



ASTERICS - H2020 - 653477

Final integral WP3 Report

ASTERICS GA DELIVERABLE: D3.17

Document identifier:	ASTERICS-D3.17.doc
Date:	7 July 2019
Work Package:	OBELICS – WP3
Lead Partner:	CNRS-LAPP
Document Status:	Final
Dissemination level:	Public
Document Link:	www.asterics2020.eu/documents/ASTERICS-D3.17.pdf

Abstract

In this report we integrated the achievements of the OBELICS work package over the full four years of the ASTERICS project. This report provides a comprehensive list of activities carried out by the work package to fulfill the objectives mentioned in the DoA.

I. COPYRIGHT NOTICE

Copyright © Members of the ASTERICS Collaboration, 2015. See www.asterics2020.eu for details of the ASTERICS project and the collaboration. ASTERICS (Astronomy ESFRI & Research Infrastructure Cluster) is a project funded by the European Commission as a Research and Innovation Actions (RIA) within the H2020 Framework Programme. ASTERICS began in May 2015 and will run for 4 years.

This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, and USA. The work must be attributed by attaching the following reference to the copied elements: "Copyright © Members of the ASTERICS Collaboration, 2015. See www.asterics2020.eu for details of the ASTERICS project and the collaboration". Using this document in a way and/or for purposes not foreseen in the license, requires the prior written permission of the copyright holders. The information contained in this document represents the views of the copyright holders as of the date such views are published.

II. DELIVERY SLIP

	Name	Partner/WP	Date
Author(s)	Jayesh Wagh	CNRS-LAPP/WP3	05/06/2019
Amendments			
Reviewed by	Rob van der Meer	ASTRON	17/06/2019
Approved by	AMST		10/07/2019

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
v1	05/06/2019	first draft	Jayesh Wagh
v2	14/06/2019	revised draft	Jayesh Wagh
v3	07/07/2019	final version	Jayesh Wagh

IV. APPLICATION 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. TERMINOLOGY

OBELICS	OBservatory E-environments Linked by common ChallengeS
ESFRI	European Forum on Research Infrastructures
HNSciCloud	Helix Nebula Science Cloud
AARC2	Authentication and Authorisation For Research and Collaboration
APPEC	Astroparticle Physics European Consortium
KM3NeT	Cubic Kilometre Neutrino Telescope
CTA	Cherenkov telescope array
ASTRI	Astrofisica con Specchi a Tecnologia replicante Italiana
ASTRON	Netherlands Institute for Radio Astronomy
INFN	Italian National Institute for Nuclear Physics
INAF	Italian National Institute for Astrophysics
IACT	Imaging Atmospheric Cherenkov Telescope
MAGIC	Major Atmospheric Gamma Imaging Cherenkov Telescopes
UCAM	University of Cambridge
UCM	Complutense University of Madrid
ADASS	Astronomical Data Analysis Software & Systems
EOSC	European Open Science Cloud
EGI	European Grid Infrastructure

A complete project glossary is provided at the following page:

<https://www.asterics2020.eu/glossary>

VI. EXECUTIVE SUMMARY

This report provides a comprehensive overview of OBELICS activities between May 2015 till April 2019. Over these four years the OBELICS work package has organized several workshops and thematic training events allowing networking and knowledge exchange between H2020 projects, astronomy consortia and industries. On the technical responsibilities the work package members successfully delivered to the needs of the astronomy community in the form of OBELICS software and scientific services repository as well as data formats, gathering together contributions from various astronomy projects to support interoperability and cross-fertilization. The report provides more information on these contributions from the OBELICS work package to the astronomy and astrophysics community.

Table of contents

I. COPYRIGHT NOTICE	1
II. DELIVERY SLIP.....	1
III. DOCUMENT LOG	1
IV. APPLICATION AREA	2
V. TERMINOLOGY	2
VI. EXECUTIVE SUMMARY	3
Table of contents.....	4
1. Introduction	5
2. Management, user engagement and data dissemination (MAUD)	7
3. Data generation and information extraction (D-GEX)	9
4. Data systems integration (D-INT)	10
5. Data analysis/interpretation (D-ANA).....	11
6. Conclusion.....	13

1. Introduction

The OBELICS work package is one of the core work packages of H2020-ASTERICS. It aims to enable interoperability and software re-use in data generation, integration, and analysis. The following objectives were set for WP3 OBELICS in the DoA.

