ASTERICS - H2020 - 653477

Repository of services

ASTERICS GA DELIVERABLE: D3.15

Document identifier:	ASTERICS-D3.15.docx
Date:	**11 September 2018**
Work Package:	**WP3 OBELICS**
Lead Partner:	**LAPP**
Document Status:	**Report**
Dissemination level:	**Public**
Document Link:	www.asterics2020.eu/documents/ ASTERICS-D3.15.pdf

Abstract

D3.15 Repository of Services are delivered in the form of series of software focusing on workflow management services. These software are released on OBELICS repository which is made publicly available on http://repository.asterics2020.eu/software. This report summarizes the highlights of these software developments.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/WP	Date
From	Jayesh Wagh	LAPP	23/07/2018
Author(s)	Jayesh Wagh	LAPP	23/07/2018
Reviewed by	Giuseppe Cimo'	ASTRON	05/09/2018
Approved by	Rob van der Meer		11/09/2018

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	23/07/2018	First Draft	Jayesh Wagh (LAPP)
2	11/09/2018	Final	Rob van der Meer (ASTRON)

IV. TERMINOLOGY

ASTERICS	Astronomy ESFRI & Research Infrastructure Cluster
A&A	Authentication and Authorization
ALMA	Atacama Large Millimeter/submillimeter Array
CADC	Canadian Astronomy Data Center
CANFAR	Consortium of Canadian university astronomers
CORELib	Cosmic Ray Event Library
CORSIKA	COsmic Ray SImulations for KAscade
CTA	Cherenkov Telescope Array
CWL	Common Workflow Language
D-ANA	Data ANAlysis /interpretation
DIRAC	Distributed Infrastructure with Remote Agent Control
DPPP	Default Pre-Processing Pipeline
E-ELT	European Extremely Large Telescope
EGI	European Grid Infrastructure
ESFRI	European Strategy Forum on Research Infrastructures
GAVO	German Astrophysical Virtual Observatory
GUI	Graphical User Interface
HERA	Hydrogen Epoch of Reionization Array
HPC	High Performance Computing
IVOA	International Virtual Observatory Alliance

KM3NeT	Cubic Kilometre Neutrino Telescope
LOFAR	The Low Frequency Array
LSST	The Large Synoptic Survey Telescope
ML	Machine Learning
OBELICS	Observatory E-environments LInked by common ChallengeS
ROAST	ROot extensions for ASTronomy
SAML	Security Assertion Markup Language
SKA	Square Kilometre Array
STOA	Script Tracking for Observational Astronomy
VO	Virtual Observatory
WMS	Workflow Management System

Table of Contents

1. Introduction

The ASTERICS/OBELICS/D-ANA task is developing software libraries for statistically robust analysis of PetaByte-scale datasets in astronomy, and testing and integrating some e-infrastructure tools (authentication & authorisation, plus workflow management systems) to ease the production of data processing pipelines.

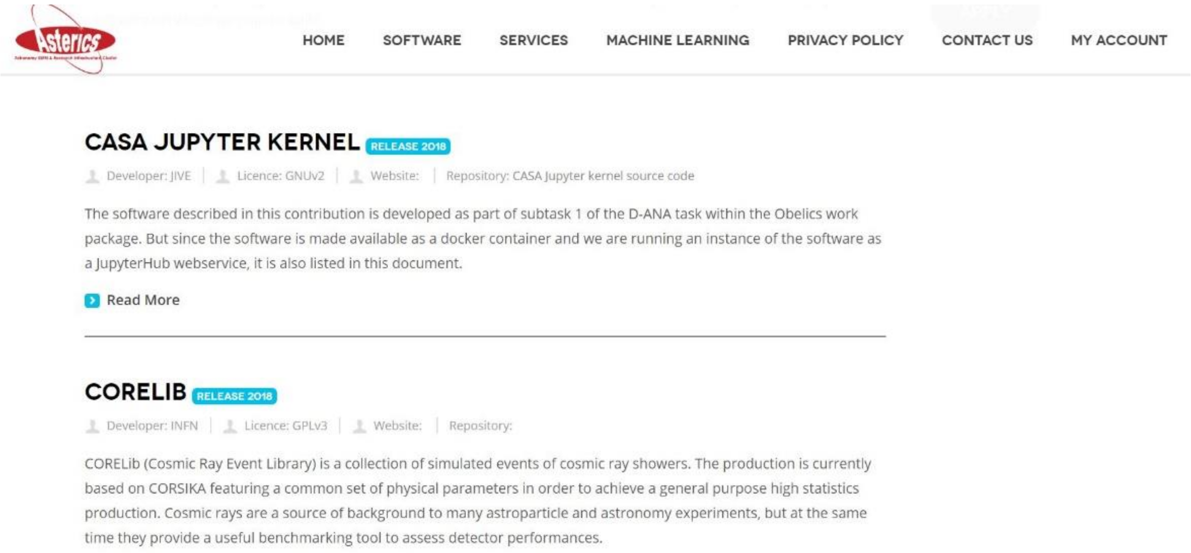
The primary outputs of this task are these software libraries which are all released as open source software on the OBELICS software repository. The primary purpose of OBELICS software repository is to act as the canonical and long-term repository for released versions of these libraries so that they remain permanently available to the public.

