Volumen 1 N. 2
Julio – Diciembre 2011
ISSN: 2027 - 7415
Pág. 101

Implementación de un Repositorio de Datos Científicos usando Dspace

Luis Alejandro Torres
Universidad Industrial de Santander (Grupo Halley de Astronomía y Ciencias Aeroespaciales) - Colombia
luis.torres@correo.uis.edu.co

Luis A. Nuñez
Universidad Industrial de Santander (Grupo de Investigación en Relatividad y Gravitación) - Colombia
lnunez@uis.edu.co

Rodrigo Torréns
Universidad de los Andes (Centro Nacional de Cálculo Científico) - Venezuela
torrens@ula.ve

Edwin Barrios
Colaboración LAGO - Large Aperture Gamma Ray Burst Observatory

Resumen

El nuevo modelo de e-colaboración científica y e-producción de conocimiento imponen el registro, catalogación, preservación distribuida y análisis de grandes volúmenes de datos. El conocimiento que de ellos emerja, se logrará mediante técnicas de minería de datos, que dependerán de los mecanismos de almacenamiento y catalogación utilizados. LAGOVirtual es un proyecto que pretende establecer una plataforma de colaboración para el Large Aperture Gamma Ray Burst Observatory *(LAGO). Este observatorio continental está diseñado para la detección de componentes de alta energía de rayos gamma, mediante la utilización de instrumentos de medición ubicados en sitios de alta montaña en Bolivia, Venezuela, México y Perú. Esta plataforma de colaboración utiliza un repositorio implementado con la herramienta DSpace adaptada para la preservación y diseminación de los datos registrados por los detectores Cherenkov para cada una de las instalaciones.*

Palabras Claves: *Repositorios de datos, preservación de datos científicos, minería de datos, e-colaboración, e-investigación, Gamma Ray Burst.*

Abstract

The new model of e-collaboration and e-science knowledge production impose registration, cataloging, preservation and analysis of distributed large volumes of data. The knowledge that emerges from them will be achieved using data mining techniques, which depend on the storage and cataloging mechanisms used. LAGOVirtual is a project that aims to establish a collaborative platform for the Large Aperture Gamma Ray Burst Observatory (LAGO). This continental observatory is designed to detect high-energy components of gamma rays, using measurement instruments at sites of high mountain in Bolivia, Venezuela, Mexico and Perú. This collaboration platform uses a repository implemented with the tool Dspace adapted for preservation and dissemination of data recorded by Cherenkov detectors for each of the facilities.

Keywords: *data repositories, preservation of scientific data, data mining, e-collaboration, e-research, Gamma Ray Burst.*

1. Introducción

A partir de los años 70 hubo un cambio en el modo de producción del sistema capitalista, de una economía industrial a una informacional. La información ha ido transformando la economía en el mismo sentido que la industria transformó la actividad económica en industrial. La materia prima de esta economía es la información. Ahora es la tecnología actuando sobre la información y no, como antes (durante la Revolución Industrial), cuando la información actuaba sobre la tecnología. La actividad científica y tecnológica no se escapa de convertirse en una e-actividad que diferirá de la que se está desarrollando hoy en día en términos metodológicos, funcionales y, sobre todo, organizacionales para crear y diseminar el conocimiento [1]-[2]. La tendencia en este uso de las Tecnologías de Información y Comunicación, TIC, por parte de la sociedad del conocimiento, proyecta a mediano plazo tener una importancia análoga a la de los servicios de agua y electricidad. De la misma forma que estos servicios impactaron la estructuración de las organizaciones sociales, la teleinformación modificará enormemente la forma como se crea y distribuye el conocimiento. Las TIC se hacen cada vez más ubicuas y de uso intuitivo por parte de una creciente comunidad de usuarios y su utilización ha ido transformado las organizaciones y el modo de vida.

Los términos “ciberinfraestructura”, “e-ciencia” y más recientemente uno más amplio, “e-investigación”, han sido acuñados para describir nuevas formas de producción y diseminación del conocimiento (Ver [3]-[4]-[5] y las referencias allí citadas). Uno de los retos que habremos de enfrentar en esta nueva manera de hacer ciencia es manejar, administrar, analizar y preservar un “diluvio de datos” [6]. Este alud de mediciones convierte a los instrumentos en herramientas informáticas y la experimentación en minería de datos. La avalancha de registros de todo tipo viene generada por experimentos de escala mundial (aceleradores de partículas, red de observatorios terrestres y satelitales e infinidad de los más variados sensores dispersos

geográficamente), los cuales desbordan toda capacidad de manejo que no sea mediante uso intensivo de las TIC.

