IV Conferencia International Biredial - ISTEC
'14
Acesso Aberto, Preservação Digital, Interoperabilidade, Visibilidade e Dados Científicos

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

CHAIN-REDS DART Challenge

Roberto Barbera[a,b], Bruce Becker[c], Carla Carrubba[b], Giuseppina Inserra[b], Salma Jalife Villalón[d], Christos Kanellopoulos[e], Kostas Koumantaros[e], Rafael Mayo-García[f], Luis A. Núñez[g], Ognjen Prnjat[e], Rita Ricceri[b], Manuel Rodriguez Pascual[f], Antonio Rubio-Montero[f], Federico Ruggieri[h,i] (On behalf of the CHAIN-REDS project)

[a] Dpt. of Physics and Astronomy, University of Catania, Viale A. Doria 6, Catania 95125, Italy

[b] National Institute of Nuclear Physics-Catania, Via S. Sofia 64, Catania 95123, Italy {roberto.barbera, carla.carrubba, giuseppina.inserra, rita.ricceri}@ct.infn.it

[c] Meraka Institute, Meiring Naudé Road, Pretoria, 0001, South Africa bbecker@csir.co.za

[d] Corporación Universitaria para el Desarrollo de Internet, Parral No. 32 Col. Condesa 06140 Mexico D.F., Mexico salmajalife@cudi.edu.mx

[e] Greek Research and Technology Network, 56 Mesogion Av., Athens, 11527, Greece {skanct, kkoum, oprnjat}@admin.grnet.gr

[f] Centro de Investigaciones Energéticas Medioambientales y Tecnológicas, Av. Complutense 40, Madrid, 28040, Spain {rafael.mayo, manuel.rodriguez, antonio.rubio}@ciemat.es

[g] Escuela de Física, Universidad Industrial de Santander, Cra 27 calle 9, Bucaramanga, Colombia lnunez@uis.edu.co

[h] Consortium GARR, Via dei Tizii 6, Rome, 00185, Italy

[i] National Institute of Nuclear Physics-Roma Tre, Via Vasca Navale 84, Rome, 00146, Italy federico.ruggieri@roma3.infn.it

ABSTRACT

CHAIN-REDS (Coordination and Harmonisation of Advanced e-infrastructure for Research and Education Data Sharing) is EU project focused on promoting and supporting technological and scientific collaboration across different communities established in various continents. Nowadays, one of the most challenging scenarios scientist and scientific communities are facing is huge amount of data emerging from vast networks of sensors and form computational simulations performed in a diversity of computing architectures and e-infrastructure. The new knowledge coming out from the interpretation of these datasets, reported on the scholar literature, is increasingly problematic to be reproducible due to the difficulty to access measured data repositories and/or computational applications that generate synthetic data through computer simulations. This paper presents CHAIN REDS approach, several tools and services, based on the adoption of standards, aimed at providing easy/seamless access to datasets, data repositories, open access document repositories and to the applications that could make use of them. All these tools and services are enclosed in what we

have called the Data Accessibility, Reproducibility and Trustworthiness (DART) challenge. This initiative allows researchers to easily find data of his interest and directly use them in a code running by means of a Science Gateway (SG) that provides access to cluster, Grid and Cloud infrastructure worldwide. In this scenario, the datasets are found by means of either the CHAIN-REDS Knowledge Base (KB) or the Semantic Search Engine (SSE), the applications ran on the CHAIN-REDS SG, accessible through an Identity Federation. The datasets can be both identified by Persistent Identifier (PID) and assigned unique number ID. Scientists can then access the data and the corresponding application in order to either reproduce and extend the results of a given study or start a new investigation. The new data (and the new paper if any) are stored on the Data Infrastructure and can be easily found by the people belonging to the same domain making possible to start the cycle again.

Keywords: Global research; Open access; Data management; Standards; Identity provision; e- Infrastructure.

INTRODUCCIÓN

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. All disciplines, either physical, life sciences and humanities are becoming increasingly data-driven and data intensive. This is happening mainly due to technological advances in information networks, computing capacity, big instruments, penetration by sensors in all areas, as well as increasingly collaborations among researchers. Big Data Science requires interdisciplinary skills in which computer scientists, statisticians and other experts combine their knowledge to create new techniques, tools and methodologies, shifting from a hypothesis-driven to a data-driven way to analyse the increasing datasets. Astronomy is the most pioneer data-driven science and its communities are early adopters and creators of multiple discovery environment incorporating strategies and tools to manipulate and analyse huge amounts of data.

Large data productions are usually carried out by global collaborations, i.e., multinational science groups that generate large volumes of data, geographically distributed and maintained only during the project life cycle. Most of these data is never published and, when the collaborations end, many is lost or stashed away in national (or international) reservoirs that have nothing to do with their origins. Production decisions, approximations and provenance are buried in a huge electronic correspondence to which no-one has access.

