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

Christoph Deil[1,a)], Catherine Boisson[3,b)], Jeremy Perkins[12], Johannes King[1], Peter Eger[1], Michael Mayer[6], Matthew Wood[13], Victor Zabalza[15], Jürgen Knödlseder[11], Tarek Hassan[10], Lars Mohrmann[5], Alexander Ziegler[5], Bruno Khelifi[4], Daniela Dorner[5], Gernot Maier[7], Giovanna Pedaletti[7], Jaime Rosado[10], José Luis Contreras[10], Julien Lefaucheur[3], Kai Brügge[5], Mathieu Servillat[3], Régis Terrier[4], Roland Walter[8] and Saverio Lombardi[14]

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

Abstract. In gamma-ray astronomy, a variety of data formats and proprietary software have been classically used, often developed for one specific mission or experiment. Especially for ground-based imaging atmospheric Cherenkov telescopes (IACTs), data and software have been 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 big IACT array, 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 writeup 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.

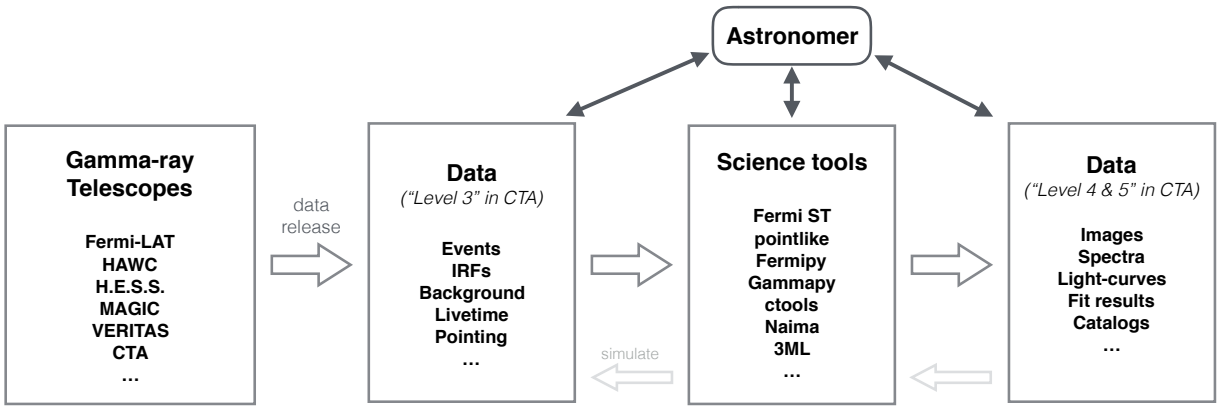

FIGURE 1. The purpose of the `gamma-astro-data-formats` effort is to encourage communication between high-level gamma-ray data producers, science tool developers and data analysts. The goal is to develop a common data model and format, to avoid duplication of efforts and confusion by astronomers working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

Introduction

Two decades ago, a coordinated multi-year, multi-mission effort took place that created common data format standards and recommendations for high-energy astrophysics:

> The HEASARC FITS Working Group, also known as the OGIP (Office of Guest Investigator Programs) FITS Working Group, has promoted multi-mission standards for the format of FITS data files in high-energy astrophysics. Its main activities took place in the mid-1990s, when it produced a number of documents and recommendations concerning the format of FITS files. Several of these recommendations have subsequently been incorporated into the FITS Standard format definition document.[1]

At that time, the goal was mostly to support X-ray and gamma-ray data from space-based missions. Today, ground-based gamma-ray astronomy is finding itself in a similar situation (illustrated in Figure 1). The existing imaging atmospheric Cherenkov telescopes (IACTs) like e.g. H.E.S.S., MAGIC, 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 of IACT, will be operated as an an open observatory, meaning that data and analysis software will be public to all astronomer. Already now, multiple open-source science tool codes for gamma-ray astronomy exist (Gammapy [1], ctools [2], pointlike [3], Naima [4], 3ML [5], Fermipy[2], Fermi ScienceTools, ...). High-level data from the Fermi-LAT space telescope is openly available, and current IACTs have started to export their high-level data (event lists and instrument response functions) to FITS formats for analysis with the existing open-source science tools.

This situation (many gamma-ray data producers and science tools, see Figure 1) 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.