- Train researchers and data scientists in the ASTERICS ESFRI and pathfinder projects to apply state-of-the-art parallel software programming techniques, to adopt big-data software frameworks, to benefit from new processor architectures and e-science infrastructures. This will create a community of experts that can contribute across facilities and domains.
- Maximise software re-use and co-development of technology for the robust and flexible handling of the huge data streams generated by the ASTERICS ESFRI and pathfinder facilities. This involves the definition of open standards and design patterns, and the development of software libraries in an open innovation environment.
- Adapt and optimise extremely large database systems to fulfil the requirements of the ASTERICS ESFRI projects. This requires the development of use cases, prototypes and benchmarks to demonstrate scalability and deployment on distributed non-homogeneous resources. Cooperation with the ESFRI pathfinders, computing centres, e-infrastructure providers and industry will be organised and managed to fulfil this objective.
- Study and demonstrate data integration across ASTERICS ESFRI and pathfinder projects using data mining tools and statistical analysis techniques on Petascale data sets. This will require adaptable and evolving workflow management systems, to allow deployment on existing and future e-science infrastructures.

The OBELICS work package was organized in the following four task groups to efficiently address these objectives.

- **Task 3.1 Management, user engagement and data dissemination (MAUD):** This task concerns overall management of the work package, user engagement & data dissemination through thematic training events and general workshops.
- **Task 3.2 Data generation & information extraction (D-GEX):** This task aims at the first stage of the scientific data flow concerning data generation and information extraction.

- **Task 3.3 Data systems integration (D-INT):** The task D-INT (3.3) is targeting the challenges in the data management of the large ESFRI infrastructures.
- **Task 3.4 Data analysis /interpretation (D-ANA):** This task addresses common challenge to assess the quality of Petascale datasets and execute automatic analysis to reduce their size by developing a collection of statistically robust and domain independent open source software libraries for data analysis and data mining on Peta-scale datasets. During the reporting period.

In this report we present the list of activities carried out by each of the task groups over the past four years to achieve the objectives listed in the DoA.

2. Management, user engagement and data dissemination (MAUD)

During the course of the project, the MAUD task group, led by CNRS-LAPP, organized four general workshops and three thematic training events.

The ASTERICS-OBELICS workshops served as a platform to bring together H2020 projects such as IndigoDataCloud, HNSciCloud, AARC2, industries, such as ATOS, NVIDIA, ORACLE Europe, Common Workflow Language, E4, OROBIX, and consortia, such as EU-T0, APPEC & EGI, and discuss the best practices in data management, data preservation and open science. The workshop also attracted participants from CERN as well as LIGO Laboratory, Caltech USA, European Gravitational Observatory, European Southern Observatory. The general workshops were key to disseminate the work package activities and achievements to the astronomy, astroparticle physics as well as particle physics community. Python-based tools are increasingly gaining acceptance and importance in astronomy. At the same time, the clear trend is towards interoperable and open data. The ASTERICS-OBELICS Pygamma workshop was organized to discuss the status and prospects for gamma-ray astronomy and related fields. ASTERICS-OBELICS workshops also introduced the OBELICS work package members with novel technologies and techniques implemented by other projects. The collaborations with industries and European Solar Telescope project were amongst the outcomes of the OBELICS workshops. Significant emphasis was given to how ASTERICS-OBELICS efforts can be directed towards the European Open Science Cloud. A set of recommendations were submitted to the EOSC-HLEG as an outcome of these discussions.



Figure 1: First ASTERICS-OBELICS Workshop, December 2016, Rome, Italy

ASTERICS-OBELICS training events addressed the vital needs in advanced software programming for the astronomy, astrophysics and astroparticle physics communities. We have received a more than expected number of requests for registration for all the three editions. Most of the participants informed us that they have never been exposed to a formal training in software programming and that they learned everything about programming

languages by self-study. These training events observed participation from all the stages in academia from masters and PhD students to postdocs and senior researchers.

