The European Science Cloud initiative: achieving open science in astrophysics and particle physics

14th June 2019

The European Union has launched the European Open Science Cloud (EOSC) initiative to support the data driven research in pursuing excellent science.

Together, astrophysics and particle physics address the open science challenges in Europe building the EOSC, argues Dr Giovanni Lamanna.

EOSC is a cloud for research data in Europe that allows universal access to data through a single online platform. EOSC will federate existing resources across national data centres, e-infrastructures, and research infrastructures, allowing researchers and citizens to access and re-use data produced by other scientists. Respective capacities and domain-based heterogeneous needs of scientific stakeholders make the EOSC implementation a real challenge.

Dr Giovanni Lamanna is the co-ordinator of 'ESCAPE – European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures'. ESCAPE is an international research cluster project consisting of 31 partner institutes. It is funded under the European Commission's H2020 scheme, and it brings together astronomy and accelerator-based particle physics that share aligned visions and expectations about EOSC in order to engage and collaborate on building it.

ESCAPE aims to address the Open Science challenges shared by next generation facilities prioritised in the European Strategy Forum on Research Infrastructures (ESFRI) and in other world-class projects, such as CTA, ELT, EST, FAIR, HL-LHC, KM3NeT, SKA, LSST as well as other pan- European research infrastructures, namely, CERN, ESO, JIV-ERIC and EGO-Virgo.

ESCAPE leverages the successful experience of a previous H2020 cluster project, ASTERICS, that brought together, for the first time, the astronomy, astrophysics, and astroparticle physics facilities encompassed within the ESFRI roadmap.

ESCAPE foundations lay on the important work conducted by ASTERICS looking towards enabling interoperability between the facilities, minimising fragmentation, and developing joint multi-wavelength/multi-messenger capabilities. Lamanna co-ordinated the OBELICS (OBservatory E-environments LInked by common ChallengeS) work programme within ASTERICS, which had a specific goal concerning the establishment of an open innovation environment to face common data-intensive challenges of ESFRI projects.

Open Science

Open science has become a pillar of the current national and international views and political programmes of governments, research funding bodies, and forums. There are three main ambitions for that:

1) Change the way citizens could perceive research and public investments for research.

Although today it might appear logical that all results from research funded with public grants should be published in open-access, free-of- charge scientific journals, this is still to be fully accomplished, but progress has been made. The particle physics community is an early precursor of making scientific literature freely available. CERN hosts the INSPIRE-HEP database collecting open access preprints that was established in the late 1960s. Astronomers and particle physicists used to self-archive their preprints in the arXiv.org repository, and NASA has constantly adopted the approach of openly shared scientific data from every space mission, and today makes its entire research library freely available in the public domain.

To stimulate vocations among young generations and to embrace an open science public engagement, major international research institutions invest in communicating science. Soon, for instance, CERN will extend its outreach with the Science Gateway building that will provide a variety of activities and exhibitions also featuring CERN's technologies and how they benefit society. Meanwhile, the European Southern Observatory's (ESO)

Supernova building at its Headquarters in Munich, Germany, is a digital planetarium that offers a unique experience through the exploitation of digital astronomical images and videos, as well as real-time events.

The way recent breakthrough results and discoveries have been publically announced and explained has also evolved, providing more resonance and enabling larger public interest in major international co-ordinated efforts in fundamental science. This is the case, for instance, with the first discovery of gravitational waves (GW) in 2015, or the 2017 LIGO-Virgo detection of GW emitted by the collapse of neutron starts and, finally, the 2019 announcement of the first ever image of a black hole taken by the Event Horizon Telescope. For all of these discoveries, a webcasted and simultaneous international multi-site announcement was accompanied by seminars, followed by press conferences which were organised to be synchronous with the publication of the results and the corresponding data.

2) Enable opportunities offered by the digital revolution to allow everybody to participate in the scientific process by accessing research data at any time.

Since the birth of the World Wide Web and with the recent beginning of the digital age, more than 4 billion Internet users have potential access to digital knowledge and grid and cloud computing platforms, as well as to a series of open-source software tools and services for scientific data exploration. The astrophysics community is a pioneer in establishing standards to publish astrophysics data and tools to enable discovery through the Virtual Observatory (VO) infrastructure for data interoperability and re-use.
NASA, ESO, and ESA are deeply involved in opening up their astronomical data products, thereby allowing citizens to have access to huge databases.

