ASTERICS - H2020 - 653477

Software Libraries

ASTERICS GA DELIVERABLE: D3.20

Document identifier:	ASTERICS-D3.20
Date:	**10 July 2019**
Work package:	**WP3**
Lead partner:	**UCAM**
Document status:	**Final**
Dissemination level:	**Public**
Document link:	www.asterics2020.eu/documents/ASTERICS-D3.20.pdf

Abstract

The Software Libraries, resulting from the activities in the OBELICS work package, are publically available through the ASTERICS Software Repository. The ultimate goal was to produce an open source catalogue of services and software produced by Astronomy ESFRI projects participating in the OBELICS work package. At the end of the ASTERICS project there are over 40 software products listed on the OBELICS repository. The repository provides on each product a short introduction and links to the software and documentation.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/WP	Date
From	Jayesh Wagh	CNRS-LAPP / WP3	28/06/2019
Author(s)	Jayesh Wagh	CNRS-LAPP / WP3	28/06/2019
Reviewed by	Rob van der Meer	ASTRON / WP1	01/07/2019
Approved by	AMST		10/07/2019

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	28/06/2019	first draft	CNRS-LAPP / WP3
2	10/07/2019	review comments included + layout	CNRS-LAPP / WP3

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. TERMINOLOGY

ASTERICS	Astronomy ESFRI & Research Infrastructure Cluster
CTA	Cherenkov Telescope Array
ELT	Extremely Large Telescope
ESFRI	European Strategic Forum for Research Infrastructures
KM3NeT	Cubic Kilometre Neutrino Telescope
OBELICS	Observatory E-environments LInked by common ChallengeS (ASTERICS WP3)
SKA	Square Kilometre Array

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

VI. PROJECT SUMMARY

ASTERICS (Astronomy ESFRI & Research Infrastructure Cluster) aims to address the crosscutting synergies and common challenges shared by the various Astronomy ESFRI facilities (SKA, CTA, KM3NeT & ELT). It brings together for the first time, the astronomy, astrophysics and particle astrophysics communities, in addition to other related research infrastructures. The major objectives of ASTERICS are to support and accelerate the implementation of the ESFRI telescopes, to enhance their performance beyond the current

state-of-the-art, and to see them interoperate as an integrated, multi-wavelength and multi-messenger facility. An important focal point is the management, processing and scientific exploitation of the huge datasets the ESFRI facilities will generate. ASTERICS will seek solutions to these problems outside of the traditional channels by directly engaging and collaborating with industry and specialised SMEs. The various ESFRI pathfinders and precursors will present the perfect proving ground for new methodologies and prototype systems. In addition, ASTERICS will enable astronomers from across the member states to have broad access to the reduced data products of the ESFRI telescopes via a seamless interface to the Virtual Observatory framework. This will massively increase the scientific impact of the telescopes, and greatly encourage use (and re-use) of the data in new and novel ways, typically not foreseen in the original proposals. By demonstrating cross-facility synchronicity, and by harmonising various policy aspects, ASTERICS will realise a distributed and interoperable approach that ushers in a new multi-messenger era for astronomy. Through an active dissemination programme, including direct engagement with all relevant stakeholders, and via the development of citizen scientist mass participation experiments, ASTERICS has the ambition to be a flagship for the scientific, industrial and societal impact ESFRI projects can deliver.

VII. EXECUTIVE SUMMARY

The objective of the OBELICS work package was to enable interoperability and software re-use for the data generation, integration and analysis of the ASTERICS ESFRI and pathfinder facilities. The ultimate goal was to make these products available as open source in a catalogue of services and software, the ASTERICS Software Repository. Such a repository did not exist before the ASTERICS project.

This document reports on the status of the software libraries at the end of the ASTERICS project and is a follow-up of D3.4 of Month 12. The Software library was then still hosted on the wiki pages of the ASTERICS project. The repository now has a better representation on a separate server. This repository web interface provides download links and all the necessary documentation on the software developed within the work-package. The repository content was widely disseminated through ASTERICS-OBELICS workshops and social media to reach out the concerned research communities.

The Software library is a central repository of the results in the OBELICS work package. It is set up as a reference site for open source multi-messenger analysis software. CNRS-LAPP will continue to maintain the repository for future use and updates, so the astronomy and astroparticle physics community will be served with the latest software tools available.