Quizá no se tenga una conciencia clara de los profundos cambios que habrán de experimentarse en la actividad académica por esa necesidad de manejar y analizar inmensos volúmenes de datos. Es tal la cantidad de información a la cual hoy tienen acceso los estudiantes, que debe plantearse una reflexión en torno a los contenidos y a las metodologías que utilizadas cotidianamente en la formación de estos futuros profesionales. La función de los docentes habrá de focalizarse en la enseñanza de los principios básicos en ciencias y humanidades, proveyendo el adiestramiento necesario para que los estudiantes puedan encontrar en la red la información pertinente y valorar su calidad. Si bien esta parece ser la tendencia, la época actual es de transición entre paradigmas y existen dificultades para la apropiación de estas prácticas por parte de la comunidad académica. Los ingentes volúmenes de datos provenientes de mediciones reales y disponibles a través de la Web, abren inmensas posibilidades para hacer una docencia productora de nuevos conocimientos y, más aún, se comienzan a ver los esfuerzos por utilizar estas herramientas y metodologías de la e-investigación [7]-[8]-[9] en la educación; no obstante, existe una resistencia bien marcada por parte de los mismos investigadores en utilizar las TIC en su docencia cotidiana [9]-[10]-[11]. Esta resistencia dificulta percibir el nuevo panorama y el ingreso a la era informacional.

Los roles tradicionales en la producción, preservación y diseminación de la información han sido trastocados por las escalas de los experimentos. Construir un instrumento de medición puede tomar, típicamente, entre 5 y 10 años; se espera que el experimento esté operativo y productivo, cuando menos, durante ese mismo tiempo. Los productores de datos no son pequeños grupos experimentales, sino colaboraciones industriales multinacionales; los volúmenes de datos son tan grandes que son distribuidos y mantenidos por el proyecto mientras este dure; nunca aparecerán publicados en los artículos que surgen de su análisis para construir o descartar teorías; cuando finalice el experimento, muchos de esos datos se perderán o serán enviados a reservorios nacionales (o internacionales) que nada tuvieron que ver con su producción; más aún, muchos de las decisiones y criterios de producción quedaron escritos en una inmensa correspondencia electrónica de la que nadie dispondrá [7]. Es imperioso planificar y construir repositorios de datos que los almacenen mientras se produzcan, que conserven la traza de las decisiones y criterios que los generaron y que los preserven en el tiempo [7]-[12].

En este trabajo se describirán algunos conceptos y estrategias para la construcción de una red de repositorios de datos para la Colaboración LAGO (por *Large Aperture Gamma Ray Burst Observatory*) en América Latina. Es la continuación de un esfuerzo por garantizar y difundir las iniciativas de acceso libre al conocimiento que hemos venido realizando durante algún tiempo [13]-[14]-[15]-[16]. En la próxima sección se describirán algunos elementos conceptuales que surgen de la reflexión sobre la necesidad de preservar el patrimonio intelectual que contienen los datos de mediciones científicas. Se hará énfasis en los conceptos de metadatos, sus estándares y el ciclo de preservación de los datos experimentales. Más adelante, se describirá la colaboración LAGO y sus alcances. Luego, se expondrá la experiencia en la creación del repositorio y de una red de repositorios de datos. Se describirán las estrategias seguidas y las adaptaciones

que se realizaron a las herramientas estándares para construir repositorios. Finalizaremos con la presentación de algunas conclusiones entre las que mencionaremos como objetivo principal del trabajo, el de construir una comunidad que tenga acceso a un Ambiente de Investigación Virtual (*Virtual Research Environment*) para la colaboración LAGO, que integre en un mismo ambiente un conjunto de herramientas que facilite la detección y el estudio de la los GRB (*Gamma Ray Bursts*); y que incluye el acceso y control de instrumentos, la realización de simulaciones de los detectores, la catalogación y preservación de los datos y la difusión libre de estos hacia toda la comunidad científica que los necesite. Se menciona finalmente como resultado de una primera etapa, el desarrollo de un prototipo del repositorio de datos usando Dspace.

2. Datos, e-investigación

La mayor parte de las investigaciones en Astronomía, Física de Altas Energías, Ecología y Medio Ambiente, Geología, Genética y Biología Molecular, por citar las áreas más relevantes productoras de datos para la e-investigación, están financiadas con fondos públicos. Por ello es de intuir que los datos provenientes de simulaciones y mediciones, y no sólo las publicaciones producidas a partir de ellos, pertenecen a todos los ciudadanos. Los datos pueden ser el comienzo (o corroboración) de las ideas y, consecuentemente, deberían ser de libre acceso para que, con su uso y reutilización se pueda seguir la cadena de producción del conocimiento: Datos - información - Conocimiento - información - Datos. La idea central es que los datos generados por financiamientos públicos son patrimonio de la humanidad y deben estar accesibles y disponibles tan amplia y directamente como se pueda [17]-[18]-[19]. Esta visión contrasta con la actitud de investigadores y grupos de investigación que consideran los datos como su patrimonio y, sobre todo, se enfrenta a la reciente posición de muchos editores, quienes comienzan a exigir los datos que respaldan las publicaciones, haciéndoles extensivo el derecho de reproducción (copyright), con la consecuente restricción para su reutilización. Quizá el acceso a los datos pueda ser limitado si su utilización arriesga la seguridad de individuos o especies, compromete derechos de confidencialidad o viola prerrogativas para su explotación temporal por quienes los recolectaron o generaron [20].

