Red Bioclimática del Estado Mérida: Repositorio Institucional de Datos Ambientales

H.Y. Contreras[2]
Departamento de Computación, Escuela de Ingeniería de Sistemas y
Centro de Simulación y Modelado, Facultad de Ingeniería,
Universidad de Los Andes, Mérida 5101, Venezuela

Z. Méndez[3] , **R. Torrens**[4]
Centro de Teleinformación Universidad de Los Andes
Corporación Parque Tecnológico de Mérida, Mérida 5101, Venezuela

y

L. A. Núñez[5]
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

Versión Mayo 2008

Resumen

Presentamos la experiencia en la construcción y desarrollo de la Red Bioclimática del Estado Mérida (Venezuela) como prueba de concepto en la utilización de las técnicas y estrategias para la captura, manejo y preservación de datos ambientales provenientes de la *e-investigación*. Se describen los elementos éticos, conceptuales y técnicos para la preservación del patrimonio de datos, haciendo énfasis en los conceptos datos libres (*Open Data*), metadatos y sus estándares. Igualmente el testimonio de la construcción del repositorio de datos bioclimáticos http://www.cecalc.ula.ve/redbc/, sus productos de información asociados y algunos resultados e impactos preliminares de su utilización. Finalizamos con algunas reflexiones y recomendaciones las cuales, quizá, permitan hacer sustentable y replicable esta iniciativa.

[2]e-mail: hyelitza@ula.ve Web: http://webdelprofesor.ula.ve/ingenieria/hyelitza/
[3]e-mail: zulay@ula.ve
[4]e-mail: torrens@ula.ve
[5]e-mail: nunez@ula.ve Web: http://webdelprofesor.ula.ve/ciencias/nunez/

1 Una nueva manera de producir, diseminar y preservar datos

Cada vez con mayor frecuencia y efectividad los investigadores de todas las naciones intercambian datos, ideas, publicaciones, referencias y artículos. Si bien una serie de emergentes y novedosos ambientes de colaboración electrónica no terminan de ser utilizados como herramientas cotidianas para estas interacciones (Coles y colaboradores, 2006; Borgman, 2006; De Roure y Frey, 2007; Collins y colaboradores, 2007), el correo electrónico, la mensajería instantánea y, sobre todo LA RED, se erigen como los apoyos y motores a éstas nuevas formas de colaboración ubicua. Pero más allá de este intercambio entre pares, transitamos por la era *postgutemberg* en la cual los productores de información (investigadores, centros de investigación y-o instituciones académicas) tienen la capacidad de publicar y difundir directamente su producción intelectual, sin intermediarios editoriales y a costos cada vez menores.

Los términos "ciberinfraestructura", "e-ciencia" y "e-investigación", han sido acuñados para describir esta nueva forma de producción y diseminación del conocimiento, donde el uso intensivo de las Tecnologías de Información y Comunicación (TIC), la distribución geográfica de los recursos de medición, procesamiento y análisis, pero sobre todo su acceso ubicuo, son sus características más resaltantes y descriptivas (ver Hey y Trefethen (2003b), Foster (2005) y Hey y Trefethen (2005), así como las referencias allí citadas). Uno de los retos que habremos de enfrentar en esta nueva manera de hacer investigación es manejar, administrar, analizar y preservar un "diluvio de datos" (Hey y Trefethen, 2003a) generado por una red de sensores a escala mundial y experimentos de grandes dimensiones (aceleradores de partículas, red de observatorios terrestres y satelitales, enormes bases de datos genéticas, por mencionar las más impactantes). Este alud de mediciones, desbordando toda capacidad para su manejo que no sea mediante las TIC, convierte a estos instrumentos en herramientas informáticas y a la experimentación en minería de datos.

