Harmonisation of standards for multi-messenger data

ACME, WP4.1

B. Khélifi (APC, Paris)

Astro-ORDAS Symposium (Lausanne)

ACME

Provision of access to a wide range of data

- Broad variety of observatories and experiments
- Multi-wavelength and multi-messenger data

→ **Big challenge of data heterogeneity**

WP4.1: Harmonisation of standards

- For facilities delivering data
- For analysis software generating data
- For web analysis platform producing data

Coordination: B. Khélifi (APC) and M. Kettenis (JIVE)

General landscape

EOSC

ACME is a EU projet, then must follow the EOCS recommendations (2023)

Strategic Research and Innovation Agenda of EOSC

FAIR principles

MD Wilkinson et al. (2016)

FAIR4RS principles

C. Hung et al. (2022)

Astronomy&Astrophysics

FITS standards (link, 1993+)

FOCUS

Differentiate the concepts of:

- Standards
- Data Model
- Data Format
- Serialisation Format

International Virtual Observatory Alliance (IVOA)

Large sets of standards and formats on data and transfer protocols (link)

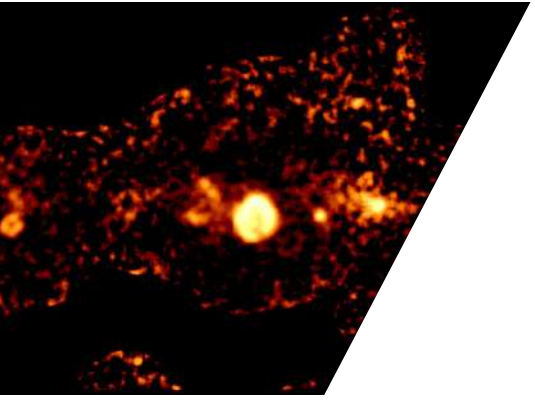
Summary

1- Survey of data formats associated to ACME

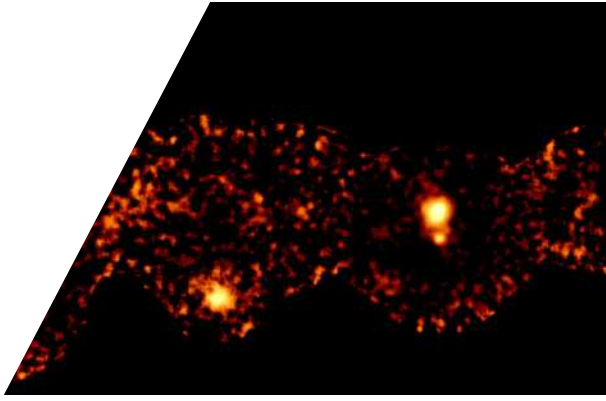
2- Survey of data access protocols and tools within ACME

3- Analysis of the data curation by the ACME facilities

4- Recommendations



Survey of data formats associated to ACME



Radio data

FITS-based data formats is widespread

- PSRFITS - A FITS Standard for Pulsar Data Storage: html
- *UVFITS* - AIPS FITS File Format, Greisen, E.W, AIPS Memo 117: pdf
- *FITS-IDI* - The FITS Interferometry Data Interchange Convention, Greisen, E.W, AIPS Memo 114: pdf
- *SDFITS* - Binary table convention for interchange of single dish data in radio astronomy
- *MBFITS* - Multi-Beam FITS Raw Data Format Convention: pdf
- *CDTS* - Casacore Table Data System: pdf
- *Zarr* - `Zarr core specification`: html
- *HDF5* - Hierarchical data format version 5: html
- *CLASS* - pdf (both data model and data format)
- *CLIC* - Continuum and Line Calibration: pdf

IR, Optical, UV data

FITS-based data formats is widespread, but not only

- *FITS* - Flexible Image Transport System: primarily for imaging, but also for spectroscopy
- *CVS* - BHTOM processes the FITS observations from 130+ telescopes to make a queryable database of high-order photometric products
- *Swift-UVOT* - own format relying on the FITS standards
- *SExtractor file format* - plain text format used for photometry
- *Lightcurves in CVS file*