IV Conferencia International Biredial - ISTEC
Acesso Aberto, Preservação Digital, Interoperabilidade, Visibilidade e Dados Científicos

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

A similar path is followed by small data producers scattered around the globe. Both large and small data producers face the same problems in knowledge cataloguing, preservation and dissemination. It is imperative to plan and build repositories that store data as they emerge and to retain the history of the decisions and criteria that generate them. Starting the century several multilateral organizations and planners in Europe and the United States generated technical reports to encourage the preservation of important scientific data collections. Recently, most of these recommendations have rooted as national and multinational initiatives for general policies concerning data curation. However, many of these recommendations have not permeated into the producing communities and/or to the collection custodians in these countries. The situation is even worse in Latin America where we are still not convinced by, or at least aware of, the new paradigms in the production and dissemination of scientific knowledge, and consequently, only a low-level use of Information and Communication Technology (ICT) awareness has been incorporated in teaching and research.

In this emerging context CHAIN-REDS (Coordination and Harmonisation of Advanced e-infrastructure for Research and Education Data Sharing) is EU project focused on promoting and supporting technological and scientific collaboration across different communities established in various continents. To do so, it is essential to promote instruments and practices that can facilitate their inclusion in the community of users, i.e. the use of standards. Then, to build on the best practices currently adopted in Europe and other continents, and promote and facilitate interoperability among different e-Infrastructures is a must.CHAIN-REDS, in accordance with several European strategies, plans to focus on including low-level services, exchanging in data infrastructures and support preservation and data exploitation services, as well as activities aimed at interoperability and data access federation and openness. Addressing basic issues such as data persistency, accessibility and interoperability will be the first general goal.

As a main issue, the efficient access, use and further analysis of Data has emerged. The number of Data Repositories (DRs), either Open access ones (OADRs) or not, and the quantity of TB they store have largely increased in the latest years. As a consequence, if CHAIN-REDS aims to allow VRCs, research groups and single researchers to efficiently use worldwide distributed resources, it is needed that the data they are employing will be interoperable as well. Otherwise, advances made on middleware interoperability will result meaningless since the computational resources will not be properly exploited.

This paper we shall describe the ecosystem of the main CHAIN REDS data tools and how this ecosystem of tools and services are used to help solving what we have called the Data Accessibility, Reproducibility and Trustworthiness (DART) challenge.

IV Conferencia International Biredial - ISTEC

Acesso Aberto, Preservação Digital,
Interoperabilidade, Visibilidade
e Dados Científicos

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

INTEGRATION OF DATA RELATED CAPABILITIES

In the current days, there have been extraordinary advances in the network and computational capacities. Just to mention a few correlated ones, academic networks have made available distributed infrastructure as grid or cloud and Infiniband links have deeply increased the parallel performances. Nevertheless, a major new challenge has arisen due to the huge amount of computational calculations and services that have been made: the management of data. Both in academic and scientific fields, the stored data have dramatically increased and, even more, their use is demanded by more and more people. This fact can be easily showed: the number of Data Repositories (DRs) and Open Access Document Repositories (OADRs) and the volume of data they store have largely increased in the latest years. As a consequence, it is necessary that the data will be easily used as well; furthermore, when these data concern datasets and publications. Otherwise, advances made on middleware interoperability will be meaningless since the computational resources will not be properly exploited. In this regard, CHAIN-REDS is promoting interoperability as a main objective and a worldwide demo has been recently shown in September 2013 covering different regions worldwide.

To achieve DART, several tools have been implemented by the project. The Knowledge Base (KB) provides information about the deployment of e- Infrastructure related topics per country and even about specific Distributed Computing Infrastructures (DCIs) by means of a Site or a Table view. During 2013, the project has been working on extending the CHAIN KB with information related to data infrastructure. To do so, it has collected both issues and best practices and has surveyed the involved regions in order to discover data repositories. The reason for that is to promote data sharing across different e-Infrastructure and continents widening the scope of the existing CHAIN KB to Data Infrastructure and to finally provide proof-of principle use-cases for data sharing across the continents.

Before describing the CHAIN-REDS tools we shall present in the next section, two important initiatives strategies closely related to link data to other type of digital contents.

PROMOTING IDENTITY PROVISION AND DATA STANDARDS

One of the major challenges in providing services is how the users access them, i.e. how they authenticate themselves and which roles are allowed to assume over that services. For many years, databases with information on users (username and password) have been provided by the service managers and more restricted solutions have been taken also into account such as that implemented to access Grid computing (personal certificates provided by an accredited Certification Authority). Nevertheless, those solutions have usually driven the users to a wide set of usernames and password pairs (with the difficulty of remembering all of them). In the latest years, the concept of Identity provision has emerged as a valid solution. Furthermore, such a concept is of importance in the

IV Conferencia International Biredial - ISTEC

**Acesso Aberto, Preservação Digital,
Interoperabilidade, Visibilidade
e Dados Científicos**

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

Academia, where every student, professor or administrative staff has his/her username and password as he/she becomes part of an institution. Thus, being a University, an R&D Centre or an NREN accredited by an Identity Provider, a huge pool of services can be accessed by a single user with only a pair of associated credentials (username and password).

