PROCEEDINGS OF SCIENCE

Data Accessibility, Reproducibility and Trustworthiness with LAGO Data Repository

H. Asorey[1,2,3], D. Cazar-Ramírez*[4], R. Mayo-García[5], L.A. Núñez[3,6], M. Rodríguez-Pascual[5], L.A. Torres-Niño[7,8], for the LAGO Collaboration[9]

[1] *Laboratorio Detección de Partículas y Radiación, Instituto Balseiro y Centro Atómico Bariloche, S.C. de Bariloche, Argentina.*
[2] *Sede Andina, Universidad Nacional de Río Negro, S.C. de Bariloche, Argentina.*
[3] *Escuela de Física, Universidad Industrial de Santander, Bucaramanga, Colombia.*
[4] *Colegio de Ciencias e Ingenierias, Universidad San Francisco de Quito, Cumbayá, Ecuador.*
[5] *Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, Madrid, Spain.*
[6] *Centro de Física Fundamental, Dept. de Física, Universidad de Los Andes, Mérida Venezuela*
[7] *Escuela de Ingeniería en Sistemas, Universidad Industrial de Santander, Bucaramanga, Colombia*
[8] *Centro de Supercomputación y Cálculo Científico, Universidad Industrial de Santander, Bucaramanga, Colombia*
[9] *lagoproject.org, see the full list of members and institutions at lagoproject.org/collab.html*
e-mail: lago-pi@lagoproject.org

Nowadays, one of the most challenging scenarios scientists and scientific communities are facing is the huge amount of data emerging from vast networks of sensors and from computational simulations performed in a diversity of computing architectures and e-infrastructures. In this work we present the strategy of the Latin American Giant Observatory (LAGO) to catalog and preserve a vast amount of data produced by the water-Cherenkov Detector network and the complete LAGO simulation workflow that characterize each site. Metadata, Permanent Identifiers and the facilities from the LAGO Data Repository are described. These initiatives allow researchers to find data and directly use them in a code running by means of a Science Gateway that provides access to different clusters, Grid and Cloud infrastructures worldwide.

The 34th International Cosmic Ray Conference,
30 July-6 August, 2015
The Hague, The Netherlands

*Speaker.

PoS(ICRC2015)672

1. Introduction

Data-intensive scientific analysis is a completely new way of doing science. How to deal with large datasets is still in evolution and has a long way to go in almost all disciplines -physical, life sciences and humanities- which are becoming increasingly data-driven and data intensive.

An increasing computing power and a deluge of connected sensor devises -now ready accessible for research communities- have driven a dramatic improved in our ability to collect, preserve and share complex multi-dimensional measured and simulated data. This emerging panorama has set new challenges around the complex discovery environments in this bursting Data Centered Science where, two concepts are identified as main key requirements for any contemporary scientific finding: *trustworthiness* and *reproducibility*.

Trustworthiness can be associated with the description and traceability of scientific protocols[1] involved in any research, and is the base of the confidence on the answers results obtained in any scientific inquiry. We trust a particular result from a researcher or a research group because their meticulousness adherent to a detail and rigorous protocol supporting the finding.

Reproducibility of a scientific discovery is a very debatable concept which ranges from strict "replication" -as in re-running a simulation in exact detail- through "reproduction" in the sense of independent reproduction of the essential aspects of the experiment [1, 2].

In this paper we shall describe the ecosystem of data tools and services and how they are implemented to help solving the Data Accessibility, Reproducibility and Trustworthiness (DART) challenge in the LAGO (for Latin American Giant Observatory formerly known as Large Aperture Gamma Ray Observatory) Project.

It is organized as follows: the coming section presents the DART Challenge; next, in section 3 the LAGO Project is introduced; LAGOData e-infrastructure and LAGO computational resources/services are discussed in section 4 and 5, respectively; we close with some remarks in 6.

2. The Dart challenge

The DART initiative was launched by CHAIN-REDS[2] (Coordination and Harmonisation of Advanced e-infrastructure for Research and Education Data Sharing): an European Commission co-project focused on promoting and supporting technological and scientific collaboration across different communities in various continents[3, 4]. This initiative provided a set of interrelated tools and services, based on worldwide adopted standards, to provide easy/seamless access datasets, data/documents repositories and the applications that could generate and/or make use of them.

Trustworthiness can be associated to data curation, particularly on the quality of the metadata describing the experimental protocol and data provenance[5], while *reproducibility* and *replicability* are closely connected to the accessibility to data sources and the possibility to manipulate/analyze data contained in them.