La e-investigación impone un manejo automático de grandes volúmenes de datos. Para ser descubiertos, accedidos y analizados, los datos deben ser fácilmente identificables. En pocas palabras, para que los datos sean útiles, deben ser asequibles y para ello, apegados a estándares acordados por las comunidades productoras. A esa información básica utilizada para describir los datos: su contenido, formato, fechas importantes, condiciones de uso, fuente, propiedad y otras características, se conoce con el nombre de metadatos. Esta información permite al usuario evaluar si determinado conjunto de datos es adecuado para sus fines y facilitar el acceso a la información. Los metadatos pueden ser o no digitales, y los datos a los que están asociados pueden existir en ambas formas. La utilización de metadatos facilita [21]-[22] la identificación y adquisición de datos para un tema determinado y para un período de tiempo o localización geográfica específica; el procesamiento, análisis y modelado automático de los datos; y la incorporación de elementos de conocimiento semántico asociado a los mismos.

Una adecuada documentación sobre el muestreo, procedimientos analíticos, anomalías y calidad de los datos, así como sobre la estructura de las colecciones de datos, ayudará a que ellos sean correctamente interpretados y reinterpretados en el futuro.

La comunidad científica que más rápidamente internalizó el modelo de e-investigación es la de Astrofísica. Tradicionalmente esta comunidad ha operado grandes instrumentos compartidos y generado campañas de observación para uso colectivo, preservando los resultados de estas mediciones en una red de repositorios de datos alrededor del mundo (una lista muy parcial de estos recursos se puede consultar AstroWeb[1]). Por años, esta comunidad ha ido construyendo una plataforma de datos y publicaciones compartidas que ha revolucionado la manera de producir, preservar y diseminar conocimientos, convirtiéndose en una referencia para otras disciplinas [23].

Por su parte, la comunidad de Física de Altas Energías ha comenzado a tomar conciencia de la importancia de preservar los datos que por décadas han registrado ingentes volúmenes de datos en costosos e irrepetibles experimentos, los cuales en gran medida se habrán perdido por falta de estrategias coherentes de preservación. Los costos que permitan el desarrollo de políticas, mecanismos y herramientas para garantizar la clasificación y el acceso a los datos en un mediano y largo plazo, no son previstos en los presupuestos de los proyectos (Ver [24]-[25] y las referencias allí citadas).

Hay casi consenso en presentar un esquema del ciclo de vida de los datos tal y como se representa en la figura 1. Este ciclo se puede esquematizar como:

- **Descubrimiento**: Son los soportes físicos donde los datos han sido almacenados originalmente. Corresponden a los archivos de medición de un determinado instrumento, el cuaderno o, sencillamente, a la bitácora del investigador.
- **Recuperación**: Una vez descubiertos los datos, deben ser normalizados en algún formato y recuperados para su análisis.
- **Análisis**: Cualquiera de los mecanismos que serán utilizados para procesar y analizar los datos para construir información y conocimiento a partir de ellos.
- **Resultados**: A partir del análisis de los datos se obtienen los resultados, generando información y conocimiento.
- **Almacenamiento**. Los datos y los resultados d**e sus anál**isis se almacenan para luego ser catalogados y diseminados.

[1] http://cdsweb.u- strasbg.fr/astroweb.html

- **Catalogación y Publicación**: Parte o todos los datos (medidos, simulados y procesados) podrán diseminarse para ser utilizados y reutilizados por los investigadores. Para ello se impone la catalogación con una metadata apropiada y acordada por la comunidad que los produce.

Figura 1. Ciclo de Vida de los Datos Científicos

Los repositorios y sus redes juegan un papel crucial en cada una de estas etapas, garantizando la preservación, la integridad y el linaje de los datos. Se pueden identificar al menos tres de estos puntos cruciales de los repositorios de datos [26]:

- **Generadores de creación**: Cada vez más, la actividad de I+D se apoya, con mayor énfasis, en reportes técnicos que emergen del modelado y remodelado de datos. Las publicaciones acabadas son vistas como un producto final luego de varios de estos reportes. Esa situación se nota con mayor frecuencia en las grandes colaboraciones. Disponer de repositorios de datos que preserven los distintos resultados del modelado se hace imprescindible.

