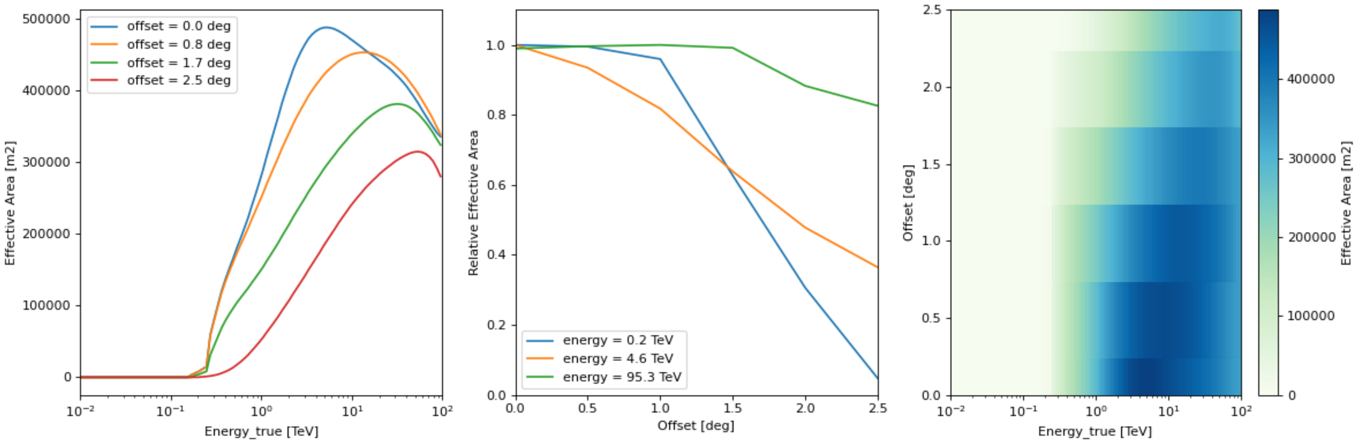


Figure 4. Example of visualisation of the effective area of a IACT and its dependency on energy and offset angle. The IRF component is read from DL3 files and displayed using `Gammapy`.

3. Projects Successfully Using the Standardised Data Format

To illustrate the maturity of the GADF standardisation effort, we review, in the following sections, projects that have successfully employed its specifications.

3.1. The H.E.S.S. First Public Test Data Release

The H.E.S.S. collaboration was the first to publicly release a test dataset in a DL3 format compliant with the GADF specifications. Few observations, amounting roughly to $\sim$50 h of observation time, gathered between 2004 and 2008 were published in the so called H.E.S.S. DL3 Data Release 1 (H.E.S.S. DL3 DR1) [38,39] to promote the standardisation effort but also to allow to test the open-source science tools in development with actual IACT data. The data release contains 30 h of observations of sources representing different galactic and extragalactic science cases, and 20 h of observations of field of views empty of known gamma-ray emitters, also labelled as *off* data, to be used for background estimation. Table 1 summarises the content of this data release.

Table 1. Content of the H.E.S.S. Data Level 3 Data Release 1.

Source	Type	Case Study	Time/h
Crab Nebula	Pulsar Wind Nebula	point-like, steady source	1.9
PKS 2155-304	Blazar	point-like, variable source	9.8
MSH 15-52	Pulsar Wind Nebula	small-extension, steady source	9.1
RX J1713.7-3946	Supernova Remnant	large-extension, steady source	7.0
off data	various	background estimation	20.7

3.2. The Joint-Crab Project

With multi-instrument analyses being one of the main objectives of the standardisation effort, after the first public release of GADF-compliant DL3 data, the next step in the format validation would have naturally been the combination of data from different experiments.

In the so-called *joint-crab* project [40], Crab Nebula observations from *Fermi*-LAT and four of the currently operating IACTs, produced in a GADF-compliant format, were combined in the first multi-instrument and fully-reproducible gamma-ray analysis. The datasets used were:

- 7 yr of *Fermi*-LAT observations, obtained in the custom high-level, DL3, format with which they are publicly released. They were reduced, before the final statistical analysis, to OGIP spectral data;
- 2 h of H.E.S.S. observations selected from the H.E.S.S. DL3 DR1 (see Section 3.1);
- 40 min of MAGIC observations produced and released specifically for this project;
- 40 min of VERITAS observations produced and released specifically for this project;
- 10 h of FACT observations from their already public data sample (see Section 2.1).