Such access has been successfully demonstrated in Grid computing, where no more personal certificates are needed and where now robot certificates can manage jobs for a long period of time, but it can be also applied to academic services for students such as those closely related to a Faculty Secretariat or to their academic record just to mention a few.

In addition to foster identity provision CHAIN-REDS has selected to promote the following standards for pursuing DART initiative:

- OAI-PMH[6] for metadata retrieval.
- Dublin Core[7] as metadata schema.
- SPARQL[8] for semantic web search.
- XML[9] as potential standard for the interchange of data represented as a set of tables.

To those, Persistent Identifiers (PID)[10] must be added as a tool to know where and how data and metadata are stored. Such a circumstance is achieved by assigning an identifier to a digital object, i.e. as d.o.i works for articles.

CHAIN-REDS TOOLS

In this Section, we will present a brief description of the CHAIN-REDS tools that are available at its website and are of interest to Data Infrastructure. They actually are the backbone that is being used by the DART challenge to achieve data trust building and will be the basis for several of the use cases coming from the regions.

THE CHAIN-REDS KNOWLEDGE BASE

The CHAIN-REDS Knowledge Base is one of the largest existing e-Infrastructure-related digital information systems. It currently contains information, gathered both from dedicated surveys and other web and documental sources, for largely more than half of the countries in the world.

In principle, the KB was implemented considering e-Infrastructure as an environment where research resources (hardware, software and content) can be readily shared and accessed where necessary to promote better and more effective research. Then, such environments integrate hard-, soft- and middleware components, networks, data repositories, and all sorts of support enabling virtual research collaborations with a final goal: to allow scientists across the world to do better (and faster) research using DCIs, independently of where they are and of the paradigm(s) adopted to achieve their goals. Thus, to better fulfill this long term milestone, the use of standards is more than an asset and, at the same time, a step forward to achieve sustainability.

The first release of the CHAIN KB presented integrated dynamically updated information about DCIs. Information about Regional and National Research and Education Networks, National Grid Initiatives, Certification Authorities, Identity Federation Providers, Regional Operation Centers, Grid sites and Applications (and already running on a Science Gateway) was available in both Country and Table views. This on-line service was a clear step forward in harmonizing the different regional infrastructure information, new capabilities should be incorporated. Once standards were identified to easily gather and access both OADRs and DRs, a demonstrator was built with them to visualize and access the repositories by means of both geo- and tab-views (as it was previously made for DCIs). Such a demonstrator was implemented with the advances carried out within the Knowledge Linking and sharIng in research dOmainS (KLIOS) project. KLIOS is based on the interconnection and the integration of scientific resources through a grid of meta-data network and provide the following services: metadata harvesting; semantic enrichment; and, linked data semantic search.

Basically, the new KB capability is composed of a multi-layer structure where two harvesters running on either Grid or Cloud search for OAI-PMH end-points from OADRs and DRs. Above them, a semantic web-enrichment layer is used to act as a previous step before the linked-data search engine, which is on the top. The process of the metadata harvesters is as follows:

- Get the address of each repository publishing an OAI-PMH standard endpoint
- Retrieve, using the OAI-PMH repository address, the related Dublin Core encoded metadata in XML format
- Get the records from the XML files and, using the Apache Jena API, transform the metadata in RDF format

Figure 1. A snapshot of the CHAIN-REDS Knowledge Base - OADR Site view (red markers refer to repositories found by automatic harvesting and yellow ones o those directly integrated by means of CHAIN-REDS activities).

Basically, the new KB capability is composed of a multi-layer structure where two harvesters running on either Grid or Cloud search for OAI-PMH end-points from OADRs and DRs. Above them, a semantic web-enrichment layer is used to act as a previous step before the linked-data search engine, which is on the top. The process of the metadata harvesters is as follows:

- Get the address of each repository publishing an OAI-PMH standard endpoint
- Retrieve, using the OAI-PMH repository address, the related Dublin Core encoded metadata in XML format
- Get the records from the XML files and, using the Apache Jena API, transform the metadata in RDF format

Save the RDF files into a Virtuoso triple store according to an OWL-compliant ontology built using Protégé.

As it has been aforementioned, as a legacy from CHAIN, it has 86 entries in its DCIs Table view, which reports on the country where the DCI is settled, the regional network, the National Research Education Network (NREN) and the National Grid Initiative (NGI) it belongs to, the Certification Authority (CA) and the Identity Federation (IdF) it relies on for accessing it, the Regional Operation Centre (ROC) it is connected to and the sites it counts on. As in the previous case, the same information can be showed searching on a world map.

