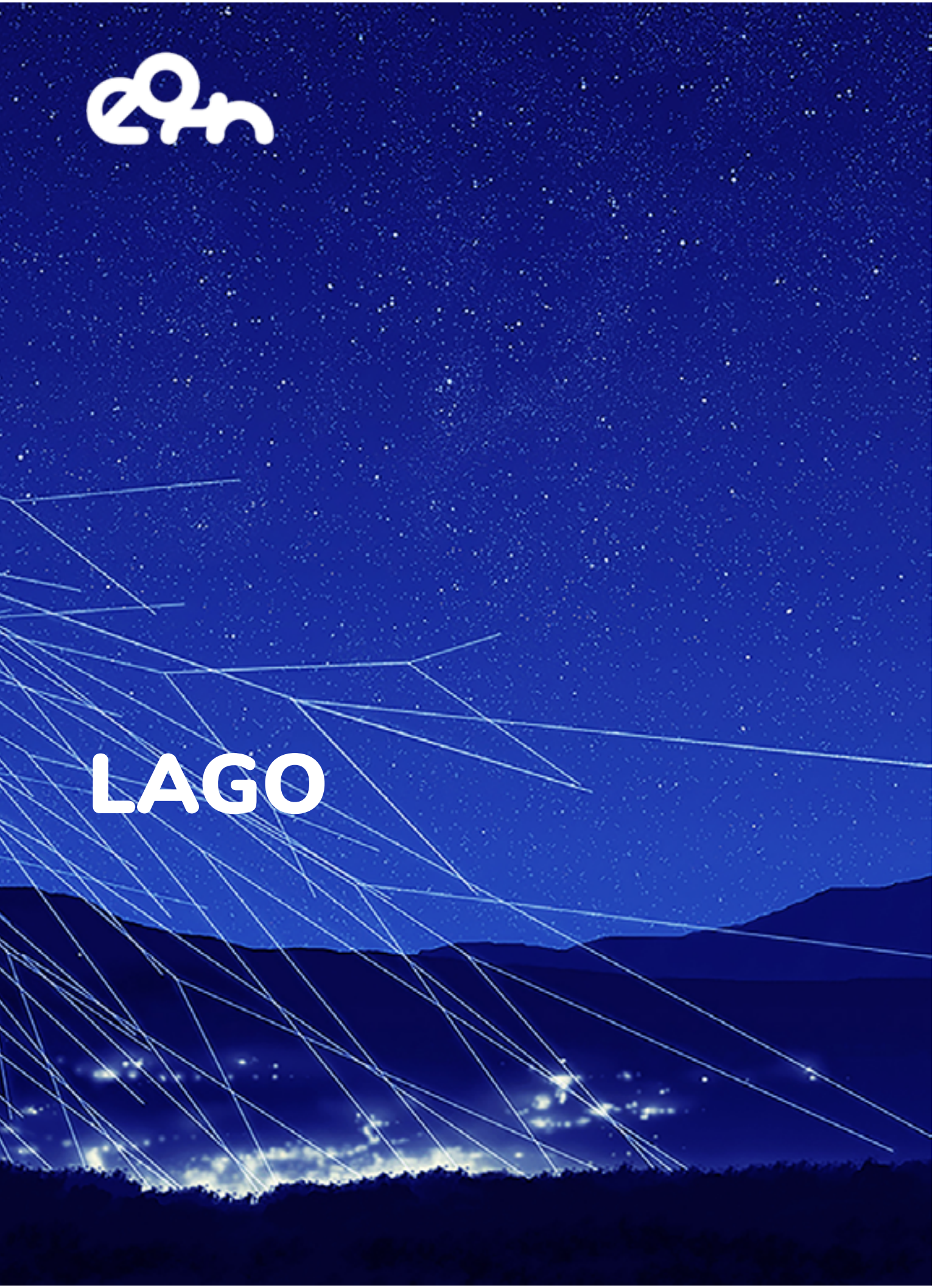
LAGO

6. LAGO

Description

The LAGO (https://lagoproject.net/) (Latin American Giant Observatory) Project is an extended astroparticle observatory at a global scale. It is mainly oriented to basic research on three branches of astroparticle physics: the extreme universe, space weather phenomena, and atmospheric radiation at ground level. Parallelly, these are the needed components of other research on high energy physics, weather forecasting, life sciences, aerospatial security or computer science. The LAGO detection network consists of single or small arrays of self-designed water-Cherenkov detectors (WCDs). These particle detectors are spanned over different sites located at significantly different latitudes (currently from Mexico up to the Antarctic region) and different altitudes (from sea level up to more than 5000 meters over sea level), covering a huge range of geomagnetic rigidity cut-offs and atmospheric absorption/reaction levels. The measurements collected from these detectors are posteriorly processed and analysed. Moreover, scientists continuously generate simulated data for arbitrary locations and weather conditions.

On the other hand, the LAGO Project is operated by the LAGO Collaboration, a non-centralized and distributed collaborative network of more than 100 scientists from almost 30 institutions in 11 countries. Additionally, several universities have incorporated LAGO studies into their curricula. Their students, especially the ones belonging to physics, electronics and computing areas, also contribute to the development of LAGO technologies. To manage this heterogeneity and take advantage of the aforementioned contributors, the LAGO Thematic Service will progressively incorporate the continuous generation of data (measurements, processing and simulations) and code into standardised mechanisms that follow the FAIR principles. This is so to guarantee the long-term curation and re-use of data as well as the dissemination or reproducibility by other communities. Therefore, the final purpose of the LAGO Thematic Service is to enable the universal profit and contribution of this research, within and outside LAGO Collaboration, through a sustainable Virtual Observatory and standardised computational model.

Architecture

To introduce the architecture of the LAGO Thematic Service, readers should understand first some basic considerations about the kind of data managed and the target community. There are two main kinds of data managed by LAGO Collaboration. The first is related to real measurements (L), and the second is to simulations (S). Thus, the measured data (raw) is pipelined for correction, obtaining the following data sub-types that corresponds with their quality level:

- **L0. Raw data.** Measurements of Water-Cherenkov detectors (WCDs).
- **L1. Preliminary data.** Low resolution but the atmospheric pressure is corrected.
- **L2: Quality for Astrophysics.** Ensures data quality to be used by experts from the astrophysics Community: fixed scalers by atmospheric parameters and the efficiency of the detector.
- **L3. Quality for the public.** Ensures high quality to be used by researchers from other subjects or the public: the histograms are also corrected.

On the other hand, users can perform their simulations of rains, generating two sub-types of data-sets:

- **S0. Plain simulations.** Plain simulated data (CORSIKA outputs managed by ARTI).
- **S1. Analysed simulations.** ARTI outputs.

There are four main premises of the collaborators:

- Officially, they are grouped into **autonomous research units** within work packages, every unit with specific responsibilities. As examples, every detector has an operator unit, every software piece has a manager, etc. External

staff should be allowed or removed by every research group for eventual contributions.

- Most are **researchers** in astrophysics and HEP **with a background in computing skills:** they are accustomed to profiting from HPC facilities and/or use control version systems such as Git.
- Although each contributor is focused on simulations, on processing or curating measurements, **they produce results of interest for any other member or external actors: whole data or code generated should be registered,** shared and published after an embargo period.
- Some **institutions** support the project **with computational resources,** such as clusters and related storage, but they generally are non-exclusive and provide a limited environment.

The design should take in mind the aforementioned tasks as well as the thematic service is focused on providing a standardised way to curate and reuse measurements, analysis and simulations. To achieve this task, the architecture follows the basic **design recommended by EGI/EOSC for cloud computing: core** intelligence **packed in Docker images**, being able to automatically check, store and publish their **results in DataHub**, with enough metadata to be **referenced by PiDs** (provided by **B2HANDLE**) and used by official harvesters **(B2FIND),** which will act **as virtual observatories**. As the whole computation is self-contained in the image, the **production** can be easily performed on cloud resources **deployed by services** such as **EC3/IM**, or even manually in private clusters.

