e-Investigación:
Una nueva manera de producir conocimiento

L. A. Núñez[2]
Centro de Física Fundamental, Departamento de Física, Facultad de Ciencias,
Universidad de Los Andes, Mérida 5101, Venezuela, y
Centro Nacional de Cálculo Científico Universidad de Los Andes (CeCalCULA),
Corporación Parque Tecnológico de Mérida, Mérida 5101, Venezuela

Resumen

Se describen algunas experiencias en e-investigación de en Física de Altas Energías, Astrofísica y Química. Luego se hace una descripción de algunas iniciativas desarrolladas por el Centro Nacional de Cálculo Científico, Universidad de Los Andes (CeCalCULA) en las áreas de e-ingeniería, e-biomedicina y e-ambiente.

Abstract

This work displays a brief description of two initiatives concerning e-research in High Energy Physics, Astrophysics and Chemistry. Later on we also describe other initiatives develop at our Centro Nacinal de Calculo Cientifico Universidad de los Andes concerning e-engineering, e-biomedicine and e-environmental Sciences.

1. E-Investigación y la nueva sociedad

Las Tecnologías de Información y Cooperación (TIC) se constituyen en el eje central de la Investigación y Desarrollo (I + D) al permitir el registro, la acumulación y el acceso a datos experimentales, facilitar el modelaje y la simulación de escenarios posibles pero, sobre todo, por promover dentro de la comunidad académica una nueva manera de relacionarse para la producción y diseminación del conocimiento científico. En países como el nuestro, el desarrollo de este tipo interacción científica tiene la ventaja adicional de generar una memoria documental de datos y productos de investigación a la cual tendrán acceso inmediato otros grupos de investigación y la sociedad en su conjunto.

A partir de los años 70 hubo un cambio en el modo de producción del sistema capitalista. Pasamos de una economía industrial a una informacional. Es decir, la información ha ido transformando la economía en informacional 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, la información actuando sobre tecnologías. La actividad Científica y Tecnológica no escapa a convertirse en e-actividad y diferirá de la que estamos desarrollando hoy en día en términos metodológicos, funcionales y, sobre todo, en la manera como se organizará la comunidad académica para crear y diseminar el conocimiento [2, 3]. Un universo de sensores recogerán una ingente cantidad de datos y los enviarán a una red centros donde serán almacenados, custodiados y estarán disponibles a través del WEB. Mediante interfaces WEB Semánticas con agentes y programas de búsqueda cada vez más inteligentes, la información será accedida, preprocesada y consolidada utilizando técnicas de representación del conocimiento y minería de datos. A partir de estos datos preprocesados, se generará, mediante programas y sistemas cooperativos distribuidos en una red de servidores, el modelaje y se simularán situaciones que habrán de predecir escenarios posibles. Estos resultados serán analizados por equipos de investigadores distribuidos geográficamente, quienes interactuarán a través de la red mediante sistemas de videoconferencias de escritorios y herramientas de colaboración electrónica. Las conclusiones y los resultados serán compartidos con la comunidad académica y diseminados a la sociedad mediante publicaciones electrónicas interactivas, en las cuales estará disponible el acceso a los datos y a las aplicaciones que generaron los resultados. El "e-lector" podrá remodelar esas situaciones y sacar sus propias conclusiones a partir de nuevas situaciones que se le ocurra. La tendencia en este uso de las tecnologías de información por parte de la sociedad del conocimiento apunta jugar, en un futuro muy cercano, el papel que

[2]e-mail: nunez@ula.ve Web: http://webdelprofesor.ula.ve/ciencias/nunez/

hoy juegan 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 creamos y distribuimos el conocimiento. Las TIC se hacen cada vez más ubicuas y de uso intuitivo por parte de una creciente comunidad de usuarios. La utilización intensiva de las TIC ha ido transformado a las organizaciones y actividades.

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 [10, 8, 11] 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 el "diluvio de datos" [9]. Esta avalancha de registros de todo tipo, viene generada por experimentos de escala mundial (aceleradores de partículas, red de observatorios terrestres y saltelitales e infinidad de los más variados sensores), desbordando toda capacidad de manejo que no sea mediante las TIC.