[1] We are considering as a scientific *protocol* the detailed specifications for carrying out an experiment, either the detector setup & configuration which condition the registered data or simulation environments that lead to synthetic datasets coming from computational applications[1].

[2] http://www.chain-project.eu

CHAIN-REDS approach to data trustworthiness and reproducibility is based on the integration of computational resources and services, with three main cornerstones:

1. adoption of standards for data discoverability, provenance and recoverability: OAI-PMHfor metadata retrieval, Dublin Core as metadata schema, SPARQL for semantic web search and XML as potential standard for the interchange of data;

2. enablement of datasets authorships and user authentication with the corresponding assignments of specific roles on data services, which can be implemented by two strategies: assignment of Persistent Identifiers (PIDs)[6] to name data in a unique and timeless manner, ensuring that future changes on URIs or internal organization of databases will be transparent to the user; and implementation of a federated identity provision, a secure, flexible and portable mechanism to access e-infrastructures worldwide, based on agreements and standards[3].

3. access to a plethora of computing power to analyze the retrieved data or to contrast them to simulations through an intuitive web-interface. Over the last years, Science Gateways[7, 8] have risen as an ideal tool to allow scientists across the world to seamlessly access different ICT-based infrastructures for research activities to support their day-by-day work and do better (and faster) research.

In other words: user identification, execution of distributed applications with a simple web interface, usage of Open Access Document Repositories/Data Repositories and reproducibility of the experiments conform the DART challenge, where the whole research cycle is covered and can be seamlessly performed by non-expert users, hiding complex processes under simple interfaces and minimizing the need for learning new tools.

3. The LAGO Collaboration

The LAGO project is a collaboration [9] of more than eighty Ibero American astroparticle researchers, motivated by the experience of the Pierre Auger Observatory [10] in Argentina and devoted to study space weather effects [11] and Gamma Ray Bursts (GRB) signals on ground-based detectors [12]. Long-term modulation and transient events can also be characterized by using the LAGO detection network, as it spans over a big area with different sites at different latitudes, longitudes and geomagnetic rigidity cut-offs.

Today it is a network of 10 ground-based Water Cherenkov Detector (WCDs) [13], located at different altitudes from Mexico through Patagonia, and other detectors are expected to be up and running by 2016 including two WCD in the Antarctica Peninsula[14]. These data are of interest for two different scientific communities:

- Gamma Astronomy Community: The LAGO WCDs installed at high altitude sites are sensitive to detect the effects of GRB. A significant number of LAGO detectors are over 3,000 m.a.s.l. and three of them are above 4,300 m.

- Space Weather: The LAGO energy range covers a plethora of phenomena related to low-energy cosmic rays physics and space weather phenomena. Nowadays, the study of such

phenomena is crucial because levels of radiation in the atmosphere and near-Earth space environment may be established.

Since the LAGO data analysis needs to take into account the influence of atmospheric effects -such as pressure or even air temperature- on the flux of particles at the detector level, each LAGO WCD is equipped with several environmental sensors. These measurements represent an opportunity to provide environmental information to other communities, like ecologists studying the high altitude environments to correlate for possible climate change and global warming effects. Additionally, since the data in the LAGO repositories is open and freely accessible, it is used as motivation to train the general public -mainly the secondary school teachers and students- in statistical data analysis and related techniques, and raising the awareness of the general public in global warming and climate change impact in everyday life. This important citizen science initiative is one of the main objectives of the LAGO collaboration and is implemented in the so-called LAGO-CS (Citizen Science) program[15].

4. LAGOData e-infrastructure

Typically, each detector generates 150 GB of data per month and the entire collaboration generates 1.5 TB/month. The LAGO dataset not only refers to those measured by WCD detectors but also to data generated by simulation of cosmic rays phenomena in the $\sim 2\,\mathrm{GeV}$ to $\sim 1\,\mathrm{PeV}$ energy range, by using CORSIKA (COsmic Ray SImulations for KAscade)[16], a software for detailed simulation of extensive air showers initiated by high energy cosmic ray particles. The CORSIKA particle flux simulations carried out generate 50GB/site and these synthetic data are also preserved in the data repository. The low energy limit depends on the geomagnetic coordinates of the site, while the high energy limit is determined by the collection area at each site and is limited by statistics as the flux becomes lower and lower at higher energies.

