Ref. Ares(2018)3354767 - 25/06/2018

Astronomy ESFRI & Research Infrastructure Cluster

ASTERICS - H2020 - 653477

Second ASTERICS-OBELICS workshop

ASTERICS GA DELIVERABLE: D3.11

Document identifier:	ASTERICS-D3.11.docx
Date:	**21 June 2018**
Work Package:	**WP3 OBELICS**
Lead Partner:	**LAPP**
Document Status:	**Final**
Dissemination level:	**Public**
Document Link:	www.asterics2020.eu/documents/ASTERICS-D3.11.pdf

Abstract

This report provides a global overview of the presentations and discussions that took place at the second ASTERICS – OBELICS workshop. This workshop was dedicated to Astronomy & Astroparticle Physics assets in building the European Open Science Cloud and took place in Barcelona from 16-18 October 2017. The workshop served as a platform for H2020-ASTERICS to present its activities under WP3-OBELICS and to bring together the ESFRI representatives and other stakeholders to discuss potential contributions from the Astronomy community to building a European Open Science Cloud. The workshop also provided an opportunity to demonstrate new technologies such as machine learning and deep learning with some use cases in astronomy.

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-Non-commercial 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
From	Giovanni Lamanna	LAPP	
Author(s)	Jayesh Wagh	LAPP	08/01/2018
Reviewed by	Rob van der Meer	ASTRON	30/04/2018
Approved by	AMST		21/06/2018

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	08/01/2018	First Draft	Jayesh Wagh, LAPP
2	21/06/2018	final version	Jayesh Wagh, LAPP

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
A & A	Authorization and Authentication
CTA	Cherenkov Telescope Array

CWL	Common Workflow Language
EOSC	European Open Science Cloud
ESDC	European Science Data Centre
ESFRI	European Strategy Forum on Research Infrastructures
EST	European Solar Telescope
FAIR	data principle: Findable, Accessible, Interoperable and Reusable
HEP	High Energy Physics
HLEG	High Level Expert Group
IVOA	International Virtual Observatory Alliance
KM3NeT	Cubic Kilometre Neutrino Telescope
LHC (HL-LHC)	Large Hadron Collider (High Luminosity LHC)
LIGO	Laser Interferometer Gravitational-Wave Observatory
LOFAR	The Low Frequency Array
LSST	The Large Synoptic Survey Telescope
OBELICS	Observatory E-environments LInked by common ChallengeS
SKA	The Square Kilometre Array
SME	Small to Medium-sized Enterprise
VO	Virtual Observatory
WLCG	World Wide LHC Computing Grid
WP	Work Package

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

VI. PROJECT SUMMARY

ASTERICS (Astronomy ESFRI & Research Infrastructure Cluster) aims to address the cross-cutting 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 second ASTERICS-OBELICS workshop focused on building bridges between ESFRI projects, concerned scientific communities, e-infrastructures, and further consortia, and to call them to action for the implementation of the European Open Science Cloud (EOSC) for data interoperability in the astrophysics and astroparticle physics related disciplines. The discussions on this subject delivered a list of recommendations on behalf of H2020-ASTERICS. In addition to the EOSC, the workshop also provided an opportunity for OBELICS members to present their achievements to external participants. The workshop invited guest speakers from machine learning and artificial intelligence to discuss some of the use cases in astronomy. The first day featured a live webcast of the official announcement of Gravitational wave detection from collision of neutron stars, a milestone achievement for multi-messenger astronomy.

Table of Contents

1. Introduction

The H2020-ASTERICS WP3 OBservatory E-environments LInked by common ChallengeS (OBELICS) aims at enabling interoperability and software reuse, software libraries for multi-messenger data and developing common solutions, sharing prototypes and best practices. Thanks to the contribution from key representatives of various ESFRI and world class projects, in the first two years of the project WP3-OBELICS has developed several solutions for data generation and information extraction, data systems integration and data analysis and interpretation. It is important for H2020-ASTERICS and WP3-OBELICS to communicate these achievements to the general public and the astronomy community and have their inputs or feedback to further improve these solutions.

The European Open Science Cloud (EOSC), an “e-initiative” from the European Commission, aims to accelerate and support the current transition to more effective Open Science and Open Innovation in the Digital Single Market. It should enable trusted access to services, systems and the re-use of shared scientific data across disciplinary, social and geographical borders. Despite the success of the European Strategy Forum on Research Infrastructures (ESFRI), fragmentation across domains still produces repetitive and isolated solutions. There was a need to bring together key ESFRI representatives from Astronomy and astroparticle physics domain to align their interests and efforts towards the implementation of EOSC.

The next generation of telescopes listed in the ESFRIs roadmap and other world-class projects will be facing a deluge of data in the near future. Traditional methods to model and analyse large volume of complex datasets have to be replaced with more advanced and powerful techniques such as machine learning. Machine learning is yet to be adopted by the astronomy community and hence it is important to learn more about these new techniques and understand their advantages over traditional methods.