To illustrate a prototypical analysis example, the Crab Nebula spectrum (Figure 5 right) was estimated, combining all the observations in an energy-dependent (or one-dimensional) joint binned likelihood. In this analysis technique, classically employed by IACT, source and background events are extracted via aperture photometry (Figure 5 left) and then an energy-dependent analytical flux model is folded with the response of the system to estimate the number of counts maximising the Poissonian likelihood describing the counts in each energy bin. The *joint-crab* project relied only on open-source software for its statistical analyses (`Gammapy`). Datasets, scripts reproducing all the analysis steps and tutorial notebooks are publicly provided on `GitHub` [41], along with a `conda` environment freezing the exact dependencies used in the paper and a docker container [42] to guarantee a long-term reproducibility. The entire package was also archived on `zenodo` [43]. Given the approach proposed and the assets openly made available, this work not only implements the first fully-reproducible gamma-ray analysis but also constitutes the first joint public release of IACT DL3 data.

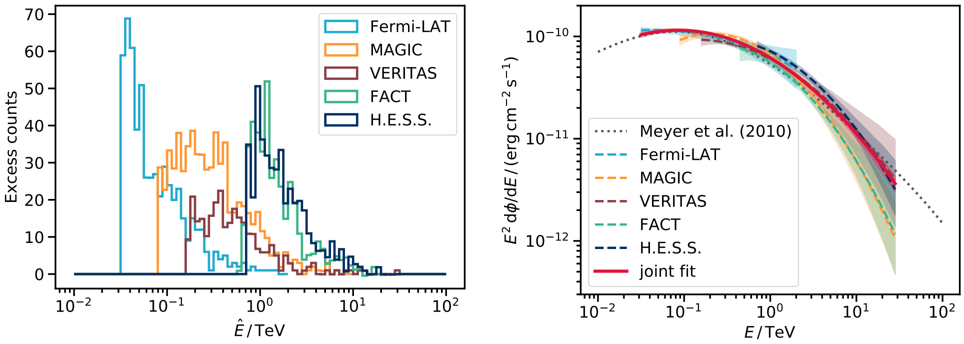


Figure 5. (**Left**): Source counts vs estimated energy extracted via aperture photometry, per each of the instrument datasets in [40]. (**Right**): Estimated flux spectrum of the Crab Nebula obtained from the individual instrument datasets (same colour code as in the figure on the left) and considering all the datasets in the same joint likelihood (red). The grey dashed line represents a bibliographic reference. In all cases the analytical flux model considered in the likelihood is a curved power-law. Figures from [40].

3.3. Analysis of the H.E.S.S. Public Data Release with Ctools

In addition to evolving in parallel with the GADF, the open-source science tools can recognise data with its specifications as input. In [44], the H.E.S.S. DL3 DR1 (Section 3.1) was used to test the capabilities of `ctools`, until then mainly used to analyse simulated CTA observations and calculate prospects for its observational capabilities. The authors presented a method to build a parametric model describing the spatial and spectral distribution of the background events in the H.E.S.S. DL3 DR1. The latter was used to perform a spectro-morphological (three-dimensional) analysis estimating the spectrum of the 4 sources included in the data release. Differently than in the one-dimensional analysis described in Section 3.2, the sources positions and morphology are included among the parameters of the model used to estimate the flux. Source and background counts are not separated, rather the background is included among the components of a model that in this case predicts the flux in the entire field of view, allowing to take into account multiple sources at a time (see [18] Section 2 for a detailed explanation). This approach has been successfully used by the *Fermi*-LAT collaboration for all its scientific publications. The results of binned and unbinned three-dimensional likelihood analyses are compared against the simpler one-dimensional binned analysis, also implemented in `ctools`, and against bibliographic references obtained from the same sources. The consistency of the results obtained with `ctools` with the different statistical methods applied and with the literature (see Figure 6 left) testifies the maturity not only of the science tool, but also of the GADF scheme that correctly encapsulates all the information needed for correct reproduction of scientific results. The paper finally illustrates the capability of `ctools`, being built on the `gammalib` library [18], to simultaneously analyse gamma-ray data with different specifications, i.e., to analyse *Fermi*-LAT data in their own high-level format (without the reduction described in Section 3.2) and IACT DL3 data compliant with the GADF specifications.

