Open high-level data formats and software for gamma-ray astronomy

Christoph Deil[1,a)], Catherine Boisson[5,b)], Karl Kosack[3], Jeremy Perkins[14], Johannes King[1], Peter Eger[1], Michael Mayer[8], Matthew Wood[15], Victor Zabalza[17], Jürgen Knödlseder[13], Tarek Hassan[12], Lars Mohrmann[7], Alexander Ziegler[7], Bruno Khelifi[6], Daniela Dorner[7], Gernot Maier[9], Giovanna Pedaletti[9], Jaime Rosado[12], José Luis Contreras[12], Julien Lefaucheur[5], Kai Brügge[2], Mathieu Servillat[5], Régis Terrier[6], Roland Walter[10] and Saverio Lombardi[16]

[a)]Corresponding author: Christoph.Deil@mpi-hd.mpg.de
[b)]Corresponding author: catherine.boisson@obspm.fr
[1]*MPIK, Heidelberg, Germany*
[2]*TU, Dortmund, Germany*
[3]*CEA/IRFU/SAp, CEA Saclay, Bat 709, Orme des Merisiers, 91191 Gif-sur-Yvette, France*
[4]*NASA/GSFC, USA*
[5]*LUTH, Observatoire de Paris, Meudon, France*
[6]*APC, Université Paris Diderot, CNRS/IN2P3, Paris, France*
[7]*FAU, Erlangen, Germany*
[8]*Humboldt University, Berlin, Germany*
[9]*DESY, Zeuthen, Germany*
[10]*Observatoire de Genève, 51 chemin des Maillettes, 1290 Sauverny, Switzerland*
[11]*Universidad Complutense de Madrid*
[12]*Institut de Fisica d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona) Spain*
[13]*IRAP, Toulouse, France*
[14]*NASA/GSFC*
[15]*SLAC National Accelerator Laboratory*
[16]*INAF, Osservatorio Astronomico di Roma, via Frascati 33, 00040 Monte Porzio Catone (Roma), Italy*
[17]*University of Leicester, UK*

Abstract. In gamma-ray astronomy, a variety of data formats and proprietary software have been traditionally used, often developed for one specific mission or experiment. Especially for ground-based imaging atmospheric Cherenkov telescopes (IACTs), data and software are mostly private to the collaborations operating the telescopes. However, there is a general movement in science towards the use of open data and software. In addition, the next-generation IACT instrument, the Cherenkov Telescope Array (CTA), will be operated as an open observatory.

We have created a Github organisation at `https://github.com/open-gamma-ray-astro` where we are developing high-level data format specifications. A public mailing list was set up at `https://lists.nasa.gov/mailman/listinfo/open-gamma-ray-astro` and a first face-to-face meeting on the IACT high-level data model and formats took place in April 2016 in Meudon (France). This open multi-mission effort will help to accelerate the development of open data formats and open-source software for gamma-ray astronomy, leading to synergies in the development of analysis codes and eventually better scientific results (reproducible, multi-mission).

This write-up presents this effort for the first time, explaining the motivation and context, the available resources and process we use, as well as the status and planned next steps for the data format specifications. We hope that it will stimulate feedback and future contributions from the gamma-ray astronomy community.