Keeping these three well identified priorities (OBELICS achievements, EOSC and machine learning) in mind, WP3-OBELICS organized the second ASTERICS-OBELICS workshop in Barcelona from 16-18 October 2017. This workshop served a threefold purpose:

- Disseminate WP3-OBELICS achievements and ongoing activities with the members of astronomy and astroparticle physics community and provide a platform for knowledge exchange.
- Bring together the ESFRI representatives and other stakeholders from Astronomy and astroparticle physics community to address European Open Science Cloud through collaborative actions.
- Discuss and demonstrate the applications of machine learning to analyse big data in astronomy.

2. Structure of the meeting

The workshop was an opportunity to bring together H2020-ASTERICS members, external participants from concerned scientific communities, e-infrastructures, and other consortia.
The first day was dedicated to present WP3-OBELICS activities to external participants to learn more about H2020-ASTERICS and in particular WP3-OBELICS, software libraries and other common solutions developed by WP3 members. Overall, 16 presentations (15 minutes each) gave a complete overview of OBELICS activities in data generation and information extraction, data systems integration and data analysis and interpretation. A summary of first day presentations is attached in appendix 1.

Figure 1: Giovanni Lamanna, CNRS-LAPP, presenting an overview of WP3-OBELICS.

On this day, we also organized a live webcast of the press conference announcing the ground-breaking detection of gravitational waves from a collision of neutron stars. LIGO, Virgo and about 70 astronomical observatories around the world were involved in this discovery. This was also a very proud moment for H2020-ASTERICS as some of the member institutes were directly involved in this discovery. This discovery was also a triumph for multi-messenger astronomy, attesting importance of work carried out by H2020-ASTERICS to bring together astronomy ESFRIs and address common challenges.

Figure 2: Live webcast of LIGO-VIRGO press conference, 16 October 2017.

The second day of the workshop was dedicated to understand the European Open Science Cloud from the perspectives of the EOSC High Level Expert Group (EOSC HLEG), EOSCpilot, EOSC Science Demonstrators, ESFRI projects, Virtual Observatory and other H2020 projects and consortia. A series of 10 presentations followed by a panel discussion addressed the European Open Science Cloud and how H2020-ASTERICS, through its ties with ESFRI projects and other research infrastructures, could contribute to the implementation of the EOSC. A summary of the discussions is attached as appendix 2.

EOSC Roadmap

Governance – (3-layered structure, mandates, working procedures)

- Member States and EC (strategic)
- Executive Board (operational)
- Stakeholders Forum (advisory)

Architecture – launching EOSC stage 1 by 2020

- Supporting EOSC components (geographical & thematic)
- Core data & services through an EOSC Hub
- Advanced services by certified actors
- Catalogues of data & Catalogues of services
- Rules of participation
- EOSC Portal as a universal entry point
- FAIR data - FAIR Action Plan

Financing – stage 1 / stage 2

- H2020 – [illegible] in particular through WP2018-20
- Developing an EOSC Business Model for long term sustainability

Figure 3: Silvana Muscella, Chair of EOSC HLEG, presenting the EOSC roadmap.

On the third day, applications and use cases of machine learning were presented. Five presentations (45 minutes each) with applications of machine learning in CTA, VIRGO, KM3NeT and LSST were presented. These presentations were followed by a panel discussion to identify steps forward to improve machine learning usage in astronomy and astroparticle physics. A summary of the presentations and discussions is attached as appendix 3.

The detailed programme of the workshop can be found on the meeting website, https://indico.astron.nl/internalPage.py?pageId=9&confId=87 .

3. Participation to the meeting

The workshop brought together 64 scientists and technologists from ASTERICS partner institutions, representatives from EOSC-HLEG, EOSCpilot, EOSC Science demonstrators, H2020 projects and industries. In addition, experts from CERN as well as the LIGO-VIRGO collaboration were amongst the participants. The complete list of participants, speakers and the presentations is available on the workshop website https://indico.astron.nl/internalPage.py?pageId=9&confId=87

During the midterm review meeting in March 2017, reviewers had suggested to invite and include representatives from the European Solar Telescope (EST). From OBELICS funding we had invited four EST participants to this workshop. This participation from EST members had led to expression of interest from EST to collaborate with OBELICS work package on machine learning activities.

We observed a few last-minute cancellations from the list of registrants due to the political situation in Barcelona. However, we managed to arrange their talks by skype.