During the project, OBELICS members also formalized collaborations with five industries to bring in their expertise in machine learning, common workflow language as well as software development to achieve innovation related objectives of the OBELICS work package.



Figure 2: Third ASTERICS-OBELICS international school, 8-12 April 2019, CNRS-LAPP, Annecy, France

OBELICS dissemination efforts were further enhanced by collaboration with Trust-IT services to improve social media presence and to reach out to the concerned astronomy and astroparticle physics communities. MAUD also formalized overall five industrial collaborations through subcontracting for various scientific projects involving machine learning, common workflow language and advanced software development. These industrial subcontracts provided OBELICS members with the state-of-the-art expertise from participating industries to address the big data challenges.

3. Data generation and information extraction (D-GEX)

The D-GEX task group was led by INAF and UCM partners. D-GEX activities were dedicated to the provision of innovative solutions that are flexible enough to be adopted by the various connected ESFRI experiments. In this section, we have listed some of the major outcomes from the activities carried out by the D-GEX task group.

- A Data format survey was performed to seek synergies between the connected ESFRI experiments with the aim of fostering the use of open data standards that would enable interoperability and software re-use.
- Significant contribution was made to a recent initiative to define specifications of an open data format within the gamma-ray community. It was found that this data format can prove to be suitable to deliver public data from KM3Net.
- The CTA raw data format is not fixed yet. Several candidates were studied by OBELICS members and they actively contributed to the development of several benchmarks to compare the raw data formats. They developed one of the candidates, hipeDATA, that allows very high-performance analysis.
- Other developments in data formats have been the adoption of the open FITS format by the INAF group for handling ASTRI/CTA data and the studies by the ASTRON group to move from the domain specific ‘casa tables’ system towards HDF5 for data products other than images and measurement sets (visibilities).
- For high performance computing a data format generator (a library of functions to create and handle a data format) adapted to vectorisation for simulation files in Imaging Atmospheric Cherenkov Telescopes such as CTA was developed.
- A converter of MAGIC data to the DL3 format to validate the CTA DL3 format with real IACT data was developed.
- The final version of the Technology Benchmark Report discussed benchmarking performed on the data formats along with the other activities performed by D-GEX task members.
- Many tests were performed on innovative hardware with low-power consumption within ASTRI for CTA and DOME for SKA projects.
- Low power computing platforms in the ASTRI/CTA pipeline in order to test first steps of the data calibration and reduction were introduced. Some benchmarks of new hardware solutions, in particular ARM processors plus GPUs (Graphics Processing Units), have been performed for this kind of analysis and compared with traditional hardware.
- Algorithms for “Fast convolutional resample” of data that are able to split the work among many processors in parallel architectures were developed.
- System-on-chips (SoC) belonging to both the Intel and ARM ecosystems were acquired and tested in collaboration with the INFN COSA project.

4. Data systems integration (D-INT)

The task D-INT (3.3) aims at studying the challenges in the data management of the large ESFRI infrastructures and it was led by CNRS-LAPP and ASTRON. In this section, we have listed some of the major outcomes from the activities carried out by D-INT task groups.

- An innovative lossless compression algorithm produced by the work package had achieved reasonable data compression ratios with comparatively small compression times.
- In the context of the Large Synoptic Survey Telescope (LSST), a specific software called Qserv has been tested by ASTERICS in order to evaluate the Qserv capabilities and performances on real data sets processed through the LSST data processing pipeline (a.k.a the "stack"), and integrate its use into the science pipelines currently developed by the Dark Energy Science Collaboration (DESC).
- The problem of reproducibility concerns every experiment supposed to be running for tens of years as history has shown that informatics systems are evolving much faster than this timescale. To run its experiments, OBELICS investigated a few alternative technologies, such as Docker, plain scripts, virtual machines, Singularity, Nix and others, finding out that while there's no single perfect solution, Docker seems to be the best choice in terms of reproducibility of physics analyses, efficiency and usability in different operating systems.
- A first prototype of a common high-level data format (DL3) for current Imaging Atmospheric Cherenkov Telescopes experiments has also been developed. This format can be used also by other gamma-ray experiments (satellites or water Cherenkov detectors) and extended for the use of any event-based instrument (e. g. neutrinos, gravitational waves).
- The INAF CTA Authentication and Authorization Infrastructure (INAF CTA AAI) developed in OBELICS provides functionalities enforcing the protection of CTA resources and digital assets by means of a role based authorization. It offers a federated authentication, based on eduGAIN inter-federation, or a centralized CTA SAML authentication service. An attribute authority (based on Grouper) is provided in order to allow a role-based authorization thanks to a set of attributes managed and agreed at consortium level.
- The OBELICS D-INT Services Repository collects several technologies enabling the integration of analysis software. Some of these technologies have been developed in the ASTERICS project, others, namely Rucio and the Dirac framework, are developed externally, but were evaluated for their use in astroparticle physics and radio astronomy.