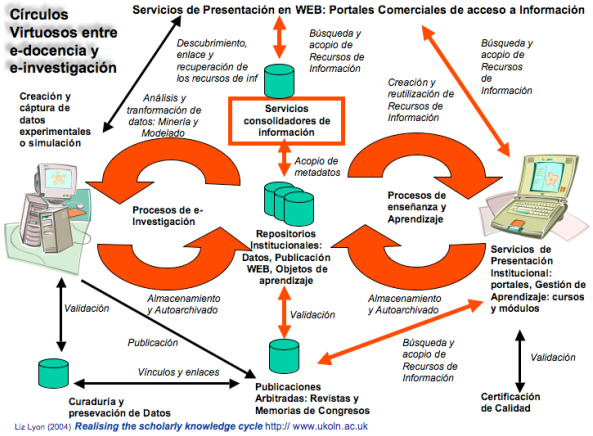

Figura 1: Circulos Virtuosos de Producción de Conocimiento entre la e-docencia y la e-investigación. Tomado y adaptado de Liz Lyon (2004) Realising the scholarly knowledge cycle http:// www.ukoln.ac.uk

Este artículo hace referencias a algunas experiencias de e-investigación en las áreas de Física de Altas energías, Astrofísica y Química, para luego describir algunos desarrollos realizados por el Centro Nacional de Cálculo Científico, Universidad de Los Andes. La próxima sección contiene las descripción de las experiencias internacionales, mientras que la Sección 3 describe los desarrollos realizados por nuestro centro.

2 Experiencias de comunidades

En esta sección exploraremos las algunas experiencias de comunidades en e-investigación. Definitivamente las comunidades de Física de Altas Energías y Astrófisica han sido los motores que han impulsado el desarrollo de herramientas computacionales para la e-investigación. Sin embargo, hoy las ciencias química muestran importantes contribuciones a la concreción de los e-Laboratorios. Todas estas experiencias comienzan a mostrar tendencias en nuevas maneras de hacer ciencia, en la transición de sistemas propietarios, centralizados de almacenamiento de datos a servicios colectivos, descentralizados, de acceso a datos, basados en estándares y de alcance global. Tal y como se muestra en la figura 1 esta nueva manera de hacer ciencia vincula el registro de datos reales provenientes de experimentos con la formación de investigadores en todos los niveles. Esta novedosa vinculación entre docencia e investigación, es quizá la más importante contribución.

2.1 Los precursores: LHC y LCG

La comunidad de Física de Altas Energías, tradicionalmente, ha venido acumulando experiencias en el almacenamiento y análisis de grandes volúmenes de datos. Es disciplina inicia la práctica de la e-investigación. Debe simular la operación de los instrumentos y luego compararla con las medidas. A partir de esa comparación tendrá idea de qué se está midiendo. Por lo tanto los volúmenes de datos provienen tanto de las mediciones como de las simulaciones de los instrumentos. Sólo la comunidad del experimento de BaBar[1] que aglutina a más de 600 físicos e ingenieros de 75 instituciones a escala mundial ha logrado consolidad una de las bases de datos orientadas a objetos más grandes del mundo con casi 900 TeraBytes para Noviembre de 2004.

La próxima puesta en marcha de mayor instrumento experimental jamás construido, el *Large Hadron Collider* (LHC) [2] ha motorizado importantes propuestas para enfrentar el almacenamiento y posterior análisis del diluvio de más de un petabyte de datos mensual. Este desarrollo que superará con creces los volúmenes de datos que vienen siendo manejados, está impulsando importantes proyectos tanto en Europa como en los Estados Unidos[3] . Para crear la infraestructura computacional para LHC ha surgido el LHC Computing Grid Project (LCG)[4] . Financiado por el Centro Europeo de Física Nuclear (CERN) es una

[1] http://www.slac.stanford.edu/BFROOT/
[2] http://lcg.web.cern.ch/LCG/New/index.html
[3] http://lcg.web.cern.ch/LCG/New/rel-proj.html
[4] http://lcg.web.cern.ch

colaboración multinacional de laboratorios de Física de Altas Energías en todo el mundo. Tiene por objetivo desarrollar ambientes distribuidos de computación que pemintan manejar esos inmensos volúmenes de datos que se esperan sean producidos por el instrumento. Esto incluye: aplicaciones para el análisis, soporte y mantenimiento de datos del LHC, siendo uno de los mayores retos la altas tasas de registro de data y el análisis de ingentes volúmenes de datos.