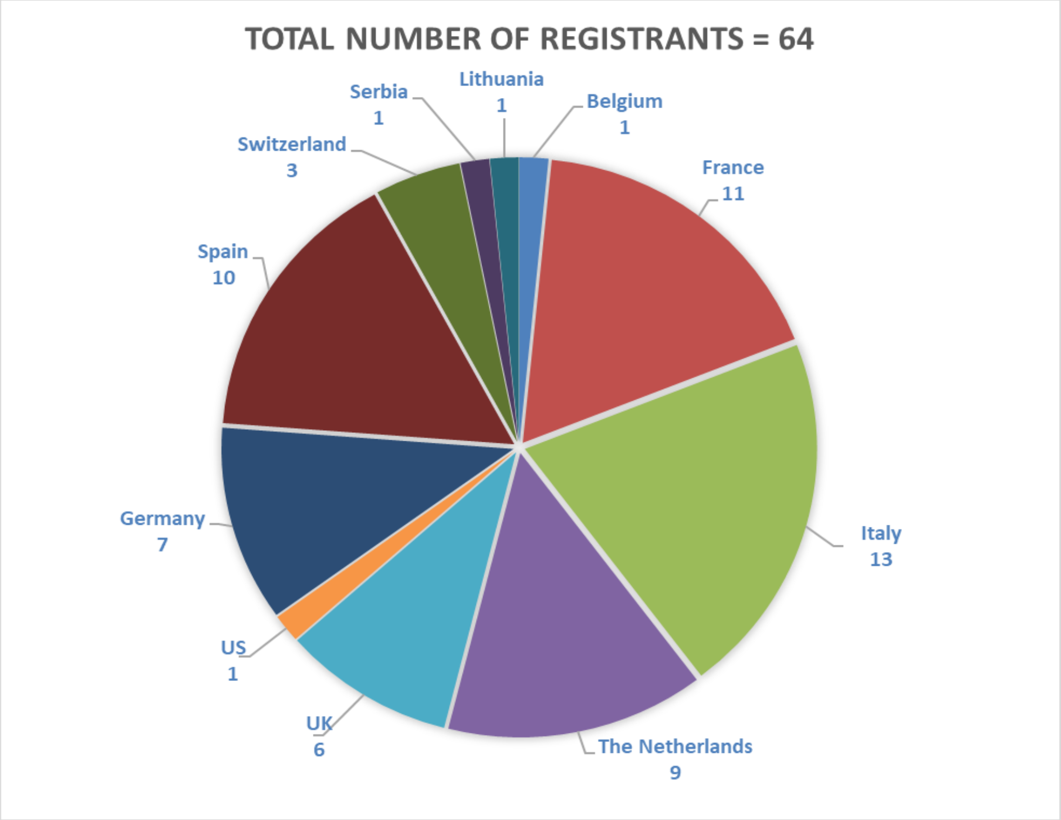


Figure 4. Workshop registrants per country.

4. Results of the meeting

Overall, the workshop was well received by the scientific community with active participation in discussions and talks. The event also produced significant impact on social media outreach of ASTERICS with 9500 tweet impressions.

In view of the next EOSC call INFRAEOSC-04, H2020-ASTERICS opened up a dialogue with representatives from ESFRI projects, other H2020 projects, HL-LHC, EOSCpilot projects and EOSC science demonstrators. H2020-ASTERICS is already developing common solutions for ESFRI projects that are facing big data challenges. SKA and CERN have signed an agreement for cooperation on big data earlier in July 2017. We believe through the workshop H2020-ASTERICS has successfully leveraged upon these synergies within ESFRI projects to address the European Open Science Cloud. The list of recommendations in appendix 2 can be considered as a starting point to develop a consortium bringing diverse expertise. This list of recommendations is one of the major outcomes of the workshop.

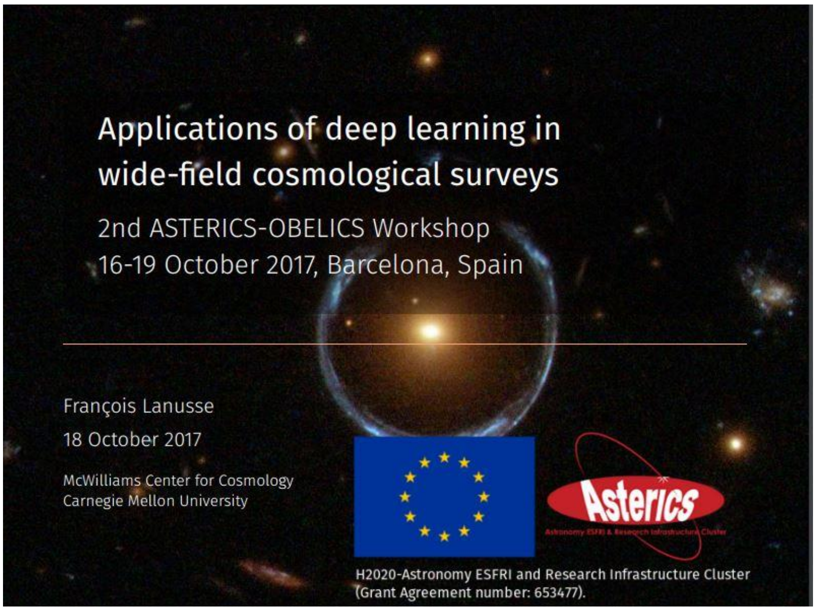
Dissemination of WP3-OBELICS activities and knowledge exchange was one of the results of the workshop. The workshop allowed WP3-OBELICS members to demonstrate the common solutions developed for astronomy ESFRI projects. The interoperability of these common solutions was discussed.

Authentication and Authorization activities presented during the workshop showed collaborative activities with links to the H2020-AENEAS, H2020 AARC2 and EGI-Engage EU projects. These links were the result of the discussions organized in the first edition of the ASTERICS-OBELICS Workshop in 2016. Some of the best practices and list of recommendations for python developers were presented on the first day. These best practices are not only well suited for CTA, but also for developers from other ESFRI projects.