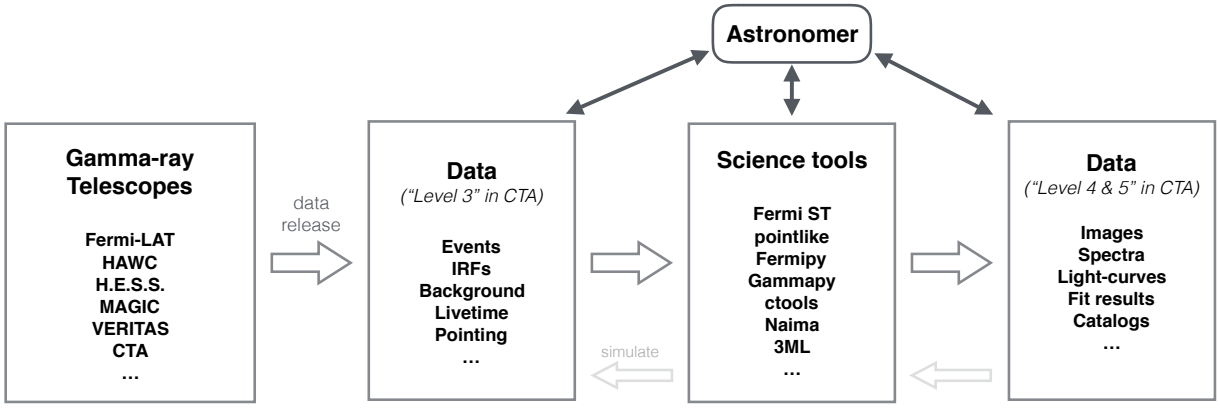

FIGURE 1. The purpose of the `gamma-astro-data-formats` effort is to encourage collaboration between high-level gamma-ray data producers, science tool developers, and data analysts. The goal is to develop common data formats to avoid duplication of efforts and confusion by astronomers working with multi-mission gamma-ray data or multiple analysis tools.

Introduction

The Flexible Image Transport System (FITS) format was created around 1980 [1] by optical astronomers. In the 1990s, the HEASARC FITS Working Group, also known as the OGIP (Office of Guest Investigator Programs) FITS Working Group, produced documents and recommendations concerning the storage of X-ray (and partly gamma-ray space telescope) data in FITS.[1] Several of these recommendations have subsequently been incorporated into the FITS standard, the latest version is FITS 3.0 from 2010 [2].

Today, very-high energy (VHE, energy > 50 GeV) gamma-ray astronomy is finding itself in a similar situation like X-ray astronomy in the 1990s (illustrated in Figure 1). The existing ground-based imaging atmospheric Cherenkov telescopes (IACTs) like e.g. H.E.S.S., MAGIC and VERITAS, have been operating independently for the past decade, using proprietary data formats and codes. Data from each IACT is stored in ROOT files containing serialised C++ objects and can only be read with the private software. The Cherenkov Telescope Array (CTA), the next generation IACT instrument, will be operated as an an open observatory, meaning that data and analysis software will be public to all astronomers. Current IACTs have started to "export" their data and instrument response functions (IRFs) to FITS, partly as a prototyping effort for CTA, but also to take advantage of the open-source science tool codes for gamma-ray astronomy (Gammapy [3], ctools [4], pointlike [5], Fermi ScienceTools[2], Fermipy[3], 3ML [6], Naima [7], ...) and to have an archival and common data data format that allows joint analysis with other astronommical multi-wavelength datasets. For science data products, the term "data level 3" (DL3) is used for event lists, IRFs and auxiliary data for analysis and provenance, "data level 4" (DL4) for higher-level science data products like images, spectra and lightcurves, and "data level 5" (DL5) for source catalogs (see Figure 1).

This situation (many gamma-ray data producers and science tools) has prompted us to start in early 2016 the `gamma-astro-data-formats` effort – an attempt to create an open forum and process to create gamma-ray data models and formats. In some cases we are using or extending the existing formats (mainly FITS and OGIP recommendations), in some cases we are creating new formats that more directly reflect our use cases. The goal is to improve collaboration between people working on this topic and to produce data format specifications to help data producers, tool developers, and astronomers working with high-level gamma-ray data.

[1] `https://heasarc.gsfc.nasa.gov/docs/heasarc/ofwg/ofwg_intro.html`
[2] `http://fermi.gsfc.nasa.gov/ssc/data/analysis/software/`
[3] `http://fermipy.readthedocs.io/`