X-rays, HE and VHE data

OGIP format, serialized into FITS files

- *OGIP.1* - FITS Rate File, 2014: html
- *OGIP.2* - Calibration files: html
- *OGIP.3* - Spectral File Format (PHA): html
- *OGIP.4* - Effective Area: pdf
- *OGIP.5* - 2D-PSF File format: pdf

INTEGRAL mission - format derived from the OGIP standards (html)

Fermi

- Own format derived from OGIP: html

IACT, WCD arrays

- *GADF* - Gamma Astro Data Format, relying on the FITS standards: html

GW, CR, ν data

Gravitational Waves

- *GWF* – raw data (strains): html
- HDF5 – Common data format: html; Public transient catalogs of data products (skymaps, posterior samples, strains): GWTC-2.1 catalog
- FITS - Skymaps on the localization of the events: html; Public transient catalogs of data products (skymaps, posterior samples, strains): GWTC-2.1 catalog

Cosmic Rays

- *JSON* - Raw data for Auger: html
- *CSV, HDF5* - Summary data specifying reconstructed parameters, monitoring data

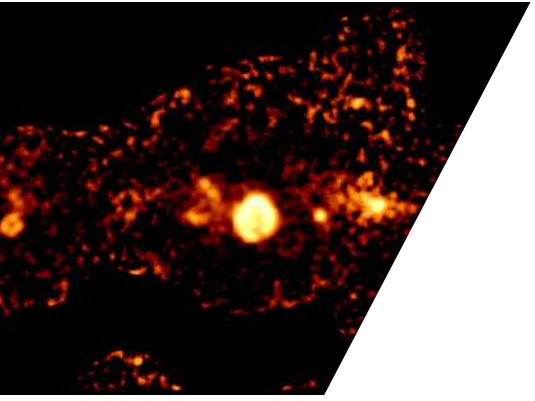
Neutrino

- Not yet standardized: CSV, HDF5, FITS, ROOT, JSON, YAML, GCN, VOTables
  - But large on-going effort about formatting and curation

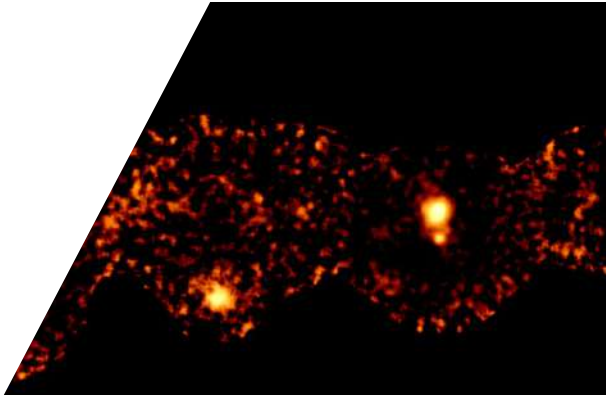
Alerts

Not traditionally been considered as standard data products.
But, in reality, they constitute a **genuine scientific data product**!!

- *VOEvent* - Sky Event Reporting Metadata: html (IVOA)
- *GCN Unified Schema* - html (NASA)
- *Parquet* - Apache Parquet format over the Apache Kafka protocol: html



Survey of Data Access Protocols and tools within ACME



Data transfer and discovery

Data transfer

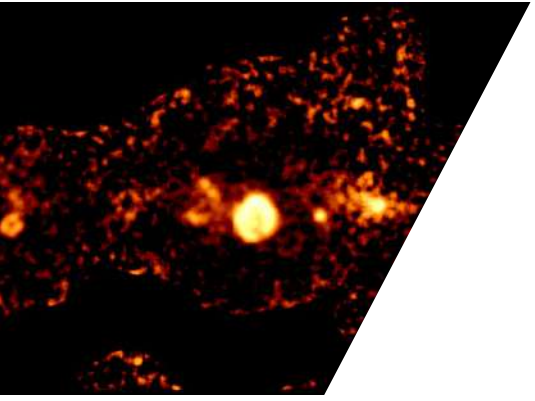
- *HTTP* – widely used
- FTP – still used, not not recommanded due to issues with stateful firewalls and network address translation (NAT) that can be hard to debug
- *S3 cloud storage protocol, CERN's Rucio/DIRAC* – Used for large data collections, that may be stored in data lakes