WP3-OBELICS is addressing 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. Numerous activities carried out on data analysis and workflow architectures under WP3-OBELICS were presented.

The algorithms developed by ASTRON for “Fast convolutional resample” of data for the LOFAR radio telescope were presented. These algorithms are able to split the work among many processors in parallel architectures. This generated code is made publicly available. These algorithms can be applicable to data processing in other observatories where one needs to resample the data and apply corrections for the instrument.

Presentation on machine learning use cases helped participants to further understand applications in astronomy and astroparticle physics. These discussions are being followed up to develop a machine learning software repository under WP3-OBELICS. The presenters as well as panellists have confirmed their support to develop this repository. During the panel discussion we realized that it takes some time for the community to fully understand and adopt new computing models that are quite complex, leading to scepticism. There is a need to agree on some sort of validation of these models. This would be essential for expanding collaborations. Certain communities have already adopted machine learning and it is being used for different applications.



Applications of deep learning in wide-field cosmological surveys

2nd ASTERICS-OBELICS Workshop
16-19 October 2017, Barcelona, Spain

François Lanusse
18 October 2017

McWilliams Center for Cosmology
Carnegie Mellon University

H2020-Astronomy ESFRI and Research Infrastructure Cluster
(Grant Agreement number: 653477).

Figure 5: Presentation on applications of deep learning in cosmology by Francois Lanusse, LSST.

5. Evaluation of the meeting

We used different channels such as newsletters, email campaigns and social media to disseminate the event. We managed to have participation from key representatives from ESFRI projects as well as EOSC related projects. We were happy to attract over 60 registrants for the event despite of the political situation in Barcelona, which caused some last-minute cancellations. We had 2 representatives from industries and this is where we would like to improve upon in our next event in 2018.

We observed that the first two days of the meeting were overloaded with series of presentations and for the next meetings the number of presentations can be reduced with more time for question and answer sessions. For the third day of the workshop we had less presentations with more time for questions. This is one of the lessons learnt from the organization point of view.

As far as the discussions and quality of the presentation was concerned, we were quite happy to address EOSC from different perspectives and have a list of recommendations in a short duration. We are already in the process of following up with these recommendations with the H2020-ASTERICS executive board. Some of these recommendations were also presented at the EOSC stakeholder forum that took place in Brussels in November 2017.

Appendix 1.OBELICS day discussions

CNRS-LAPP is the lead institute for WP3-OBELICS and it is also responsible for the overall work package management, user engagement and dissemination of WP3 activities. In addition to the primary responsibility of internal coordination within the work package, OBELICS management has carried out various activities to achieve its objectives of user engagement and dissemination with continued support from H2020-ASTERICS members.

Some of the achievements that were presented included the first ASTERICS-OBELICS workshop addressing science data cloud & computing models in astronomy and astroparticle physics. This workshop was a success as it gathered over 87 scientists and technologists from industries, academia and other H2020 projects to discuss their experiences and best practices in Authorization and Authentication, Data Storage, Transfer & Preservation, Large Databases, and Workflow management & Interoperability. These interactions provided participants as well as the organizers some directions for future collaborations at individual level as well as project level. During this workshop a call for industrial collaboration was also launched. The OBELICS management is in process of formalizing collaboration with one of the applicants, Orobix Srl, an Italian company with expertise in the field of machine learning.

The first ASTERICS-OBELICS international school on advanced software programming for astrophysics and astroparticle physics was also a highlight for WP3-OBELICS management. This thematic training event brought together around 80 PhD students and senior researchers from astronomy and astroparticle physics domain to learn and upgrade their programming knowledge through series of very well-designed theory and hands-on sessions. The school was a success from a WP3 management point of view, as it helped them recognize the needs of astronomy community. The feedback from participants provided organizers with some directions and suggestions for the organization of the next thematic training event.

WP3-OBELICS has also collaborated with Trust-IT Services through industrial subcontracting to further strengthen its dissemination efforts to a wider community.

1. Data Generation and information extraction (D-GEX)

This session presented technical activities carried out by OBELICS partners to address the first stage of the scientific data flow, that is, the Data Generation and information extraction (D-GEX). Although these presentations were very specific for ESFRI experiments, each presentation discussed how the results obtained can be applicable for other ESFRI projects.

One of the examples are algorithms developed by ASTRON for “Fast convolutional resampling” of data for the LOFAR radio telescope. These algorithms are able to split the work among many processors in parallel architectures. This generated code is made publicly available. These algorithms can be applicable to data processing in other observatories where one needs to resample the data and apply corrections for instrument. IFAE and UCM have been investigating the use of common data formats for event based observatories. These data formats are the HDF5 for low level data and the CTA DL3 for high level data. High level formats for one observatory constitute the basis for integrating their results with those of other facilities. CNRS-LAPP presented its activities in the field of compression algorithms. This work on compression algorithms and optimized data arrangements were developed for the Cherenkov Telescope Array. INFN has been testing the implementation of the BeeGFS on a low power implementation.

2. Data Systems Integration (D-INT)

This session presented the efforts from OBELICS members to address the challenges in the data management phase of the large ESFRI infrastructures.