IV Conferencia International Biredial - ISTEC

Acesso Aberto, Preservação Digital,
Interoperabilidade, Visibilidade
e Dados Científicos

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

THE CHAIN-REDS SEMANTIC SEARCH ENGINE

Even when the KB is searchable in one of the four topics it contains (country, name, domain, organization), the CHAIN-REDS consortium has decided to semantically enrich the OADRs and DRs gathered in the KB and build a search engine on the related linked data. The CHAIN-REDS Semantic Search Engine has been the result of such an effort.
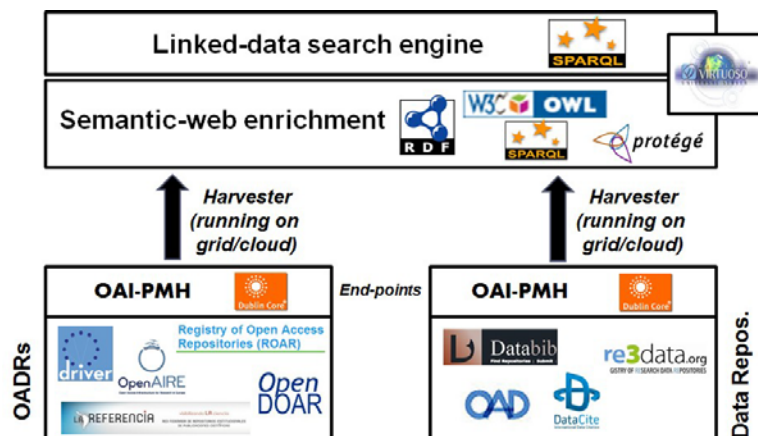
Figure 2. Architecture of the Semantic Search Engine

The multi-layered architecture of this engine is sketched in Figure 2 where both the official and *de facto* Semantic Web standards and technologies adopted are described by small logos.

Using it, visitors can either enter a keyword and submit a SPARQL query to the Virtuoso triple store or select a language and get, on the left side of the page, the list of subjects available in that language with the indication, between parentheses, of the number of records available for that particular subject (see Figure 3).

The results of a given query are listed in a summary view directly displayed on the webpage. For each record found, the title, the author(s) and a short description of the corresponding resource are provided. Clicking on the "More Info" link, visitors can access the detailed view of the resource. In the "Dataset information" panel users get the link to the open access document and, if existing, to the corresponding dataset. Clicking on the "Graphs" tab, which appears at the top of the summary view, users can select one or more of the resources found and get a graphic view of the semantic connections among Authors, Subjects and Publishers, as shown in Figure 4. In this way, if new links appear, connecting different resources (as shown in the lower left corner of the figure), users can infer new relations among resources, thus discovering new knowledge.

IV Conferencia International Biredial - ISTEC
Acesso Aberto, Preservação Digital, Interoperabilidade, Visibilidade e Dados Científicos

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

The technological description of how this process can be made follows. The first pillar is the harvester procedure. Then, each Resource Description Framework (RDF) file retrieved and saved in a Virtuoso-enabled triple store is mapped onto a Virtuoso Graph that contains the ontology expressly developed for the search engine. The ontology, built using Dublin Core and FOAF standards, consists of:• Classes that describe the general concepts of the domain: Resource, Author, Organization, Repository and Dataset (where Resource is a given open access document);

- Object properties that describe the relationships among the ontology classes; the ontology developed for the service described in this paper has several specific properties such as hasAuthor (i.e., the relation between Resources and Authors) and hasDataSet (i.e., the relation between Resources and Datasets)
- Data properties (or attributes) that contain the characteristics or classes parameters.
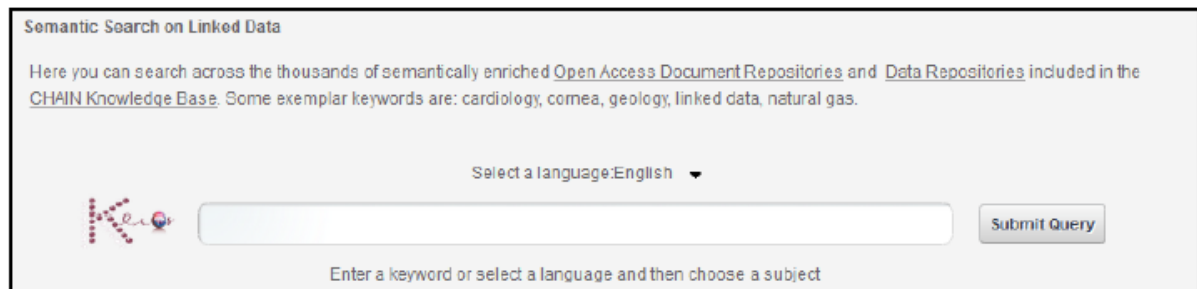

Figure 3. Schema of the ontology used for the Semantic Search Engine

The highest-level, component is the Search Engine itself. Using it, visitors can either enter a keyword and submit a SPARQL query to the Virtuoso triple store or select a language and get, on the left side of the page, the list of subjects available in that language with the indication, between parentheses, of the number of records available for that particular subject. The results of a given query are listed in a summary view directly displayed on the webpage. For each record found, the title, the author(s) and a short description of the corresponding resource are provided. Clicking on the "More Info" link, visitors can access the detailed view of the resource.