Los grandes productores de datos son colaboraciones mundiales, industriales y multinacionales, las cuales generan ingentes volúmenes de datos que deben ser distribuidos geográficamente y mantenidos por esos proyectos mientras duren. Muchos de esos datos nunca aparecerán publicados y cuando finalice la colaboración, muchas de las medidas experimentales se perderán o serán enviadas a reservorios nacionales (o internacionales) que nada tuvieron que ver con su producción. Gran parte de las decisiones y criterios para generarlos quedarán sepultados en una inmensa correspondencia electrónica que nadie dispondrá (Gray y Szalay, 2002). Igual suerte correrán los datos producidos por multitud de pequeños grupos de investigación quienes, atacando problemas similares, se encuentran distribuidos por toda la geografía mundial. Todos ellos, grandes y pequeños productores de datos enfrentarán los mismos problemas de catalogación, preservación y diseminación de datos y del conocimiento que a partir de éstos surge. Es imperioso planificar y construir repositorios de datos que los almacenen mientras se produzcan y que conserven la traza testimonial de las decisiones y criterios que los generaron (Gray y Szalay, 2002; Karasti y colaboradores, 2006; Borgman y colaboradores, 2007; Murray-Rust, 2008).

Si bien hace más de una década hubo algunos esfuerzos pioneros por generar un marco de recomendaciones que guiaran la preservación y diseminación de bases de datos científicos (Dozier y colaboradores, 1995), es recientemente cuando organizaciones multilaterales y organismos planificadores de Europa y los Estados Unidos, comienzan a generar reportes técnicos, cada vez más detallados, con lineamientos respecto a como fomentar la preservación de colecciones de datos científicos (Lord y Macdonald, 2003; Arzberger y colaboradores, 2004b; Simberloff y colaboradores, 2005; Pilat y Fukasaku, 2007; Lyon, 2007; Clark, 2007). Inclusive, algunos países como Canadá y China comienzan a fijar posición a ese respecto (Sabourin y Dumouchel, 2007; Xu, 2007). Sin embargo, muchas de estas recomendaciones no han permeado hacia las comunidades productoras y/o custodios de las colecciones de datos en esos países. Aún peor es la situación en nuestra región, donde todavía no estamos convencidos del cambio que experimentó la producción y diseminación del conocimiento y no nos hemos apropiado de un uso mínimo de las TIC en nuestra docencia e investigación.

Esta creciente consciencia por preservar y compartir datos, y a partir de su uso y reuso, generar nuevos conocimientos, ha impulsado una nueva faceta del Movimiento de Acceso Libre al Conocimiento (simplemente OA por sus siglas de *Open Access* en Inglés). La preservación y diseminación libre a los datos científicos comienza a ser considerada, debatida y normada en distintos escenarios (Arzberger y colaboradores, 2004a; Lyon, 2007; Pilat y Fukasaku, 2007; Clark, 2007; Sabourin y Dumouchel, 2007; Xu, 2007). Heredera de la reflexión y fundamento conceptual del OA, el Movimiento Datos Libres (OD por *Open Data* en Inglés) es un término reciente acuñado para indicar los estándares y procesos para que los datos de investigaciones puedan ser utilizados y reutilizados sin barreras ni costo. La comunidad académica percibe que los datos publicados les pertenecen como patrimonio. Sin embargo, muchos editores comienzan a demandar derechos de reproducción (*copyright*) lo cual, sin duda, es uno de los mayores impedimentos que habrá de enfrentar la e-investigación científica (Murray-Rust, 2008).

Este artículo pretende una doble finalidad. Por un lado divulgar la reflexión que se viene dando en distintos foros internacionales refrente al OD con el objeto de sensibilizar a los productores de datos para que asumamos la responsabilidad de difundir/preservar nuestras mediciones y datos, libremente y, con ello poder impulsar la creación de nuevos conocimientos. Por otro lado, queremos mostrar un ejemplo un repositorio de datos libres que viene operando desde hace más de un lustro. Es una prueba de concepto de un repositorio de datos libres ambientales, el cual, no sólo colectiviza su recolección sino que comparte su custodia y curaduría. Describiremos algunos conceptos y estrategias para la construcción de ese repositorio. Se presenta también una evaluación preliminar de su uso e impacto. Es la continuación de un esfuerzo por garantizar y difundir libremente el conocimiento producido por los grupos de investigación de nuestra Universidad de Los Andes (Núñez, 2002; Dávila y colaboradores, 2006b; Dávila y colaboradores, 2006a).

Organizaremos la presentación de la siguiente manera. En la próxima sección expondremos algunos elementos de la reflexión general de OD, su relevancia y relación con el contexto del OA. Seguidamente, en la Sección 3 se presentan referencias técnicas relacionadas a los metadados y su importancia para la difusión/preservación del patrimonio de mediciones, datos sintéticos y registros históricos. Luego, en la Sección 4 mostramos algunos ejemplos de colecciones de datos. La Sección 5 describe la experiencia en la creación y puesta en operación

de la Red Bioclimática del Sur de Lago de Maracaibo y algunos de sus primeros resultados. Finalizaremos con la Sección 6 en la cual presentamos algunas conclusiones y recomendaciones.