Extracting meaningful data products from large, heterogeneous sets of multiple radio observations can be labour intensive and difficult. Script Tracking for Observational Astronomy (STOA), a process management system that runs scripts on multiple sets of data, each time with different parameters and a different environment (containerisation) was demonstrated with a use case on ALMA. This work was carried out by the University of Cambridge. QServ is a database solution and its integration into science pipelines was presented by CNRS-LAPP. This solution is being developed for the Large Synoptic Survey Telescope (LSST), an instrument designed to make high precision images of the whole accessible sky in 4-D (x, y, z, t). QServ tests and its results on real science cases in LSST were discussed. QServ and other LSST software are open source and available on github. In the context of the Cherenkov Telescope Array (CTA), python wrapper performances were presented. Performance benchmark results obtained by comparing Numpy Python libraries and a mathematical Python library wrapped on optimized C/C++ code developed by CNRS-LAPP were presented. Along with benchmark performance, some of the best practices and recommendations that could be useful for python developers involved in other experiments were also discussed.

3. Data analysis and interpretation

WP3-OBELICS is addressing 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. This session highlighted numerous activities carried out on data analysis and workflow architectures.

INAF members presented Authorization and Authentication (A&A) activities carried out in the context of the Italian Astrophysical research infrastructure project (IA2) and SKA. An IA2 use case was presented in a step by step demo on the Remote Authentication Portal (RAP). A&A architecture developed for CTA was also highlighted with its links to the H2020-AARC2 project.

CORELib, a cosmic ray event library for open access has been developed by INFN members. Some of the advantages of this library are:

- Flexibility: several models used to provide simulations
- Plug-and-play: common data formats, immediate usage
- Open-access: SFTP with common user/password
- Extensibility: we provide the full set of parameters, so if other Collaborations or institutions want to add datasets, they can do with/without overlap

This library was presented in detail with its applications for the KM3NeT project as well as its potential application for CTA.

UCAM members have produced a source detection/characterisation code, building on an in house code at Cambridge (BayeSys). 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 and we have already got some promising results, applying it to ALMA data. This code is available on https://bitbucket.org/PeterHague/basc .

Astronomy is entering the big data era and facing new challenges such as data volumes and data rates. To address these challenges, high performance computing is being used. CNRS-LAPP members presented some of the high-performance computing software and algorithms such as PLIBS software library which is developed to address big data challenges of CTA.

CASA is a widely-used software package for processing astronomical data. A Jupyter kernel was created for CASA by JIVE members. The presentation highlighted integration of CASA with Jupyter for efficient remote processing. JIVE members mentioned that they provide both Docker and Singularity images for easy deployment. UCAM also produced a task-based parallelisation framework for Astronomy with focus on the CASA package. The architecture of a task-based parallelisation system for CASA built on top of the SWIFT/T framework was presented. This parallelisation system is now in use in Cambridge for imaging-based processing and calibration of data from the Hydrogen Epoch of Reionization Array (HERA) telescope, an official SKA precursor telescope.

CPPM partners presented the performance of a package for Likelihood-Based Fitting Methods describing probability density functions and C++ structure.

Appendix 2.EOSC day summary

1. Understanding EOSC from the experts

The second EOSC High Level Expert Group (EOSC HLEG) has been mandated to support the European Commission in the set-up of a data-driven infrastructure that builds on what already exists, that caters for the whole scientific community and brings together the building blocks developed through different categories of stakeholders to provide actionable options for the design of the EOSC governance and services. Some of the key publications and events organized by the EC to define the steps to implement EOSC include:

- EOSC Declaration, which is a key input for the EOSC roadmap.
- EOSC Stakeholders Forum November 2017, which aims at endorsing the EOSC declaration and discussing the EOSC roadmap.
- EOSC Roadmap, which will be announced in December 2017 providing information on the EOSC governance structure, architecture and core services.

The EOSC HLEG would like to consider introducing some EOSC in Practice use cases in the Interim Report (to be published by February 2018) which considers the following items discussed by members of the second expert group:

a) Recognition/Reputation, b) Trusted Services, c) Data Certification, d) GDPR compliant, e) Vouchers / Cloud Coins, f) Multiple member state collaboration, g) Cross disciplinary research.

The EOSC HLEG are inviting members from research infrastructures or e-Infrastructure communities who could produce a short use case paragraph, for the interim report, that describes these elements in practice.

2. Insights from other projects and research infrastructures

H2020 EOSC-hub mobilises providers from the EGI Federation, EUDAT CDI, INDIGO-DataCloud and 18 major research e-infrastructures offering services, software and data for advanced data-driven research and innovation. The project will create the Hub for the integration and management system of the future EOSC.

EOSCHub will address 2 main challenges:

- How to make large data more easily accessible? We already have some experience with previous projects but other communities have their data managed by third party.

- Policy and financial problems. There is no easy way to provide access to services that can be used by different communities.

The scientific, technical and cultural challenges for EOSC include:

- Scientific challenges: Deploying the EOSC to deliver Open Science.
- Technical challenges: Developing technical solutions to meet scientific needs.
- Cultural challenges: Adopting new and more open ways of working.

