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 exist, often developed for one specific mission or experiment. Especially for ground-based imaging atmospheric Cherenkov telescopes (IACTs), data and software have been so far mostly private to the collaborations operating the telescopes. However, there is a general movement in science towards open data and software and 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). The hope is that 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.

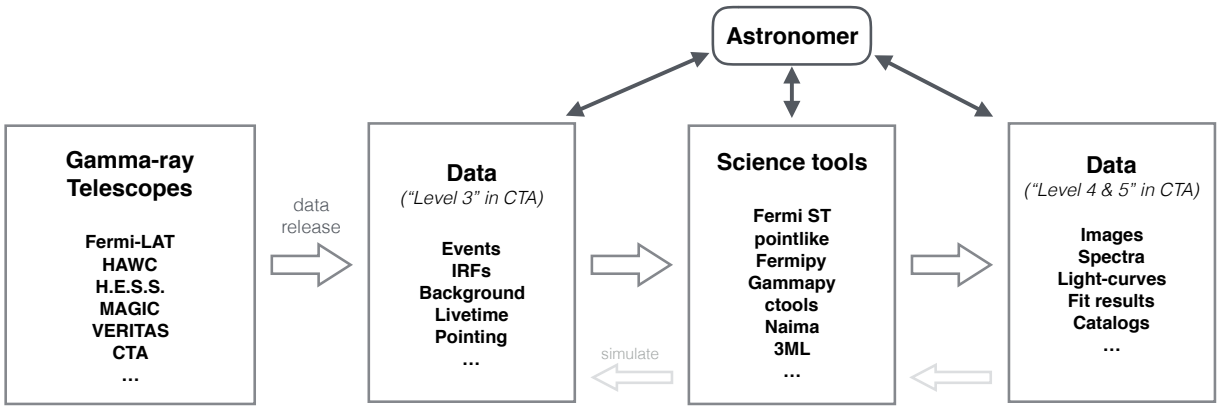

FIGURE 1. The purpose of the `gamma-astro-data-formats` effort is to encourage communication between high-level gamma-ray data producers, science tool developers and data analysts. The goal is to develop and agree on some common data models and formats, to avoid duplication of efforts and confusion by astronomer that want to work with gamma-ray data from different telescopes or try different tools.

Introduction

Science is aiming towards open and reproducible research. In gamma-ray astronomy, especially for ground-based imaging atmospheric Cherenkov telescopes (IACT), data and software have so far been mostly private to the collaboration operating the telescopes. The Cherenkov Telescope Array (CTA), the next generation of IACT, will be operated as an an open observatory, meaning that data and analysis softwares will be public at the end-user, science tools level. Momentum was thus given to the development of open data format and open-source software for gamma-ray astronomy, leading to synergies between experiments, ground based and space missions.

This is illustrated in Figure 1: there are many gamma-ray data producers and science tools to work with gamma-ray data, so agreeing on common data models and formats that cover most common analysis cases makes things simpler for data producers, tool developers and users.

Open source analysis tools for VHE gamma-ray astronomy have emerged. They all meet on common ground of using FITS files for data transfer. The current IACTs (H.E.S.S., MAGIC and VERITAS) all use their own software based on the ROOT library, which makes it impossible to analyze the respective data with anything else than the corresponding software. Current open source analysis tools provide alternative analysis techniques compared to the present standard in VHE astronomy. It is assumed that these techniques (e.g. 3D likelihood analysis such as already implemented for Fermi-LAT) improve the sensitivity of IACTs by roughly 20%. The data model for CTA is currently being developed but still work in progress. A main point of defining the high level data format for CTA is to understand the instrument including its systematics and IRF dependencies. In addition the needs from a user perspective have to be taken into account to create a solution as simple as possible. Having agreed on a common data format for files and on a way how to store and access those files (folder structure), makes mid-level (event energies, positions) and high-level (source position, morphology, spectrum) checks between the different chains, algorithms and open-source tools possible. This will also ease interoperability with other codes (e.g. to check results, combine results in one plot, ...). Currently two open-source science tools packages are being designed for current IACT and CTA data analysis, Gammapy ([1]) and Ctools ([2]). Gammapy is an in-development Astropy-affiliated package. Ctools is based on the GammaLib analysis framework, which is mainly written in C++.

We have created a Github organization where we are developing high-level data format specifications, in accordance with astronomy standards.