Professional researchers volunteer to make their scientific products available to the public. Zooniverse is the largest platform for people-powered research where citizens, independent on their own background, can participate in studying authentic objects of interest gathered by researchers, such as images of faraway galaxies, historical records and diaries, or videos of animals in their natural habitats.

“For instance, with the support of the ASTERICS cluster, the ESCAPE precursor, a project was implemented providing gamma-ray astronomy images from Cherenkov telescopes. Members of the public are invited through a web interface to visually scan these images and separate them in sub-samples of images potentially produced by different particles, namely photons, muons, and protons,” Lamanna said. “After only the first five days that the project was on-line, we counted more than 1.3 million citizen classifications of images inspected on the web. It is impressive!”

Particle physicists consider that the complexity of their analysis would not leave room to any advantageous citizen-based contributions. However, this does not limit their desire to engage with society for outreach. At CERN, every year, thousands of high-school students become particle physicists for a day. The CERN Open Data web portal provides tailored projects for managing high-level data products of the LHC experiments. They are traditionally used in high-school ‘Masterclass’ exercises, training students to event-display and easy particle tagging analysis.

3) Accelerate the discoveries and increase scientific value by sharing data and by transferring knowledge within scientific communities.

The sharing of data is a successful long-term tradition in astronomy, and recent progress in particle physics towards this end has now been achieved, with the first works about experiments at the LHC being published based on open data. The orientation of consortia of researchers committed to the construction of big science facilities to open up the access to their data, implies a sociological evolution, since some members of the consortia running the facilities could feel as though they are aiding other researchers to make progress towards the next discovery, rather than themselves.

“The current generation of researchers demand re-using, reproducing and combining scientific data from different sources to increase their experimental knowledge and accelerate the scientific process,” Lamanna said. “Future generations will make use of open-access advanced IT technologies and open-source software to potentially reveal new results by mutually transferring knowledge amongst scientists and teams. Science and scientific results will not be communicated simply through international journal

publications, but via the systematic access to data and software platforms to reproduce the results behind them."

The combination of data from the astronomy-related ESFRI projects and the accelerator-based particle physics ESFRI facilities will together open new complementary paths towards the understanding of the Universe. The recipe for this is the implementation of an effective 'open-data paradigm'.

Open Data

'Open data' does not mean 'free data' and, furthermore, depending on what 'data' is being referred to, conditions governing its access and reuse will always be applied.

Open Science and the EOSC initiative support an uptake of a set of guiding principles about the way to plan and produce scientific data. These are the 'FAIR' Data Principles, where 'FAIR' stands for 'Findable, Accessible, Interoperable, and Reusable'. Those principles should be applied through the entire research process, implying rules of engagement and methodological practice from scientists and facilities, and investments to increase support infrastructure of FAIR data- publishing, analytics, computing, workflow services. But sharing data is not enough.

The data-analysis of every ESFRI project encompassed within ESCAPE requires specific software and auxiliary data and metadata to extract information by performing further data reduction/correlation, elaborate intermediate results and data-selection, followed by statistical data analysis and, potentially, the assessment of the agreement between the final measurements and the theoretical model predictions. Therefore, the full chain, including datasets and analysis code, needs to be accessible in order to enable the precise repetition of the published results and potentially for enhanced discovery. "In such a way 'open data' assumes the meaning of 'Data FAIRness for science reproducibility'. For most astrophysics and particle physics", says Giovanni Lamanna.

For most astrophysics and particle physics projects, the main issue is defining for each-project their specific data preservation policy and reproducibility goals early on in their

own preparatory phase, thereby ensuring the future reusability of the scientific results to come.

If EOSC is the infrastructure for open science consumption, a series of implications is then raised:

• Establish a reference federated digital repository for the sustainability of scientific data preservation independently on the lifetime of each single research infrastructure
• Apply rules and operate an authority to inquire adoption of data management plan and certified open archival model (such as ISO OAIS) by each new facility/data producer
• Adopt a governance system that relies on the steering power of open scientific collaborations for peer review, transparency, and efficiency
• Define and support credit systems, acknowledgement, standard licences, and certification for the results of all researchers engaged in co-operative work that allow all to reap the benefits of open science.

ESCAPE

The new facilities encompassed within ESCAPE will significantly extend our multi-messenger observational capabilities across the electromagnetic spectrum, neutrinos, high-energy gamma rays, and gravitational waves. The advent of the new generation of high-energy and high-intensity particle accelerator facilities will enable forefront research on matter constituents, the search for dark matter candidates, studies of the origin of the matter-antimatter asymmetry in the Universe, and investigations into the structure and dynamics of matter under extreme conditions, thereby also providing new insights into the evolution of the Universe and the nucleosynthesis in stars and star explosions.