Webforms

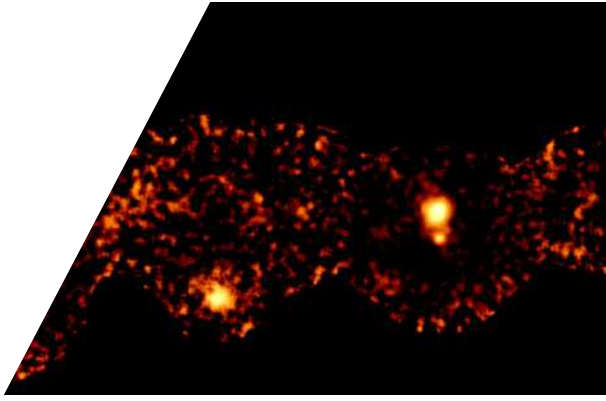
- Still many experiments and archives provide a search engine via a web form
- Not recommanded because it forces users to visit many web pages by hand (no API)

VO protocols

- *VOSI* – minimum interfaces of a web service to work with VO: html
- *Table Access Protocol via query in ADQL* - access protocol for tabular data: *TAP*
- *Simple Cone Search* - discovery protocols for catalogues: html
- *Simple Image Access and Simple Spectrum Access* - discovery/access protocol for any multidimensional datasets (spectrum, 2D images, data cubes and event more : SIA and SSA
- *Server-side Operations for Data Access* - access protocol for cut-outs: SODA
- *Simple Application Messaging Protocol* – enables astronomy software tools to interoperate and communicate: SAMP



Analysis of the data curation by the ACME facilities



1- Formats of data produced by ACME facilities

Self-analysis of FAIR-compliance

Within the WP4.1, facilities representatives realized a self-analysis to assess the FAIR-compliance of their own products.

Creation of a large table with: [Facility, Format(s), FAIR-compliance, Note]

The FAIR-compliance is evaluated for each principle:

- e.g “F, A, I, R-” or “A,I”

Here are given some global features

About Data Distribution

- All data accessible via web portals, w or w/o API
- Data filtering options not uniform

Desirable: greater harmonization in selection criteria (IVOA provides recommendations)

- Only some facilities employ VO-dedicated servers

Strong recommendation: investment in the deployment of VO servers (contact us to get insights)

About Data Formats

- Frequent confusion between data format and serialization format
- All are using well-established and well-documented file standards
- Data formats used within these files are highly diverse; and their documentation are complicated to find, when existing

Strong recommendation: provision of complete documentation of data and metadata formats, include as metadata the documentation URL

- FITS used for small datasets, HDF5 or cloud-native formats (Zarr, Parquet) of use for large datasets
- However, major international observatories (ESO, NASA, etc) keep FITS

Recommendation: keep use FITS (while progressively exploring newer)

About Metadata and Provenance

- 4 generic types of metadata:
  - Descriptive metadata: identify and characterise the data
  - Structural metadata: describe the composition and organisation of a dataset
  - Administrative metadata: include information such as licensing, creation date, and authorship
  - Specific metadata: associated to a facility's data model

Recommendation: be as comprehensive as possible when recording metadata, in order to maximise data transparency, discoverability, and reuse

- Open Science objectives of reliability, quality, and trustworthiness → provision complete information about the entities, activities, and people involved in data production: "Provenance"

Minimal recommendation: include a metadata with an URL that points to a comprehensive provenance description

About Referencing

- Too few datasets are referenced with a DOI
- Even if data are 1 of the 3 pillars of Open Science (with the software)

Strong Recommendation: creation of DOIs for each dataset

2- Science web platforms involved in ACME

Online tools used in ACME for:

- Data distribution: selection, filtering,
- Data processing: preliminary analysis, combinaison of data (SED, light-curves)

Self-analysis of FAIR-compliance

Within the WP4.1, facilities representatives realized a self-analysis to assess the FAIR-compliance of their own products.