This raw data collected by the LAGO detectors and produced by the simulation tools developed are shared trough LAGOData[17], a platform conceived to promote data curation and sharing among LAGO collaborators, which is part of a more ambitious project, LAGOVirtual[18] a working environment which ensure access to the data recorded in all LAGO Sites and facilitate the analysis of such data.

4.1 Data curation through DSpace

Dspace is an open source software that enables sharing of many types of content, it is generally used for institutional repositories, providing basic functionality for saving, storing, and retrieving of digital content. DSpace was adopted for the LagoDATA repository, because it hosts Dublin Core metadata with a straightforward adaptability for non-native metadata schemes. It also supports two important interoperability protocols: OAI-PMH (Open Archive Initiatives Protocol for Metadata Harvesting[3]) and SWORD (Simple Webservice Offering Repository Deposit[4]). The OAI-PMH protocol at the LAGO repository allows the CHAIN-REDS Knowledge Base search engine to navigate into LAGO curated data.

[3]http://www.openarchives.org/pmh/

[4]http://swordapp.org/

We have overcome one of the most important DSpace limitations: its inability to upload/download multiple records. Dspace offers the possibility to upload the corresponding metadata through a command line option via an *import tool* by using *simple archive format* and including it in a separate way. We have developed a script to ingest data profiting from the above mentioned DSpace capability and found that, the assignment of the PID ID to each dataset become a significant overload to the ingest process and we have solved this difficulty by lowering the number (≈ 10) of datasets to be ingested in each group.

4.2 LAGOData metadata

The Dublin Core metadata element set is a standard for cross-domain information resource description, it is elaborated and sponsored by DCMI (Dublin Core Metadata Initiative ,the implementation of which makes use of XML. Dublin Core is Resource Description Framework based and comprises fifteen metadata elements: Title, Subject, Description, Source, Language, Relation, Coverage, Creator, Publisher, Contributor, Rights, Date, Type, Format, and Identifier. Despite this functionality is mostly centered on the Dublin Core metadata scheme, the additional non-native metadata can be configured as custom fields which are also stored, searched and displayed as the native ones.

The datasets are classified into two different types: simulated and measured data, with their corresponding associated metadata:

measured metadata scheme: **data** corresponds to the version/type of the Digital/Analog electronic board; **site** contains the *name*, *latitud*, *longitude*, *altitude*, *type*, *geometry* of the detector and *orcid-id* of the person in charge of the site operations; **voltage** indicating the voltage of polarization of each one of the independent PMTs controlled by the electronic board; **trigger** with the level of trigger and subtrigger of each signal channel; and **sensor** describing type of sensor and calibration constant for each other physical variables measured on board the detector.

simulated metadata scheme: **primary**, corresponding to the CORSIKA `.dbase` output file; **site** with the *latitude*, *longitude* and *altitude* of the site; **libraries** indicating the compiling CORSIKA options and libraries, and the version of the LAGO scripts used for this particular calculation; and **computation**, describing the computational environment given by the standard unix commands: `uname -a`, `lsb_release -a`, `free` and `gcc -v`.

4.3 PID and LAGOData

The main interface to register and manage PID services for European Research Communities is EPIC (European Persistent Identifiers Consortium[5]) which is based on the Handle System [6] for the allocation and resolution of persistent identifiers. There are several compatible "flavors" of

[5] http://pidconsortium.eu

[6] The Handle System (http://www.ietf.org/rfc/rfc3650.txt) provides an efficient, extensible, and secure resolution mechanism for unique and persistent identifiers of digital objects. The Handle System includes an open set of protocols (http://hdl.handle.net/4263537/4086), a namespace (http://hdl.handle.net/4263537/4068) and a reference implementation (http://handle.net/download.html) of the protocols.

PIDs. The most common is DOI PID[7], which is the most frequently used for publications while the EPIC PIDs cover a wider range of Digital objects.

The GRNET PID service enables the allocation, management and resolution of PIDs and has been employed to ensure the data persistence and reproducibility of the experiments. It supports the use of part identifiers as they are provided by the Handle system. Part identifiers can compute an unlimited number of handles on the fly, without requiring registering each separately.

4.4 SWORD and LAGOData

The SWORD (Simple Web-service Offering Repository Deposit), based on the Atom Publishing Protocol (AtomPub), was first developed to standardize a deposit interface to digital repositories but now, it has been further extended to support the whole deposit lifecycle, i.e. deposit, update, replace, and delete resources [19]. Many interfaces of laboratory equipment allow automatic capture of results in an information system and SWORD permits to upload data directly into a repository, without human intervention, tagging as metadata the conditions in which the data was collected.