La primera de las fases del proyecto comenzó en el 2002 y el objetivo era crear un prototipo de organización de servicio y desarrollar algunas herramientas computacionales (*mdidleware*) con el objeto de ganar experiencias las cuales alimentarían el Reporte de Diseño Técnico (TDR por sus siglas en inglés de *Technical Design Report*). La segunda fase (2006-2008) arrancar con la primera preoperación de cómputo y almacenamiento a gran escala.

El LCG ha adoptado las herramientas computacionales y ha creado grupos de evaluación y prueba conjunta con el Proyecto Europeo de DataGrid [5] . La organzación para la operación del LCG es jerárquica compuestas con *Tiers*, donde la raíz de esta jerarquía está en el CERN (Tier 0), luego nodos regionales (Tier 1) en IN2P3 Lyon-Fancia, PIC Barcelona-España, CNAF en Boloña-Italia, PPARC en Rutherford-Inglaterra, por mencionar los principales nodos europeos; luego seguirían nodos menores de servicos. Cada uno de estos nodos deberá realizar labores de curaduría de datos y proveer capacidades de cómputo para el análisis de esos resultados (para una reciente revisión de los proyectos de Grid en Física de Partículas pueden consultar [14] y las referencias allí citadas).

2.2 Los Observatorios Virtuales

La comunidad que más rápidamente internalizó el modelo de e-investigación ha sido la comunidad de Astrofísica. Tradicionalemente esta comunidad opera grandes instrumentos compartidos y genera campañas de observación para uso colectivo, mantienendo los resultados de estas mediciones en una red de repositorios de datos alrededor del mundo (una lista muy parcial de estos estos recursos está disponigle en la iniciativa de AstroWeb [6]))

A finales de la década de los 90 surge la idea de crear un ambiente de trabajo web que permitiera manejar grandes volúmenes de datos complejos, procedentes tanto de simulaciones numéricas como de mediciones de distintos instrumentos terrestres y espaciales. Este ambiente debería permitir no sólo conservar y catalogar las mediciones, sino que además proveyera un conjunto de herramientas para su tramiento y análisis. El objetivo inicial de los Observatorios Virtuales (OV) era mejorar y unificar el acceso a data astronómica, y principalmente proveer servicios de análisis de estos datos tanto a astrónomos profesionales como al público en general. Rápidamente integró como servicio la vinculación con el conocimiento. Cada objeto astrófisico, no solo tienen asociado las mediciones que se han realizado desde distintas perspecivas (longitudes de onda, sensibilidad, observaciones indirectas) sino que, adicionalmente, al vincularse con la bases de datos publicaciones[7] se le asocia también al objeto las publicaciones que sobre ese objeto se han realizado.

Replicando las tendencias de esta nueva era informacional, hoy existe una alianza internacional de más de una docena de observatorios virtuales, los cuales federan esfuerzos a escala mundial. La práctica de esta nueva manera de hacer "observaciones virtuales" a través de los OV ha traído como consecuencia un acceso igualitario a la data astronómica. En el mismo espíritu que el WEB, nadie regula cuál data es buena y cuál no. Sencillamente, cualquiera puede publicar sus datos y la comunidad evaluará su calidad por el uso. Otra contribución a esta nueva práctica se basa en los esfuerzos colectivos. Cada vez más son menos las propuestas individuales para observar un determinado objeto astrofísico y son más los esfuerzos colectivos por realizar campañas de mediciones, en distintas longitudes de onda en grandes regiones del cielo. Para más detalles de la estructura y servicios de los OV pueden consultar [7] y las referencias allí citadas.