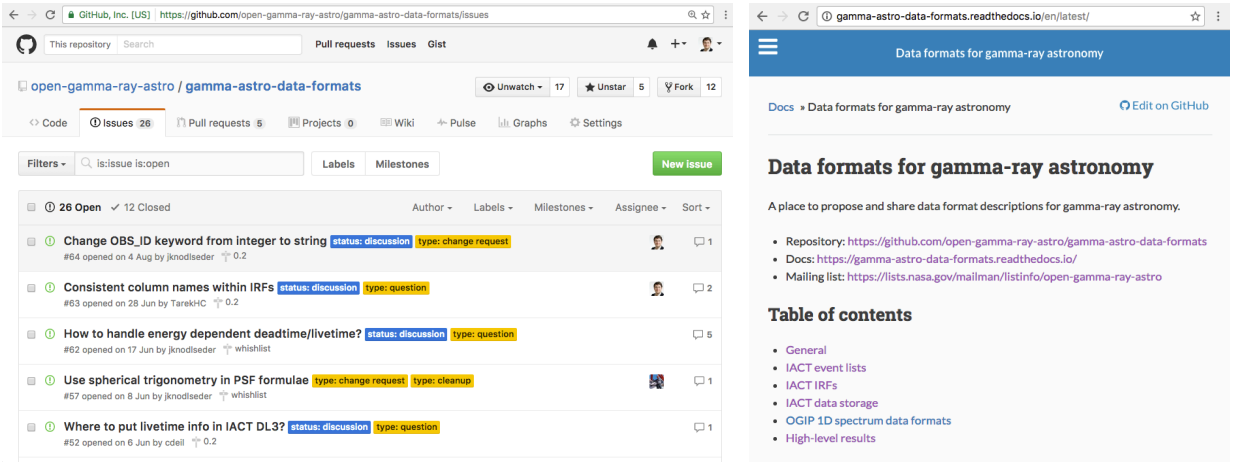

FIGURE 2. *Left:* `gamma-astro-data-formats` Github issue tracker with ongoing discussions. *Right:* latest version of the `gamma-astro-data-formats` specifications on Read the Docs (PDF and older tagged versions also available).

Resources, Process, Work Product

The goal of the `gamma-astro-data-formats` effort is to enable efficient collaboration on gamma-ray data formats and codes. To this end, we have set up the following resources that are open to anyone interested in the topic:

- A mailing list (currently 75 members, including people from all major gamma-ray collaborations) with this official description: "This group is organized for the discussion of software and data formats for the gamma-ray astronomy community. If you are interested in open and common data and software formats for space- and ground-based instruments you are encouraged to join.": `https://lists.nasa.gov/mailman/listinfo/open-gamma-ray-astro`
- A Github organisation for online collaboration on data format specifications via issues and pull requests: `https://github.com/open-gamma-ray-astro/gamma-astro-data-formats`
- Our main work product, the data format specifications, are available online at: `https://gamma-astro-data-formats.readthedocs.io/`
- We hold monthly tele-conferences and plan to hold roughly bi-yearly face-to-face meetings. The first one (Meudon, France in April 2016) was focused on IACT DL3, future meetings will be a bit broader in scope: `https://github.com/open-gamma-ray-astro/2016-04_IACT_DL3_Meeting/`

Our main work product will be a set of data format specifications for gamma-ray data. Each format usually specifies the names and semantics of data and metadata (a.k.a. "header") fields. The scope, status, ongoing discussions, and plans for the data format specifications are presented in the next section. The development of open-source tools and libraries as well as export of existing gamma-ray data to these proposed formats is highly encouraged. However, that work is mainly done by members of the collaborations and software projects mentioned in Figure 1, who then make suggestions for additions or improvements to the existing specifications.

Currently the process of specification writing is informal and the data format specifications currently written should be seen as proposals, not final standards. We are following the "release early and often" philosophy, hoping for feedback and contributions from the larger gamma-ray astronomy community. This approach was motivated by the lack of progress in the past five years on IACT DL3 formats. Although work has begun within CTA on the development of a DL3 format, CTA doesn't produce DL3 data yet. Current IACTs were starting to export their data to FITS format and analyzing them with the current science tools, and many slightly different ways to store the same information in FITS files appeared. Our hope is that this more open format development, making adoption and contributions easy (sending a comment to the mailing list, or making an issue or pull request on Github), will help accelerate the process. Achieving format stability and dealing with "requests for enhancement" after a first stable version of the format specifications is released will be discussed at future meetings.