These challenges can be addressed with coordinated efforts to:

- Bring the existing research infrastructures together.
- Bring e-infrastructure projects together (GEANT , PRACE..).
- Interoperate between their services and data.
- Allow new resources to be added (HPC providers, Cloud providers, Data providers…).

The main objective for the **EOSCpilot Architecture and the services** is to propose a governance framework for EOSC. There are 10 Science Demonstrator from different domains at present and 5 more will be added at the end of November 2017. EOSCpilot is also defining who the users of EOSC are: data scientists, researchers, managers, service developers, service providers, and e-infra providers. In the architecture framework we should consider where all these stakeholders fit.

The **EOSC Pilot Science Demonstrators for LOFAR** aims to improve the user experience both for power users and non-power users. The project will be built upon existing knowledge and by combining existing tools to show the complete path from facility to user, to demonstrate it can be done and to demonstrate the capacity and shortcomings, or challenges for the future.

The project plan for this demonstrator includes

- SURFsara providing connectivity
- Common Workflow Language to standardize existing pipelines
- User Settings to build frontends
- Use of Zenodo platform for persistent storage DOI

The final objective of this demonstrator is to demonstrate a complete system can be built from existing tools.

EVN as cloud version is a pilot from the JIVE institute, to make the **European VLBI Network (EVN)** data accessible through the cloud. The EVN archive has everything that a large archive has, except the size. Providing a cloud version would be a good test case for storage and data reduction in the cloud.

The goal of the **H2020-AENEAS** project is to design a distributed **European Science Data Centre (ESDC) for SKA,** to support the pan-European astronomical community in achieving the scientific goals of the SKA. It is very important that the design of the ESDC runs parallel to the emergence of the EOSC and on the way learning from each other.

The work carried out in **H2020- HNSciCloud** is a contribution towards the development of the EOSC as well as in view of the next Work programme 2018-2020 call INFRAEOSC.

In order to share data, we need to store the data in a Trustworthy Digital Repository (TDR). Researchers must be certain that data held in archives remain useful and meaningful in the future. In a multidisciplinary environment it is necessary to understand and implement FAIR (Findable, Accessible, Interoperable and Re-usable) Data principles. In case of High Energy Physics, **LONG TERM DATA PRESERVATION** refers to documentation, software and the environment in which it runs. As per e-Infrastructure Reflection Group, **Long Term** stands for a period of time long enough for there to be concern about the loss of integrity of digital information held in repositories, including deterioration of storage media, changing technologies, support for old and new media and data formats (including standards), and a changing user community. **From the experience of EOSC Pilot HEP LTDP we understand that even at “modest” scale (100TB), HEP data formats and long-term needs mean that a “generic” TDR is unlikely to work.**

The upgrades of the LHC and its large detectors, planned for the middle of the next decade (the HL-LHC project), will pose significant new challenges for the computing and data infrastructure. HL-LHC will produce several Exabytes of scientific data per year, and will require some 20 M cores of processing power. The Worldwide LHC Computing Grid (WLCG) community is investigating potential changes to the computing models and distributed computing infrastructure that will be needed to meet those challenges over the coming years. In particular, it will address how to create a data infrastructure at the Exabyte scale, that is able to manage and process data in an effective way. The High Energy Physics community has produced a community white paper (CWP) through HSF meeting. The main themes of the CWP are:

- Allow/help countries or regions to flexibly manage compute and storage resources internally

- Investigate the "data-lakes" concept – keep bulk data (down to derived Analysis object data or AODs) in a cloud-like realm (data-lake). Plug in processing via traffic-managed networks, bulk processing close to the data.

A prototype of the **"data-lakes"** concept would represent a valid asset to build EOSC . This concept is of interest for other projects such as CTA, SKA.

The **Square Kilometre Array** is an international project to build the world's largest radio telescope. The project is in its design phase (2013-2018). About 100 organisations across about 20 countries are involved in this phase. The construction phase 1 is expected to start in 2019. SKA will generate 300PB/year from each of its SDP systems. Hence the computing challenges are similar to WLCG. Both HL-LHC and SKA will be Exabyte-scale scientific experiments on a 10-year timescale. **Recognising this commonality, on 13th July 2017, CERN-SKA signed an agreement for cooperation on computing and big data management.**

Initial topics for collaboration include:

- Bi-annual collaboration meetings
- A position paper/roadmap to be produced quickly to focus on Exabyte-scale data infrastructure needs
- Explore collaboration opportunities on common aspects of networking, storage, computing, etc.
- Investigate work with industry via CERN Openlab
- Joint projects to demonstrate/prototype concepts for regional centres and computing models

A long tradition of international collaboration has been made to build telescopes and instruments, since 1977 where the Data format FITS was developed, which is still used today. Astronomical Interoperability Framework was created in 2012 and defines the development of interoperability standards. In Astronomy, there are open interoperable resources that have a centralized access point in the Virtual Research Environment. **Virtual Observatory (VO)** can build blocks reused by other data providers. Researchers will expect to find the data services used in everyday life? in the same EOSC system of system. At a national level ESFRI should be considered as a building block towards the EOSC development. Light Interoperability layer has been one of the keys to success as experienced by the International Virtual Observatory Alliance (IVOA), which is already included in the Registry of Resources in EUDAT. Astronomy online services and the VO should be included "in" EOSC as they are used in practice by users.