- **Conectores de Comunidades**: Una red de repositorios genera conexión intra y entre comunidades. Los ambientes de preservación reflejan el tipo de investigación que se está desarrollando y metadatos informan sobre el tipo y calidad de las medidas. Cada vez más la interrelación entre distintas fuentes de datos, proveniente de distintas disciplinas, se convierte en el centro de la actividad para la producción de conocimiento.

- **Curaduría de Datos**: Las redes de repositorios se convierten en bancos de preservación de datos. Las volátiles y frágiles bitácoras de laboratorios o los archivos en sistemas de medición, son transportados y clasificados a sistemas robustos desde donde pueden ser accedidos mucho tiempo

después de que el experimento haya finalizado, e inclusive, de que el grupo de investigación que lo generó se haya disuelto.

3. La colaboración LAGO

El proyecto LAGO (por *Large Aperture Gamma Ray Burst, Observatory*[2]) es una activa colaboración de casi 40 investigadores en 15 instituciones en América Latina, quienes mediante la utilización de detectores de agua Cherenkov, WCD, en alta montaña, buscan detectar destellos gamma, GRB, (figura 2). Esta comunidad, derivada de las experiencias del Observatorio Pierre Auger, viene colaborando desde hace más de seis años, plantando WCD en varias regiones de nuestra América Latina.

Figura 2. La Colaboración LAGO

[2] http://particulas.cnea.gov.ar/experiments/ lago/

La colaboración LAGO dispone todas las características para desarrollar una comunidad virtual, geográficamente distribuida, que coopere en torno a un proyecto de e-Astronomía en América Latina. Es una experiencia de colaboración emergente, surgida de la asociación de investigadores latinoamericanos, que disponen de un instrumental distribuido en 6 países (Argentina, Bolivia, Colombia, México, Perú y Venezuela). La dispersión geográfica de los detectores y lo esporádico de los GRBs, impone que se deban correlacionar las mediciones de las distintas instalaciones LAGO. Por lo tanto, los datos registrados deben ser catalogados, preservados y, sobre todo, compartidos entre los distintos grupos de la colaboración. Adicionalmente, es necesaria para la operación de los WCD, la simulación y registro de datos tanto de la geometría de los tanques como de las condiciones de la atmósfera, para cada una de las instalaciones que son parte del proyecto. Estas simulaciones son muy particulares y corresponden a cada tanque en cada instalación geográfica, para que con estas se pueda identificar con claridad el origen de las trazas registradas por los WCD. Estos datos sintéticos, generados por aplicaciones estándares como Aires (por *AIRshower Extended Simulations*[3], [27]) y GEANT4 (por *GEometry ANd Tracking*[4], [28]), también deben ser catalogados y almacenados para ser luego contrastados con los registros.

3.1 La red de repositorios LAGODATOS

La red de Repositorios LAGODATOS consiste en una infraestructura diseñada para proveer a los investigadores de la colaboración LAGO de una herramienta para preservar, catalogar y difundir los datos (registrados y simulados) generados en cada una de las instalaciones, así como difundir publicaciones generadas por cada uno de los grupos de trabajo de LAGO. Esta red, actualmente formada por 6 países de Latinoamérica, genera una gran cantidad de datos a partir de los WCD conectados al sistema. Su plataforma de funcionamiento está basada en la plataforma para repositorios DSpace, a la cual se le han realizado adaptaciones con el fin de convertirlo en la herramienta de gestión de las colecciones de datos generados.

LAGODATOS funciona actualmente con un servidor principal que se encarga de catalogar y almacenar los datos adquiridos en cada nodo del sistema, permitiendo la e-colaboración de los diferentes países y proporcionándole a cada uno de ellos acceso inmediato a los eventos (GRBs) que puedan ser detectados en cada zona geográfica. Cada nodo tiene su propio sistema de adquisición en donde los datos son almacenados y posteriormente enviados al servidor principal, por lo que cada país miembro mantiene su autonomía y un respaldo de los datos generados por su infraestructura.