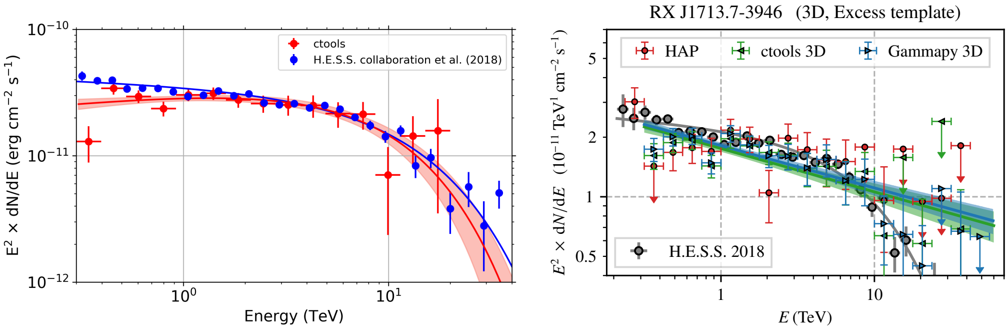


Figure 6. Flux spectrum of the extended gamma-ray source RXJ1713.7-3946. (**Left**): Comparison of the result obtained from the H.E.S.S. DL3 DR1 using `ctools`' three-dimensional unbinned likelihood analysis (red) against the literature (blue). Figure from [44]. (**Right**): Comparison of the result obtained from the H.E.S.S. DL3 DR1 using `ctools`' (green) and `Gammapy`'s (blue) three-dimensional likelihood analyses against a result obtained on the very same data sample with the H.E.S.S. private analysis chain (red), performing a one-dimensional analysis. A bibliographic reference, the same as in the figure on the left, is given in grey. Figure from [45].

3.4. *Validation of Open-Source Science Tools and Background Model Construction in γ-ray Astronomy*

Expanding on the project described in Section 3.3, ref. [45] aims at testing both `Gammapy` and `ctools` using the H.E.S.S. DL3 DR1. The results of the one-dimensional and three-dimensional analyses provided by both science tools are validated against each other. For the three-dimensional analysis, a novel background model is used, not parameterised from the *off* sources within the H.E.S.S. DL3 DR1, but built using ∼4000 h of H.E.S.S. private observations. For this work, the results of the science tools are validated not only against the literature, but also against the results obtained with one of the closed-source analysis chains of the H.E.S.S. collaboration, performing a classical one-dimensional analysis on the exact same observations included in the H.E.S.S. data release (see Figure 6 right). The agreement of the results of the different science tools among them and with the private analysis chain represents a landmark in the analysis tools and data formats validation for future VHE gamma-ray analyses.

3.5. *Open and Standardised Formats for γ-Ray Analysis Applied to HAWC Observatory Data*

The GADF specifications were primarily developed by and for the IACT community. However, due to their generality, it is possible to use them to format data from WCD, such as the HAWC observatory, as shown by [46]. In this work the authors presented the first GADF-compliant production of event lists and instrument response functions for a ground-based wide-field instrument. These data products were then used to reproduce with excellent agreement the published spectrum of the Crab Nebula as measured by HAWC. This result, shown in Figure 7, was obtained using the open-source software `Gammapy`. As highlighted by Section 3.2, a common data format and shared analysis tools allow multi-instrument joint analysis and effective data sharing. This synergy between experiments is particularly relevant given the complementary nature of pointing and wide-field instruments. This will be specially relevant for the joint scientific exploitation of future observatories such as SWGO and CTA.