TODO: also mention other codes: pointlike [3], Naima [4], 3ML [5], Fermipy[1], Fermi ScienceTools.

TODO: Is there anything we can cite for CTA data challenge 1 or data challenge 2 plans? TODO: Mention website and some info on how CTA is releasing IRF files at the moment (prod2)?

[1] `https://github.com/fermipy/fermipy`

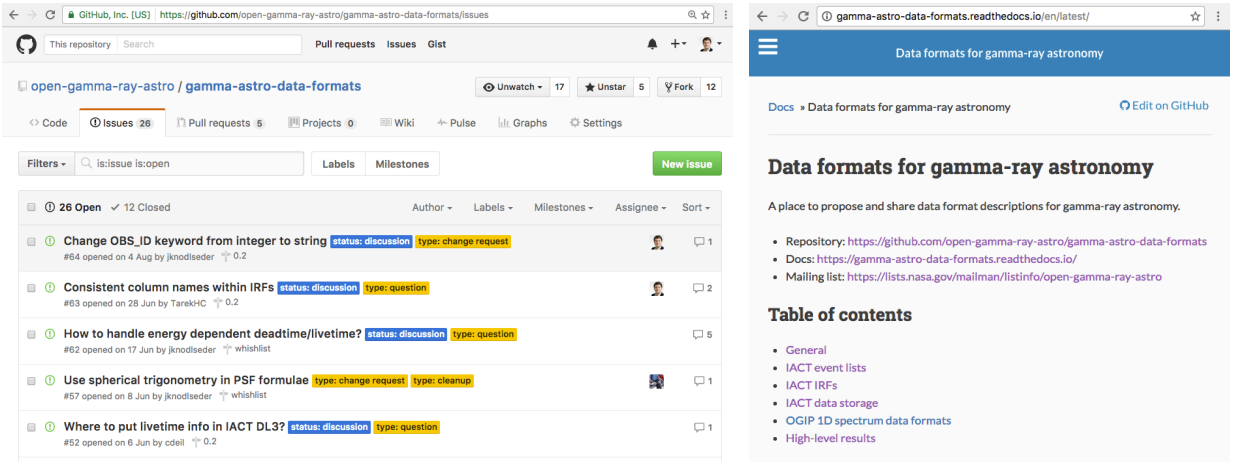

FIGURE 2. *Left:* `gamma-astro-data-formats` Github issue tracker with ongoing discussions. *Right:* latest version of the `gamma-astro-data-formats` specifications on Read the Docs (PDF and older tagged versions also available).

Resources and Process

The specifications of a given data level format defines the names and semantics of data and header fields. Such specifications are made easy to understand. Specifications of this format are currently written that can form the basis for prototyping for data producers (mainly existing IACTs and simulated CTA data) and consumers (mainly science tool codes). We include example files and some explanations, in addition to the detailed specifications for a given format.

The scope is high-level data, starting with event lists and instrument response functions (IRFs), what is called ”data level 3” (DL3) in CTA. The first stable release (archived on Zenodo with a DOI) is coming soon.

If you want to contribute, it is simple:

- Use the existing format and give feedbacks. Propose additions and changes.
- Join the mailing list (see next section). Send an e-mail with an idea or proposal.
- Create a Github account. File an issue with a correction or make a pull request proposing additions.

No formal approval process is in place yet as this is a very recent effort.

- Mailing list for announcements and high level discussions (75 members, including people from all major gamma-ray collaborations : `https://lists.nasa.gov/mailman/listinfo/open-gamma-ray-astro`
- Github issues and pull requests are used for detailed discussions: `https://github.com/open-gamma-ray-astro/gamma-astro-data-formats`
- Data format specifications in HTML and PDF format, including example files: `https://gamma-astro-data-formats.readthedocs.io/`

The specs are written in a markup format called ”restructured text” (RST), which gets transformed by Sphinx to HTML or PDF, the latest rendered version is always available online (see Figure 2).

Data models and formats

This section gives an overview of the status and plans for the gamma-ray data model and formats. As mentioned before, this effort was only started very 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 images, 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 that gives precise definitions of common quantities, such as precise definitions of time scales as well as coordinates 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 questions we are discussing about where to be specific or flexible in our format specifications. One example is whether our data format specifications should be tied to FITS (and e.g. say the data type of a column is `1E`, which is the data type code for 32-bit float in FITS), or whether it would be better to only say that the data type is "float", implicitly allowing the storage of this column as 32-bit or 64-bit float, and being able to store the data e.g. in text ECSV files in addition to binary FITS files. Another contentious point is whether physical units should be fixed (e.g. "MeV" or "TeV"), or whether the units should be flexible and only serialization formats that support declaration of units (such as FITS or VOTABLE or ECSV) should be supported and science tools are expected to process the unit information correctly.