[3] http://www.fisica.unlp.edu.ar/auger/aires/
[4] http://geant4.cern.ch/

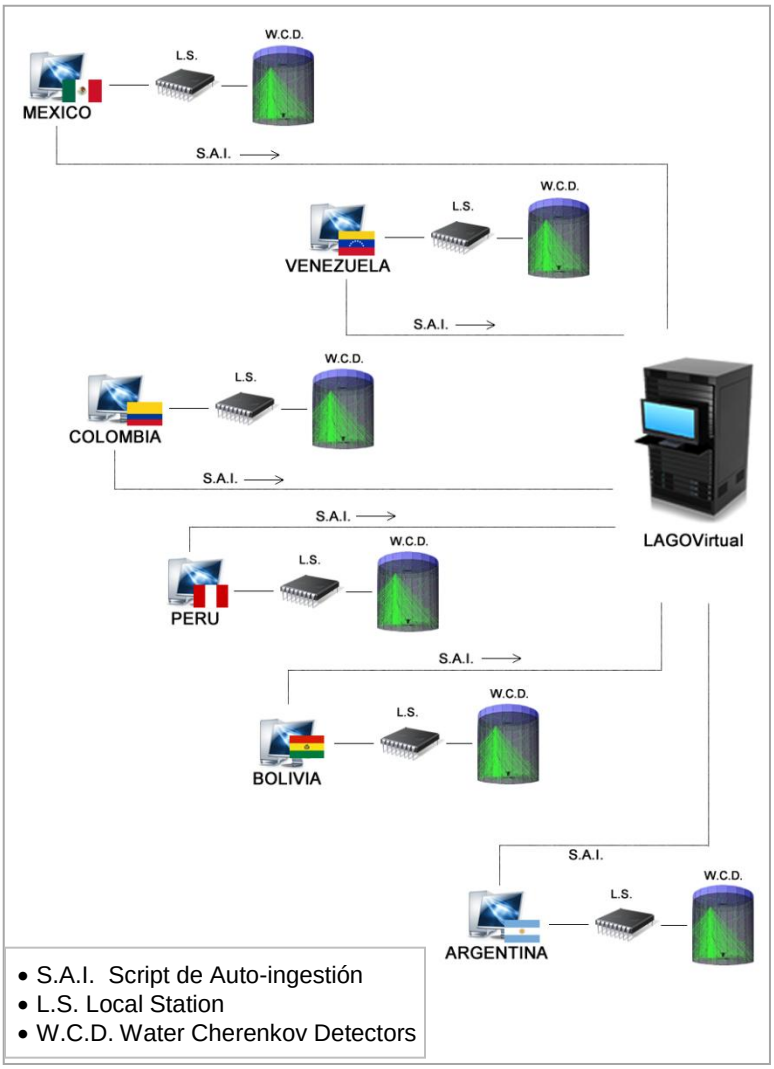

Figura 3. Arquitectura de la Red de Repositorios LAGODatos

La plataforma de código abierto DSpace fue elegida como la solución base para el desarrollo de la Red de Repositorios LAGODATOS, debido al soporte que presta a una gran variedad de datos y a su sistema organizacional que refleja los niveles jerárquicos necesarios para el desarrollo de la red. Otro requerimiento fundamental fue tener un sistema capaz de administrar un gran volumen de datos. Dspace soporta el uso de los Sistemas Manejadores de Bases de Datos PostgreSQL y ORACLE para la gestión de los datos, ambo s sistemas capaces de manejar volúmenes de datos considerables. La plataforma DSpace, fue programada

usando el lenguaje JAVA, y fue modificada y adaptada a los diferentes requisitos iniciales que se presentaron en el desarrollo de los primeros prototipos del servidor de LAGODATOS, implementado en CECALCULA (Centro Nacional de Cálculo Científico, Universidad de Los Andes).

Para LAGODATOS se diseñó una estructura de contenidos que comienza con el país miembro de la colaboración (comunidad), seguido de la institución (subcomunidad), luego los diferentes tipos de datos y experimentos manejados por el proyecto (colecciones), bajo los cuales se almacenan todos los datos y documentos. En la figura 3 se observa una captura de pantalla del prototipo con parte de la estructura comunidad-colección-item desplegada para el país Venezuela.

Figura 4. Captura de pantalla de la interfaz de LAGODATOS, mostrando la estructura de comunidades y colecciones.

Para LAGODATOS la información se clasificó en cinco tipos principales a saber: datos de calibración de los instrumentos, colecciones de datos capturadas por los instrumentos WCD o simuladas por AIRES, datos de

investigadores y documentos asociados o generados por cada institución miembro del proyecto (artículos, presentaciones, etc.).

Para lograr la preservación a largo plazo y la reutilización efectiva de los datos generados por la colaboración LAGO, es indispensable el uso de estándares de descripción de estos datos. Para los datos de calibración y los capturados por los instrumentos de LAGO, se generó una estructura de metadatos basada en el estándar de metadatos Dublin Core[5] en el modelo de metadatos para colecciones de datos científicos propuesto por el CCLRC (*Council for the Central Laboratory of the Research Councils*[6]). En Dspace se configuraron los elementos de metadatos necesarios para la descripción de las colecciones de datos, usando elementos Dublin Core calificados (*qualified Dublin Core*) adaptados a los elementos del estándar CCLRC que eran útiles para los datos generados por LAGO. Adicionalmente, la herramienta Dspace soporta los protocolos de interoperabilidad OAI-PMH (*Open Archive Initiatives Protocol for Metadata Harvesting*[7]) y SWORD (*Simple Webservice Offering Repository Deposit*[8]), que permiten el depósito e intercambio de datos y metadatos entre servidores, lo que facilita tareas como replicación de contenidos y creación de metabuscadores que abarquen varios repositorios, entre otros servicios de valor agregado a los contenidos, relevantes para la creación de redes de contenidos.