2.3 Los e-Laboratorios

Las ciencias químicas proveen un ejemplo de vincular los datos al protocolo de medición. Las mediciones siempre van acompañadas de informción del experimento y de las condiciones experimentales. Además de mediciones existe una importante información secundaria respecto a las razones de la selección de un conjunto particular de datos, el anásis estadístico, modelos y métodos de simulación vinculados con las mediciones. En definitiva, el cuaderno de laboratorio de los científicos

[5] http://eu-datagrid.web.cern.ch/eu-datagrid/

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

[7] El sistema de datos Smithsoniano/NASA http://www.adsabs.harvard.edu/ o el sistema de archivos de preprint arXiv.org http://lanl.arxiv.org/

experimentales registra informaciones adicionales que debe ser accesible para garantizar la repetibilidad de las mediciones.

El proyecto CombeChem[8] muestra una clara experiencia que vincula el manejo remoto de un equipamiento, la vinculación de los datos con el protocolo de medida y la asociación de la publicación al conjunto de datos registrados en el laboratorio. Vinculados a otros importantes proyectos de uso y reuso del conocimiento como eBank[9] and CoAKTinG[10] , CombeChem busca vincular los datos a las publicaciones que referencian las propiedades de estructuras cristalográfiacas. Se estima que cerca de un 1.5 millones de estructuras cristalográficas se determinan al año y sólo el 20% de esa data está disponible a los investigadores[6].

3 La experiencia de CeCalCULA

En esta sección enumeraremos algunas de las experiencias que en el desarrollo de ambientes de e-ciencia, hemos desarrollado en el Centro Nacional de Cálculo Científico Universidad de los Andes (CeCalCULA) algunas de estas experiencias han sido descritas en recientes publicaciones [4].

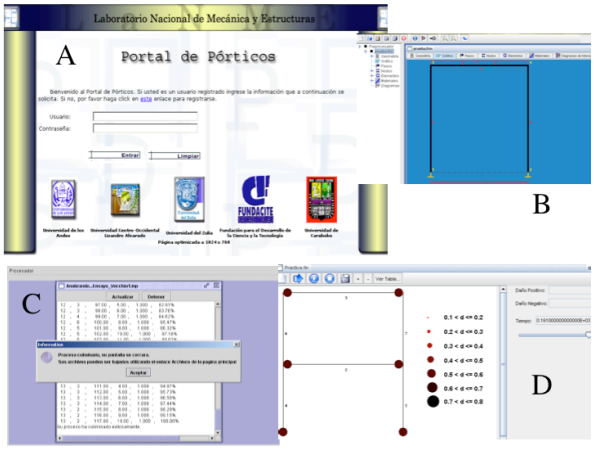

Figura 2: Portal de Daño. Cuadrante A: Página principal Portal de Daño http://portaldeporticos.ula.ve/. Cuadrante B: Pre-procesador. Plate C: Visualización del análisis a través del portal. Cuadrante D: Post procesador Gráfico

[8] http://www.combechem.org/
[9] http://www.ukoln.ac.uk/projects/ebank-uk/
[10] http://www.aktors.org/coakting/

3.1 e-Ingeniería

Uno de los primeros desarrollos ha sido el Portal de Daño[11] . Este ambiente de trabajo web, está basado en el análisis estructural por elementos finitos[13]. Este ambiente permite simular numéricamente los procesos de agrietamiento y colapso de estructuras aporticadas sometidas a sobre cargas producto de movimientos sísmicos. El sistema está constituido por varios módulos (Java y Fortran) que se intercambian entre el servidor y el cliente.

El simulador de daño se basa en la teoría del daño concentrado [5] calculados mediante elementos finitos no lineales. El ambiente muestra la cuantificación de las densidad y ubicación de agrietamiento de la estructura, en una escala que varía entre cero (ningún daño) hasta uno (colapso total).

Esta portal ha sido utilizado con éxito en la evaluación del riesgo estructural de edificaciones educativas[16, 15].