This repository is for public releases of the libraries, the ongoing software development work is managed separately. In this report we present and discuss highlights of these developments as well as their applications to astronomy ESFRI projects.

2. Weblink for OBELICS repository of services

OBELICS software repository is publicly available on following url

http://repository.asterics2020.eu/software



CASA JUPYTER KERNEL RELEASE 2018

Developer: JIVE | Licence: GNUv2 | Website: | Repository: CASA Jupyter kernel source code

The software described in this contribution is developed as part of subtask 1 of the D-ANA task within the Obelics work package. But since the software is made available as a docker container and we are running an instance of the software as a JupyterHub webservice, it is also listed in this document.

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.

Figure 1 Web interface for OBELICS software repository

3. List of software

The Data Analysis/Interpretation (D-ANA) task is steadily progressing towards its completion. The completion of this task is expected in April 2019. Under D3.15 Repository of services, following software developments focusing on workflow management services were carried out. The software are still preliminary and will be updated through to the final milestone in April 2019. In cases where the original software is yet to be released within D-ANA, we have currently posted the pre-existing open-source software on which we plan to build. In this section, we briefly present these software developments under OBELICS with their respective applications.

3.1 Yabi, a Workflow Management System and Workflow Engine for Data Reduction Pipeline

Yabi is a 3-tier application stack to provide users with an intuitive, easy to use, abstraction of compute and data environments. It provides users with an easy to use tool to perform custom data reduction on private and public data. If astronomy ESFRI project data centres adopt YABI, the end user will not have to worry about software installation, hardware configuration, or retrieving huge data from complex archive. The users only have to address scientific analysis. During the development, it was observed that the lack of a standard language for workflow management systems might cause a problem. We believe that common workflow language can address this issue. The developers also recommend system administrators to take into account interoperability while configuring their workflow environment.

3.2 Technologies for Authentication and Authorization in an Astronomical Data Centre

Software and technologies to manage A&A in an astronomical data centre must satisfy a number of requirements. In particular they must allow:

- access to private and public data according to data policy
- data sharing among users
- access to software and computational resources according to data policy
- data computing through workflow applications

INAF has developed a set of technologies to address these requirements. These technologies provide users and data centres with a robust Authentication and Authorization system. To access data centre services, the user has to memorize multiple passwords. There is also risk to forget these passwords. Thanks to the technologies developed by INAF for A&A, the user can now access all the data centre services with only one credential. These technologies also provides Data centre administrators with a single tool to manage all the authorizations of all users. It may seem that it is not worth to configure all data centre services with a common A&A system, since it is faster and easier to deploy each service with its own A&A system. However, this solution has a huge cost in maintenance, and a single A&A system for all the data centre services must be preferred.

3.3 INAF CTA - Workflow Management System API

The INAF CTA science gateway aims at providing a web instrument for high-energy astrophysics. It leverages open source technologies giving web access to a set of tools and software widely used by the CTA community. An extended (though not exhaustive) list of tools provided by this technology embrace XANADU software package, GammaLib & ctools, Fermi Science Tools, Aladin, IRAF. The gateway is based on the Liferay platform. It provides a Workflow Management System (WMS) with a customizable graphical web user interface and a web-desktop environment. The integrated WMS is based on WS-PGRADE/gUSE that seamlessly enables the execution of astronomical and physics workflows (and jobs) on major platforms such as DIRAC (Distributed Infrastructure with Remote Agent Control) INTERWARE, ARC, Globus, gLite, UNICORE(Uniform Interface to Computing Resources), PBS as well as web services or clouds.

