ASTERICS - H2020 - 653477

Data Format Survey

ASTERICS GA DELIVERABLE: D3.3

Document identifier:	ASTERICS-D3.3-draft-v0.pdf
Date:	**01-04-2016**
Work Package:	**OBELICS – WP3**
Lead Partner:	**UCM**
Document Status:	**INCOMPLETE DRAFT**
Dissemination level:	**INTERNAL**
Document Link:	https://www.asterics2020.eu/dokuwiki/lib/exe/fetch.php?media=intra:wp3:task3.2:ASTERICS-D3.3-draft-vo.pdf

Abstract

<<write the abstract here>>

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/WP	Date
From	Jaime Rosado	UCM/WP3	01/04/2016
Author(s)			
Reviewed by			
Approved by			

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
V0	01/04/2016	Incomplete draft	J. Rosado/UCM

1			
2			
3			
4			

IV. APPLICATON AREA

This document is a formal deliverable for the GA of the project, applicable to all members of the ASTERICS project, beneficiaries and third parties, as well as its collaborating projects.

V. DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors. The procedures documented in the ASTERICS "Document Management Procedure" will be followed: https://wiki.asterics2020.eu/wiki/Procedures

VI. TERMINOLOGY

A complete project glossary is provided at the following page: http://www.asterics2020.eu/about/glossary/

VII. PROJECT SUMMARY

TBD

VIII. EXECUTIVE SUMMARY

TBD

Table of contents

Acronym list

ANTARES	Astronomy with a Neutrino Telescope and Abyss environmental RESearch
CASA	Common Astronomy Software Applications
CTA	Cherenkov Telescope Array
E-ELT	European Extremely Large Telescope
EGO	European Gravitational Observatory
ESFRI	European Strategy Forum on Research Infrastructures
ESO	European Southern Observatory
e-VLBI	Electronic Very-Long-Baseline Interferometry
EVN	The European VLBI Network
FITS	Flexible Image Transport System
HDF	Hierarchical Data Format
H.E.S.S.	The High Energy Stereoscopic System
IACT	Imaging Atmospheric Cherenkov Telescope
IGWD	Interferometric Gravitational Wave Detector
JIVE	Joint Institute for VLBI in Europe
KM3NeT	Cubic Kilometre Neutrino Telescope
LIGO	Laser Interferometer Gravitational Wave Observatory
LOFAR	The Low Frequency Array
LSST	The Large Synoptic Survey Telescope
MAGIC	Major Atmospheric Gamma-Ray Imaging Cherenkov
SKA	The Square Kilometre Array

1. Introduction

TBD

2. Document goal

TBD

3. ESFRI projects and related pathfinders

ASTERICS will benefit research infrastructures identified in the ESFRI Roadmap (CTA, KM3NeT, SKA and E-ELT) as well as other major international projects including precursor experiments. This data format survey encompasses the 13 projects listed below. The table includes the field of application of each experiment and the type of data produced as identified in section 4.

Project	ESFRI	Field	Type of data
CTA	Yes	Cherenkov telescope arrays for gamma astronomy	Events
H.E.S.S.	Pathfinder for CTA	Cherenkov telescope array for gamma astronomy	Events
MAGIC	Pathfinder for CTA	Cherenkov telescope array for gamma astronomy	Events
KM3NeT	Yes	Neutrino telescope	Events
IceCube	Pathfinder for KM3NeT	Neutrino telescope	Events
ANTARES	Pathfinder for KM3NeT	Neutrino telescope	Events
E-ELT	Yes	Ground-based optical/near-infrared telescope	Images
LSST	No	Optical telescope	Images

EUCLID	No	Satellite mission to map the dark Universe	Images
SKA	Yes	Radio telescope arrays	Signals
e-EVN	Pathfinder for SKA	e-VLBI network for radio astronomy	Signals
LOFAR	Pathfinder for SKA	Radio interferometric array	Signals
Advanced LIGO	No	Gravitational wave detectors	Signals
Advanced Virgo	No	Gravitational wave detector	Signals

Table 1: Projects included in this data format survey. Experiments are classified according to the type of data that they produce.

4. Types of data

Data produced by the experiments included in this survey can be classified in three groups: event-based, image-based and signal-based data. Each type of data has different needs for processing, storage and accessibility, hence formats should be chosen accordingly.

Synergies between experiments producing similar type of data are more natural and easier to identify. Therefore software reuse and common standards are expected to be attained to much more extent in experiments of the same group. In the next sections, these three groups of experiments classified according to the type of data that they produce are reviewed and their corresponding data formats (both in use and envisaged) are discussed.