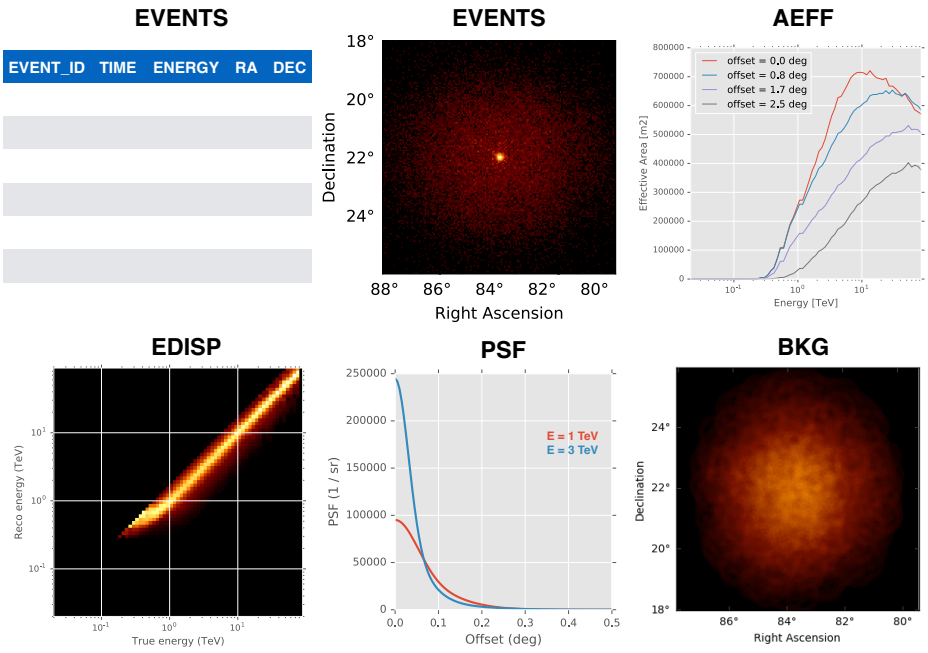

FIGURE 3. Illustration of major components of IACT DL3 data (using a H.E.S.S. 1 Crab nebula observation). The `EVENTS` are stored as a table with the most relevant parameters shown. To derive spectra and morphology measurements of astrophysical sources, instrument response functions (IRFs) are used: effective area (`AEFF`), energy dispersion (`EDISP`), and point spread function (`PSF`). Sometimes background (`BKG`) models are also created and released as part of DL3 data (as an additional IRF component), and other times they are derived at the science tools level. Note that this picture is not complete, see the "IACT DL3" section.

Data models and formats

This section gives an overview of the current status and plans for the gamma-ray data model and formats. As mentioned before, this effort was only started recently and none of the formats should be considered stable. The next two sections will describe the effort to define an event data model and format (DL3) and higher-level formats for sky-maps, spectra, and lightcurves (DL4), i.e. a content split as already illustrated in Figure 1. In the data specification document we have created a "general" section where common quantities are defined, such as precise definitions of time scales as well as coordinate systems. There are some general topics still under discussion, e.g. there is no consensus on how specific or flexible the format specifications should be. E.g. some people prefer to be very specific (data must be stored in FITS files, data types and units fixed), others would prefer to be flexible (only define header keywords and column names, but data can be stored in other file formats as well, e.g. text-based formats like ECSV).

Data level 3 specifications

The interface between low-level (calibration, shower reconstruction, gamma-hadron separation) and high-level (science tools) analysis for gamma-ray data is usually represented by an event list, where at a minimum the `EVENT_ID`, `TIME`, as well as the reconstructed `ENERGY` and sky position (`RA`, `DEC`) is given for every event. In addition, instrument response functions (IRFs) as well as auxiliary technical information such as telescope configuration options, good time intervals (GTIs), live-time, and pointing information (collectively called `TECH` in the CTA context) are needed by the science tools to compute exposures, effective resolutions (PSF and EDISP), and ultimately fluxes to compare the data with sky models. This DL3 data, illustrated in Figure 3, is similar for all gamma-ray telescopes (and other event-recording instruments like e.g. neutrino telescopes). One major difference that affects data formats and analysis tools is whether the gamma-ray telescope was operated in a pointed observation mode (like IACTs most of the time) or in a slewing mode (like HAWC or Fermi-LAT most of the time).