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:

[1] `https://heasarc.gsfc.nasa.gov/docs/heasarc/ofwg/ofwg_intro.html`
[2] `https://github.com/fermipy/fermipy`

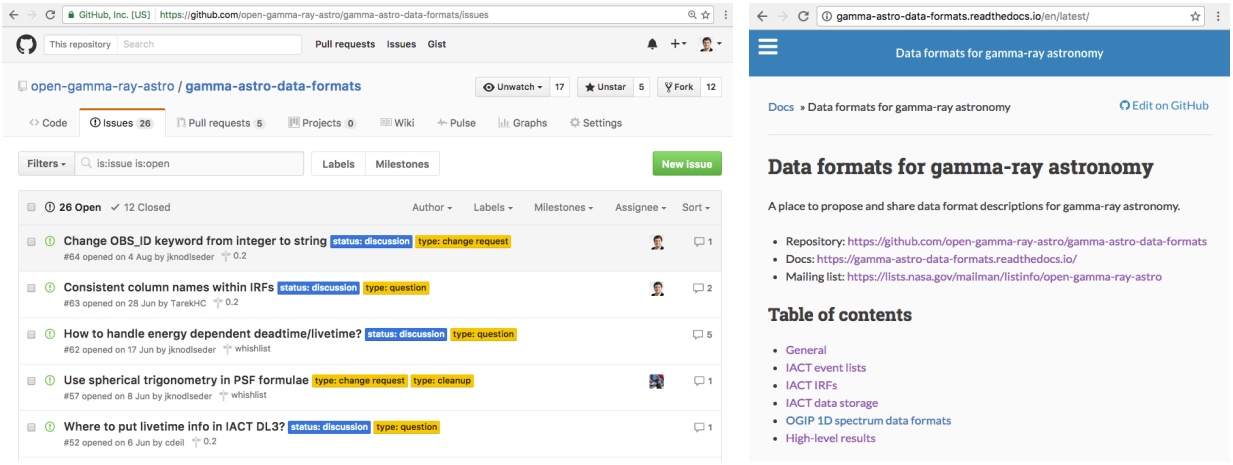

FIGURE 2. *Left:* `gamma-astro-data-formats` Github issue tracker with ongoing discussions. *Right:* latest version of the `gamma-astro-data-formats` specifications on Read the Docs (PDF and older tagged versions also available).

- 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 first suggestions, not final standards. In a sense we are following the "release early and often" philosophy, hoping for feedback and contributions from the larger gamma-ray astronomy community. To a certain degree this was motivated by the lack of progress in the past five years on IACT DL3 formats – in CTA people were starting to work on this, but CTA doesn't produce DL3 data yet, and 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, and 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. The need to achieve stability and how to deal with "requests for enhancement" after a first stable version of the format specifications will be discussed at future meetings.

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. 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. One example is a precise definition of `AZIMUTH` and `ALTITUDE`. We define `AZIMUTH` to be oriented east of north, and `ALTITUDE` to be relative to the zenith direction (not the horizon plane or a reference earth ellipsoid) and without applying a refraction correction.

There are some general topics still under discussion: There is no consensus on how specific format specifications should be (for instance, fixing the data type and precision of each variable, or allowing certain flexibility) or if other data formats (e.g. text ECSV) should be also supported apart from the initially proposed FITS format.

Data level 3 specifications

The interface between low-level (calibration, and shower reconstruction pipeline) and high-level (science tools) analysis for gamma-ray data is usually represented by an event list, where at a minimum the `EVENT_ID`, observation `TIME`, as well as the reconstructed `ENERGY` and sky position (`RA`, `DEC`) is given for every event. In addition, 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 CTA) 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 (and also neutrino) telescopes, with certain different details, e.g. some instruments mostly perform pointed observations (like IACTs) while others do slewing observations (like Fermi-LAT or HAWC).

This work presents the development of a data model and format specification for IACT DL3 data. As a starting point, we wrote down the existing formats used by H.E.S.S. and partly also VERITAS and MAGIC, that are mostly supported by the exiting science tool prototypes (Gammapy and ctools). H.E.S.S. is planning to release a small test dataset in the current format consisting of roughly 50 hours of H.E.S.S. 1 observations on a few sources in fall 2016.

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 (see `https://github.com/open-gamma-ray-astro/2016-04_IACT_DL3_Meeting/`). 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`?
- Pointing and live time information
- IRF axis specification and validity ranges
- FoV coordinates
- 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 [6]. 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 or lightcurves (data level 4) or catalog (data level 5).

- For 2-dimensional images, the FITS standard (including world coordinates systems WCS) 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 meta information like the energy band or provenance parameters used to make the image.