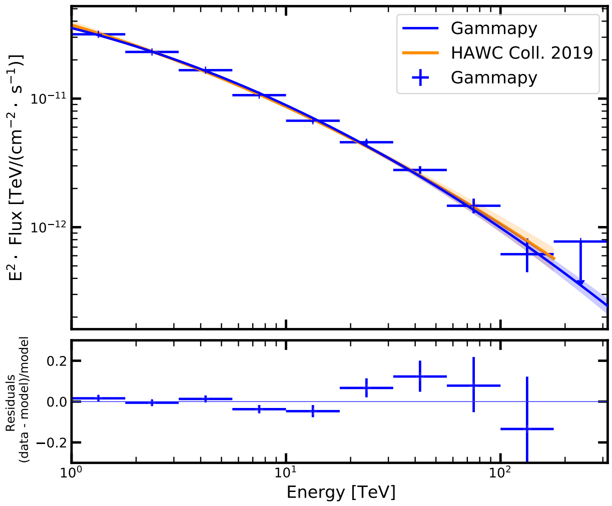


Figure 7. Estimated flux spectrum of the Crab Nebula derived from the GADF data production using `Gammapy` (blue) compared against the reference HAWC spectrum from [47] (orange). The bottom panel shows the residual comparison of the obtained flux points with the reference spectrum. Figure from [46].

4. Discussion

The future of data formats in gamma-ray astronomy will very likely be linked to the future of the GADF initiative. As discussed over the text, this community-driven initiative has proposed the first available set of specifications for high-level data for the current and next generation of ground-based gamma-ray instruments. In this section, we will discuss the main limitations affecting current specifications, as well as foreseeable ways in which they will evolve over the next decade.

One of the main drivers of the evolution and improvement of the GADF will be the target requirements imposed by the future ground-based observatories. These will impose high-level data (and especially, the IRFs) to be described and parameterised in more complex ways, directly benefiting also the current generation of instruments. Possible extensions of the format to meet these requirements could include: a better field of view binning approach, removing the assumption of radial symmetry; inclusion of time dependency in the IRF components; distinguishing between different event types based on the hardware, reconstruction or analysis settings. Mature format specifications will be crucial for defining and testing current instruments legacy data, as they face the challenge of digesting decades of data (taken by instruments with evolving capabilities) and ensuring their proper use and interpretation.

In order to confront these challenges and to ensure the long-term feasibility of the GADF specifications, a more formal governance structure is needed. For this reason, a body of representatives from the high-energy ground-based community will be defined to act as a coordination committee. This governance definition effort, currently in progress, will inherit from the evolution of similar community-driven initiatives (for instance, the Astropy Project role responsibilities [48]).

Even if the GADF specifications were inspired by high-energy satellites and primarily developed by and for the IACT community, they are able to represent high-level data products from other event-based high-energy astrophysical instruments. As shown in

Section 3.5, other high-energy gamma-ray observatories, such as WCD (like HAWC or the future SWGO) naturally fit the GADF specifications, allowing the use of available open-source data analysis tools. In the coming years, the inclusion of other observatories will be explored, especially in the context of high-energy multi-messenger astronomy: allowing the inclusion of data from neutrino or even gravitational wave observatories would require some changes to the specifications, but at the same time would naturally allow the use of common science tools for joint multi-messenger analyses.

5. Conclusions

This review presented an outlook on the evolution of the data format in VHE gamma-ray astronomy from private and diverse specifications to the open and standardised ones proposed under the GADF initiative. The GADF initiative is presented as a community-driven effort to provide a common and open high-level data format for gamma-ray instruments. The specifications proposed within the GADF refer to high-level data products that would allow the production of scientific results: they are independent of the particular detection technique, thus allowing to accommodate data from different telescopes (e.g., IACT and WCD). The format definition was driven by the requirement to operate the next generation of gamma-ray instruments (such as CTA) as open observatories, with the consequent need of providing non-expert external users with open data products that are easy to interpret. Another aspect of this demand was the development of open-source gamma-ray data-analysis tools, whose evolution is now also linked to the data standardisation effort.