Data level 3 (DL3) specifications

The interface between low-level (calibration, reconstruction and gamma-hadron separation pipeline) and high-level (science tool) 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, instrument response functions (IRFs) as well as auxiliary information such as telescope configuration options, good time intervals (GTIs), livetime and pointing information (collectively called `TECH` in CTA) are needed by the science tools to compute exposures as well as effective resolutions (PSF and EDISP) and ultimately fluxes and to compare the data with sky models. This DL3 data, illustrated in Figure 3 is similar for all gamma-ray telescopes (and also neutrino telescopes), although in detail it is different e.g. for telescopes that mostly do pointed observations (like IACTs) or that do slewing observations (like Fermi-LAT or HAWC).

We have started to develop 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, e.g.

- 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 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 images, spectra or lightcurves (data level 4) or catalog (data level 5).

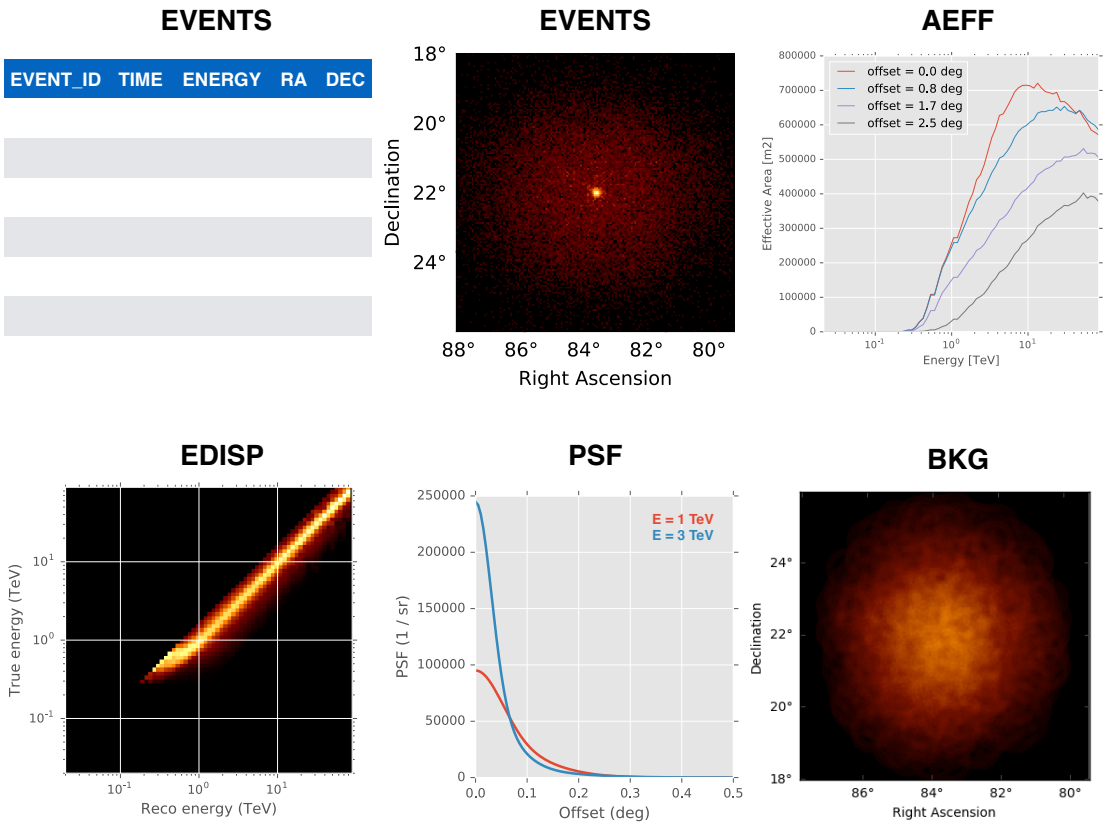

FIGURE 3. Illustration of major components of IACT ("imaging atmospheric Cherenkov telescope", here a H.E.S.S. 1 Crab nebula observation) DL3 ("data level 3") data. The `EVENTS` are stored as a table with the most important parameters shown. To derive spectra and morphology measurements of astrophysical sources, IRFs ("instrument response functions") 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 a DL3 data release, and can be treated like an IRF, and sometimes deriving background models is left to the science tools. Note that this picture is not complete, see the "IACT DL3" section.