3.4 INAF CTA - Authentication and Authorization Infrastructure API

The INAF CTA Authentication and Authorization Infrastructure provides functionalities enforcing the protection of CTA resources and digital assets by means of a role based authorization and by allowing both a federated authentication (based on eduGAIN inter-federation) and a centralized authorization managed at consortium level. The infrastructure offers a proper environment for enforcing accountability allowing maintenance and audit of logs for relevant events.

The current implementation of the INAF-CTA Authentication and Authorization Infrastructure provision more than 1000 consortium SAML identities and is releasing a persistent and non-reassignable ID as requested by CTA user requirements. Authentication of Observers and Guest Users can be achieved also using eduGAIN inter-federation. The authorization is performed through a dedicated Attribute Authority which grants the definition, management and provisioning of roles based on groups and subgroups. Web interfaces are provided for group administration and web services allows access to those functionalities in a service oriented architecture.

3.5 INAF Cloud Science Platform

The INAF Cloud Science Platform explores a possible technological solution for large projects, to implement a new integrated approach to data access, manipulation and sharing. In particular, in case of worldwide distributed collaborations that needs to share distributed infrastructures. It is a hybrid cloud built of three main blocks: the CANFAR e-infrastructure, the EGI Federated Cloud and a cloud gateway site INAF-OATs.

The INAF-OATs cloud site is a gateway between the two infrastructures: it is compliant with the EGI Federated Cloud and interoperable with the CANFAR e-infrastructure thanks to the IVOA standards implementation, so short-circuiting the interoperability. It offers an integrated set of astronomical services to Astronomy and Astrophysics community allowing data and applications to work in the same way regardless the infrastructure and to implement the ability of data, VMs and software to be easily moved and reused in the two cloud environments.

The interoperable Science Platform is possible thanks to the gateway site at INAF-OATs deployed exploiting and extending a set of software APIs, open source released by the Canadian Astronomy Data Center (CADC). This software is used to set up in Europe the set of IVOA standards based services interoperable with the CANFAR one.

Software services include storage, authentication, delegation, data access authorization. These services are provided as EGI Community Services and based on IVOA standards:

- the Access Control Service is the Authentication and Authorization service implementation managing users and groups.
- the VOSpace service is a distributed storage implementation;
- the Credential Delegation Service is the IVOA Credential Delegation Protocol implementation.

3.6 STOA - A Workflow Management System for Radio Astronomy

STOA (Script Tracking for Observational Astronomy) is developed by UCAM partners. It is a workflow management system with both command line and web interfaces that permits the efficient handling of large, heterogeneous data sets, and provides a fast run-test-rerun work cycle for these situations.

STOA emerged as a way to manage a series of analysis programs being executed across multiple targets in the ALMA archive. In the first instance, it was a command line utility that ran the same program in multiple directories and stored the results in a database. This has evolved into a more complex system with a web interface that permits remote operation and multiple user collaboration on a single project. Fine control over the execution of each instance is now possible, with a hierarchical system of default inputs minimising the amount of labour required to do this. The ability to tag results and add comments has been added to facilitate group work. The STOA web interface allows two way communication with SAMP enabled applications such as Topcat and DS9. Tasks in STOA are managed using the Common Workflow Language (CWL) standard, and the files produced can be utilised by any other program using this standard.

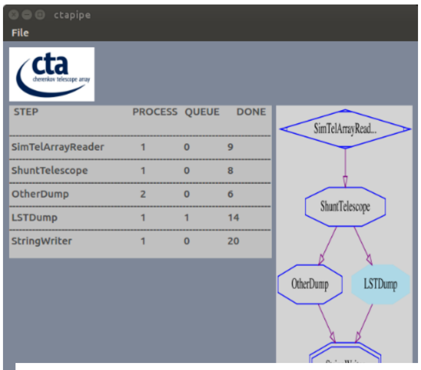
3.7 SWIFTCASA

SWIFTCASA developed by UCAM partners addresses the processing need of SKA and SKA precursor instrument "Hydrogen Epoch of Reionization Array" (HERA). Here there is a high-degree of task-based parallelism based on partitioning of the input data in time and frequency, which is, however difficult to expose in traditional software. SWIFTCASA enables such parallelization while maintaining easy use by astronomers through a simple and clear "scripting" language. SWIFTCASA is directly interoperable with CASA and with standard HPC environment such as SLURM scheduling, MPI libraries and LUSTRE filesystem.