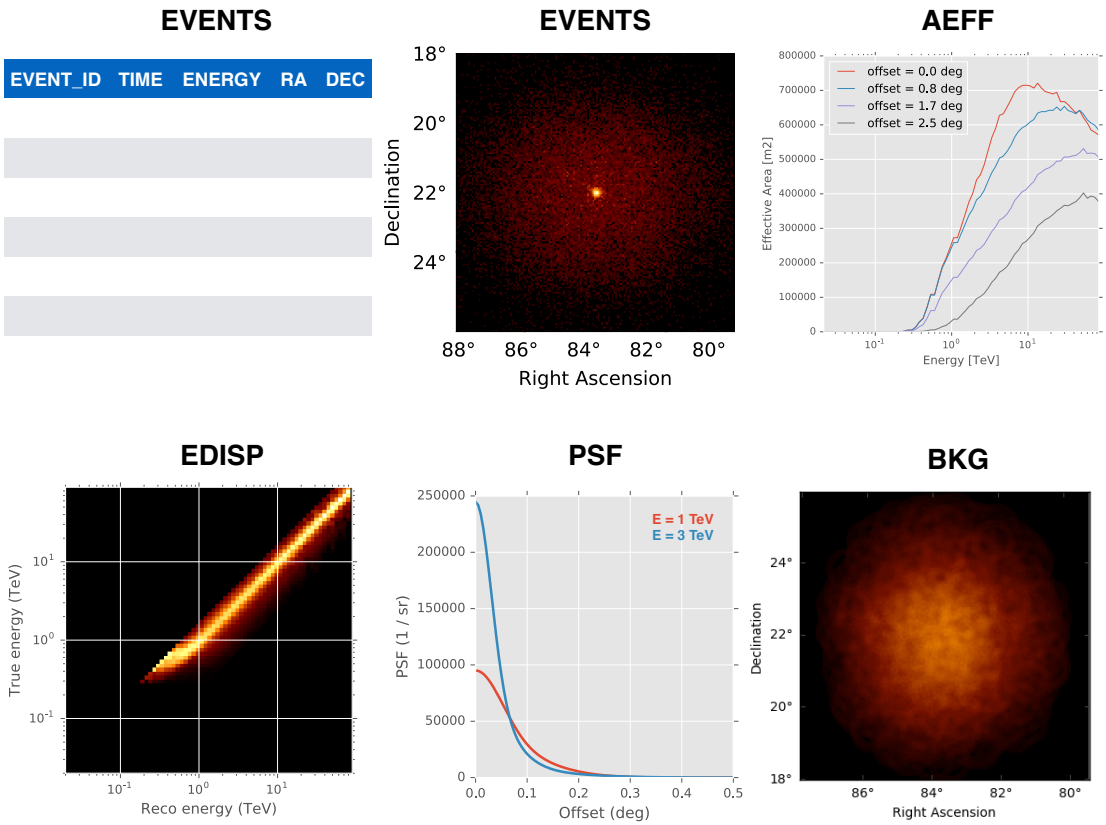

FIGURE 3. Illustration of major components of IACT (here a H.E.S.S. 1 Crab nebula observation) DL3 data. The `EVENTS` are stored as a table with the most relevant parameters shown. To derive spectra and morphology measurements of astrophysical sources, IRFs are used: the effective area (`AEFF`), energy 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.

- 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 a separate `BINTABLE HDU` instead 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). Even if this format has been widely used in gamma-ray astronomy for a long time, a specification at `gamma-astro-data-formats` defining the exact semantics of how the energy axis should be interpolated and integrated would be welcome.
- For all-sky maps and cubes, HEALPIX 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 [3]. These gamma-ray specific extensions are not standardized, and a specification at `gamma-astro-data-formats` would be welcome.
- 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 [7] and a pull request with discussions for a lightcurve specification at `gamma-astro-data-formats` is ongoing.