Creation of a large table with: [Facility, Format(s), FAIR-compliance, EOSC integration, VO integration Note]

The FAIR-compliance is evaluated for each principle:

- e.g “F, A, I, R-” or “A,I”

Here are given some global features

Learnt Lessons

- VO is not the main source of data provision for the web platforms
  - In contrary to the alert brokers

Strong recommendation: use the VO tools and APIs!

- Data format of the generated data are most of the case custom, and FITS is rarely used!
  - Very weak interoperability of these new products

Strong recommendation: generation of FAIR data and use IVOA standards

- Provenance is rarely captured for the output data (problem of traceability for the user)

Desirable: capture and write the provenance associated to output data

- Most of them are not in EOSC, have no open source code, no DOI, not following the FAIR4RS principles

Desirable: Follow the FAIR4RS for the webportal project

3- Used libraries in ACME

Any data is produced by a software
A data is FAIR only IF it is coded into a software

ACME is supporting some improvements of libraries towards this direction, and also to provide more features for MWL and MM interoperability

Self-analysis of FAIR-compliance

Within the WP4.1, facilities representatives realized a self-analysis to assess the FAIR-compliance of their own products.

Creation of a large table with: [Name, Facility, Description, FAIR4RS-compliance, DOI, SoftwareHeritage, (FA)IR-compliance of output data, Note]

The FAIR-compliance is evaluated for each principle:

- e.g “F, A, I, R-” or “A,I”

Here are given some global features

Learnt Lessons

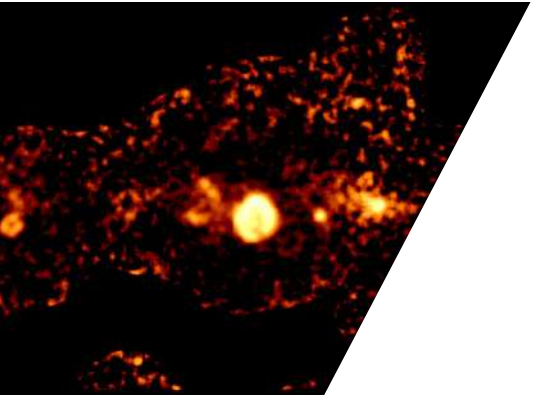
- Compliance to the FAIR4RS principle very heterogeneous
  - Due to s/w birth date, funding, etc
  - Metadata, DOI, SWHID

Desirable: go by steps towards FAIR4RS compliance

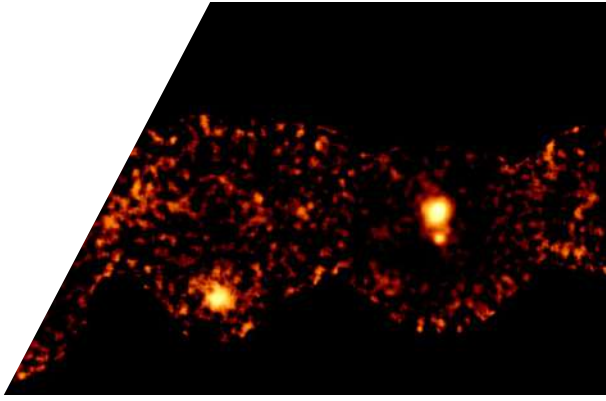
- Documentation to be improved, **always**
  - User's feedback is essential – organize schools!

Strong recommendation: rich documentation

- **Insert metadata and provenance into generated data**, to permit Interoperability and Reusability
  - Software used to create the data: version, URL towards the project or its documentation, DOI
  - Provenance of the used data: for each, DOI, documentation of the data release, documentation of the used data format, exact dataset
  - Data processing: at least a reference describing how the used data are processed or used to generate the final product
  - Context metadata: who has generated the final product, contact information



Conclusions and Recommandations



WP4.1 main summary

Harmonisation of Data Format and Content

- **Serialization Formats**: Adopt widely recognised and standardised data formats (e.g., FITS format)
- **Controlled Vocabularies and Ontologies**: data categories, terms, and definitions are consistent and universally understood for the data format and their metadata
- **Version Control**: reproducibility and citation
- **Identifiers**: citation, access
- **Licensing**: clarify how users can reuse the data