The full development report is available on
https://www.mrao.cam.ac.uk/~bn204/publications/2017/2017-08-casaswift.pdf

3.8 Flow-based framework (flow)

ctapipe.flow is a Python implementation of the flow-based programming paradigm for ctapipe framework developed by CNRS-LAPP. The main objective behind this development was to accelerate data processing and maximize the use of available CPUs. Initial plan was to achieve this objective by parallelization of the algorithms to process data. This parallelization is presented in a schematic diagram in figure 2.



STEP	PROCESS	QUEUE	DONE
SimTelArrayReader	1	0	9
ShuntTelescope	1	0	8
OtherDump	2	0	6
LSTDump	1	1	14
StringWriter	1	0	20

Figure 2 : Pipeline run in ctapipe.flow. Each algorithm is running on a separate core simultaneously .OtherDump is running on 2 cores.

In flow-based programming, applications are defined as networks of black-box components that exchange data across predefined connections. These components can be reconnected to form different applications. ctapipe-flow executes ctapipe processing modules in a sequential or multiprocess environment and user implements steps in Python class. It is important to note that externals libraries must be selected carefully to obtain maximum feedback from users during such development to avoid any unforeseen delay.

Main highlight of this development is that processing a pipeline on CTA Monte Carlo data has become faster approximately by factor of 4 using the ctapipe.flow on a CPU with 4 cores compare to running it on a single core. We expected to develop a multicore and multi system parallelization system. During the course of development we decided to discard multisystem parallelization, as it does not suit to CTA requirements at the present.

3.9 A high performance computing data format generator

The new generation research experiments will introduce huge data surge to a continuously increasing data production by current experiments. Data to be efficiently accessed and processed by High Performance Computing (HPC) codes require the optimization of their model and format. Developing a HPC data format is a complex and challenging task that results in a time consuming process. Codes generators have been introduced to provide flexibility and to hopefully reduce the development time. Furthermore, they can provide larger set of functionalities, format-versioning and languages compatibility.

Well known existing data-format generators like Protocol Buffer, Avro or Thrift are able to perform versioning and serialization, but their use are complex even for computer scientists. CNRS-LAPP intended to provide data formats intuitively usable by physicists. They developed the first version of the data format generator especially for HPC purposes.

After the update of the CTA data requirement (based on our data format generator capabilities) we developed a second data format generator to achieve all the desired properties of the CTA data format. With this new version, we added features like streaming, compression, textual configuration for both C++ and Python (provided through wrapper). We also used a new development paradigm based on internal representation modification which is far better than inheritance and offer more flexibility (the more functionalities, the less code).

The generator language has been designed to be as simple as possible. It creates C + +, Python and wrapped Python data formats. The generated code is fully human-readable compared to other generators. Up-to-date documentation of the generated code is also released. This simplicity enables potentially this generator to be easily applicable in many research domains. Moreover, a fast Python interface can be generated on the user demand. This combines the HPC data format speed and the versatility of Python with a minimal wrapper impact on performances. Some programs and libraries like Swig of PyBind11 provide generated Python wrappers but they are not as optimized as the user needs.

3.10 Machine learning algorithms for transient signal classification for gravitational wave astronomy

Gravitational wave observations are limited by a background of transient signals from instrumental and environmental origin. This package includes a set of machine learning tools that allow to classify those transient signals, in order to better characterize their large population, give hints about their source, and provide new ways for mitigating this background. The algorithms take a labelled set of transients, extract features from the time series, and learn a classifier (neural network) using standard ML libraries. While this toolbox is intended for gravitational wave data specifically, it is general enough to be adapted to problems in other fields. This package is expected to be released in September 2018.

3.11 Default Pre-Processing Pipeline

The LOFAR telescope relies on calibrating the instrument accurately, while dealing with large data volumes. Because of the large data volume, it is beneficial to perform the calibration in a streaming fashion, avoiding writing data to disk.

The LOFAR Default Pre-Processing Pipeline (DPPP) software is developed by ASTRON partners. It reads and writes radio-interferometric data in the form of Measurement Sets. It goes through visibilities in time order and contains standard operations like averaging, phase-shifting and flagging bad stations. Between the steps in a pipeline, the data is not written to disk, making this tool suitable for operations where I/O dominates. As part of the ASTERICS/OBELICS project, gain calibration was implemented as a part of DPPP, using the StefCal algorithm. Furthermore, a move towards open standards for calibration tables was implemented, by adopting the HDF5 file format.