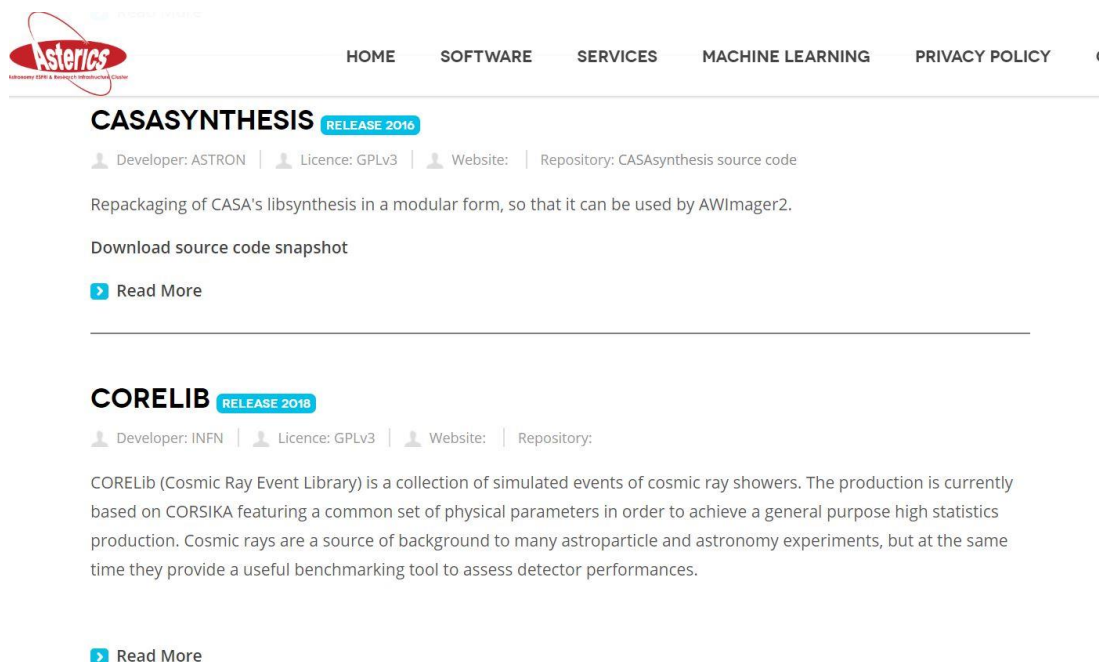
5. Data analysis/interpretation (D-ANA)

The task D-ANA addresses common challenges to assess the quality of Petascale datasets and execute automatic analysis to reduce their size by developing a collection of statistically robust and domain independent open source software libraries for data analysis and data mining on Peta-scale datasets. D-ANA task was led by UCAM and INAF. In this section, we have listed some of the major outcomes from the activities carried out by D-ANA task groups.