It should be noted, however, that these groups of experiments are not completely bounded and overlap each other. This opens the possibility to look for common solutions for a wider range of experiments.

5. Event-based data

This type of data is produced by gamma-ray observatories (i.e., CTA [CTA web], H.E.S.S. [HESS web] and MAGIC [MAGIC web]) as well as neutrino observatories (i.e., KM3NeT [KM3NeT web], IceCube [IceCube web] and ANTARES [ANTARES web]). Cosmic-ray experiments (e.g., the Pierre Auger Observatory [Auger_web]) also belong to this group, although they are not included in table 1.

A gamma-ray observatory comprises several imaging atmospheric Cherenkov telescopes (IACTs) with hundreds or thousands of camera pixels each, which involves an enormously high data rate. The data acquisition system builds "events" in real time by grouping data from different pixels and telescopes based on minimal topological and temporal criteria. Other data (i.e., telescope calibration data, weather, pixel voltages, etc.) are associated to the event. Reconstruction of shower parameters is performed by combining all this complex information. As a consequence, permanent data storage must be done with relatively little filtering to allow a complete shower reconstruction afresh. Whereas the H.E.S.S. and MAGIC observatories have few Cherenkov telescopes, CTA will operate two arrays, each one having many tens of telescopes. This will translate into a raw data rate of around 10 GB/s and a storage requirement of 27 PB/year [CTA_DataManag].

Cosmic neutrino telescopes also comprise thousands of optical sensors spread over a large volume. Likewise, event building and shower reconstruction rely on the temporal coincidence of signals and the stereoscopic technique, implying requirements for data handling similar to gamma-ray observatories. For instance, the data acquisition system of the ANTARES detector, which is deployed in the deep waters of the Mediterranean Sea, is based on the "all-data-to-shore" concept, meaning that all signals that pass a certain threshold are sent to shore for real-time processing [ANTARES_DAQ]. In the case of the KM3NeT project, the total data rate will amount to 25 GB/s [KM3NeT_TDR].

Event-based data are organized hierarchically. In the case of CTA, the raw data model may be seen as a class structure with the following levels: sub-array, telescope, monitor unit, and pixel [CTA_DataManag], where "sub-array" refers to a set of telescopes used for a given observation. In principle, this hierarchical structure conditions the data model to be specific for each experiment. Consequently, proprietary data formats have been developed for this purpose. For instance, in MAGIC, raw data are stored in binary form and then translated to ROOT format [ROOT] in a preprocessing step (see, e.g., [MAGIC_ThesisAleksic]). ROOT-based formats are also the selected option for reduced event data in H.E.S.S. [HESS_DAQ] and ANTARES [ANTARES_DAQ]. It is also worth mentioning that IceCube provides publicly event data in an open format based on python [IceCube_PublicData].

The data model must also take into consideration that events are usually grouped into "runs" with constant observation conditions. In addition, it is necessary to define the metadata containing the information needed to locate and handle event data as well as the way in which calibration and other auxiliary data are associated to them. Note that these auxiliary data may be related to an individual PMT or the whole observatory. Several approaches are followed for this purpose. For instance, the archival and data exchange system of IceCube creates XML files following a metadata specification called "DIF Plus" [IceCube_web] compliant with the format required by the Office of Polar Programs for all experiments funded by the National Science Foundation.

The creation of standards for raw event-based data with the above features would facilitate interoperability and software reuse for these experiments. In particular, it could be conceived an event-data structure general enough to be applied to any experiment belonging to this group. Objects as “telescope” and “optical module” may be seen as instances of generic hierarchal levels of such a structure. As an alternative to existing ROOT-based formats, HDF5 [HDF5] would be very suitable for this kind of data.

The data flow in the experiments of this group is organized into a series of steps that transform the raw event-data into high level scientific data products. The first steps are performed on site, whereas higher data levels are processed off site. At a certain stage of the reduction chain, data consist in a list of well reconstructed events along with associated instrumental response characterizations and any technical data needed for science analysis. Event data of this level only contain parameters such as the type of particle, energy, arriving direction, etc. and should be nearly independent of the particularities of the detector. Guaranteed access to this data level will be provided in the CTA archive to basic users [CTA_DataManag]. Efforts are being made to provide open standards for this data level, especially in FITS format [FITS]. Remarkably, an initiative has been created to describe current data formats for this and higher levels in gamma ray astronomy [gamma_astro_data_formats]. Open specifications should be easily extended to neutrino and cosmic ray communities too. For example, the Pierre Auger Collaboration has already made publicly available information on a fraction of reconstructed events [Auger_PublicData], but this collaboration may profit from these open formats if they go one step further.