2 Los Datos: un patrimonio intelectual colectivo

La mayor parte de las investigaciones en Astronomía, Física de Altas Energía, 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 lo tanto, es de intuir que los datos provenientes de simulaciones y mediciones, y no sólo las publicaciones producidas a partir de ellos, nos 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 re-uso podamos seguir la cadena de producción del conocimiento. Esto es: *Datos –Información – Conocimiento – Información - Datos*. De su análisis, en variados contextos y desde distintas perspectivas, surgen los nuevos conocimientos (Lessig, 2004). 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 puedan (Arzberger y colaboradores, 2004a; Alonso y Valladares, 2006). 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 de revistas, quienes comienzan a exigir los datos que respaldan y soportan ideas afirmaciones y propuestas contenidas en las publicaciones cientoficas, haciéndoles extensivoa los datos el derecho de reproducción (*copyright*) con la consecuente restricción para su reutilización.

El paralelo de la actitud de los editores con los datos y lo crítico de su re-utilización para la producción de nuevos conocimientos, ha hecho surgir el Movimiento de Datos Libres. Con argumentos similares a quienes defendemos el acceso libre a las publicaciones periódicas, pero con las implicaciones mucho más contundentes que los datos tienen para la generación de nuevos conocimientos (ver los ejemplos provenientes de las ciencias Químicas descritos en Murray-Rust, 2008), este movimiento promueve el acceso irrestricto a los datos. Quizá el acceso pueda ser limitado, si su utilización arriesga la seguridad de individuos o especies, compromete derechos de confidencialidad, o viola algunas prerrogativas para su explotación temporal por parte de quienes los recolectaron o generaron. Adicionalmente, el hecho que hayan sido producidos por proyectos o investigadores financiados con fondos públicos no elimina la posible comercialización de productos o servicios que de ellos puedan derivarse.

La razón por la cual muchos grupos de investigación e investigadores individuales (financiados o no con fondos públicos) tienden a considerar los datos como su patrimonio se debe, principalmente, a la falta de claridad de los entes financieros y de las instituciones a las cuales pertenecen los investigadores y sus grupos de investigación. Las ventajas (y de allí la imperiosa necesidad) de preservar y difundir de forma libre los datos y mediciones no son (aún) apreciadas (Uhlir y Schröder, 2007). La disponibilidad de acceso libre a los datos, además de incentivar la curiosidad y diversidad de análisis científicos, promueve nuevas áreas de investigación transdisciplinaria, al facilitar mecanismos automáticos de minería de datos. Adicionalmente, el disponer de datos reales para su análisis, fomenta el surgimiento de nuevas estrategias para la formación y entrenamiento de investigadores. Estas nuevas generaciones, al

estar expuestas a novedosas herramientas de análisis de datos reales, prefigurarán nuevas maneras y mecanismos de medición, sepultando las sempiternas formas experimentales con las cuales fueron formados sus predecesores.

Para favorecer el uso compartido de los datos es imprescindible que se tome conciencia de la importancia de los metadatos. Es el único modo de hacer que los datos sean comprensibles y puedan ser utilizados en el futuro no sólo por el investigador quien los generó, sino por otros investigadores del área. Los metadatos es una información adicional que describe el contenido, la calidad, la estructura y la accesibilidad de una serie de datos. En la próxima sección desarrollaremos con un poco de detalle la importancia de este concepto.

3 Datos, metadatos y la *e-investigación*

Podemos formular una idea simplificada de dato como aquella representación de información, capaz de ser procesada, interpretada, trasmitida y preservada. Los datos surgen de mediciones de sensores (datos observacionales o experimentales), de resultados de simulaciones (datos sintéticos o matemáticos, proveniente de modelos) o de registros históricos (datos históricos). Los datos se pueden considerar crudos si provienen directamente de las mediciones o modelos, o procesados cuando han sido sometidos a algún tipo de filtraje mediante algún criterio. Pero, además habrá todos los matices y complejidades que se puedan imaginar. Lo que pudieran ser considerados datos procesados para algunas instancias, serán datos crudos para otras y los datos históricos podrán emerger de modelos matemáticos que generaron su registro.