A Java SWORD Client has been developed -and is part of the firmware of the WCD electronics- to send each recorded dataset to the DSpace Repository, stamping the basic set of metadada describing the status of the WCD at this moment, i,e. data, site, voltage, TrigLevel, and sensor.

5. LAGO advanced computing e-infrastructure

In addition to CORSIKA, LAGO uses intensively other important codes: MAGNETOCOSMIC, GEANT4, ROOT and specific self-designed statistical codes for data analysis, focusing most of the collaboration activities (research & outreach) on a data repository.

Thanks to the cooperation of CHAIN REDS, the advanced computational e-infrastructure for the LAGO project was implemented, creating a fully dedicated Virtual Organization called *lago-project* and integrating it into the European Grid Infrastructure (EGI) activities. The Grid implementation of CORSIKA was deployed in two "flavors" being able to run by using GridWay Metascheduler [20] or through a Catania Science Gateway interface[8]. We have crosschecked the performance of both procedures and found that GridWay approach seams to be more appropriated for massive systematic running while CORSIKA Science Gateway implementations is useful for quick exploratory or specific studies./

In the Science Gateway approach a user can seamlessly run a code on different infrastructures by accessing a unique entry point with an identity provision. He/she only has to upload the input data or use a PID to reference it and click on the run icon. The final result will be retrieved whenever the job will be ended. The underlying infrastructure is absolutely transparent to the user and the system decides on which sites and computing platform the code is performed. More than 21,000 CPU hours have been consumed from Oct 2014 by CORSIKA-GRID jobs executed via the this last interface.

[7]The Digital Object Identifier (DOI) System (http://www.doi.org/) is a service, operated by the International DOI Foundation (IDF), which provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers on digital networks.

6. Conclusions

LAGO is an experiment that could handle, with reasonable scale, a distributed community, collaborating across Latin America, building a network of data repositories through the continent, using computational intensive applications and developing an outreach program to promote Data Science as a Citizen Science initiative.

LAGO aims to study cosmic rays in the energy range $\sim 1\,\text{GeV}$ to $\sim 100\,\text{TeV}$. In this energy range there emerge phenomena related to the physics of low-energy cosmic rays, and also to solar activity and space weather. Nowadays it is crucial to study of these effects because it may establish levels of radiation in the atmosphere and near-Earth space environment. Thus the data repository (and the network of data repositories) will be of interest not only for the LAGO or even the cosmic ray community but useful for the solar physics and space climatology communities. It is then expected that its adoption of the DART Challenge will have important impact, as it will extend the use of LAGO resources (computational and data) for a much wider community.

As has been demonstrated, the implementation of DART applications and methodology has lead to a clear improvement in LAGO robustness, usability and visibility, leading to a wide adoption of the new tools/services by the final research users.

Acknowledgments

The LAGO Collaboration is very thankful to the Pierre Auger Collaboration for its continuous support. We are also in debt to the EC co-funded project CHAIN-REDS (GA 306819) and to the GRNet Technical staff for the implementation of EPIC PID. One of us (LAN) gratefully acknowledges the financial support from CDCHT-ULA project C-1598-08-05-A and Vicerrectoría Investigación y Extensión Universidad Industrial de Santander. This work was also partially supported by Universidad San Francisco de Quito and CEDIA through CEPRA IX research grants.

References

[1] J. Cooper, J. O. Vik, and D. Waltemath. A call for virtual experiments: accelerating the scientific process. *Progress in biophysics and molecular biology*, 117(1):99–106, 2015.

[2] R.D. Peng. Reproducible research in computational science. *Science (New York, Ny)*, 334(6060):1226, 2011.

[3] R. Barbera, B. Becker, C. Carrubba, G. Inserra, S. Jalife-Villalón, C. Kanellopoulos, K. Koumantaros, R. Mayo-García, L.A. Núñez, O. Prnjate, R. Ricceri, M. Rodríguez-Pascual, A. Rubio-Montero, F. Ruggieri, and CHAIN-REDS Project. A chain-reds solution for accessing computational services. In *Actas Cuarta Conferencia de Directores de Tecnología de Información, Gestión de las TICs para la Investigación y la Colaboración*, volume TICAL2014, pages 15–24, 2014.

[4] R. Barbera, B. Becker, C. Carrubba, G. Inserra, S. Jalife-Villalón, C. Kanellopoulos, K. Koumantaros, R. Mayo-García, L.A. Núñez, O. Prnjate, R. Ricceri, M. Rodríguez-Pascual, A. Rubio-Montero, F. Ruggieri, and CHAIN-REDS Project. Chain-reds dart challenge. In *ANAIS DAS SESSÕES TEMÁTICAS E PÔSTERS*, page 166, 2014.