The current specification contains a very preliminary proposal of a data model and formats for IACT DL3 data

that is based on "observations" (with an OBS_ID) and assumed stable IRFs during the observation. This proposal was inspired by existing formats used by H.E.S.S. and partly also VERITAS and MAGIC, that are mostly supported by the existing science tool prototypes (Gammapy and ctools). A dedicated two-day face-to-face meeting on IACT DL3 data was held in April 2016 in Meudon, France, with 16 participants from all major existing IACTs and CTA.[4] The use cases and status of efforts to export and archive their data in FITS was presented, as well as the ongoing prototyping in science tools. Many important points were discussed:

- What is an observation? Good time interval? Response time interval?
- How to link EVENT and IRF? (naming conventions, header references, index tables)
- Pointing and live time information
- Exact definition of field of view (FoV) coordinates
- IRF axis specification, validity ranges, errors
- How to support multiple EVENT classes and types?

A major result of the face-to-face workshop was to agree to focus on IRF formats that use the multi-array convention and FITS BINTABLE to store the IRF data and axis information, where previously a second format was being developed and prototyped for CTA [8]. The prototyping of IACT DL3 is continuing in the different IACT collaborations and in Gammapy/ctools, with communications online via Github, monthly joint tele-conferences, and a planned face-to-face follow-up meeting in fall 2016. So far the focus is set on pointed gamma-ray observations. Contributions and involvement from people working on slewing telescopes (e.g. Fermi-LAT or HAWC and also IACTs) or non-gamma-ray telescopes with similar data (e.g. neutrino telescopes) are welcome. The largest stakeholder for the IACT DL3 work is CTA.

Data level 4 & 5 specifications

Another topic in the gamma-astro-data-formats specifications is the development of formats to store high-level data products such as sky-maps, spectra, and lightcurves (data level 4) or source catalogs (data level 5). Here we list DL4 and DL5 format specifications that are currently included or under consideration:

- For 2-dimensional images, the existing FITS and world coordinate system (WCS) standard provides a solution that works for gamma-ray sky-maps as well. If something gamma-ray specific were to be added, it would likely be specifications on how to store parameters of interest for analysis or provenance in the header.
- For 3-dimensional cubes, where the third dimension is ENERGY, commonly 3-dimensional FITS IMAGE extensions are used. However, due to either the complexity or missing features in the FITS WCS model, the energy axis information is not represented in the FITS header, but in a separate BINTABLE HDU called ENERGY (if the cube represents quantities at given energies, like exposure or flux), or EBOUNDS ("energy bounds", if the cube represents integral quantities like e.g. counts). A specification at gamma-astro-data-formats can document the exact semantics for storing the energy axis and how interpolation and integration should be performed by science tools (e.g. for exposure or diffuse model flux cubes).
- For all-sky maps and cubes, HEALPix[9] is commonly used in gamma-ray astronomy (e.g. by Fermi-LAT). While 2-dimensional HEALPix images are standardized, extensions have been developed to represent cubes, as well as to store sparse data or images that don't cover the whole sky [5]. These gamma-ray specific extensions are not standardized, and a specification at gamma-astro-data-formats would be welcome.
- The common method for 1-dimensional spectral analysis in X-ray astronomy [10], as well as the corresponding file formats (e.g. ARF for effective area, RMF for energy dispersion) are also used in VHE gamma-ray astronomy. In the current specification we have added a section referencing the relevant OGIP documents and explained how the formats are commonly used in gamma-ray astronomy (e.g. using a "reconstructed energy" axis instead of the "pulse height channels" axis used in X-ray astronomy).
- For 1-dimensional spectra, a format to store flux points and upper limits, as well as full likelihood profiles, is available at gamma-astro-data-formats (see Figure 4 left panel). It was first developed in Fermipy and applied to Fermi-LAT analyses, and is now being adopted for IACT spectra.
- No format specification for light curves (see Figure 4 right panel for an illustration) is available yet. Previously a format has been proposed in [11] and a pull request with discussions for a lightcurve specification at gamma-astro-data-formats is ongoing.