Comprehensive Metadata Standards

- **Descriptive Metadata**: information about the data's content, structure, context, and meaning; use of established metadata standards (e.g., DataCite Metadata Schema).
- **Machine-Readable Metadata**: machine-readable check (e.g., using schema.org, RDF, or JSON-LD) → aids in automating data discovery and integration processes, as well safe processing of data by libraries

Provenance Information

- **Documentation of Data Provenance**: describe the data’s origins, manipulations, transformations, processing steps, and link the data with their provenance document via metadata
- **Access to Source Data**: link to and provide access to original or raw datasets

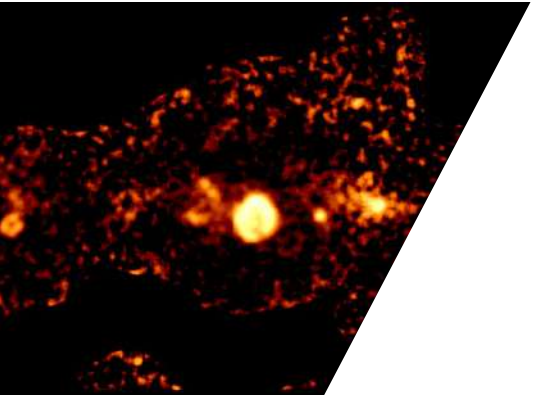
WP4.1 main summary (2)

For Software and Libraries

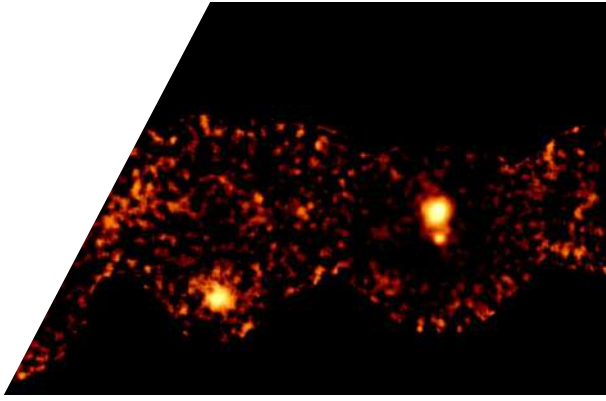
- **FAIR4RS principles**: of paramount importance. A significant portion of scientific data's value derives from the software's underlying logic and algorithms
- **Metadata and Provenance handling**: ensure coherent and secure processing, and 'metadata reduction' to generate FAIR-compliant output data. Software is responsible for the quality, reusability and interoperability of the data it generates
- **Serialization formats**: Avoid using formats that prevent users or libraries from accessing metadata and provenance information. Bad examples: JPEG or PNG
- **Software Documentation**: enables users to fully understand the data provided.

#FAIR4RS

ASTRO-ORDAS
OSCARS-FIESTA
ACME-ASTRO.eu
ACME
Astrophysics Centre for
Multimessenger studies in Europe