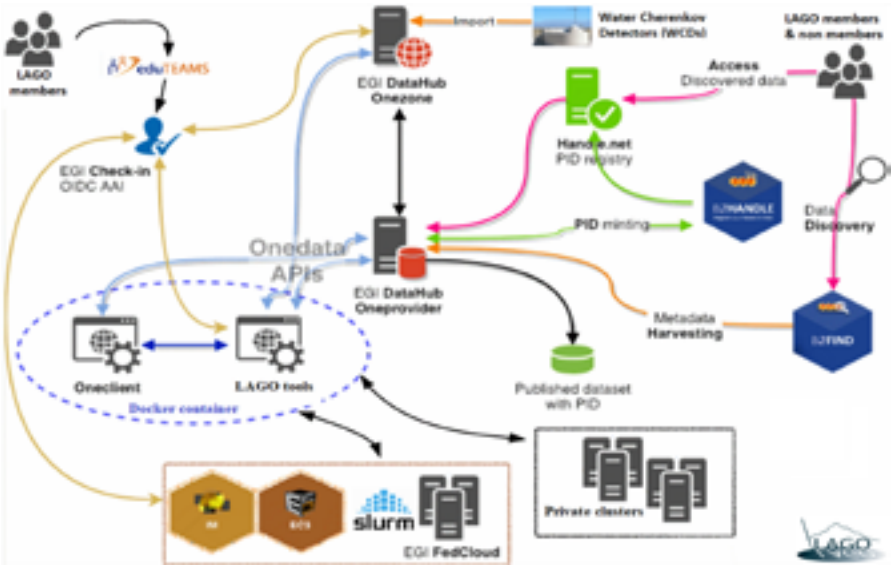

Figure 20 - The architecture of the LAGO thematic service

Therefore, besides the integration of these services and the creation of a new Virtual Organisation, the core contributions of the LAGO thematic service are focused on the definition of metadata and its generation by LAGO tools deployed in Docker images. Note that only the **L0, S0 and S1** types of data will be covered in EOSC-SYNERGY. Therefore only related metadata and LAGO tools will be migrated to Docker during the project, although the architecture will be maintained when adding the rest of computation in the future. Additionally, for this **prototype**, only S0 simulations are supported, being the process that consumes more computational resources and it implies to overcome more difficulties in its deployment. Its Docker instance, named **OnedataSim** (https://github.com/lagoproject/onedataSim), encapsulates ARTI and CORSIKA software, generates the data and metadata and stores them in DataHub. It is currently available for the whole LAGO community.

6. LAGO

EOSC Services

LAGO thematic has selected and it is integrating the following services listed in the EOSC marketplace:

- EGI Check-in (through EduTeams Perun at GEANT): it is needed for accessing any EOSC service, in particular for obtaining a OneData token. However, managing the VO with Perun at GEANT was considered because of flexibility, certain independence from EU Framework projects and long-term support to Latin American users. Perun provides the needed flexibility allowing several roles and permissions over the data, such as conventional users (allowed seeing whole data, restricted write), research group chiefs (allowed enrolling their researchers by their own), robots, main administrators, etc. On the other hand, the sustainability of the VO is guaranteed by the support of RedClara (associated with GEANT), allowing extending users and resources beyond EOSC.
- EGI DataHub: OneData allows researchers several ways to access the data and metadata of their interest. Collaboration members can directly explore the directory tree at https://datahub.egi.eu or mount it on their PC's. Meanwhile, the public will get published data through B2FIND. On the other hand, OneData eases storing results without modifying simulation/processing codes, as well as maintaining usable replicas around the world. Currently, DataHub is storing S0 simulations with metadata and L0 raw without metadata, taking up over one TB.
- The EOSC Cloud services (IM and EC3): simulations are arbitrarily performed by researchers running the dockers in EOSC Cloud services. To perform these tasks, they dynamically deploy individual virtual machines or batch clusters through IM or EC services. Although users can create any kind of cluster, only Slurm workload manager is supported for now, because it is commonly used by LAGO collaborators.
- B2FIND and B2HANDLE: currently under development, will be adopted in the coming months because we expect that the integration will be straightforward since we use standard metadata. In the case of B2FIND, we do not discard to additionally use other CKAN repositories in the future, such as the ones used in the EU Joint Research Centres and other government repositories, to completely benefit from the linked-metadata in JSON-LD + DCAT2-AP format.