Inicialmente se hicieron algunas adaptaciones al funcionamiento de la plataforma Dspace; unas realizadas para mejorar la usabilidad de la interfaz y del sitio Web del repositorio; otras para implementar la estructura de metadatos de las colecciones de datos y facilitar su manejo; y por último, algunos cambios que mejoran algunos aspectos de configuración de Dspace.

Debido a la cantidad de información que es generada por cada WCD, se advirtió que la alimentación manual de las colecciones de datos no es viable para manejar grandes volúmenes de información y archivos. Para sortear las dificultades que conlleva la generación de los metadatos y la alimentación de cada archivo de datos de forma individual, se desarrolló un *script* basado en la ingestión masiva de datos por medio de una consola de comandos, funcionalidad que está implementada de forma predeterminada en DSpace. Este *script* añadido a la plataforma, crea de forma automática lo que se conoce como el *formato simple de archivo de DSpace*, que es una estructura de carpetas y archivos que contiene toda la información necesaria de cada ítem a incluir en el repositorio. Se incluyen en esa estructura los archivos de datos a almacenar junto con su índice y un archivo XML que contiene los metadatos (manteniendo el esquema Dublin Core). Después de generado el formato simple de archivo, el *script* se encarga de la ingestión de los ítems al repositorio de datos. Este sistema de auto-ingestión fue configurado junto con el primer servidor funcional de la Red de Repositorios.

[5] http://dublincore.org/

[6] http://epubs.cclrc.ac.uk/bitstream/485/
[7] http://www.openarchives.org/pmh/
[8] http://swordapp.org/

La finalidad principal del desarrollo de la Red de Repositorio LAGODatos es facilitar a los investigadores ubicados en cada nodo (país e institución) de la comunidad el proceso de captura, clasificación, documentación, ingestión y difusión de los datos, por lo cual no sólo se desarrolló el *script* de auto-ingestión sino que además se está desarrollando un sistema para que la red funcione de forma automática en todo sentido. Este sistema (en fase de desarrollo) distribuye los datos desde cada nodo de la red hasta el servidor principal que se encarga de las tareas de ingestión. El sistema permite a la Red de Repositorios LAGODatos funcionar en la nube, permitiendo acceder y analizar los datos desde cualquier lugar que preste una conexión a internet. Actualmente, el servidor principal de LAGODATOS, ubicado en la Universidad Industrial de Santander, almacena los datos adquiridos por los países miembros de LAGO y se encuentran accesibles para la comunidad en general.

4. Conclusiones

Ha sido presentado un segundo reporte preliminar para la construcción de la comunidad a través del desarrollo de un Ambiente de Investigación Virtual (*Virtual Research Environment* [29]) para la colaboración LAGO. El fin último es desarrollar e integrar bajo un mismo ambiente de trabajo web, un conjunto de herramientas en línea que facilite la detección y el estudio de la los GRB, las trazas de partículas provenientes de eventos solares y las eventuales correlaciones de las variaciones del geomagnetismo terrestre relacionadas con eventos sísmicos. Este ambiente de colaboración se encuentra esquematizado en la figura 2 y permitirá:

- Acceder/controlar el instrumental de forma remota, definiendo y teniendo acceso remoto a algunos parámetros importantes de los tubos fotomultiplicadores, como la línea de base, la ganancia y el acceso a canales específicos de detección.
- Realizar simulaciones de la operación de los detectores (atmósfera con AIRES y traza en el volumen de agua con GEANT4) para determinar el tipo de partícula asociada con las trayectorias registradas.
- Preservar y catalogar los datos (registrados y sintéticos) en cada una de las instalaciones.
- Compartir datos y publicaciones generadas por cada uno de los grupos LAGO.
- Disponer de una plataforma de colaboración en tiempo real (chat y videoconferencia) que permita desarrollar encuentros virtuales para realizar seminarios y reuniones de trabajo de la colaboración.

Figura 4. Esquema de Lago Virtual

En una primera etapa se desarrolló un prototipo del repositorio de datos. Para ello, se diseñó un modelo de metadatos para la catalogación de los registros provenientes de los WCD, así como también de los resultados de las simulaciones. Seguidamente, se modificó el sistema DSpace para preservar y compartir los datos en las distintas instalaciones LAGO. Este prototipo de repositorio ha sido descrito en [30].