Higher level data are binned data products (e.g., spectra and sky maps) and legacy observatory data (e.g., survey sky maps and source catalogues), representing a very small fraction of the total data produced by an experiment. At least for future projects, these data are meant to be integrated within the International Virtual Observatory Alliance (IVOA) infrastructure using FITS format. The final goal is to allow scientists to do multi-messenger astronomical studies by combining data from different facilities in a single analysis. Significant progresses have already been made to adopt IVOA standards in the very-high-energy astronomy community [MAGIC_Public, SLF_gamma] in view of the CTA project, which is intended to work as a “standard” astronomical observatory.

6. Image-based data

Images are the main type of data produced by optical/near-infrared telescopes. The three projects included in table 1 that belong to this group are E-ELT [E-ELT_Web] , LSST [LSST_web] and Euclid [Euclid_web], which are under construction or design phase. E-ELT and LSST are ground-based telescopes, whereas Euclid is a satellite mission.

These experiments will stream large amounts of data. For example, E-ELT will produce approximately 15 TB per night of raw imaging data [LSST_DAQ] and Euclid will deliver about 110 GB of compressed data per day, which is an unprecedented large volume of data for a space mission [Euclid_DSR]. As a consequence, real-time processing is a challenge for these projects. One of the requirements of E-ELT is that the raw scientific data, including calibrations, shall arrive at the ESO Science Archive Facility not later than 1 h after they have been obtained [E-ELT_req]. The data management systems of these experiments will also generate alerts of transient events with the implication that timely follow-up is of interest. For this purpose, the VOEvent format [VOEvent] is broadly accepted and used by the astronomical community.

Processed data and science-ready data products (e.g., calibrated images, spectra and catalogues) from these experiments will be archived and made public. The data management system of LSST will reprocesses annually the accumulated survey data to form a static, self-consistent data release, the contents of the final data release alone being expected to be around 70 PB [LSST_DAQ]. In the case of E-ELT, raw data, including calibrations, will also be made available to the worldwide community at the end of a proprietary period [E-ELT_req]. All these data will be in a standard format, namely FITS format [FITS]. Science-ready data products shall contain metadata compliant with Virtual Observatory standards and be made available to Virtual Observatory tools.

7. Signal-based data

In radio interferometric arrays, signals from all the individual telescopes are brought together and processed by the so-called correlator, which combines the signals to form an image of the observed radio source. Among the projects listed in table 1, SKA [SKA_web], EVN [EVN_web] and LOFAR [LOFAR_web] belong to this category.

The European VLBI Network (EVN) comprises 21 very large and sensitive reflector telescopes (dishes) spread throughout Europe and beyond. In addition to “traditional” VLBI sessions in which the data are first recorded at the telescopes on tapes or discs and then physically delivered to the processor, the EVN consortium is conducting a development programme called e-EVN, which aims at creating an e-VLBI network where data is transferred and processed in real time. All data correlated at the EVN Data Processor at JIVE (raw FITS format data, processed images, calibration tables, etc.) are placed into the EVN archive and made public once the 12-month proprietary period has expired [EVN_web].

The LOFAR array, on the other hand, consists of about 7000 simple omni-directional antennas organised in stations containing local computing resources to perform beam-forming [LOFAR_paper]. All these beam-formed data (about 19 GB/s for the entire array) are sent via a high-speed fibre network to the central processing facility, where they are pre-processed on-line. In interferometric imaging mode, pre-processed data are stored in the standard format for the CASA package [CASA], whereas the HDF5 format is used for pulsar processing. Then, several off-line processing steps are performed to obtain either images or pulsar data products in PRESTO format [PRESTO]. The final scientific data are transferred to the LOFAR long-term archive for cataloguing and distribution to the community [LOFAR_doc]. This archive is expected to grow by up to 5 PB per year.

The above two projects are recognised as pathfinders for SKA, which will operate both an array of dishes and an array of antennas grouped into stations. The SKA project is designed in two phases and, when both phases are complete, the observatory will consist of many thousands of connected radio telescopes. For the first phase, the summed data rate is estimated to be about 3 TB/s and the expected archive data volume is 100 PB per year [SKA1]. The SKA project will profit from the lessons learned on data management in existing radio interferometric arrays, although the file formats to be used to store data is still to be defined.