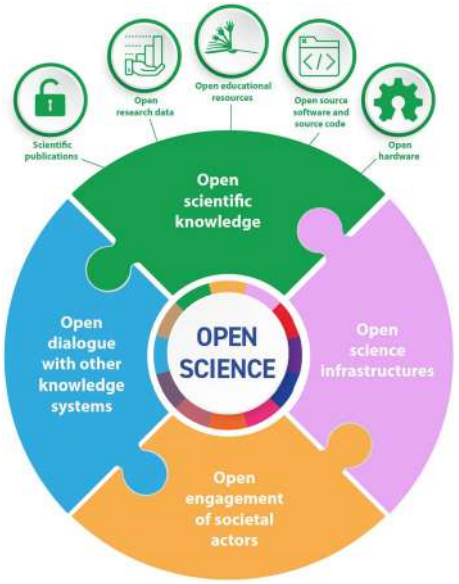


Annexes

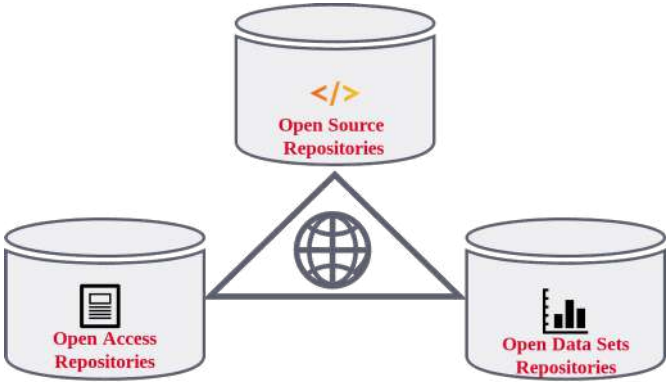


The emergence of Open Science

« Open Science – Making science more accessible, inclusive and equitable for the benefit of all »
UNESCO



The Open Science pillars



Open software are difficult to separate from open data...

The project Gammapy

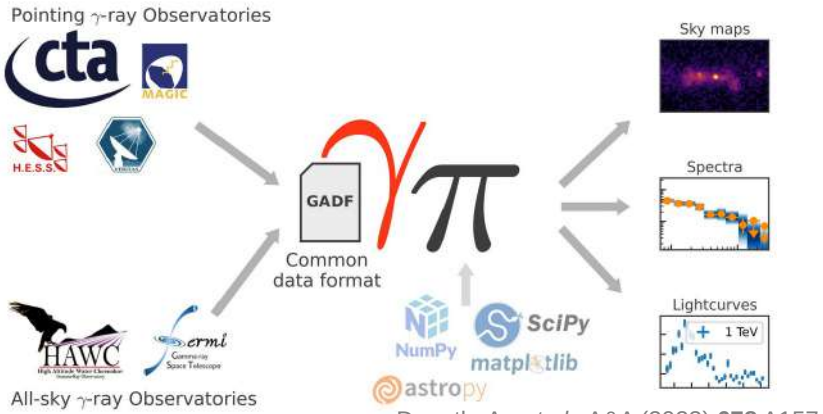
A Python package for gamma-ray astronomy

Open Python analysis library

- Uses data written in the GADF format
- Inserted into the Python ecosystem
- Respecting the FAIR4RS principles
- Making multi-instrument joint analyses

Open Research Software

- Open contributions within an open organisation with an open governance
- Reference library for the VHE gamma-ray astronomy
  - Selected as core library of the open Science Analysis Tool of CTAO
- Well recognized : jury's prize of the CoSO (2022)



Donath, A., *et al.*, A&A (2023) **678** A157

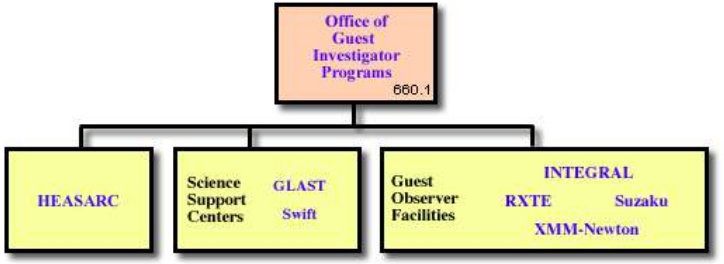
FAIR4RS principles

As with the FAIR Guiding Principles, the FAIR4RS Principles (2022) are intended to be aspirational. The application of the FAIR4RS Principles is the responsibility of the owners (who are often the creators) of the software, not the users.