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 poder ser descubiertos y para ello deben ser descritos apegados a estándares acordados por las comunidades productoras. A esa información básica utilizada para describir los datos, como 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. Obviamente, los metadatos pueden ser o no digitales y, los datos a los cuales están asociados pueden existir en ambas formas. La utilización de metadatos facilita (Michener y colaboradores, 1997):

- 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.
- la incorporación de elementos de conocimiento semántico asociado a los datos

Una adecuada documentación sobre el muestreo, procedimientos analíticos, anomalías y calidad de los datos, y estructura de las colecciones de datos ayudará a que esos datos puedan ser correctamente interpretados y reinterpretados en el futuro.

Conseguir la estandarización de los metadatos es muy importante porque permite definir una terminología común, permite llevar a cabo la entrada, validación, acceso, integración y síntesis de los datos de manera automatizada y asegura una documentación completa y precisa del contenido de los datos. Existen muchos y variados estándares de metadatos disponibles (*Dublín Core*,

Darwin Core, *Content Standard for Digital Geospatial Metadata*, ISO 19115 *Geographic information metadata, Ecological Metadata Language* etc.), la razón de que existan tantos estándares es que los metadatos se emplean para diversas aplicaciones (Bibliotecología, Biología, Geología, Ecología, Cartografía, etc.) y cada una de estas comunidades acuerda un conjunto de informaciones indispensables para identificar y manipular sus datos.

Las estructuras de metadatos están compuestas por elementos asociados a definiciones semánticas descriptivas de algunos de los posibles atributos del dato. Esas estructuras pueden ser arbitrariamente simples o complejas, y la información que contienen puede ser muy variada dependiendo del tipo de dato y de las necesidades que para su uso impone la comunidad que lo generó. Cada comunidad puede, dentro de su modelo de metadato, definir de manera diferente una propiedad o elemento. Por ejemplo, la iniciativa *Dublin Core* especifica un conjunto base de 15 elementos mientras que el modelo de metadato de *Learning Objects Metadata*, que está siendo desarrollado por la IEEE y otras organizaciones para describir recursos en ambientes de enseñanza-aprendizaje tiene cerca de 100 elementos (Torrens, 2003).

Obviamente la incorporación de los metadatos demanda una inversión de tiempo y esfuerzo por parte de quienes generan y requieren preservar/compartir los datos. Es imperioso tomar en cuenta el tiempo que requiere la definición del modelo de metadatos, su aprendizaje y, posteriormente su implantación. Luego vendrán los costos del mantenimiento a corto, mediano y largo plazo del sistema. Para que la implementación de un sistema basado en metadatos resulte exitosa debe existir un compromiso institucional del equipo de trabajo para la catalogación y preservación de los datos a largo plazo. Esto es es imperioso involucrar en estas a actividades al personal técnico de campo, los investigadores, estudiantes, personal técnico de informática y técnicos de laboratorio de la institución productora de datos.

Todo este esfuerzo se da de manera natural en grandes experimentos, en la "gran ciencia", sin embargo, es necesario que los grupos de investigación y los investigadores individuales tomen también consciencia de la importancia que tiene la catalogación y preservación de los datos. Sólo así será posible garantizar su futuro uso y re-uso por parte de las generaciones que nos seguirán (Borgman y colaboradores, 2007).

4 Ejemplos de sistemas de manejos de colecciones de datos

La creciente preocupación por lograr la preservación, el acceso y el análisis de datos científicos de forma automática y en línea, ha llevado a desarrollar repositorios especializados en el manejo de colecciones de datos. Una colección de datos es una serie de observaciones, simulaciones o registros, recolectados y-o generados con una misma metodología. En estas colecciones los metadatos definen los protocolos y mecanismos estandarizados para su identificación, localización, uso y re-uso. De esta forma, los datos son preservados y pueden ser ubicados, accedidos, recuperados y compartidos, garantizando así su utilización a largo plazo y contribuyendo a preservar el patrimonio intelectual, no sólo de nuestras instituciones académicas sino también de organismos públicos y privados que tengan que ver con la producción de datos científicos.