In the "Dataset information" panel users get the link to the open access document and, if existing, to the corresponding dataset. Clicking on the "Graphs" tab, which appears at the top of the summary view, users can select one or more of the resources found and get a graphic view of the semantic connections among Authors, Subjects and Publishers. In this way, if new links appear, connecting different resources, users can infer new relations among resources, thus discovering new knowledge.

A programmable use of the CHAIN-REDS SSE is also possible due to the development of a RESTful API that has been created on purpose; now, it is possible to get and/or re-use the many millions of open access resources contained in the CHAIN-REDS KB and stored in a Virtuoso RDF-compliant database by calling the Semantic Search Engine from a common website or even mobile application.

Now it is possible to perform either single or parallel semantic searching[13] [26]. By passing the mouse over the "Semantic Search" link of the CHAIN-REDS webpage, any user can see a sub-menu with several items; the first two are:

- Single: the usual semantic search service described above; and,

- Parallel: the new parallel semantic search service that allow users to search in parallel (i.e., at the same time) across the millions of resources contained in the CHAIN-REDS Knowledge Base and in the ENGAGE Platform [14] . Parallel semantic search engines have been made available also in the SGs of some (collaborating) projects, enhancing and extending in this way the solutions proposed by CHAIN-REDS. This parallel semantic search can be found at:

- agINFRA, here the user can search in parallel across the millions of resources contained in the CHAIN-REDS Knowledge Base and in the OpenAgris[15] repository; and,

- DCH-RP, here the user can search in parallel across the tens of millions of resources contained in the CHAIN-REDS Knowledge Base and in the Europeana[16], Cultura Italia[17] and Isidore[18] repositories.
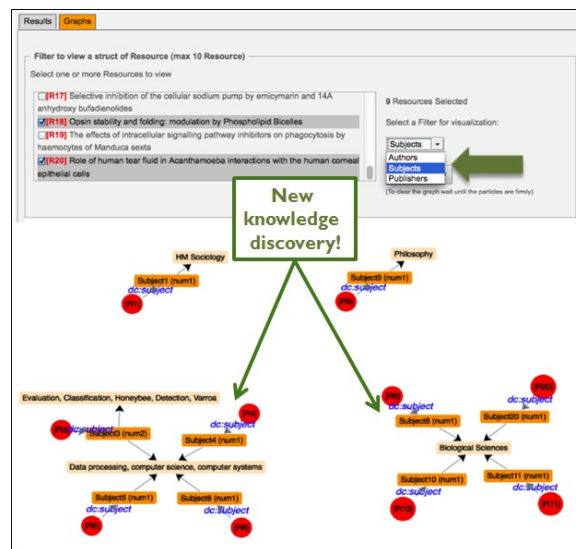

Figure 4. Graphic connections among records found by the Semantic Search Engine

Another two extensions have been the programmable use of the Semantic Search Engine by a RESTful API and the implementation of the engine on a mobile app for both Apple store and Google Play. Any user can access them also by passing the mouse over the "Semantic Search" link of the CHAIN-REDS webpage.

IV Conferencia International Biredial - ISTEC
Acesso Aberto, Preservação Digital, Interoperabilidade, Visibilidade e Dados Científicos

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

THE DART CONCEPT

The vision of the emerging Data Centric Sciece is that a researcher, of a given scientific domain, could find publications and being automatically redirected to the data used to produce those papers and to the applications used to produce those data. Or, alternatively and simply, access raw data of interest to be lately used as input in applications.

Researchers can then access that data and the corresponding application in order to reproduce and extend the results of a given study. The new data (and the new paper if any) are stored on the Data Infrastructure and can be easily found by the people belonging to the same domain making possible to start the cycle again (see the diagram below in Fig 5).The requirements that are needed and not directly managed by CHAIN-REDS are related to intellectual properties issues and unique identifiers (PID) referring to papers, data and applications. Nevertheless, CHAIN-REDS is supporting the assignment of PID to digital objects by means of the service provided by the partner GRNET[19].