- For 2-dimensional images, the FITS standard (including world coordinates systems WCS) provides a solution that works for gamma-ray images 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 other analysis or provenance parameters used to make the image.
- 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). 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 [2]. 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

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

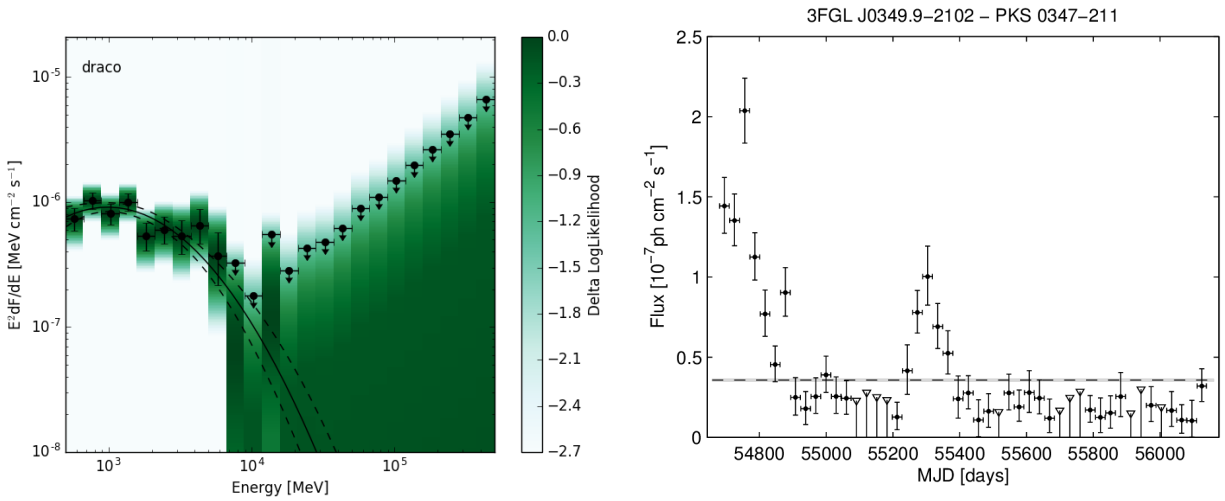

FIGURE 4. Gamma-ray "data level 4" examples. *Left:* spectral energy distribution (SED) likelihood profiles (green), with flux points and upper limits as well as a best-model fit overplotted. *Right:* Lightcurve of 3FGL J0349.9-2102 from the third Fermi-LAT catalog.

applied to Fermi-LAT analyses, and is now being adopted for IACT spectra.

- Lightcurves (see Figure 4 right panel). Description of light curves is another topic open for discussion. Previous work to define a gamma-ray lightcurve format: [7]
- 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 come 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

This is just a draft for an outline.

TODO: write a paragraph or two for the conclusions.

- 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.

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, C. Deil, M. P. Arribas, J. King, E. Owen, R. Terrier, I. Reichardt, J. Harris, R. Bühler, and S. Klepser, ArXiv e-prints September (2015), arXiv:1509.07408 [astro-ph.IM] .

[2] J. Knödlseder, M. Mayer, C. Deil, J.-B. Cayrou, E. Owen, N. Kelley-Hoskins, C.-C. Lu, R. Buehler, F. Forest, T. Louge, H. Siejkowski, K. Kosack, L. Gerard, A. Schulz, P. Martin, D. Sanchez, S. Ohm, T. Hassan, and S. Brau-Nogué, 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, R. J. Lauer, P. Younk, L. Tibaldo, J. M. Burgess, H. Ayala, P. Harding, M. Hui, N. Omodei, and H. Zhou, ArXiv e-prints July (2015), arXiv:1507.08343 [astro-ph.HE] .

[6] J. E. Ward, J. Rico, T. Hassan, and f. t. CTA Consortium, ArXiv e-prints August (2015), arXiv:1508.07437 [astro-ph.IM] .

[7] M. Tluczykont, E. Bernardini, K. Satalecka, R. Clavero, M. Shayduk, and O. Kalekin, AAP **524**, p. A48December (2010), arXiv:1010.5659 [astro-ph.HE] .