Figure 1: Iniciativas de repositorios de datos: **Cuadrante I:** Interfaz global de búsqueda de datos geoespaciales de GSDI. http://clearinghouse1.fgdc.gov/; **Cuadrante II:** Iniciativa NASA de servicio de datos de ciencias de la tierra http://gcmd.nasa.gov/; **Cuadrante III:** Red de Investigación Ecológica a Largo Plazo *Long Term Ecological Research* (LTER) http://www.lternet.edu/; **Cuadrante IV:** Red del Conocimiento para la Biocomplejidad http://knb.ecoinformatics.org/

Los *clearinghouses* de datos representan, quizá, la mayor experiencia en el uso de los repositorios de datos científicos dentro del ámbito geoespacial. Estos, constituyen una red distribuida de productores y usuarios de datos los cuales permite encontrar y acceder a metadatos y datos geográficos o espaciales. El registro de *clearinghouses* es mantenido por el grupo u organización que promueve su uso dentro de comunidades con intereses compartidos. Tal es el caso, por ejemplo, del *clearinghouse* de datos del "Comité Federal de Datos Geográficos" (FGDC *Clearinghouse*), y de la "Infraestructura Global de Datos Geoespaciales" de los EEUU (GSDI *The Global Spatial Data Infrastructure*) que se ilustra en la Figura 1 Cuadrante I.

Por otro lado, la NASA también ha lanzado una iniciativa para reunir y distribuir gran cantidad de datos sobre el medio ambiente mundial, para el estudio de los efectos de los cambios climáticos a escala global. A través de este sitio Web, se puede obtener datos de diferentes tipos organizados de forma temática, por su localización, por el tipo de instrumentación usada para obtenerlos, o por proyectos (ver Figura 1 Cuadrante II). Desafortunadamente los cambios caprichosos de políticas que apuntan a viajes interplanetarios, han afectado un número importante de proyectos de medición, catalogación y preservación de datos climáticos y las colecciones de datos sobre el calentamiento global de nuestro planeta comienzan a verse afectadas.

Más recientemente, las comunidades de ciencias ecológicas, ambientales y biológicas han venido desarrollando protocolos de preservación y la utilización de estándares de metadatos para el intercambio de información en colecciones de datos. Tal es el caso de la Red de Investigación Ecológica a Largo Plazo (LTER por el acrónimo inglés, *Long Term Ecological Research*). El

programa LTER fue diseñado desde su concepción para considerar e incorporar el manejo y gestión de datos como una componente integral de la investigación científica (ver Figura 1 Cuadrante III). Dentro de su objetivo global de dar facilidad a la investigación se dedicaron significativos esfuerzos en desarrollar métodos para manejar su documentación, y estudiaron todo lo que tiene que ver con formatos y custodia de datos, desarrollo, y mantenimiento de códigos y aplicaciones informáticas, etc. En los 90 se hicieron desarrollos basados en los avances anteriores, utilizando a su vez las tecnologías asociadas a Internet que estaban en rápida expansión para el desarrollo de sistemas de información accesibles globalmente (Torrens, 2003).

Conjuntamente con LTER, el proyecto Red del Conocimiento para la Biocomplejidad (KNB por las siglas *Knowledge Network for Biocomplexity*) desarrolló el Lenguaje EML para el intercambio de metadatos ecológicos y el Metacat como un servidor de almacenamiento centralizado de colecciones de datos, que puede conectarse con otros servidores que usan los mismos mecanismos de comunicación e intercambio (ver Figura 1 Cuadrante IV). Hoy día, la KNB está desarrollando un sistema de almacenamiento y cómputo de datos a través de un GRID (Ecogrid), partiendo de la infraestructura de información que tenían para el momento.

En América Latina es importante mencionar el Programa Brasileño de Investigaciones Ecológicas de Larga Duración, PELD (por *Pesquisas Ecológicas de Longa Duração*). Este programa gubernamental busca organizar y consolidar el conocimiento existente sobre los principales ecosistemas del Brasil. Comprende una docena de sitios que cubren selvas, lagos, ríos y manglares. Las mediciones incluyen datos físicos y químicos de aguas y suelos, densidad de especies y muestras de ADN, y caracterización temporal y espacial de mucha de esta información. La complejidad y diversidad de los datos, conjuntamente con la importancia que para Brasil tiene su medio ambiente ha permitido concretar una biblioteca digital de datos, única en su tipo, con la cual los investigadores brasileños preservan y comparten estos datos (Barros y colaboradores, 2007)