CHAIN-REDS has started working on this DART challenge for providing proof-of principle use-cases for data sharing across continents. The first prototype has been successfully tested. It counts on a couple of datasets stored in ZENODO[20] and DataCite[21] and applications that could make use of them. Thus, datasets related to molecular cross sections hosted by the Max-Planck Institute [22] can be downloaded from ZENODO and data on genes to be compared across species hosted by Ensembl [23] can be downloaded from DataCite. The former can be used to obtain molar absorption coefficients by using an application devoted by CIEMAT already included in a Science Gateway portlet called Molon[24] and the latter can be taken as input for several applications, like for example jModelTest2[25], to obtain models of nucleotide substitution.
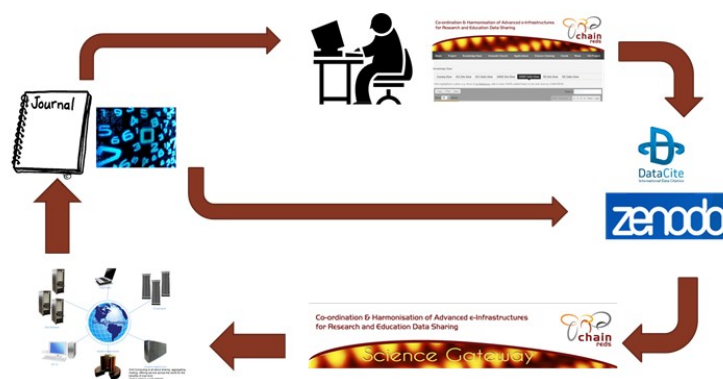

Figure 5. The DART challenge.

This proof-of-concept test is going to be shown and proposed to the CHIAN-REDS collaborative communities for their own use with their own datasets and applications. These communities come from the Agriculture, e-Government, Earth Science, Cultural Heritage and Astroparticle domains. This concept is also being promoted in the CHAIN-REDS targeted regions in order to find success stories; Latin America the LAGO (Latin American Giant Observatory) Collaboration has been identified, and will be described below. At the same time, DART has been proposed to the Workflows Working Group of the EUDAT initiative[26] and now CHAN-REDS representatives are part of it.

LAGO AS A DART EXAMPLE

The Latin America Giant Observatory (LAGO) (first known as *Large Aperture Gamma Ray Observatory*[27]) project is a recent collaboration that comes from the association of more than 80 Latin American astroparticle researchers, keeping a close collaboration with researchers at IN2P3 in France and INFN in Italy. This collaboration was motivated by the experience of the Pierre Auger Observatory, and the idea to install Water Cherenkov Detector (WCDs) in 9 Latin American countries: Argentina, Bolivia, Brazil, Colombia, Ecuador, Guatemala, Mexico, Peru and Venezuela It started in 2005 and it was originally designed to survey the high-energy component of GRBs. Today it is a network of ground-based WCDs, located at different altitudes from Mexico through Antarctica (see Figure 6 below), devoted to study space climatology effects and GRBs signals on ground-based detectors by measuring the variations of the flux of secondary particles at ground level. 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. Presently LAGO collaboration has 10 up/running WCD and it is planned double them in the next two years, with five more new detectors installed in 2014 and other five in 2015. Typically, each detector generates 150 GB of data per month and the entire collaboration generates 1.5 TB/month. This experimental data is preserved locally and shared trough a data repository based on DSpace[28].

Additionally, at each site the particle flux over the ground detector has to be simulated and correlated to the signals emerging from the detectors. The particle flux simulation is carried out using CORSIKA [29] (COsmic Ray SImulations for KAscade) a software for detailed simulation of extensive air showers initiated by high energy cosmic ray particles. CORSIKA is extensively used by the astroparticle community. Particularly, CORSIKA is used by the Auger Observatory [30] collaboration and more recently the HAWC [31] project. Typically CORSIKA simulations generate 10GB/site. These synthetic data is also preserved in the data repository. The collaboration also uses GEANT4[32] to evaluate the response of the instrumentation of the WCD to the crossing of particles through the water volume of the detector.

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.

Inspired by the debutante, the Open Data Movement, LAGO has been developing four main initiatives to deploy an open network of curated data repositories, namely:

1. To preserve, curate and share the data registered by WCDs array, now through data repository[33] (DR) and in the near future across a network of DR
2. To generate a toolkit of scripts and algorithms to detect the proper operation of the detectors. This toolkit will part of the next generation of firmware and will allows the system to reconfigure some of it parameters in order to minimize some of diagnosed malfunctioning.
3. To offer a computational infrastructure that allows the collaborating members to ubiquitously analyse the curated data efficiently
4. To open part of the data and use it to outreach Data Science to university students.

These initiatives aim to pave the way to openly share the data recorded by LAGO collaboration with any other domain disciplines and to Citizen Science initiatives.
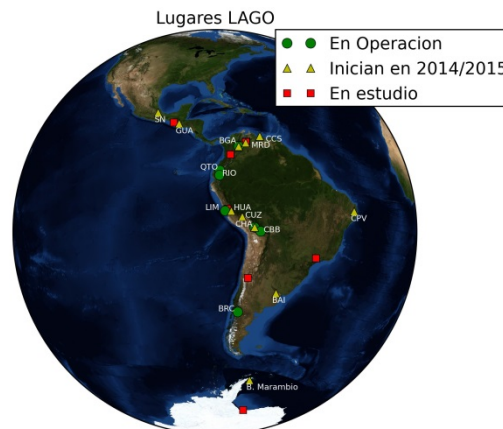

Figure 6. Lago Sites: Green in operation, starting operation on 2014/2015 and red under study/consideration.