Table of Contents

I. COPYRIGHT NOTICE 1
II. DELIVERY SLIP 1
III. DOCUMENT LOG 1
IV. APPLICATON AREA 2
V. TERMINOLOGY 2
VI. PROJECT SUMMARY 2
VII. EXECUTIVE SUMMARY 3
Table of Contents 4
1. Introduction 5
2. Plan 5
3. Content 5
4. Repository Services v.s. Software 6
5. Deviations from the plan 7
6. Results 7
7. Next steps 7
8. Conclusion 7
ANNEX 1 – CORELib - 2019 Release 8
ANNEX 2 – BaSC 12

1. Introduction

The objective of the OBELICS work package was to enable interoperability and software re-use for the data generation, integration and analysis of the ASTERICS ESFRI and pathfinder facilities. The ultimate goal was to make these products available as open source in a catalogue of services and software, the ASTERICS Software Repository. Such a repository did not exist before the ASTERICS project.

This document reports on the status of the software libraries at the end of the ASTERICS project and is a follow-up of D3.4 of Month 12. The Software library was then still hosted on the wiki pages of the ASTERICS project. The repository now has a better representation on a separate server. This repository web interface provides download links and all the necessary documentation on the software developed within the work-package. The repository content was widely disseminated through ASTERICS-OBELICS workshops and social media to reach out the concerned research communities.

2. Plan

The OBELICS work package facilitated the creation of an open innovation environment for establishing open standards and software libraries for multi-wavelength/multi-messenger data. Furthermore, we concentrated on development of common solutions for streaming data processing and extremely large databases. Study of advanced analysis algorithms and software frameworks for data processing and quality control was the third focus area.

The development of all products was done at the OBELICS partners connected to the various ESFRI facilities in the ASTERICS project. The OBELICS meetings and workshops facilitated the collaboration and the exchange of information between the partners.

The results will be publically shared with all researchers through a central repository.

3. Content

For most of the software products we have provided all the necessary information on the repository pages. As an example, we use the following two developments:

- http://repository.asterics2020.eu/content/corelib-2019-release
  The information on the page shows on 01 July 2019:
  - Developer: INFN
  - Licence: GPLV3

  - Page hits: 47
  - link to download of latest release
  - link to the GitHub Repository
  - Detailed information provided on the page is shown in Annex 1.

- http://repository.asterics2020.eu/content/basc-%E2%80%93-bayesian-source-characterisation
  The information on the page shows on 01 July 2019:
  - Developer: UCAM
  - Page hits: 33
  - link to download of latest release
  - link to the GitHub Repository
  - Detailed information provided on the page is shown in Annex 2.

All the information available on the repository has been cleared for release by the ESFRI project representatives in OBELICS.

4. Repository Services v.s. Software

The OBELICS repository is split over two instances:

- Services: http://repository.asterics2020.eu/services
- Software: http://repository.asterics2020.eu/software

The repository of services concerns data integration software, whereas software libraries and repository concerns data analysis software. This is mentioned on the repository site pages to explain what is what and why are they used? What is the difference?

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.

The ASTERICS/OBELICS/D-ANA task is developing software libraries for statistically robust analysis of PetaByte-scale datasets in astronomy. The primary outputs of this task are these software libraries, which are all released as open source software. The primary purpose of this page is to act as the canonical and long-term repository for released versions of these libraries so that they remain permanently available to the public.

The term repository and library are basically referring to the same catalogue of software.

5. Deviations from the plan

There were no deviations from the plan in the Grant Agreement.

6. Results

At the end of the ASTERICS project there are over 40 software products listed on the OBELICS repository. The repository provides on each product a short introduction and links to the software and documentation.

7. Next steps

The repositories will be available in the future and will be maintained by the CNRS-LAPP. Software developed within OBELICS shall be merged with the ESCAPE-824064 project repository with clear acknowledgment of H2020-ASTERICS. In addition, updates on information and links to new releases provided by the designers, even outside the H2020 projects, will be maintained.

8. Conclusion

The Software library is a central repository of the results in the OBELICS work package. It is set up as a reference site for open source multi-messenger analysis software. CNRS-LAPP will continue to maintain the repository for future use and updates, so the astronomy and astroparticle physics community will be served with the latest software tools available.

ANNEX 1 – CORELib - 2019 Release

<Copy on -1 July 2019 of the information on the software repository page of CORELib>

CORELib (Cosmic Ray Event Library) is a collection of simulated events of cosmic-ray showers. The production is currently based on the CORSIKA software 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.

Especially in the high-energy region of the spectrum, simulation is very demanding in terms of CPU time spent per each event. A reference set of simulations such as CORELib can be used "as-is" in several cases, hence saving computing resources (wall time, CPU time and energy) and providing a common playground for reconstruction algorithms and detector performance comparison; when custom running parameters are needed, it can serve as a benchmark to save debugging and fine tuning.