3.2 e-Biomedicina

Una de las herramientas más comunes en el análisis de alineamiento de nucleótidos y proteínas es BLAST [12] (por su acrónimo inglés de *Basic Local Alignment Search Tool*). Esta herramienta computacional permite encontrar regiones similares entre secuencias, las compara con las secuencias existentes en base de datos genética y calcula, estadísticamente el grado de similitud entre las secuencias. Este proceso de buscar la homología es computacionalmente exigente no tanto por la búsqueda de una secuencia particular, sino que normalmente son cientos de secuencias que se buscan simultáneamente.

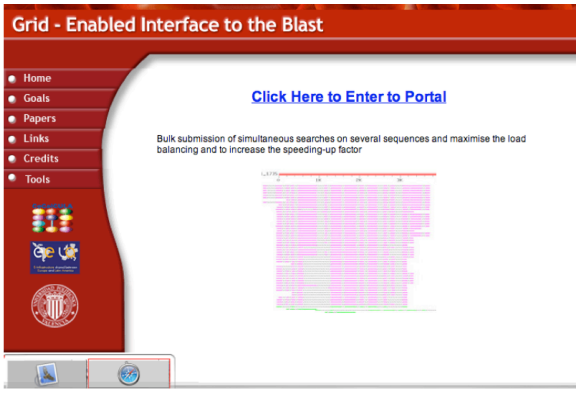

Figura 3: Blast2EELA Poral Biomédico
http://www.cecalc.ula.ve/blast

[11] http://portaldeporticos.ula.ve
[12] http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

La comunidad de bioinformática utiliza mayormente instalaciones locales o servidores públicos como NCBI[13] o gPS@[14] , sin embargo este ambiente se hace ineficiente, dado el limitado número de búsquedas simultáneas que se pueden hacer. Adicionalmente estas las bases de datos se actualizan frecuentemente por lo que es importante rechequear las secuencias y su homologías. En el marco del proyecto EELA (por su acrónimo de E-Infraestructure Share between Europe and Latin America) como una de las aplicaciones biomédica[1] se ha desarrollado el portal Blast2EELA[15] . Este portal se basa en la implantación del código BLAST paralelo mpiBLAST [12] y permite el envío de múltiples secuencias simultáneas

3.3 e-Medio Ambiente

Otra de las iniciativas para el apoyo a la e-investigación la constituye La Red de Estaciones Bioclimáticas del Estado Mérida (REDBC)[16] . El acceso a los conjuntos de datos ambientales producidos por los organismos e instituciones en Venezuela se dificulta por varias razones: disponibilidad de datos sólo en papel, altos costos para la adquisición de datos, datos incompletos o descontinuados, etc. A ello, se suman problemas relativos a los procedimientos de captura y a la confiabilidad de esos datos.

Figura 4: Portal de la Red de Estaciones Bioclimáticas del Estado Mérida http://www.cecalc.ula.ve/redbc Se muestra el portal principal, la ubicación de las estaciones Meteorológicas, las colecciones de datos y los organismos que colaboran con la red.

[13] http://www.ncbi.nlm.nih.gov/
[14] http://gpsa.ibcp.fr/
[15] http://www.cecalc.ula.ve/blast
[16] http://www.cecalc.ula.ve/redbc

Por estas razones desarrollamos la iniciativa de la Red Bioclimática. Consiste en un proyecto piloto que contempla la captura, procesamiento y difusión de información climática, ambiental y ecológica mediante la implementación de un sistema de información y mecanismos de comunicaciones. Actualmente se documentan, se ordenan y se publican en la red los datos crudos (sin procesar) de seis (6) estaciones meteorológicas: Chama, Mucujún, La Hechicera, Ciplat, Mucujún, Santa Rosa y San Juan. Adicionalmente se suministra información didáctica sobre el manejo de datos y metadatos a los integrantes de la red (ver Figura 4). De esta manera se garantiza la custodia y distribución de los datos generados, se asegura la permanencia de los datos por períodos largos del tiempo y se crean mecanismos y herramientas que faciliten el acceso a los datos generados por cualquier estación participante en la red o científicos que los necesiten. La Red de estaciones Meteorológicas usa estándares internacionales (de formato y contenidos) para estructurar y almacenar los datos (para permitir intercambio con otros sistemas).