The two Advanced LIGO detectors in EEUU [LIGO_web] and the Advanced Virgo detector in Italy [Virgo_web] included in table 1 are the first ones of a network of very sensitive interferometric detectors of gravitational waves. These detectors also produce signal-based data, but they are very different to those generated by radio interferometric arrays. The data produced by these experiments are stored in “frames” with thousands of channels, where the gravitational-wave strain channel only represents a small fraction of data and all the other channels are used to auxiliary instrumental and environmental monitoring. Each interferometer has a data rate of a few tens of MB/s [LIGO_data_management, Virgo_advanced].

A standard data format, called IGWD Frame format, was already established by a LIGO-Virgo agreement in 1997 [IGWD_frame]. Both Virgo and LIGO plan to make data publicly available in IGWD Frame format as well as in other standard formats easier to use for most open data users (e.g., HDF5). In addition, it is proposed to begin a program that uses the Virtual Observatory VOEvent Transport Protocol to communicate real-time alerts to follow-up observers.

8. Synergies

TBD

9. Risks

TBD

10. Conclusions

TBD

References

[CTA_web] https://www.cta-observatory.org/

[HESS_web] https://www.mpi-hd.mpg.de/hfm/HESS/

[MAGIC_web] https://magic.mpp.mpg.de/

[Auger_web] https://www.auger.org/

[KM3NeT_web] http://www.km3net.org/home.php

[IceCube_web] https://icecube.wisc.edu/

[ANTARES_web] http://antares.in2p3.fr/

[CTA_DataManag] G. Lamanna et al., Cherenkov Telescope Data Management, PoS (ICRC2015) 947. Available at: http://arxiv.org/abs/1509.01012

[MAGIC_ThesisAleksic] J. Aleksić, Optimized Dark Matter Searches in Deep Observations of Segue 1 with MAGIC, PhD Thesis (2013).

[MAGIC_DAQ_II] J.A. Coarasa et al., The Data Acquisition of the MAGIC II Telescope, 29th ICRC (2005).

[HESS_DAQ] A. Balzer et al., Astropart. Phys. 54 (2015) 67.

[KM3NeT_TDR] KM3NeT Technical Design Report. Available at: http://km3net.org/TDR/TDRKM3NeT

[ANTARES_DAQ] J.A. Aguilar et al., NIM A 570 (2007) 107.

[IceCube_DAQ] R. Abbasi et al., NIM A 601 (2009) 294.

[IceCube_OpenData] http://icecube.umd.edu/PublicData/

[ROOT] https://root.cern.ch/

[FITS] W.D. Pence et al., A&A 524 (2010) A42.

[gamma_astro_data_formats] http://gamma-astro-data-formats.readthedocs.org/en/latest/

[Auger_PublicData] http://auger.colostate.edu/ED/

[HDF5] https://www.hdfgroup.org/HDF5/

[MAGIC_Public] http://magic.pic.es/ (see section "Public").

[SLF_gamma] https://www-zeuthen.desy.de/multi-messenger/GammaRayData/

[E-ELT_Web] https://www.eso.org/sci/facilities/eelt/

[LSST_web] http://www.lsst.org/

[Euclid_web] http://www.euclid-ec.org/

[LSST_DAQ] M. Jurić et al., The LSST Data Management System. Available at http://arxiv.org/abs/1512.07914

[Euclid_DSR] Euclid Definition Study Report. Available at http://arxiv.org/abs/1110.3193

[E-ELT_req] E-ELT Programme – Observatory Top Level Requirements. Available at https://www.eso.org/sci/facilities/eelt/

[VOEvent] VOEvent: Sky Event Reporting Metada. Available at http://www.ivoa.net/documents/VOEvent/

[SKA_web] https://www.skatelescope.org/

[EVN_web] http://www.evlbi.org/

[LOFAR_web] http://www.lofar.org/

[LOFAR_paper] M.P. van Harlem et al., A&A 556 (2013) A2.

[SKA1] SKA1 System Baseline v2. Available at https://www.skatelescope.org/key-documents/

[CASA] https://casa.nrao.edu/

[LOFAR_doc] http://www.astron.nl/radio-observatory/astronomers/technical-information/lofar-technical-information

[PRESTO] https://prestodb.io/

[LIGO_web] http://www.ligo.org/index.php

[Virgo_web] http://www.virgo-gw.eu/

[EGO_web] https://www.ego-gw.it/

[LIGO_data_management] LIGO Data Management Plant. Available at https://dcc.ligo.org/LIGO-M1000066/public

[Virgo_advanced] F. Acernese et al., Class. Quantum Grav. 32 (2015) 024001.

[IGWD_frame] Specification of a common data frame format for interferometric gravitational wave detectors. Available at http://dcc.ligo.org/cgi-bin/DocDB/RetrieveFile?docid=329