IV Conferencia International Biredial - ISTEC

**Acesso Aberto, Preservação Digital,
Interoperabilidade, Visibilidade
e Dados Científicos**

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

LAGO has developed a prototype of data repository, LAGOData[34] as part of a more ambitious project, LAGOVirtual[35] oriented to develop a working environment to have access and analyse data recorded in all LAGO Sites. In LAGO repository data is classified into three types: instrument calibration data, WCD data sets and simulated data sets. In the near future we want the members of the collaboration to use this repository also to preserve papers, thesis, Labs Notes and Technical Reports related to the project. The idea is to link documents to all the data sets analysed in it.

Each data file is tagged by a metadata set specifically adapted to LAGO. The existence and implementation of a scientific metadata standard model will allow an uniform access to data for all the LAGO collaboration members, the interoperability between scientific information systems and also will contribute to the data preservation and its usability in time. The metadata model the collaboration uses for LAGOData is an adaptation of the model raised by the Council for the Central Laboratory of the Research Councils[36].

LAGO repository exposes data/metadata through the Open Archives Initiatives Protocol for Metadata Harvesting[37]. This protocol is used by external systems to collect the data and metadata and create aggregated value services like meta- searchers.

We expect to release a second version of LAGOData by the end of 2014 which will have an implementation of the SWORD (for Simple Web-service Offering Repository Deposit) protocol which became a way to address the need for a standardised deposit interface to digital repositories[38].
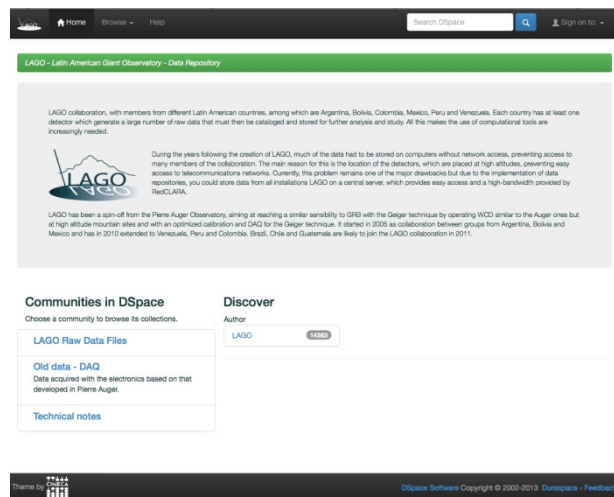

Fig. 2. Screen shot of LAGOData repository http://halley.uis.edu.co/LAGOVirtual/

IV Conferencia International Biredial - ISTEC

Acesso Aberto, Preservação Digital,
Interoperabilidade, Visibilidade
e Dados Científicos

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

An important Data Science outreach activity is carried out by using LAGO Data Repository. LAGO Colombia is developing an intense outreach program on Data Science, teaching statistical data analysis to university students. This initiative is supported by ColCiencias (Colombian Science and Technology funding agency)

Presently, LAGO collaboration has COSIKA and GEANT4 installed at the Universidad de Santander Supercomputing Center[39] and in a small dedicated cluster of a six Workstation (Quad Core Intel Xeon E5520, 4 GB RAM/node and 2 TB DD/node) with NVIDIA Quadro NVS 420 which controls a visualization wall of 16 monitors of 24 inches, capable to generate a resolution of 32 MP. It is foreseen to have access to the CORSIKA grid version which is available in other Astroparticle Observatories and which is supported by the ROC-LA.

CONCLUSIONS

CHAIN-REDS has as a major goal to propose a model for accessing and managing data. To achieve such an objective, several tools based on standards have been implemented and access to these services through identity providers has been relied on. Now, any user can perform a whole cycle of searching data and documents, retrieve the raw data, use them as an input of a service, and obtain final results susceptible of being stored under the same standards format.

For the specific case of Latin America, CHAIN-REDS is closing collaborating with RedCLARA. Thus, it is supporting the SCALAC service and integrating major repositories as La Referencia into the project Knowledge Base. Also, success stories are being aware of the CHAIN-REDS developments and will benefit from them.

In principle, scientific cases have been identified, but all the services promoted by CHAIN-REDS can be adopted by the Academia and the NRENs. In this sense, CHAIN-REDS is much interested in collaborating with this kind of institutions to support them in the use of its services.