3. Some recommendations for the way forward

- Today there is already a good network in the Astronomy and Astrophysics community based on agreements between different infrastructures. We have to guarantee that these networks are maintained in the future. What we will need is to consolidate the workflows, the access to the archive data and a collaboration between a central EOSC point and the new infrastructures. The Astronomy and Astroparticle physics community could contribute to the EOSC design while it is in the design phase. Some of these EOSC design aspects can be workflow technologies and pipelines that allow users to change settings and inputs. As an example, easy Access to LOFAR data and knowledge extraction through the Open Science Cloud can be achieved.

- EOSC could be defined as a system of systems made of existing and emerging RIs, e-infrastructures, data repositories, registries etc. The system of systems approach should present certain features; the most important one is that EOSC will maintain an operational and managerial independence. EOSC is not going to take over an existing infrastructure but to add value to it.

- **EOSC should be built upon the existing infrastructure** and by using knowledge and requirements of current large archives and compute facilities, and mapping a scale increase of one to two orders of magnitude. It will stretch the capacity of any cloud or existing infrastructure to the limit.

- ESFRI projects can participate to the 2018-2020 work programme making use of commercial services and engage with EOSC. The ESFRIs could be used to ensure that there is commitment for Member States in the development of EOSC. It should leverage on the ESFRI contributions with support and inclusion of new models and services such as the **"data lakes"**.

- The next EU H2020 call clearly states "*Research Infrastructures, such as the ones on the ESFRI roadmap and others, are characterised by the very significant data volumes they generate and handle. These data are of interest to thousands of researchers across scientific disciplines and to other potential users via Open Access policies*." Hence **Synergies between ESFRIs** is important to identify and use FAIR data management platforms. ESFRI representatives attending the panel discussion mentioned the need of cluster actions such as H2020-ASTERICS for the EOSC implementation.

- At present HL-LHC and SKA are the research infrastructures facing the Exascale challenges. Other infrastructures such as CTA, LSST, Virgo-LIGO are facing challenges close to the order of Exascale as well. Therefore a common and coordinated approach to access HPC and open science resources is a current requirement. Other ESFRI and world class projects such as CTA, LSST, Virgo-LIGO have similar requirements and constraints. In order to form such a consortium, it is important to understand the commonality and complementarity. It would make sense to work with EOSC as a single group of aligned interests.

- If science drives, the cooperative actions with e-infrastructures and service providers to implement open science would be more effective and reliable. **EOSC cannot replace the ESFRI communities** but it can support the ESFRI communities shared interest by helping the Cluster activities. Regarding the Long Tail of Science, it can support the high bandwidth needed by the ESFRI projects data centres and also interoperability and preservation goals. The EOSC cannot be only about Cloud. EOSC should be a framework to bring together software tools, services, communities. EOSC should be sufficiently open to accept solutions developed or adopted or pre-evaluated in the Cluster activities or other communities.

- EOSC should move towards the federation of existing e-infrastructures. EOSC should be explored in little steps/issues to make them interoperable with others. Consortium of European Social Science Data Archives (CESSDA) has a huge amount of Catalogue of Services. Being an ERIC, they realized that now they have to change their communication so what they were offering before is now clearer. The fact that EOSC can be promoting services that have been developed by other infrastructures, could be more interesting for funding agencies, because in this way, being in EOSC would work as an **incentive for the infrastructures which deployed services used in EOSC.**

- **H2020-ASTERICS** seems to be the only project connecting ESFRIs together to address data interoperability and software reuse and this experience can play a key role for implementing the EOSC even from the point of view of other scientific domains. The ESFRI communities can contribute to data interoperability and software reuse by bringing together people towards common grounds to work together and find solutions/services that can be used across different communities.

- **EOSC for ESFRIs:** Each RI produces data at a continuous rate, and performs basic processing (calibration, geometric registration, etc.). For that purpose, it is expected to have its own dedicated e-infrastructure. The result are archives of persistent data. EOSC should provide data-computing interoperability mechanisms to allow processing on archived data at the archive location (data centre).

- **ESFRIs for EOSC:** Each ASTERICS RI provides an archive of datasets in physical units (i.e. reusable); if the IVOA standards are used, data are FAIR. Additional services could be provided (cut-out service, catalogues of astro-objects, classification, etc.) at the discretion of the data centre. Additionally, a repository of software tools, pipelines, etc., could also be provided.

- **Policy Access:** Datasets are open and public, after a (usually short) proprietary period; metadata are always public except for peculiar cases (e.g. search for extrasolar planets); software is usually public as well –in these cases registering users is not necessary (even undesired). For the use of additional services requiring computing users should be authorised (A&A necessary).