5 Red Bioclimática

En Venezuela, el uso de repositorios para el almacenamiento de datos ambientales, documentados por metadatos, accesible de forma libre a través de Internet es incipiente. Los pocos casos que existen prácticamente se limitan al *clearinghouse* de datos geoespaciales del Instituto Geográfico de Venezuela Simón Bolívar (IGSM) y al Sistema de Información Geográfico del Instituto Venezolano de Investigaciones Científicas (ECOSIG-IVIC http://ecosig.ivic.ve), y, a las iniciativas de la Universidad de Los Andes como son: la Red Bioclimática de Mérida (RedBC) y la Red Venezolana de Estaciones de Investigación Ecológica a Largo Plazo (EcoRed Venezuela).

La EcoRed Venezuela fue establecida en 1997 con el apoyo del FONACyT, la Red Americana de Estaciones de Investigación Ecológica a Largo Plazo (LTER; http://www.lternet.edu/) e instituciones científicas nacionales. Desde entonces, se han hecho grandes esfuerzos para lograr que diferentes comunidades realicen la representación, manejo e intercambio eficiente de datos y colecciones de datos relacionados con las ciencias ecológicas, a través del uso de estándares internacionales y TIC. Con el apoyo del Centro Nacional de Cálculo

Científico Universidad de Los Andes, CECALCULA, la EcoRed-Venezuela ha venido trabajando en el diseño y operación de una plataforma para el manejo de datos provenientes de proyectos de investigación ecológica a largo plazo.

Figura 2: Portal de la Red Bioclimática

5.1 Organización y actores de la RedBC

La RedBC es una iniciativa de el diseño y construcción de ambientes de *e-investigación* y *e-colaboración*, desarrollada por CECALCULA. Consiste en un proyecto piloto, una prueba de concepto de un repositorio de datos para la captura, procesamiento y difusión de información bioclimática (ver portal de entrada al repositorio en la Figura 2). De esta manera, se garantiza la preservación, catalogación, custodia y distribución libre de los datos generados por el proyecto y por distintas instituciones que se han venido sumando a esta iniciativa. Está concebido dentro de un esquema demostrativo y cooperativo donde, instituciones o individuos que dispongan de datos pueden catalogarlos y enviarlos a CECALCULA para su preservación y difusión libre.

Se inicia a raíz de un proyecto denominado Sistema de Información Bio-climático para el Sur de Lago de Maracaibo y la meseta de Mérida (SIBILA). Este proyecto consistió en instalar una red de estaciones telemétricas meteorológicas, construir e instalar un conjunto de colectores de esporas y desarrollar un sistema de información a través del WEB. La herramienta computacional permite el acceso de investigadores, productores y organismos relacionados con la actividad del agro, a la información producida por esta red de estaciones. Se mostró la factibilidad de enviar los datos a centros de acopio en el Sur del Lago y desde allí CECALCULA. Para una segunda fase, una vez acumulados un conjunto mínimo de datos que permitiera determinar correlaciones entre los datos climáticos y las incidencias de algunas afecciones que registran las plantaciones,

se podrían generar, de forma automática, boletines informativos a los productores. Igualmente, a partir de los datos generados por las distintas estaciones, los investigadores podrían formular modelos microclimáticos y diseñar medidas fitosanitarias que permitieran el manejo efectivo de posibles desastres naturales. Recientemente comienzan a concretarse iniciativas de este tipo (Freitez, 2007) y ya nuestro repositorio contiene los primero datos de la incidencia de sigatoka negra en plantaciones de plátanos en el Sur del Lago de Maracaibo (ver sección 5.2 abajo).

5.2 Estaciones climatológicas

En estos momentos son colaboradores de la RedBC el Centro Internacional del Plátano (CIPLAT), el Instituto Nacional de Investigaciones Agrícolas-Estación Chama (INIA-Chama), La Universidad del Sur de Lago, Jesús María Semprúm (UNESUR); La Universidad de Los Andes a través de: el Instituto de Investigaciones Agropecuarias y el Laboratorio de Geofísica; el Parque Tecnológico a través del Centro Nacional de Cálculo Científico y el Instituto de Meteorología e Investigación Climática de la Universidad de Karlsruhe Alemania (*Institut für Meteorologie und Klimaforschung*, IMK).