CORELib consists of a "pilot production" (approximately 0.6 TB) in which only proton-induced showers are simulated; energy spectrum has a power law with spectral index equal to -2. The "full scale production", instead, consists of several kinds of primary cosmic rays: protons and Heavy Nuclei (He, C, N, O, Fe) induced showers are simulated (see Table 1). For proton-induced showers, two productions are available: with and without Cherenkov radiation (see Table 2). In addition, to increase the statistics a flat energy spectrum (only for proton-induced showers) is evaluated (see Table 2).

Heavy Nuclei (He-CNO-Fe) induced showers						
High Energy Model	Low Energy Model	Option		Cherenkov Radiation		STATUS
		TAULEP	CHARM	with	without	α = - 2
QGSJET01	GHEISHA		x		x	DONE
QGSJET01	GHEISHA	x			x	DONE
QGSJET -II	GHEISHA	x			x	DONE
EPOS	GHEISHA	x			x	DONE

Proton induced showers							
High Energy Model	Low Energy Model	Option		Cherenkov Radiation		STATUS	
		TAULEP	CHARM	with	without	α = - 2	α = 0
QGSJET01	GHEISHA		x	x	x	DONE	DONE
QGSJET01	GHEISHA	x		x	x	DONE	DONE
QGSJET -II	GHEISHA	x		x	x	DONE	DONE
EPOS	GHEISHA	x		x	x	DONE	70% DONE

Table 1 (on the left): Summary of the heavy nuclei induced showers produced. Actually, only the production without contribution of Cherenkov radiation is completed.
Table 2 (on the right): Summary of the proton induced showers produced. The production with α = -2 is completed while the flat spectrum production is ongoing.

In order to simplify the access to the library, the information about secondary cosmic rays are extracted from CORSIKA output files and put in separated ASCII files (EM, Hadrons+Tau, Muons, Neutrinos). Both productions are stored at CNAF (see below).

The CORELib production is now using more than 1000 CPU cores, allocated to the KM3NeT Collaboration, but its technical specifications exceed the needs of KM3NeT, in the spirit of serving the community of astroparticle experiments and potentially also other fields of applications. The angular range extends from 0° (vertical) to 89°. Energy bins are populated as shown in Table 3:

Energy range (GeV)	**Number of events**
200-100	10000000
1E3-1E4	10000008
1E4-1E5	1000002
1E5-1E6	100000
1E6-1E7	10000
1E7-1E8	1000
1E8-1E9	100

Table 3: Number of simulated events for each energy bin.

Several high-energy interaction models are available (QGSJET01, QGSJETII-04, EPOS LHC) in combination with GHEISHA for the low-energy interaction model, and with TAULEP/CHARM options.

The first productions (about 0.6 TB) are stored and available via SFTP in a local server hosted at the University of Salerno: **corelib@193.205.188.227 (password: Asterics2020)**

The whole production completed so far (April 2019, about 30 TB) is stored at CNAF, the Information Technology National Centre of INFN (Italian Institute for Nuclear Physics). It can be downloaded through gridFTP with .X509 certificate using the endpoint:

gsiftp://gridftp-plain-virgo.cr.cnaf.infn.it:2811/storage/gpfs_data/corelib/

prior to the administrator authorization (lucia.morganti@cnaf.infn.it).

Beside the CNAF repository, the KM3NeT Virtual Organization members can access to the whole production via GRID using the storage endpoints:

srm://recas-km3netse01.na.infn.it:8446/srm/managerv2?SFN=/dpm/na.infn.it/home/km3net.org/user/bspisso/corsika-75000/

srm://recas-km3netse01.na.infn.it:8446/srm/managerv2?SFN=/dpm/na.infn.it/home/km3net.org/user/smstellacci/corsika-75000/

Progress update

CORELib is a library of simulated cosmic ray events. The physics of cosmic ray-induced particle showers is a source of background for many experiments seeking rare phenomena and high-energy particles from astrophysical and cosmological origin. Conversely, secondary cosmic rays such as high-energy muons provide the tool for investigation in application fields such as muography (i.e. using muons to image the interior of volcanic edifices and faults, buildings that are part of the cultural and historical heritage, nuclear waste depots and reactors, etc.). Simulations of primary cosmic ray interactions in the atmosphere and secondary fluxes are expensive in terms of setup, computation time, electrical power and data storage. CORELib can be used as a ready-made source of data as well as a reference benchmark to other simulations.