Agradecimientos

Los autores agradecen los financiamientos de RedCLARA, bajo el programa de ComCLARA2010 y de la Vicerrectoría de Investigación y Extensión de la Universidad Industrial de Santander, Bucaramanga (Colombia), bajo el proyecto GridUIS2 5541.

Referencias

[1] M. Castells. *The Rise of the Network Society*. Cambridge, MA, USA: Blackwell Publishers, Inc., 2000.

[2] M Castells. *The Internet Galaxy*. Oxford UK: Oxford University Press, 2001.

[3] T. Hey and A. E. Trefethen. *e-science and its implications*. Phil. Trans. R. Soc. Lond. A, 361:1809–1825, 2003.

[4] I. Foster. Service-oriented science. *Science*, 308:814–817, May 2005.

[5] T Hey and A. E. Trefethen. Cyberinfrastructure for escience. *Science*, 308:817–821, May 2005.

[6] T. Hey and A. Trefethen. The data deluge: An escience perspective. In Fran Berman, Geoffrey Fox, and Tony Hey, editors*, Grid Computing: Making the Global Infrastructure a Reality*, pages 809–824. John Wiley & Sons Ltd, 2003.

[7] J. Gray and A. Szalay. The world-wide telescope. *Commun*. ACM, 45(11):50–55, 2002.

[8] M. Bardeen, E. Gilbert, T. Jordan, P. Nepywoda, E. Quigg, M. Wilde, and Y. Zhao. The quarknet/grid collaborative learning e-lab. *Future Gener. Comput*. Syst., 22(6):700–708, 2006.

[9] C Borgman. *What can studies of e-learning teach us about collaboration in e-research? some findings from digital library studies*. Computer Supported Cooperative Work (CSCW), 15(4):359–383, August 2006.

[10] P. Wouters. What is the matter with e-science? – thinking aloud about informatisation in knowledge creation. *THE PANTANETO FORUM*, July 2006.

[11] N. F. Foster and S. Gibbons. Understanding faculty to improve ir content recruitment. *D-Lib Magazine*, 11(1), January 2005.

[12] H. Karasti, K. Baker, and E. Halkola. Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (lter) network. *Computer Supported Cooperative Work (CSCW)*, 15(4):321–358, August 2006.

[13] L.A. Núñez. La reconquista digital de la biblioteca pública. *Interciencia*, 27(4):195–201, 2002.

[14] J.A. Dávila, L.A. Núñez, B. Sandia, and R. Torréns. Los repositorios institucionales y la preservación del patrimonio intelectual académico. *Interciencia*, 31(1):22–29, 2006.

[15] J.A. Dávila, L.A. Núñez, B. Sandia, J. G. Silva, and R. Torréns. www.saber.ula.ve: un ejemplo de repositorio institucional universitarioi. *Interciencia*, 31(1):29–37, 2006.

[16] H.Y. Contreras, Z. Méndez, R. Torréns, and L.A. Núñez. Desarrollo de la red bioclimática del estado mérida, venezuela: Estrategias de captura, manejo y preservación de datos ambientales. *Interciencia*, 33(11):795, 2008.

[17] P. Arzberger, P Schroeder, A Beaulieu, and G Bowker. Science and government: An international framework to promote access to data. *Science*, 303:1777=1778, Jan 2004.

[18] L. Lessig. *Free Culture*. New York: THE PENGUIN PRESS, Feb 2004.

[19] B. Alonso and F. Valladares. Bases de datos y metadatos en ecología: compartir para investigar en cambio global. *Ecosistemas*, 6(2):410, Jan 2006.

[20] P Murray-Rust. *Open data in science.* precedings.nature.com, 2008.

[21] W Michener, J Brunt, J Helly, T Kirchner, and SG Stafford. Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1):330–342, Jan 1997.

[22] R. Torréns. *Desarrollo de sistemas de información bio-climática*. Tesis de Maestría, Ingeniería de Sistemas, Facultad de Ingeniería, Universidad de Los Andes, Mérida, Venezuela, 2003.

[23] R. Norris, H. Andernach, G. Eichhorn, F. Genova, E. Griffin, R. Hanisch, A. Kembhavi, R. Kennicutt, and A. Richards. Astronomical data management. In K.A. van der Huch, editor, *Highlights of Astronomy. Proceedings of Special Session SPS6 (Astronomical Data Management) at the IAU GA*, volume 14, 2006.

[24] DPHEPORG. Data Preservation in High-Energy Physics. *Technical report, International Committee for Future Accelerators*, dec 2009.

[25] D. M. South. Data Preservation in High Energy Physics. ArXiv e-prints Proceedings of plenary talk given at the *18th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2010)*, January 2011.