Service Endpoint

Collaborators in LAGO commonly use the command-line shell to run many scientific applications only available on Linux, which are needed for their research. Additionally, they like to inspect configurations, code and results for debugging. Moreover, they usually take advantage of remote HPC facilities. Thus, it is reasonable to offer a similar environment for them.

For these reasons, neither a customised submission web page nor portal (i.e. Galaxy) is built. However, the whole infrastructure is offered as the grid computing fashion: every user can deploy his cluster. Researchers only need a guide (https://lagoproject.github.io/DMP/howtos/) to enrol into the LAGO VO, to build a cluster in the cloud, to run standardised Docker instances in the batch system, and finally to register the results in DataHub. Additionally, every standardised Docker has a specific guide at their code repository, as it is the case of this prototype, onedataSim (https://github.com/lagoproject/onedataSim).

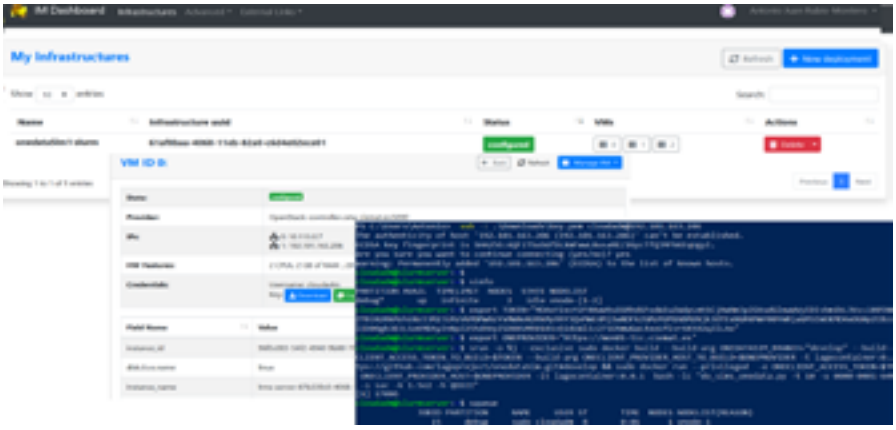

Figure 21 - An example of running onedataSim in the Slurm deployed through the IM service

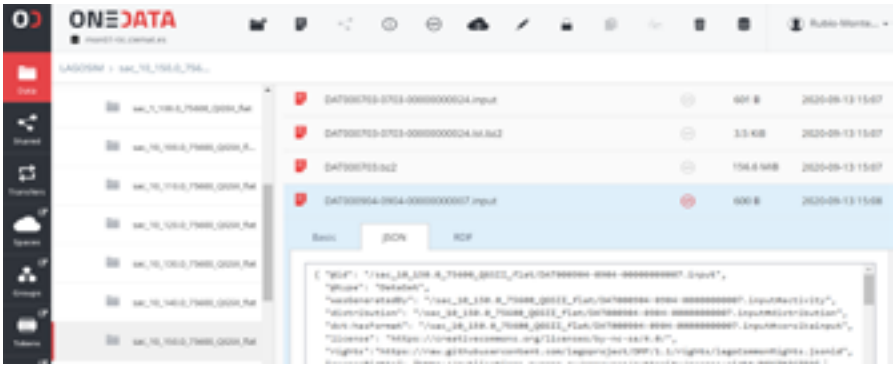

Figure 22 - S0 metadata and data stored in DataHub by simulation shown in figure 21

Demonstration Video

The demonstration video describes the LAGO project and the final purpose of the thematic service associated with the present deliverable.

The video includes an explanation on how the EOSC-SYNERGY project is helping LAGO to incorporate the continuous generation of data (measurements, processing and simulations) and codes into standardised mechanisms that follow the FAIR principles. The video describes the different types of data generated by the virtual astroparticle observatory and the computational models and presents the chosen technology solutions and architecture of the management plan. Additionally, includes instructions on how to join the LAGO organization and how to access the data repositories. Finally, the video includes a tutorial on how to access the infrastructure manager and the description of the tools for ARTI simulation and analysis on OneData.

Click the image to view the video