The cluster built by ESCAPE partners endorses common approaches for open data management in ESFRI projects, leveraging two major complementary excellences in data stewardship:

• The astronomy Virtual Observatory infrastructure extending its standards and methods according to FAIR principles to a larger scientific context as well as connecting to the EOSC
• The long-standing expertise of the particle physics community in large-scale distributed

computing and Big Data management for setting up a data infrastructure beyond the current state-of-the-art, linking it to EOSC.

On the one hand, the co-ordination and interoperability between ESCAPE-participating ESFRI projects shall produce versatile solutions to make data more accessible to growing scientific communities as they collaborate and interact with of new generation scientists conducting fundamental science research through a global multi-probe approach. On the other, as the complexity of all next generation facilities in ESCAPE grows rapidly, the data volume produced by them is seeing a tremendous increase and the software to analyse the data is becoming more and more complex. Projects such as SKA and HL- LHC will generate data at multi-Exabyte scale, followed by other data-intensive projects such as CTA, FAIR at GSI, LSST, KM3NeT, Virgo, and others. The ability to curate and serve data at all scales up to these unprecedented needs goes far beyond the current technologies.

Data lake

The lessons learned from the successful GRID computing model introduced for the LHC and extended to other disciplines by the EGI infrastructure, the proven high reliability of high- bandwidth regional and international networks together with the increasing offer of large-scale heterogeneous CPU-only resources (grid, HPC, public and commercial cloud, volunteer computing) led to the proposal of a ‘data lake’ model to support a large worldwide scientific community in achieving the scientific goals of the ESCAPE ESFRI projects and to extract science from their data.

The ESCAPE data lake aims at designing, implementing, and operating a prototype concept for a distributed federated infrastructure. The data lake concept enables the large, reliable, national research data warehouse centres to work together, where CPU and storage resources are no longer always co-located as it is in the GRID. Higher quality and security of service management for replication and access latency, redundancy and storage of data is a major investment. The data lake is built as a distributed storage system in which metadata can be stored centrally in a reference data centre for a specific ESFRI project and separately from data that, on the contrary, are archived together with their replicas in national centres (supporting more than one facility).

Depending on the specific pipeline of a project, on the users' access popularity of datasets at different levels of the data processing, the data caching versus data pre-staging, is evaluated and optionally chosen to minimise data movement within the lake and to reduce latency. ESCAPE looks forward to proposing the data-lake prototype as a backbone of the EOSC federated cloud of national data centres.

Making the seamless connection between ESFRI astronomical facilities to the EOSC through the Virtual Observatory and connections to the necessary computing resources is one main task of the cluster. This implies the need to scale the VO framework to the biggest data sets that will be produced. The data lake infrastructure aims at associating the VO archives of ESO observatories that would become part of such a backbone, making accessible high-level astrophysical products.

However, if fostering a common approach among a series of major Big Science projects leverages economies of scale to decrease cost in development and implementation, but the infrastructures are exploited according to EOSC mandate for open data reuse, collection and exposure of new analysis results and so on, then the operation costs increase and the assessment on whom they will be charged on, among the funding agencies, the legal entities operating the facilities and the EOSC governance body, is still an open question.

Open-source software

Open-source software is the easiest and oldest declination of the global open-science movement and began historically as a reaction to a monopole of proprietary operating systems and services (i.e. GNU/Linux). Today, it is clear that data isn't limit to the digitally-encoded bits and also includes the main analysis software involved to produce scientific results. Its preservation and open access enable actual internal or external community reuse of research outputs and open reproducibility of science.

Particle physicists are pioneer adapters of open and shared software services for statistical data analysis, visualisation, and data management, although until recently the software underlying an analysis has not been easily shared. Next generation astronomical

experiments, being operated as observatories, commit to providing open access to data and analysis software: radio astronomers adopt the CASA software stack, Gamma-ray Cherenkov telescopes develop intermediate level reconstructed data instrument response functions with software stack for analysis (i.e. ctools, Gammapy and GammaLib). LIGO and Virgo gravitational wave interferometers are, in 2019, running a co-ordinated run of observations: astrophysics alerts as well as platforms including data sets and software for workflow analysis are openly provided, etc. This implies that EOSC finally would play a key role in adding value to the data research by supporting and labelling community-based co-operation.