A first production was made with spectral index -2. Simulation of Cherenkov radiation for proton-induced showers was completed in this period. In parallel, data production was completed including heavy nuclei. The spectral index of -2 is relatively close to the real dependency of the flux on energy, but leaves small statistics at very high energy. In view of the usage of CORELib products to develop algorithms and train machine-learning models to recognise rare high-energy events, another data production with a flat logarithmic spectrum for proton primaries was set up.

CORELib extends to from 0° to the inclination of 89°, which makes it suitable for different applications from underwater neutrino telescopes to air showers and muography of large objects above the sea level. The efforts required included designing the simulation jobs for the CORSIKA 7.5000 framework, managing a large production over thousands of computation nodes on the GRID, checking data quality and writing code to cast the output in a shape that enhances and simplifies data access and usage, in particular splitting the particles produced in different categories.

In the next future, simulation of Cherenkov radiation is planned also for heavy nuclei-induced events. Contacts are ongoing to involve a larger community from several ESFRIs in joint efforts. Not only this version of CORELib will stay as a reference production, but it is also promoting cross-fertilization and cooperation among researchers with different interests and cultural background.

Motivation for the contribution

We wanted to provide a reliable set of simulated cosmic ray events for users that cannot or do not want to afford the time and cost for a large scale simulation. We did it by identifying a suitable version of CORSIKA as generator and paying a large but sustainable effort in computing resources.

Highlights of success and how it will benefit the astronomy ESFRI project/projects overall

Unlike usual simulations, the library of events produced features a particularly hard spectrum, extending to very high energies. This allows high statistics for very rare events, which are the most interesting for some classes of problems (e.g. diffuse ultra-high energy neutrino flux).

Changes with respect to initial plan for software development

We found it was useful to simulate a flat log-E spectrum, whereas the first production had spectral index -2, resulting in fewer high-energy events.

Lessons learnt during the development

In addition, the lower end of the energy spectrum needs more statistics. It would be useful to simulate several kinds of atmosphere and different heights of observation levels.

Previous version

CORELib - 2018 Release (http://repository.asterics2020.eu/content/corelib)

ANNEX 2 – BaSC

<Copy on -1 July 2019 of the information on the software repository page of BaSC>

BaSC is an advanced Bayesian source finding library that uses a likelihood model that is mathematically proven in terms of visibilities but which can operate using only a dirty map; thus gaining accuracy without large computational expense.

Introduction

Source finding that uses CLEANed maps is subject to the loss of information necessarily caused by that algorithm. In fact, no matter the quality of such algorithms, the best that they can hope to recover is the CLEAN model – not the actual sources. BaSC works on dirty maps, which still have the messier PSF that is produced by interferometers but do not suffer a loss of information. We apply an MCMC method alongside a likelihood function on the map that is proven equivalent to a (much more expensive) function for the visibilities. The result is more accurate source positions and fluxes, better ability to discriminate sources near the level of instrument resolution, and a more transparent provenance for the source list produced.

Installation and use

BaSC is provided as a Python 3 library. Download from the Git repository above and, for the case of Linux type 'make' or for Mac type 'make mac'. Windows is not natively supported at this time. A user manual is included in the repository.

Future Development

This code will continue development at Cambridge once ASTERICS is concluded. Focus will be on included more functionality within BaSC – at the moment generation of the dirty maps has to be done externally with programs such as CASA, but there is an advantage of being able to process the visibilities directly within BaSC before applying the source finder.

Optimisation of the MCMC process is also being looked at; at present, the source finder can take a long time to run in the case of many (>30) sources due to degeneracies in the parameter space. We aim to improve this. Wide field images can cause issues in BaSC due to the change in the PSF across the field. There are a number of ways to address this which are being investigated

Progress update

This time period was concerned with the development and testing of BaSC. The software is now in a useable form, publicly available on github. We have produced a paper (Hague et al. 2019) that outlines the results of the tests we have done, and confirms that the software is

superior at source discrimination tasks than the usual pathway for radio astronomy of CLEAN/SExtractor and approaches the mathematically optimal resolution limit.

Challenges at the beginning of the project

At the beginning of this project, the challenge was correctly packaging an older MCMC code with a python wrapper and a new likelihood function. There was also the challenge of clustering the output.

Innovative solutions produced over last 12 months

We were able to calculate the optimal performance for a constrained task (discriminating between two nearby points) and then construct appropriate testing sets. The design of the experiment was critical to confirming that BaSC did indeed work as expected. The main effort was in programming BaSC itself, and creating realistic test observations that permitted the experiment .Production of the paper and the meeting the referees requirements also took up time.

Future activities planned

More features for analysis of radio interferometry images will be included in the package

Contact

Haoyang Ye and Peter Hague