[26] R.E Luce. No Brief Candle: Reconceiving Research Libraries for the 21st Century, volume 142, *chapter A New Value Equation Challenge: The Emergence of eResearch and Roles for Research Libraries*, pages 42–51. Council on Library and Information Resources, 2008.

[27] S. J. Sciutto. AIRES: A system for air shower simulations (version 2.2.0). *Technical report, Universidad Nacional de la Plata*, La Plata, Argentina, 1999.

[28] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, F. Behner, L. Bellagamba, J. Boudreau, L. Broglia, A. Brunengo, H. Burkhardt, S. Chauvie, J. Chuma, R. Chytracek, G. Cooperman, G. Cosmo, P. Degtyarenko, A. Dell'Acqua, G. Depaola, D. Dietrich, R. Enami, A. Feliciello, C. Ferguson, H. Fesefeldt, G. Folger, F. Foppiano, A. Forti, S. Garelli, S. Giani, R. Giannitrapani, D. Gibin, J. J. Gómez Cadenas, I. González, G. Gracia Abril, G. Greeniaus, W. Greiner, V. Grichine, A. Grossheim, S. Guatelli, P. Gumplinger, R. Hamatsu, K. Hashimoto, H. Hasui, A. Heikkinen, A. Howard, V. Ivanchenko, A. Johnson, F. W. Jones, J. Kallenbach, N. Kanaya, M. Kawabata, Y. Kawabata, M. Kawaguti, S. Kelner, P. Kent, A. Kimura, T. Kodama, R. Kokoulin, M. Kossov, H. Kurashige, E. Lamanna, T. Lampèn, V. Lara, V. Lefebure, F. Lei, M. Liendl, W. Lockman, F. Longo, S. Magni, M. Maire, E. Medernach, K. Minamimoto, P. Mora de Freitas, Y. Morita, K. Murakami, M. Nagamatu, R. Nartallo, P. Nieminen, T. Nishimura, K. Ohtsubo, M. Okamura, S. O'Neale, Y. Oohata, K. Paech, J. Perl, A. Pfeiffer, M. G. Pia, F. Ranjard, A. Rybin, S. Sadilov, E. Di Salvo, G. Santin, T. Sasaki, N. Savvas, Y. Sawada, S. Scherer, S. Sei, V. Sirotenko, D. Smith, N. Starkov, H. Stoecker, J. Sulkimo, M. Takahata, S. Tanaka, E. Tcherniaev, E. Safai Tehrani, M. Tropeano, P. Truscott, H. Uno, L. Urban, P. Urban, M. Verderi, A. Walkden, W. Wander, H. Weber, J. P. Wellisch, T. Wenaus, D. C. Williams, D. Wright, T. Yamada, H. Yoshida, and D. Zschiesche. G4–a simulation toolkit. *Nuclear Instruments and Methods in*

Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 506(3):250 – 303, 2003.

[29] A. Borda, J. Careless, M. Dimitrova, J. Fraser, M. Frey, P. Hubbard, S. Goldstein, C. Pung, M. Shoebridge, and N. Wiseman. Report of the working group on virtual research communities for the ost e-infrastructure steering group. *Technical report, Office of Science and Technology*, London UK, 2006.

[30] 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 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 Conferencem Choroní Venezuela, Madrid España*, 2009. EELA2, CIEMAT.

Sobre los autores

Luis A. Núñez
Licenciado en Física, Doctor en Ciencias, especialidad en Astrofísica Relativista, Física Teórica y Física Computacional. Tiene una significativa experiencia en la gestión de proyectos teleinformáticos institucionales. Ha sido uno de los proponentes y gestores de la Red de Datos de La Universidad de Los Andes (RedULA), fundador de la Escuela Latinoamericana de Redes (EsLaRed) y la Escuela Latinoamericana de Paralelismo y Computación de Alto Rendimiento (ELPCAR).

Luis Alejandro Torres
Ingeniero de Sistemas. Desarrollador de software Comunidad LAGO. Grupo Halley de Astronomía y Ciencias Aeroespaciales. Universidad Industrial de Santander (Colombia).

Rodrigo Torrens
Ingeniero de Sistemas, Maestría en Computación (Universidad de Los Andes, Venezuela). Desarrolla proyectos relacionados con la preservación y libre difusión del patrimonio intelectual producido en la Universidad de Los Andes, Venezuela y en la región latinoamericana. Fundador y responsable del Repositorio Institucional de la Universidad de Los Andes (SABER-ULA) desde el año 2000.

Edwin Barrios
Licenciado en Física (Universidad de Los Andes Venezuela). Participante en la Colaboración LAGO como desarrollador de software para el repositorio de datos.