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

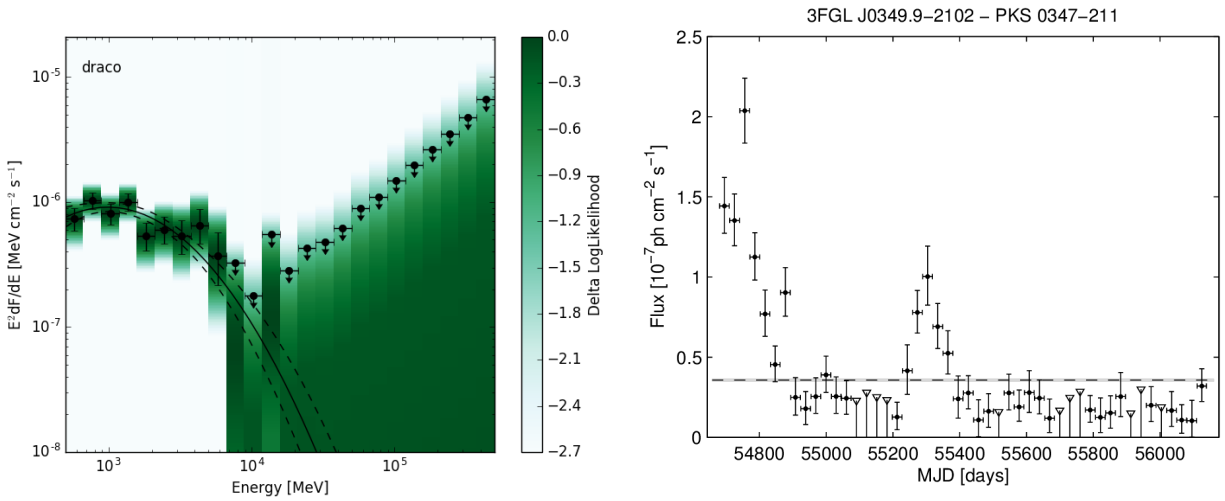

FIGURE 4. Gamma-ray "data level 4" examples. *Left:* spectral energy distribution (SED) likelihood profiles (green), with flux points and upper limits as well as a best-model fit overplotted. *Right:* Lightcurve of 3FGL J0349.9-2102 from the third Fermi-LAT catalog.

- 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. Whether it makes sense to try and specify a common catalog format for gamma-ray astronomy remains to be discussed. Probably at least adopting the spectrum and lightcurve formats mentioned before could be useful.

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 and software formats for space- and- ground based gamma-ray instruments. This is similar to the OGIP FITS

> The HEASARC FITS Working Group, also known as the OGIP (Office of Guest Investigator Programs) FITS Working Group, has promoted multi-mission standards for the format of FITS data files in high-energy astrophysics. Its main activities took place in the mid-1990s, when it produced a number of documents and recommendations concerning the format of FITS files. Several of these recommendations have subsequently been incorporated into the FITS Standard format definition document.[4]

- High-level gamma-ray data (DL 3 and up) from different telescopes is very similar, there's always event lists and IRFs, plus some extra info like good time interval and pointing information.
- We have started the first effort to define open data models and data formats for gamma-ray astronomy.
- The motivations for this are the development of open-source tools (Gammapy, ctools, Fermi ST, Fermipy, point-like, emcee) to analyze gamma-ray data, and that IACTs are starting to produce data mostly in FITS format that these tools shall consume. So "many tools" and "many telescopes" makes a common data model and formats useful.
- We have chosen an open process (Mailing list, Github, monthly tele-conferences, bi-yearly f2f meetings).
- There are useful preliminary specs, we encourage everyone to have a look now and give feedback and contribute. Many important questions are under discussion (refer back to section listing those).
- But there's no stable version or "standard" yet. Especially for CTA the process will have to be more formalized if CTA data is to be released in those formats.

[4] `https://heasarc.gsfc.nasa.gov/docs/heasarc/ofwg/ofwg_intro.html`

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 and spending the effort to transform their existing data to the common formats defined here.

We would also like to thank the following services make this form of collaboration possible: `Github` for collaboration, `Sphinx` as documentation system and `Read the docs` for building and hosting the HTML and PDF version of the spec.

REFERENCES

[1] A. Donath *et al.*, ArXiv e-prints September (2015), arXiv:1509.07408 [astro-ph.IM] .
[2] J. Knödlseder *et al.*, AAP **593**, p. A1August (2016), arXiv:1606.00393 [astro-ph.IM] .
[3] M. Kerr, "Likelihood methods for the detection and characterization of gamma-ray pulsars with the Fermi large area telescope," Ph.D. thesis, University of Washington 2010.
[4] V. Zabalza, ArXiv e-prints September (2015), arXiv:1509.03319 [astro-ph.HE] .
[5] G. Vianello *et al.*, ArXiv e-prints July (2015), arXiv:1507.08343 [astro-ph.HE] .
[6] J. E. Ward *et al.* for the CTA Consortium, ArXiv e-prints August (2015), arXiv:1508.07437 [astro-ph.IM] .
[7] M. Tluczykont *et al.*, AAP **524**, p. A48December (2010), arXiv:1010.5659 [astro-ph.HE] .