Está concebido como un proyecto demostrativo y cooperativo, donde instituciones o individuos que dispongan de datos pueden catalogarlos y enviarlos a nuestro centro para su preservación. Esta enfoque voluntario ha sido exitoso por cuanto se comenzó con 3 estaciones, 2 en la zona sur del Lago de Maracaibo y una en la meseta de Mérida. Ahora se disponen de las seis estaciones antes mencionadas, la mayor parte de ellas en la meseta.

Agradecimientos

Este trabajo ha sido apoyado financieramente por el Fondo Nacional de Investigaciones Científicas y Tecnológicas bajo los proyectos S1-2000000820 y F-2002000426.

Referencias

[1] M. Cárdenas, V. Hernández, R. Mayo, I. Blanquer, J. Pérez-Griffo, R. Isea, L. Núñez, H. R. Mora, and M. Fernández. Biomedical aplications in eela. *Stud Health Technol Inform*, 120:397–400, 2006.

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

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

[4] J. L. Chaves, G. Díaz, V. Hamar, R. Isea, F. Rojas, N. Ruíz, R. Torrens, M. Uzcátegui, J. Flórez-López, H. Hoeger, L. Núñez, and C. Mendoza. e-science initiatives in venezuela.

In R. M. y. R. M. J. Casado, editor, *Spanish Conference on e-Science Grid Computing*, Madrid España, 2007. CIEMAT.

[5] A. Cipollina, A. López-Inojosa, and J. Flórez-López. A simplified damage mechanics approach to nonlinear análisis of frames. *Comp. & Struct., Struct.*, 54(6):1113 – 1126, 1995.

[6] S. Coles, J. Frey, M. Hursthouse, M. Light, A. Milsted, L. Carr, D. DeRoure, C. Gutteridge, H. Mills, K. Meacham, et al. An e-Science environment for service crystallography-from submission to dissemination. *Journal of Chemical Information and Modeling*, 46(3):1006–1016, 2006.

[7] S. G. Djorgovski and R. Williams. Virtual observatory: From concept to implementation. *Preprint arXiv:astro-ph/0504006*, 2005.

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

[9] T. Hey and A. Trefethen. The data deluge: An e-science perspective. In F. Berman, G. Fox, and T. Hey, editors, *Grid Computing: Making the Global Infrastructure a Reality*, pages 809–824. John Wiley & Sons Ltd, 2003.

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

[11] T. Hey and A. E. Trefethen. Cyberinfrastructure for e-science. *Science*, 308:817–821, May 2005.

[12] Y. J. Kim, A. Boyd, B. D. Athey, and J. M. Patel. miblast: scalable evaluation of a batch of nucleotide sequence queries with blast. *Nucleic Acids Res.*, 33(13):4335–4344, 2005.

[13] M. E. Marante, L. Suárez, A. Quero, J. Redondo, B. Vera, M. Uzcategui, S. Delgado, L. R. León, L. Núñez, and J. Flórez-López. Portal of damage: a web-based finite element program for the analysis of framed structures subjected to overloads. *Advances in Engineering Software*, 36(5):346 –358, May 2005.

[14] J. Marco de Lucas. A Grided World? *Journal of Physics: Conference Series*, 53:397–412, 2006.

[15] A. Moreno. Influencia del factor de reducción de respuesta en el daño estructural de pórticos de concreto armado sometido a solicitaciones sísmicas. Master's thesis, Facultad de Ingeniería, Universidad del Zulia, 2005.

[16] M. Torres, P. Gonzales, and J. Mujica. Evaluación de la vulnerabilidad estructural de edificaciones esenciales: caso hospital rotario. Trabajo especial de grado, Universidad Centro Occidental Lisandro Alvarado, Barquisimeto Venezuela, 2007.