Current GADF specifications have proven to be robust by several publications analysing GADF-compliant data with these open-source science tools, validating their results against those obtained with the established closed-source software in use by current collaborations. These publications confirmed not only the correctness of the information incorporated in the format specifications but, at the same time, the capabilities of this new generation of open-source science tools. Other publications have instead proven the feasibility of multi-instrument and fully-reproducible analyses once the common format and open software are used. Even if future instruments are driving the open data and software development, the current generation can significantly benefit from their advancement. Their adoption ensures a larger user and maintainer base for the legacy data of current instruments, and, eventually, more sophisticated data storage and analysis techniques. The H.E.S.S. collaboration already pioneered a first public release of GADF-compliant data. All currently operating VHE gamma-ray experiments are nowadays also able to produce GADF-compliant data products, though for the moment they have mostly been used internally. Multi-instrument scientific projects using these data products are on their way, sharing data among collaborations through the use of memoranda of understanding.

The standardisation effort remains open to the inclusion not only of more gamma-ray instruments but also of telescopes observing the universe with other messengers. With the initiative being community-driven, high-energy astrophysicist in need of new extensions to the format are able to propose them. The recent efforts reviewed in this issue successfully employing GADF-compliant data and open-source analysis tools will surely foster their usage for further scientific projects. The GADF does not represent an isolated effort and aims at maintaining compatibility with other established standards in high-energy astronomy, such as the OGIP (on which the GADF largely draws), or those used for high-level products within the virtual observatory [49]. Promoting the use of open-source analysis tools, as well as common open data formats will distinguish high-energy astrophysics in the future as one of the few branches of modern science unconcerned by the reproducibility dilemma affecting many other disciplines [50].

Author Contributions: C.N. conceptualized the paper and wrote the original draft. All authors contributed equally to the review of the relevant literature, to writing and editing the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Commission's Horizon 2020 Program under grant agreement 824064 (ESCAPE—European Science Cluster of Astronomy and Particle Physics ESFRI Research Infrastructures), by the the ERDF under the Spanish Ministerio de Ciencia e Innovación (MICINN, grant PID2019-107847RB-C41), and from the CERCA program of the Generalitat de Catalunya.

Acknowledgments: We are grateful to Gernot Maier, Maximilian Nöthe, Bruno Khélifi, Catherine Boisson and Karl Kosack for their suggestions and for reviewing the manuscript.