- A complete report on the benchmarking activities on A&A technologies performed in the framework of D-ANA is available on the open ASTERICS wiki at “Authentication & Authorisation Technology Benchmarking Report” .
- For imaging, a new method for putting measured correlations on a grid to perform an FFT was developed. This method, Image Domain Gridding (IDG), has proven to work well on both CPUs and GPUs, and is being ported to software for imaging LOFAR data.
- A library for fast, vectorised, array reductions and moment calculations written in C++ with Python binding was developed.
- A Statistical Analysis planning tool (StatPLAN) & Multi-wavelength inference (MW-INFERENCE) developed within OBELICS allow combination of data from different facilities, different wavelengths and potentially multi-messenger astronomy.
- A source detection/characterisation code, building on an in house code at Cambridge (BayeSys) was produced by OBELICS members. Compared to many other source finding methods, it produces a more detailed description of the shapes of sources and does not require maps to be run through CLEAN first. It is still in development, but we have already got some promising results applying it to ALMA data.
- A complete report on the benchmarking activities performed on A&A technologies “Authentication & Authorisation Technology Benchmarking Report” was produced.
- A Minimal Recomputation framework for astronomy was developed and this work was presented at ADASS, Santiago, 2017.
- Deep learning methods were investigated in the context of the Cherenkov Telescope Array. This work was carried out under the GammaLearner subcontracting project in collaboration with Orobix, an Italian sme.
- CTA - Workflow Management System API (INAF CTA-WMS API) developed within OBELICS interfaces with the WMS of the INAF CTA science gateway and it is implemented through web services, which exposes the whole life cycle of the workflow management through SOAP and REST APIs.
- The Common Workflow Language (CWL) was also investigated, since it is a specification for describing workflows in a way that makes them portable and scalable across a variety of software and hardware environments, including workstations, cluster, cloud infrastructure, and high performance computing environments.
- General tools for wavelet cleaning and the analysis chain prototype in the context of the Cherenkov Telescope Array were developed. These developments included reducing the number of scales in the wavelet cleaning at the minimum, i.e. two scales: the first scale

correspond to the uncorrelated background charge (one pixel); the second scale (2 pixels) concentrates most of the shower signal.

- Foreground separation and signal extraction methods for the Cosmological HI intensity mapping experiments were developed.
- Cosmic Ray Event Library (CORELib), a collection of simulated events of cosmic ray showers was also developed and updated during the project.
- CASA Jupyter kernels were updated to the latest CASA releases (CASA 5.3, CASA 5.4.1).
- A proof-of-concept single-detector low-latency classifier for astrophysical signals (inspiral and mergers of binary black holes) and local detector artifacts (glitches) based on deep learning, using minimally-processed data (time series) were developed.
- All the OBELICS developments were made open source and open access on the OBELICS software repository with necessary documentations, tutorials and download links.



ASTERICS Astronomy EPR & Beyond's Infrastructure Cluster

HOME SOFTWARE SERVICES MACHINE LEARNING PRIVACY POLICY

CASASYNTHESIS RELEASE 2016

Developer: ASTRON | Licence: GPLv3 | Website: | Repository: CASAsynthesis source code

Repackaging of CASA's libsynthesis in a modular form, so that it can be used by AWImager2.

Download source code snapshot

[Read More](#)

CORELIB RELEASE 2018

Developer: INFN | Licence: GPLv3 | Website: | Repository:

CORELib (Cosmic Ray Event Library) is a collection of simulated events of cosmic ray showers. The production is currently based on CORSIKA featuring a common set of physical parameters in order to achieve a general purpose high statistics production. Cosmic rays are a source of background to many astroparticle and astronomy experiments, but at the same time they provide a useful benchmarking tool to assess detector performances.

[Read More](#)

Figure 3: OBELICS software and services repository (<http://repository.asterics2020.eu/>)

6. Conclusion

Overall, all the four task groups have achieved their planned objectives. MAUD has successfully provided the management support to the largest work package of the H2020-ASTERICS project, while ensuring dissemination, user engagement and communication to the collaborating other H2020 projects, consortia, as well as industries. The ASTERICS-OBELICS thematic training events addressed the needs of the astronomy and astrophysics community with maximum possible participants in each of the editions.

Along with the developments of new data formats, D-GEX members have performed various tests for data format benchmarking as well as low power computing platforms. D-INT and D-ANA activities have provided the astronomy and astroparticle physics community with a repository of scientific services and software with documentation. This open source OBELICS repository serves as a one-stop web address for astronomers and general public to learn about OBELICS developments. This repository is one of the major contributions of OBELICS to the astronomy and astrophysics community and it will be merged with the EOSC catalogue in the near future.