- **F** Software, and its associated metadata, is easy for both humans and machines to find
- **A** Software, and its metadata, is retrievable via standardized protocols
- **I** Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
- **R** Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software)

X-rays → GeV: OGIP

Office of Guest Investigator Programs conventions



15 specific format recommendations (link) Corcoran *et al*., 1995 **ADASS**

- None on the final astrophysical products

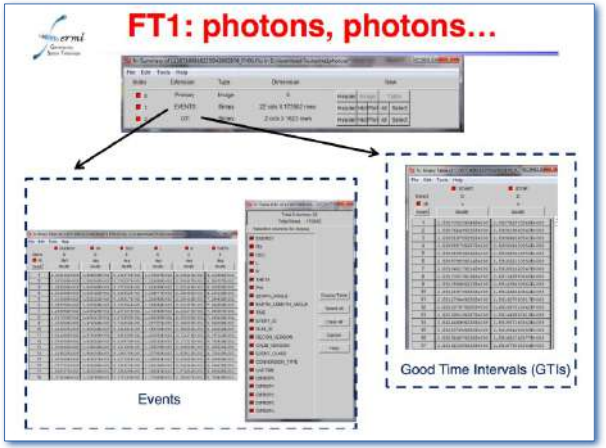
- **Event list:** well described, except errors, no detailed metadata
- **Instrument Response Files** (IRFs):
  - PHA (XSPEC compatible): ie RMF+ARF, ie Edisp+Area, includes statistical errors and systematics, link to the background file, GTI, metadata (e.g. creation, history)
  - PSF (radial or 2D): includes statistical errors, metadata (e.g. creation, history)

Name	Units	Description
TIME	`s' or `d'	The time associated with the event
RAWX	`pixel'	Raw telemetry X position of the event
RAWY	`pixel'	Raw telemetry Y position of the event
DETX	`pixel'	Linearized X position of the event on the detector
DETY	`pixel'	Linearized Y position of the event on the detector
X	`pixel'	Projected X position of the event on the sky
Y	`pixel'	Projected Y position of the event on the sky
PHA	`chan'	Pulse height analyzer' energy channel

HE γ-rays: Fermi-LAT

Formats regulated very early, and afterwards updated

- Project Data Management Plan (PDMP), 2007 (link)
- Science Data Products File Format Document (FFD), 2019 (link)



Definition of Data levels (L0 → L3)

- like for X-rays (HEASARC continuity)
- (raw data)
- **L0**: 'cleaned raw data'
- **L1**: starting point for scientific analyses
- **L2**: result of science analysis tools
- **L3**: catalogs or compendia of DL2 data, including e.g. flux history, sce identification

In addition, **ancillary data**: diffuse galactic interstellar, extragalactic emission models, pulsar ephemerides

Data format

- based on OGIP + specificities (e.g. interstellar emission model, BAT data, LAT LLE)
- specific format for L2 (LC, spectrum) and L3 (catalogs)
- Metadata: s/w name and version, but not the release name, no provenance in data (but in web pages)

Almost FAIR
Light compliance to IVOA standards

VHE γ-rays: GADF

Up to mid-10's, VHE community worked in a totally competitive and closed mode

- All was private
- Except few MoUs around scientific projects

Better results
Interoperability between instruments
Respect of the FAIR principles

Some 'dreamers' worked towards the opening of the VHE astrophysics

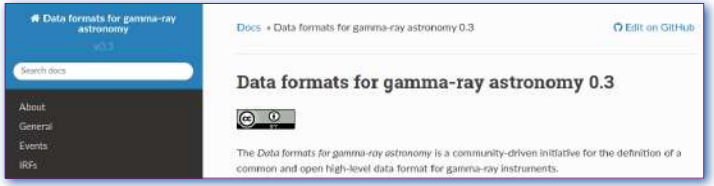
Deil, C., *et al.*, ASTERICS 2016 (link)

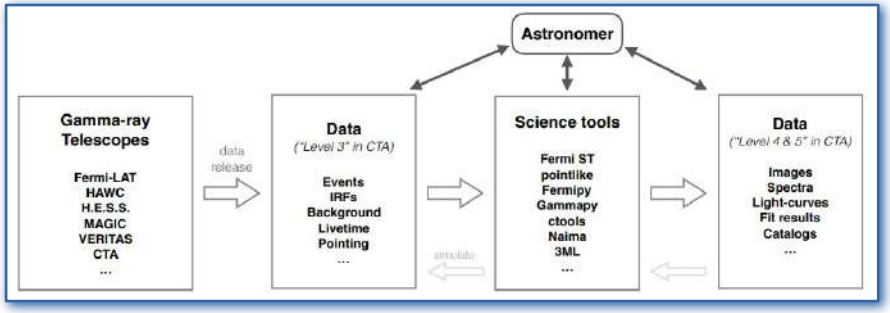
- Data format standardization: open initiative 'Open Gamma-Ray Astro'
- Open Science Tools: Gammapy