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

References

1. Funk, S. Ground- and Space-Based Gamma-ray Astronomy. *Annu. Rev. Nucl. Part. Sci.* **2015**, *65*, 245–277. [CrossRef]
2. Thompson, D.J. Space detectors for gamma rays (100 MeV-100 GeV): From EGRET to Fermi LAT. *Comptes Rendus Phys.* **2015**, *16*, 600–609. [CrossRef]
3. de Naurois, M.; Mazin, D. Ground-based detectors in very-high-energy gamma-ray astronomy. *Comptes Rendus Phys.* **2015**, *16*, 610–627. [CrossRef]
4. Cherenkov Telescope Array Consortium. *Science with the Cherenkov Telescope Array*; World Scientific Publishing Co. Pte. Ltd.: Singapore, 2019; [CrossRef]
5. Bai, X.; Bi, B.Y.; Bi, X.J.; Cao, Z.; Chen, S.Z.; Chen, Y.; Chiavassa, A.; Cui, X.H.; Dai, Z.G.; della Volpe, D.; et al. The Large High Altitude Air Shower Observatory (LHAASO) Science White Paper. *arXiv* **2019**, arXiv:1905.02773.
6. Albert, A.; Alfaro, R.; Ashkar, H.; Alvarez, C.; Alvarez, J.; Arteaga-Velázquez, J.C.; Ayala Solares, H.A.; Arceo, R.; Bellido, J.A.; BenZvi, S.; et al. Science Case for a Wide Field-of-View Very-High-Energy Gamma-ray Observatory in the Southern Hemisphere. *arXiv* **2019**, arXiv:1902.08429.
7. National Aeronautics and Space Administration (NASA). *Fermi* Data Policy. Available online: https://fermi.gsfc.nasa.gov/ssc/data/policy/ (accessed on 28 September 2021).
8. Pittori, C.; The Agile-SSDC Team. The AGILE data center and its legacy. *Rend. Lincei Sci. Fis. Nat.* **2019**, *30*, 217–223. [CrossRef]
9. National Aeronautics and Space Administration (NASA). *Fermi* Science Support Center. Available online: https://fermi.gsfc.nasa.gov/ssc/ (accessed on 28 September 2021).
10. Space Science Data Center (SSDC). AGILE Data Center. Available online: https://agile.ssdc.asi.it/ (accessed on 28 September 2021).
11. Zanin, R.; Carmona, E.; Sitarek, J.; Colin, P.; Frantzen, K.; Gaug, M.; Lombardi, S.; Lopez, M.; Moralejo, A.; Satalecka, K.; et al. MARS, The MAGIC Analysis and Reconstruction Software. In Proceedings of the 33rd International Cosmic ray Conference (ICRC2013), Rio de Janeiro, Brasil, 2–6 July 2013; Volume 33, p. 2937.
12. Khelifi, B.; Djannati-Ataï, A.; Jouvin, L.; Lefaucheur, J.; Lemiere, A.; Pita, S.; Tavernier, T.; Terrier, R. HAP-Fr, a pipeline of data analysis for the HESS-II experiment. In Proceedings of the 34th International Cosmic ray Conference (ICRC2015), The Hague, The Netherlands, 30 July–6 August 2015; Volume 34, p. 837.
13. Maier, G.; Holder, J. Eventdisplay: An Analysis and Reconstruction Package for Ground-based Gamma-ray Astronomy. In Proceedings of the 35th International Cosmic ray Conference (ICRC2017), Busan, Korea, 12–20 July 2017; Volume 301, p. 747.
14. Brügge, K.; Nöthe, M.; Buß, J.; Temme, T.F.; Bulinski, M.; Mueller, S.A.; Bockermann, C.; Linhoff, L.; Freiwald, J.; Neise, D.; et al. Fact-tools. *Zenodo* **2019**. [CrossRef]
15. FACT Collaboration. FACT Open Data. Available online: https://fact-project.org/data/ (accessed on 28 September 2021).
16. HAWC Observatory. Public Datasets. Available online: https://data.hawc-observatory.org/ (accessed on 28 September 2021).
17. Lamanna, G.; Antonelli, L.A.; Contreras, J.L.; Knödlseder, J.; Kosack, K.; Neyroud, N.; Aboudan, A.; Arrabito, L.; Barbier, C.; Bastieri, D.; et al. Cherenkov Telescope Array Data Management. In Proceedings of the 34th International Cosmic ray Conference (ICRC2015), The Hague, The Netherlands, 30 July–6 August 2015; Volume 34, p. 947.
18. Knödlseder, J.; Mayer, M.; Deil, C.; Cayrou, J.-B.; Owen, E.; Kelley-Hoskins, N.; Lu, C.-C.; Buehler, R.; Forest, F.; Louge, T.; et al. GammaLib and ctools. A software framework for the analysis of astronomical gamma-ray data. *Astron. Astrophys.* **2016**, *593*, A1. [CrossRef]
19. Deil, C.; Zanin, R.; Lefaucheur, J.; Boisson, C.; Khélifi, B.; Terrier, R.; Wood, M.; Mohrmann, L.; Chakraborty, N.; Watson, J.; et al. Gammapy—A prototype for the CTA science tools. In Proceedings of the 35th International Cosmic ray Conference (ICRC2017), Busan, Korea, 12–20 July 2017; Volume 301, p. 766.
20. Brun, R.; Rademakers, F. ROOT—An object oriented data analysis framework. *Nucl. Instrum. Methods Phys. Res. A* **1997**, *389*, 81–86. [CrossRef]
21. Cogan, P. VEGAS, the VERITAS Gamma-ray Analysis Suite. In Proceedings of the 30th International Cosmic Ray Conference (ICRC2007), Mérida, Yucatán, Mexico, 3–11 July 2007; Volume 3, pp. 1385–1388.
22. Nigro, C. Study of Persistent and Flaring Gamma-ray Emission from Active Galactic Nuclei with the MAGIC Telescopes and Prospects for Future Open Data Formats in Gamma-ray Astronomy. Ph.D. Thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, Berlin, Germany, 2019; [CrossRef]