[4] https://github.com/open-gamma-ray-astro/2016-04_IACT_DL3_Meeting/

[5] https://github.com/tburnett/Fermi-LAT/blob/master/pointlike_document/Data\%20Format.ipynb

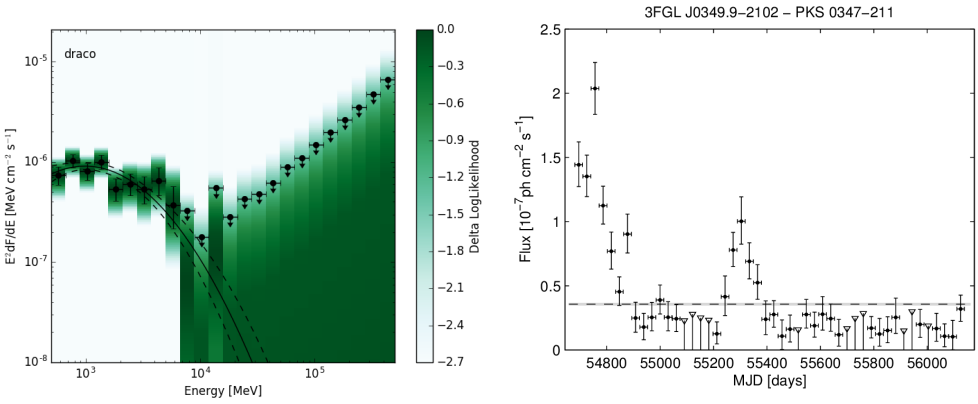

FIGURE 4. Gamma-ray "data level 4" examples. *Left:* spectral energy distribution (SED) likelihood profiles (green), with flux points, upper limits and best-fit model shown. *Right:* Lightcurve of 3FGL J0349.9-2102 from the third Fermi-LAT catalog.

- No format specifications have been proposed for catalogs (data level 5, DL5) yet. So far each catalog (Fermi-LAT, upcoming H.E.S.S. and HAWC) is unique (but all similar) and some science tools have per-catalog code to produce corresponding sky models.

Conclusions

In early 2016, we have started the `gamma-astro-data-formats` effort to create an open forum (mailing list, Github, meetings) and eventually open and common data formats for space- and ground-based gamma-ray instruments. This effort is similar to the HEASARC FITS working group from the 1990s, but this time driven mainly by the movement of ground-based gamma-ray observatories toward producing high-level gamma-ray data in FITS format (IACT DL3 data). We invite everyone interested in this topic to join the mailing list, regular meetings and to contribute or give feedback on how the current formats could be improved to support your use cases.

Acknowledgements

We would like to thank everyone that has contributed to or supported this effort, be it directly via contributions to the format specification, or indirectly via feedback or adopting the existing formats, spending the effort to transform their existing data to the common formats defined here, or by giving people time or travel money to work on this.

We would also like to thank the following services: `NASA` for hosting the `open-gamma-ray-astro` mailing list, `Github` for making this way of online collaboration possible, `Sphinx` as documentation system and `Read the docs` for building and hosting the HTML and PDF version of the specification.

REFERENCES

[1] D. C. Wells, E. W. Greisen, and R. H. Harten, AAPS **44**, 363–+, June (1981).
[2] W. D. Pence, L. Chiappetti, C. G. Page, R. A. Shaw, and E. Stobie, AAP **524**, A42+, December (2010).
[3] A. Donath *et al.*, ArXiv e-prints September (2015), arXiv:1509.07408 [astro-ph.IM] .
[4] J. Knödlseder *et al.*, AAP **593**, A1, August (2016).
[5] M. Kerr, Ph.D. thesis, University of Washington 2010.
[6] G. Vianello *et al.*, ArXiv e-prints July (2015), arXiv:1507.08343 [astro-ph.HE] .
[7] V. Zabalza, ArXiv e-prints September (2015), arXiv:1509.03319 [astro-ph.HE] .
[8] J. E. Ward *et al.* for the CTA Consortium, ArXiv e-prints August (2015), arXiv:1508.07437 [astro-ph.IM] .
[9] K. M. Górski *et al.*, ApJ **622**, 759–771, April (2005).
[10] J. E. Davis, APJ **548**, 1010–1019, February (2001).
[11] M. Tluczykont *et al.*, AAP **524**, A48, December (2010).