Appendix 3.Machine learning day discussions

Machine learning has been widely used for big data analysis. Considering the deluge of data that will be produced by astronomy ESFRI and other world class facilities, it is necessary to deploy machine learning. The third and final day of the workshop was dedicated to discuss and present some of the machine learning use cases in astronomy and how ASTERICS-OBELICS can introduce these machine learning solutions for ESFRI projects.

The advancements in artificial intelligence and deep learning that occurred over the last few years have opened new avenues for building next generation data analysis pipelines. The GammaLearn project aims to leverage these methodologies to address event discrimination as well as energy and direction estimation for the Cherenkov Telescope Array. This project is a result of cooperation between CNRS-LAPP and Orobix Srl through the first OBELICS call for industrial cooperation. The talk provided an overview of the approaches that will be put into action, along with examples of lessons learnt from prior experiences on manufacturing and medical imaging applications. The talk also elucidated the ongoing activities in the GammaLearn project, the unique challenges posed by the analysis of IACT data and the new opportunities created by the possibility of pushing data analysis at the edge. E4 Computer Engineering also intends to collaborate with OBELICS partner INAF Rome for the development of a new data analysis technique for IACTs, based on machine learning and using Deep Neural Networks (DNNs), for analysing images and classify within a software-hardware integrated system adopting new hardware architectures. This project with E4 is yet to be formalized.

Machine learning use cases in Gravitational wave detections were presented. Noise of non-astrophysical origin contaminates science data taken by the Advanced Laser Interferometer Gravitational-wave Observatory and Advanced Virgo gravitational-wave detectors. Characterization of instrumental and environmental noise transients has proven critical in identifying false positives in the first LIGO observing runs. The European Gravitational Observatory (EGO) has investigated new ways to identify and classify signals using machine learning techniques. They used unsupervised algorithms for unlabelled transient signals and started using more efficient methods using supervised methods as Deep Learning on labelled training data sets. They also explored image classification techniques based on GPU technology, which can be used for pattern recognition. Apache Spark is a fast and general-purpose cluster computing system used for Scalable, efficient analysis of Big Data. A machine learning pipeline based on Apache Spark will be investigated by EGO as a next step.

The next generation of cosmological surveys such as the ones conducted by LSST, Euclid and SKA will bring unprecedented constraints on the nature of dark matter and dark energy. They also entail new challenges, in particular from the sheer volume of data they will produce. Dr Francois Lanusse from McWilliams Center for Cosmology at Carnegie Mellon University presented some exciting applications of Deep Learning to address these challenges at different levels, from image processing to modelling galaxy physics. His talk focused in particular on the problem of automated strong gravitational lens finding (see https://goo.gl/TnnTLE), a typical image classification problem, to illustrate how Deep Learning can have a profound impact on a science analysis pipeline, in this case by dramatically reducing (and maybe even eliminating) the need for human visual inspection. As a point of reference, it was estimated that previous methods would have required around one million volunteers participating in a citizen science initiative to classify the whole LSST survey in a matter of weeks. The main takeaway here is that automated lens finders based on machine learning are faster and more reliable than humans and this can eliminate the need for visual inspection. For cosmology, the machine learning can model and analyse large volumes of complex datasets, open new and powerful ways to look at the data and help control systematics in conventional analyses

Neutrino astronomy often deals with high backgrounds and low signal statistics. Shallow and deep machine learning techniques are allowing KM3NeT to address the tasks such as Up/Down classification, Track/Shower classification, Selectfit and Particle identification. It was hardly possible to tackle these challenging tasks before using the deep learning techniques. These applications of deep learning in KM3NeT along with the algorithms were discussed in detail.

Some of the discussion points for the steps forward included

- We should leverage upon existing systems and standards in machine learning instead of pushing for developing on our own.
- In order to avoid vendor lock-in situations, we should develop our environments with our own tools. We need to be flexible and not linked to a specific technology. As far as the software are concerned, there is very little or no possibility of vendor lock-in situations, as most of the ESFRI experiments already are using different software. However, for the hardware components such as GPUs we are already in a vendor lock-in situation.
- Documentation of machine learning libraries and algorithms respecting the FAIR principles should be given priority instead of the technological needs. Such documentations should essentially include training, tutorials and relevant communications. This would allow the end users to compare with alternative tools they already know or are using.
- It takes some time for the community to fully understand and adopt the models that are quite complex, leading to scepticism. We should maybe agree on some sort of validation of these models, this would be very essential for increased collaboration. In certain communities machine learning is being used for different applications

In addition to machine learning, Common Workflow Language (CWL) was also discussed. Workflow management systems (CWL is lighter in code lines in respect to python or other languages) are universally used to facilitate science and data analysis.

CWL has tiered view of standards in a research context: thinner high level standards and data formats that are applicable across disciplines (like CWL & researchobject.org) and community specific ontologies, data formats, and common vocabularies that specialize/build-upon the general standards to the needs of that particular community. For example: http://edamontology.org from bioinformatics plugs well into CWL.

There are more than 130 workflow management systems. A sort of standard is needed. The main features for the Astrophysics community include:

- CWL tool descriptions can self-describe the shape of the computational resources required: this can simplify and most important make data analysis faster.
- Platforms that understand CWL can use its identifiers to send compute to where or near the location data
- REANA project at CERN has used CWL.