23. Hillas, A.M. Cerenkov Light Images of EAS Produced by Primary Gamma rays and by Nuclei. In Proceedings of the 19th International Cosmic ray Conference, San Diego, CA, USA, 11–23 August 1985; Volume 3, p. 445.
24. Abeysekara, A.U.; Albert, A.; Alfaro, R.; Alvarez, C.; Álvarez, J.D.; Arceo, R.; Arteaga-Velázquez, J.C.; Ayala Solares, H.A.; Barber, A.S.; Bautista-Elivar, N.; et al. Observation of the Crab Nebula with the HAWC Gamma-ray Observatory. *Astrophys. J.* **2017**, *843*, 39. [CrossRef]
25. Vianello, G.; Lauer, R.; Burgess, J.M.; Ayala, H.; Fleischhack, H.; Harding, P.; Hui, M.; Marinelli, S.; Savchenko, V.; Zhou, H. The Multi-Mission Maximum Likelihood framework (3ML). *arXiv* **2015**, arXiv:1507.08343.
26. Deil, C.; Boisson, C.; Kosack, K.; Perkins, J.; King, J.; Eger, P.; Mayer, M.; Wood, M.; Zabalza, V.; Knödlseder, J.; et al. Open high-level data formats and software for gamma-ray astronomy. In Proceedings of the 6th International Symposium on High Energy Gamma-ray Astronomy, Heidelberg, Germany, 11–15 July 2016; American Institute of Physics Conference Series, Volume 1792, p. 070006. [CrossRef]
27. Data Formats for Gamma-ray Astronomy. Online Documentation. Available online: https://gamma-astro-data-formats.readthedocs.io/ (accessed on 28 September 2021).
28. Data Formats for Gamma-ray Astronomy. `GitHub` Repository. Available online: https://github.com/open-gamma-ray-astro/gamma-astro-data-formats (accessed on 28 September 2021).
29. Deil, C.; Wood, M.; Hassan, T.; Boisson, C.; Contreras, J.L.; Knödlseder, J.; Khelifi, B.; King, J.; Mohrmann, L. Data Formats for Gamma-ray Astronomy—Version 0.2. *Zenodo* **2018**. [CrossRef]
30. Contreras, J.L.; Satalecka, K.; Bernlöhr, K.; Boisson, C.; Bregeon, J.; Bulgarelli, A.; de Cesare, G.; de los Reyes, R.; Fioretti, V.; Kosack, K.; et al. Data model issues in the Cherenkov Telescope Array project. In Proceedings of the 34th International Cosmic ray Conference (ICRC2015), The Hague, The Netherlands, 30 July–6 August 2015; Volume 34, p. 960.
31. Wells, D.C.; Greisen, E.W.; Harten, R.H. FITS - A Flexible Image Transport System. *Astron. Astrophys. Suppl. Ser.* **1981**, *44*, 363.
32. National Aeronautics and Space Administration (NASA). High Energy Astrophysics Science Archive Research Center. Available online: https://heasarc.gsfc.nasa.gov/ (accessed on 28 September 2021).
33. Corcoran, M.F.; Angelini, L.; George, I.; McGlynn, T.; Mukai, K.; Pence, W.; Rots, A. The OGIP FITS Working Group. In *Astronomical Data Analysis Software and Systems IV*; Shaw, R.A., Payne, H.E., Hayes, J.J.E., Eds.; Astronomical Society of the Pacific Conference Series; 1995, Volume 77, p. 219. Available online: https://www.adass.org/adass/proceedings/adass94/corcoranm.html (accessed on 26 September 2021).
34. National Aeronautics and Space Administration (NASA). FTOOLS: A General Package of Software to Manipulate FITS Files. Available online: https://heasarc.gsfc.nasa.gov/ftools/ (accessed on 28 September 2021).
35. Arnaud, K.A.; George, I.M. The OGIP Spectral File Format, OGIP Memo OGIP/92-007. Available online: https://heasarc.gsfc.nasa.gov/docs/heasarc/ofwg/docs/spectra/ogip_92_007/ogip_92_007.html (accessed on 28 September 2021).