[5] Y.L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *ACM Sigmod Record*, 34(3):31–36, 2005.

[6] J. Hakala. Persistent identifiers: an overview. *KIM Technology Watch Report*, 2010.

[7] J. Alameda, M. Christie, G. Fox, J. Futrelle, D. Gannon, G. Hategan, M. Kandaswamy, G. von Laszewski, M.A. Nacar, M. Pierce, E. Roberts, C. Severance, and Thomas M. The open grid computing environments collaboration: portlets and services for science gateways. *Concurrency and Computation: Practice and Experience*, 19(6):921–942, 2007.

[8] R Barbera, M. Fargetta, and R. Rotondo. A Simplified Access to Grid Resources by Science Gateways. In *The International Symposium on Grids and Clouds and the Open Grid Forum*, Taipei, Taiwan, March 2011.

[9] H. Asorey, S. Dasso, and the LAGO Collaboration. Lago: the latin american giant observatory. In *The 34th International Cosmic Ray Conference*, volume PoS(ICRC2015), page 247, 2015.

[10] The Pierre Auger Collaboration. The Pierre Auger Cosmic Ray Observatory. *Submitted to Nucl. Inst. Meth. A. arXiv:1502.01323.*, 2015.

[11] H. Asorey, S. Dasso, L.A. Núñez, Y. Pérez, C. Sarmiento-Cano, M. Suárez-Durán, and for the LAGO Collaboration. The LAGO space weather program: Directional geomagnetic effects, background fluence calculations and multi-spectral data analysis. In *The 34th International Cosmic Ray Conference*, volume PoS(ICRC2015), page 142, 2015.

[12] H. Asorey, P. Miranda, A. Núñez-Castiñeyra, L.A. Núñez, J. Salinas, C. Sarmiento-Cano, R. Ticona, A. Velarde, and the LAGO Collaboration. Analysis of background cosmic ray rate in the 2010-2012 period from the LAGO detectors at chacaltaya. In *The 34th International Cosmic Ray Conference*, volume PoS(ICRC2015), page 414, 2015.

[13] I. Sidelnik and the LAGO Collaboration. The sites of the latin american giant observatory. In *The 34th International Cosmic Ray Conference*, volume PoS(ICRC2015), page 665, 2015.

[14] S. Dasso, A.M. Gulisano, J.J. Masías-Meza, H. Asorey, and the LAGO Collaboration. A project to install water-cherenkov detectors in the antarctic peninsula as part of the LAGO detection network. In *The 34th International Cosmic Ray Conference*, volume PoS(ICRC2015), page 105, 2015.

[15] H Asorey, L.A. Núñez, and C Sarmiento-Cano. Exposición temprana de nativos digitales en ambientes, metodologías y técnicas de investigación en la universidad. *arXiv preprint arXiv:1501.04916*, 2015.

[16] D. Heck, J. Knapp, J.N. Capdevielle, G. Schatz, and T. Thouw. Corsika : A monte carlo code to simulate extensive air showers. Technical Report FZKA 6019, Forschungszentrum Karlsruhe GmbH, 1998.

[17] L.A. Torres, L.A. Núñez, R. Torréns, and E.H. Barrios. Implementación de un repositorio de datos científicos usando dspace. *E-Colabora*, 1(2):101–117, 2011.

[18] R. Camacho, R. Chacón, G. Díaz, C. Guada, V. Hamar, H. Hoeger, A. Melfo, L. A. Núñez, Y. Perez, C. Quintero, M. Rosales, R. Torrens, and LAGO Collaboration. LAGOVirtual: A collaborative environment for the large aperture grb observatory. In R. Mayo, H. Hoeger, L. Ciuffo, R. Barbera, I. Dutra, P. Gavillet, and B. Marechal, editors, *Proceedings of the Second EELA2 Conference Choroní Venezuela*, Madrid España, 2009. EELA2, CIEMAT.

[19] S. Lewis, P. de Castro, and R. Jones. SWORD: Facilitating Deposit Scenarios. *D-Lib Magazine*, 18(1-2), January 2012.

[20] E. Huedo, R.S. Montero, and I.M. Llorente. The gridway framework for adaptive scheduling and execution on grids. *Scalable Computing: Practice and Experience*, 6(3), 2001.