The Gamma Astro Data Format

DOI 10.5281/zenodo.7304668

- Strongly influenced by the Fermi-LAT format (and OGIP) and serialization into FITS
- Same type of data levels: DL3, DL4, DL5
- Full description of the DL3: event list and IRFs





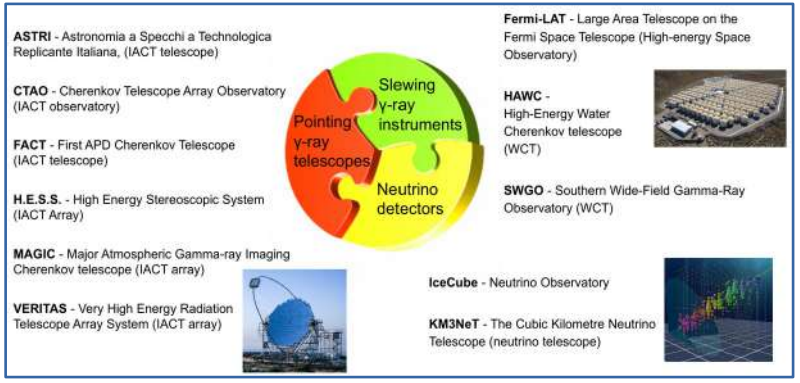
Deil, C., *et al.*, Proc. of Gamma 2016

VHE data: VODF

Open Initiative ‘Very-high-energy Open Data Format’ (link)

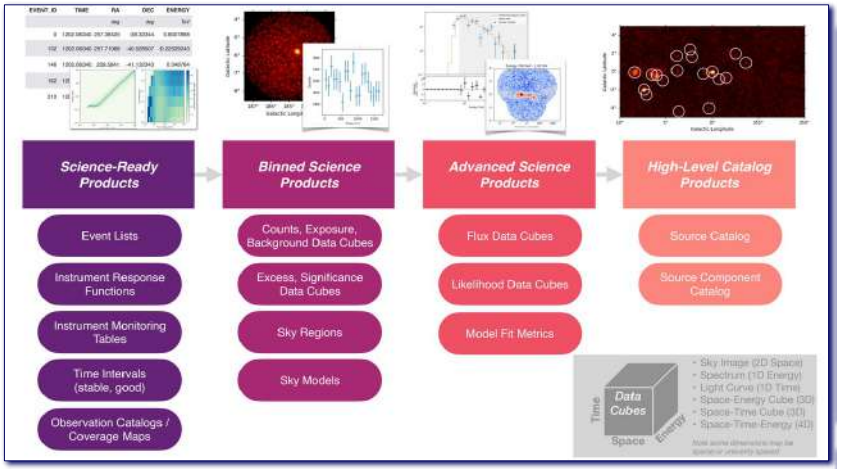
- Aims to format VHE data (gamma and neutrino)

- Officially supported by 11 experiments



- Structured with a project organization

  Coordination Committee, Conveners: R. Zanin, B. Khélifi
  Lead Editors: K. Kosack, L. Olivera-Nieto, J. Schnabel



Khélifi, B., *et al.*, Proc. of 38th ICRC (2023)

```
/*****************************************************************************
/* FITS FILE:
/*    VODF Level-1 Event Data
/*
/* EXTENSIONS SUMMARY:
/*  IDX  NAME                VER CLASS                TYPE
/*  ---  ------------------- --- -------------------- ----------------
/*    0. EVENTS                0 OGIP.EVENTS          [TableExtension]
/*    1. SOI                   0                      [TableExtension]
/*****************************************************************************

/ ############################################################################
/ HDU: EVENTS
/ DESCRIPTION:
/    VODF Level 1 Event List
/ ############################################################################
XTENSION = BINTABLE
EXTNAME  = EVENTS
EXTVER   = 0
```

© Kosack, K. and Khélifi, B.