36. Freeman, P.; Doe, S.; Siemiginowska, A. Sherpa: A mission-independent data analysis application. In *Astronomical Data Analysis*; Starck, J.L., Murtagh, F.D., Eds.; Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series; 2001; Volume 4477, pp. 76–87. Available online: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/4477/1/Sherpa-a-mission-independent-data-analysis-application/10.1117/12.447161.short (accessed on 26 September 2021).
37. Doe, S.; Nguyen, D.; Stawarz, C.; Refsdal, B.; Siemiginowska, A.; Burke, D.; Evans, I.; Evans, J.; McDowell, J. Developing Sherpa with Python. In *Astronomical Data Analysis Software and Systems XVI*; Shaw, R.A., Hill, F., Bell, D.J., Eds.; Astronomical Society of the Pacific Conference Series; 2007, Volume 376, p. 543. Available online: https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwj66tiD-cbzAhVoQfUHHbBnCSUQFnoECAUQAQ&url=https (accessed on 26 September 2021).
38. H.E.S.S. Collaboration H.E.S.S. first public test data release. *arXiv* **2018**, arXiv:1810.04516.
39. High Energy Stereoscopic System (H.E.S.S.). H.E.S.S. First Public Test Data Release. Available online: https://www.mpi-hd.mpg.de/hfm/HESS/pages/dl3-dr1/ (accessed on 28 September 2021).
40. Nigro, C.; Deil, C.; Zanin, R.; Hassan, T.; King, J.; Ruiz, J.E.; Saha, L.; Terrier, R.; Brügge, K.; Nöthe, M.; et al. Towards open and reproducible multi-instrument analysis in gamma-ray astronomy. *Astron. Astrophys.* **2019**, *625*, A10. [CrossRef]
41. Towards Open and Reproducible Multi-Instrument Analysis in Gamma-ray Astronomy. `GitHub` Repository. 2021 Available online: https://github.com/open-gamma-ray-astro/joint-crab (accessed on 28 September 2021).
42. Towards Open and Reproducible Multi-Instrument Analysis in Gamma-ray Astronomy. `DockerHub` Container. Available online: https://hub.docker.com/r/gammapy/joint-crab (accessed on 28 September 2021).
43. Nigro, C.; Deil, C.; Zanin, R.; Hassan, T.; King, J.; Ruiz, J.E.; Saha, L.; Terrier, R.; Bruegge, K.; Noethe, M.; et al. The Joint-Crab Bundle. *Zenodo* **2018**. [CrossRef]
44. Knödlseder, J.; Tibaldo, L.; Tiziani, D.; Specovius, A.; Cardenzana, J.; Mayer, M.; Kelley-Hoskins, N.; Venere, L.D.; Bonnefoy, S.; Ziegler, A.; et al. Analysis of the H.E.S.S. public data release with ctools. *Astron. Astrophys.* **2019**, *632*, A102. [CrossRef]
45. Mohrmann, L.; Specovius, A.; Tiziani, D.; Funk, S.; Malyshev, D.; Nakashima, K.; van Eldik, C. Validation of open-source science tools and background model construction in γ-ray astronomy. *Astron. Astrophys.* **2019**, *632*, A72. [CrossRef]
46. Olivera-Nieto, L.; Joshi, V.; Schoorlemmer, H.; Donath, A. Standardized formats for gamma-ray analysis applied to HAWC observatory data. *PoS* **2021**, *ICRC2021*, 727. [CrossRef]

47. Abeysekara, A.U.; Albert, A.; Alfaro, R.; Alvarez, C.; Álvarez, J.D.; Camacho, J.R.A.; Arceo, R.; Arteaga-Velázquez, J.C.; Arunbabu, K.P.; Rojas, D.A.; et al. Measurement of the Crab Nebula Spectrum Past 100 TeV with HAWC. *Astrophys. J.* **2019**, *881*, 134. [CrossRef]
48. The Astropy Project. Astropy Team. Available online: https://www.astropy.org/team.html (accessed on 28 September 2021).
49. Louys, M.; Bonnarel, F.; Schade, D.; Dowler, P.; Micol, A.; Durand, D.; Tody, D.; Michel, L.; Salgado, J.; Chilingarian, I.; et al. Observation Data Model Core Components, its Implementation in the Table Access Protocol Version 1.0. *arXiv* **2011**, arXiv:1111.1758. doi:10.5479/ADS/bib/2011ivoa.spec.1028T.
50. Baker, M. 1500 scientists lift the lid on reproducibility. *Nature* **2016**, *533*, 452–454. [CrossRef]