DPPP now contains a streaming calibration step, which will reduce time spent on calibration. This will make it possible to spend more time on other steps in the calibration scheme. We expanded the calibration scheme to also incorporate direction dependent calibration (needed because the field of view of LOFAR can be so large that the instrument effectively looks through different patches of ionosphere). Furthermore, we implemented a form of constrained calibration, which makes it possible to reduce the amount of free parameters. The HDF5 standard is well suited for storing calibration solutions; designing a flexible constraint framework allows for easy expansion of the effects that can be calibrated for.

3.12 CASA Jupyter kernel

CASA Jupyter kernel is developed by JIVE. The software is made available as a docker container and we are running an instance of the software as a JupyterHub webservice.

The size of astronomical datasets has increased dramatically over the years; terabyte sized datasets are no longer an exception. This trend will only accelerate; the SKA is expected to produce nearly 1 TB of archived data each day. This means that it will no longer be feasible for astronomers to download these huge datasets and perform the data reduction on their own machines, as is currently the practice. Instead the data reduction is likely to be done close to where the data is archived in central data processing centres, with the astronomer operating remotely on the data.

One way of facilitating this is through Jupyter notebooks. Jupyter is a web-based application which allows users to create interactive notebooks which can include annotated text and graphics as well as executable code. Currently Jupyter supports more than 40 different programming languages, including Python, R, and Matlab. Jupyter is designed be extended and makes it easy to add additional languages.

We have created a Jupyter kernel for CASA, a widely-used software package for processing astronomical data. The kernel allows all CASA tasks to be run from inside a Jupyter notebook, albeit non-interactively. Tasks which normally spawn a GUI window are wrapped so that their output is saved to an image instead, which is then displayed inside the notebook.

The notebook format also has the great advantage that all steps of the data reduction are preserved inside the notebook. This means that the whole data reduction process is self-documenting and fully repeatable. It also allows users to very easily make changes to their pipeline and then rerun the pipeline steps affected.

3.13 CORELIB

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. This library is being developed by INFN partners.

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

Unlike usual simulations, the library of events produced 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).

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. Also 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.

3.14 ROAst

The ROOT analysis framework is one of the most used software for the analysis and indeed it is the “de facto” standard for high-energy physics. The goal of the ROAst library (ROot extensions for ASTronomy) is to extend the ROOT capabilities adding packages and tools for astrophysical research. This library is being developed by INFN partners. We identified a set of missing features in ROOT, for which we have written original code.

ROAst comes with three feature sets:

- access to astronomical catalogues;
- coordinate conversion tools;
- high-precision Moon and Sun position models relative to the Earth;
- graphical tools to produce commonly used plots (general and partial skymaps).

ROAst provides seamless access to the following catalogues: UCAC4, URAT1 (local), the VizieR online catalogue repository, VO (Virtual Observatory) catalogues (such as Euro-VO, MAST, GAVO, VAO, IVOA). An intermediate abstraction layer makes the addition of more catalogues easy to implement. Catalogue querying can be done extracting regions of various shapes (rectangles, circles, ellipses) around each object. ROAst supports equatorial, galactic, ecliptic, horizontal astronomical coordinates (using Lat-Long and UTM as geographical coordinates) and allows coordinate transformations. Plots can be obtained in “flat” and “Aitoff” projection, in equatorial, galactic and horizontal coordinate systems (some combinations of projection and coordinate systems are not allowed).

The main highlight of this development is that accessing astronomical catalogues and transforming coordinates from software born in the context of high-energy physics is not trivial. ROAst provides a convenient and easy way to reuse in the high-energy physics community tools from the astrophysics community.

4. Intended Audience

The intended audience for this repository of services are scientists working data integration activities in astronomy, astroparticle physics field as well members of the public who may have a need or interest in such services, for commercial or non-commercial applications.

5. Citing request

All of the developments mentioned in this deliverable are publicly available. We would request external users to acknowledge H2020-ASTERICS and/or include a link to http://repository.asterics2020.eu/software in the documentation and/or any derived publications so that the readers are able to locate the original software easily. This will also help us gauge the public interest and use of these services. The current url is the permanent url of the repository.