Indeed the complexity of collection, processing, and deployment of data produced and handled by the concerned ESCAPE facilities demands innovative solutions and calls for cross-fertilisation actions. The development of multi-messenger and multi- probe data analysis practices promotes activities to maximise software re-use and co-development. Today, scientists are also software producers and are embracing agile development methods; they are able to guarantee a continuous development of more sensible analysis methods which require access to more complex and deeper levels of data (closer to raw). Although, even from VO high-level products, the generation of value-added content in the ESO archive services in preparation for ELT is expected through the development of cutting edge algorithmic approaches in ESCAPE.

The costs for each ESFRI project for the continuous maintenance of open-source tools for analysis, together with services and computing resources to access its data, would constrain the openness to only a limited level of data complexity. If supported by EOSC in federating efforts and sharing results through a foundation approach, researchers will be able to go further than the minimal open science commitment of a research facility. This approach is actually a partial shift of reasonability for access to quality-certified scientific data from the facility to the community on a longer term. It implies an increase of costs that will have to be supported by EOSC transversally and a guarantee of sustainability.

Vision and aims

“ESCAPE’s main objectives,” Lamanna explained, “are: support a community-based approach for continuous development, deployment, exposure, and preservation of domain-specific open-source scientific software; enable open science interoperability and software re-use for the data analysis; create an open innovation environment for establishing open standards, common regulations, and shared software libraries for multi-probe data. This integration of both the data and necessary tools fosters additional synergies between the different ESFRIs as common approaches can be identified and exploited.”

Such an approach was already applied in ASTERICS with significant success. Although focused on the use-case of any specific ESFRI project, the ASTERICS community has developed and deployed open-source software solutions of potential transversal application. As an example, machine learning and other high performance programme software methods were developed for real time multi-messenger astrophysical alert follow ups and offline more sensible analyses. Other services for pipelines access and exploitation were also finalised.

All these open science sub-projects and related products are made available in a dedicated ASTERICS-OBELICS repository to be included in a new ESCAPE catalogue that will be part of the global EOSC marketplace of scientific services.

“The next goal in ESCAPE is the building of a credit system to acknowledge the current and future results that researchers from our community produce co-operatively for the benefit of open science and associate digital object identifier to data and software to enhance data discoverability,”Lamanna said.

In ESCAPE, central regulations compliant with the EOSC global ones will ensure the software and service quality via audits. A help desk and training activities will provide support to the ESFRIs to devise regulations and adhere to them.

An effective virtual research environment

In order to make EOSC an effective virtual research environment for open science, a change of the data analysis paradigm for scientists is envisaged. ESCAPE partners believe

that the new generation of scientists need to be 'uploaded' into the EOSC environment and in order to achieve this a flexible science platform for data analysis is required. The platform will allow EOSC researchers to identify and access existing data collections (in the data lake infrastructure) for analysis, tap into a wide-range of software tools and packages developed by the ESFRIs (and exposed by ESCAPE in its dedicated catalogue), bring their own custom workflows to the platform, and take advantage of the underlying HPC and HTC computing infrastructure to execute those workflows.

The ESCAPE approach is to provide a set of functionalities from which various communities and ESFRIs can assemble an analysis platform geared to their specific needs. Technological progress in information technology is now helping researchers to capture the full research process with innovative tools that help to encapsulate a pipeline for the preservation, reproducibility and sharing purposes of workflow results. Current evaluations are about web applications to access collaborative software-development frameworks, to create and share documents, live code, equations, enabling data cleaning and transformation, modelling, and visualisations etc.; tools to connect to and interact modifying analysis workflows in a live; services for access to open data behind a scientific publication.

In conclusion the astronomy and particle physics community supporting ESCAPE embraces open science, acknowledges data sharing and data stewardship as critical issues for the next generation ESFRI facilities. ESCAPE vision for EOSC is:

• A data infrastructure commons serving the needs of scientists, providing functions delegated to community level and federating resources
• Researchers should contribute to define the main EOSC common functionalities needed in their own scientific domain; therefore, they will not be merely EOSC users but continuous digital innovators, data (and software) producers, and consumers
• A continuous dialogue to build trust and agreements among funders, scientists, and service providers is necessary for data- research sustainability.

Dr Giovanni Lamanna
Laboratoire d'Annecy de Physique des Particules
+ 33 450091601

lamanna@lapp.in2p3.fr

https://www.projectescape.eu