ACKNOWLEDGEMENTS

REFERENCES

1. See a classical discussion of this problem of how small labs and institutes are facing the data deluge problems in C Borgman, J Wallis, and N Enyedy.(2007) **Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries** *Int J Digit Libr,* **7,** 17-30.

2. Arzberger, P., Schroeder, P., Beaulieu, A., et al. (2004) **Science and government: an international framework to promote access to data**. *Science*, **303**:1777–1778; Simberloff, D., Barish, B. C., Droegemeier, K. K., et al. (2005) **Long-lived digital data collections: enabling research and education in the 21st century**. Technical Report NSB-05-40, National Science Foundation, Washington DC, USA; and Lyon, L. (2007) **Dealing with data: roles, rights, responsibilities and relationships**. Consultancy Report, UKOLN, University of Bath, UK.

3. See **Report of The European Commission Public Consultation on Open Research Data** (July 2013) http://ec.europa.eu/research/science-society/document_library/pdf_06/report_2013-07-open_research_data- consultation.pdf and **Digital Research Data Sharing and Management** Committee on Strategy and Budget National Science Board NSF Technical Reports (2011) http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

4. CHAIN-REDS interoperability demo, http://science-gateway.chain-project.eu/demo-status and also CHAIN-REDS D4.2, available at http://www.chain-project.eu/deliverables

5. CHAIN-REDS Knowledge Base, http://www.chain-project.eu/knowledge-base

6. OAI-PMH, http://www.openarchives.org/pmh/

7. Dublin Core, http://dublincore.org

8. SPARQL, http://www.w3.org/2001/sw/wiki/SPARQL

9. XML, http://www.w3.org/XML/

10. PID, http://www.pidconsortium.eu/

11. KLIOS, http://klios.ct.infn.it

12. The programmable use of the Semantic Search Engine, http://www.chain-project.eu/semantic-search-api

13. The parallel search in the CHAIN-REDS KB, http://www.chain- project.eu/parallel-semantic-search

IV Conferencia International Biredial - ISTEC

**Acesso Aberto, Preservação Digital,
Interoperabilidade, Visibilidade
e Dados Científicos**

Universidade Federal do Rio Grande do Sul
15 - 17 de octubre de 2014

14. ENGAGE, http://www.engagedata.eu/

15. The OpenAgris repository, http://aims.fao.org/openagris

16. The Europeana repository, http://www.europeana.eu/

17. The Cultura Italia repository, http://www.culturaitalia.it/

18. The Isidore repository, http://www.rechercheisidore.fr/

19. GRNET PID Service, http://epic.grnet.gr/

20. The ZENODO repository, http://zenodo.org/

21. The DataCite initiative, http://www.datacite.org/

22. The MPI-Mainz UV/VIS Spectral Atlas, http://satellite.mpic.de/spectral_atlas

23. Ensembl Gene Trees and Homologues,
http://www.ensembl.org/info/website/tutorials/compara.html

24. Molon portlet, http://science-gateway.chain-project.eu/molon

25. M. Loureiro et al. Grid selection of models of nucleotide substitution. Studies in Health
Technology and Informatics 159, 244-248 (2010) and also in jModelTest2 portlet,
http://science-gateway.chain-project.eu/jmt

26. EUDAT, http://www.eudat.eu/workflows

27. Full author/institution list at http://www.lagoproject.org

28. An open source software that enables open sharing of many types of content, generally used
for institutional repositories http://www.dspace.org

29. https://web.ikp.kit.edu/corsika/

30. http://visitantes.auger.org.ar/index.php/el-observatorio/colaboracion-
internacional.html

32. http://www.inaoep.mx/~hawc/

33. The standard a toolkit to simulates the passage of particles through matter. See http://geant4.cern.ch

34. http://halley.uis.edu.co/LAGOVirtual/

35. 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,
36. 2011

37. R. Camacho, R. Chacón, G. Díaz, C. Guada, V. Hamar, H. Hoeger, A. Melfo, L. A. Núñez, Y. Pérez, C. Quintero, M. Rosales, and R. Torréns. Lagovirtual. a
38. 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,
39. *Proceedings of the Second EELA2 Conferencem Choroní Venezuela*, Madrid
40. España, 2009. EELA2, CIEMAT.

41. Shoaib Sufi and Brian Mathews. Cclrc scientific metadata model: Version 2.
42. *Final report. Council for the Central Laboratory of the Research Councils. Report No: DL-TR-2004-001*, 2004.

43. http://www.openarchives.org/pmh/ also Herbert Van de Sompel, Michael L

44. Nelson, Carl Lagoze, and Simeon Warner. Resource harvesting within the oai- pmh framework. *D-lib magazine*, 10(12):1082–9873, 2004

45. Stuart Lewis, Pablo de Castro, and Richard Jones. SWORD: Facilitating

46. Deposit Scenarios. *D-Lib Magazine*, 18(1-2), January 2012.

47. http://sc3.uis.edu.co