Hasta el presente contribuyen con datos climáticos las siguientes estaciones:

- **Estación Chama- INIA**. Estación automatizada localizada en el Km. 41, vía Santa Bárbara. Sur del Lago de Maracaibo.
- **Estación CIPLAT, Sur del Lago de Maracaibo**. Estación automatizada localizada en Pueblo Nuevo El Chivo. Sur del Lago de Maracaibo.
- **Estación Mucujún-ULA, El Vallecito**. Estación automatizada perteneciente al Centro de Investigaciones Atmosféricas y del Espacio (CIAE) de la Universidad de Los Andes.
- **Estación La Hechicera-ULA, Mérida**. Estación automática, perteneciente al Instituto de Ciencias Ambientales y Ecológicas de la Universidad de Los Andes (ICAE).
- **Estación Santa Rosa, Mérida**. Estación convencional localizada en el Instituto de Investigaciones Agrícolas de la ULA, Santa Rosa, Mérida.
- **Estación San Juan-ULA, San Juan de Lagunillas, Mérida**. Estación convencional localizada en San Juan de Lagunillas. Pertenece al Instituto de Investigaciones Agrícolas de la ULA.
- **Estación Pico Espejo-MARS, Mérida**. Estación automatizada localizada en la Estación Humboldt, Pico Espejo Mérida, Parque Nacional Sierra Nevada. La estación forma parte de un convenio entre Venezuela (Universidad de Los Andes) y el Gobierno de Alemania.
- **Estación La Glorieta-UNESUR**. Estación automatizada localizada en Santa Bárbara del Zulia. Pertenece a la Universidad Nacional Experimental Sur del Lago "Jesús María Semprúm" (UNESUR).
- **Estación La Chiquinquirá-UNESUR**. Estación automatizada localizada en Santa Bárbara del Zulia. Pertenece a la Universidad Nacional Experimental Sur del Lago.

Cada una de las estaciones se muestran en la Figura 3 se encuentran registradas en la Web, con sus coordenadas geográficas y ubicación física, sus especificaciones técnicas, las variables que registra y el contacto responsable de la operación de la estación.

Figura 3: Ubicación geográfica de las estaciones climatológicas.

5.3 Procesos, Datos y Metadatos

La RedBC funciona a través de un Sistema de Manejo Centralizado de Datos en el que la unidad inicial de coordinación es CECALCULA. Tal y como puede apreciarse en la Figura 4, el suministro de datos se realiza vía mensajería electrónica (correo electrónico o mensajería instantánea) o a través de envío en medios digitales (*diskette*, CD, DVD). El propietario de una estación o un investigador puede publicar sus datos fácilmente o, si lo desea, sólo publica los metadatos. En CECALCULA, personal de soporte y apoyo a la RedBC revisa los archivos de información: la organiza, genera los metadatos y verifica que éstos se ajusten a los estándares predefinidos. En este nivel se almacenan los datos provenientes de las estaciones (temperatura, precipitación, etc.) en un formato que permite su fácil consulta, recuperación y actualización. Los datos y metadatos son de libre acceso y se publican en el portal de la RedBC. De esta manera pueden ser consultados por los miembros de la red o por cualquier persona o institución que los necesite.

Figure 4: Esquema de procesos de envío y preservación de los datos bioclimáticos. El suministro de datos se realiza vía mensajería electrónica (correo electrónico o mensajería instantánea) o a través de envío en medios digitales (*diskette*, CD, DVD)

5.4 Colecciones de Datos

Los datos se organizan en colecciones, con la finalidad de facilitar el uso de estándares de metadatos para su documentación. Se presentan en archivos clasificados por estación y por año. Las estaciones automáticas registran los datos por horas en archivos texto delimitados por un carácter (tabulación, espacio en blanco o coma). Las estaciones convencionales registran los datos por día. Si bien no es una norma estricta, se recomienda a los colaboradores que los archivos a ser enviados a la RedBC posean datos crudos, recabados en campo, en lugar de datos procesados. La estadística sumaria, figuras y otros comentarios se pueden anexar en un archivo separado en la documentación. Con esta información se construye otra colección derivada.

Las colecciones de datos disponibles a través del portal del la RedBC son:

- Datos climatológicos Estación Chama período 2001-2005
- Datos climatológicos CIPLAT período 2002-2003
- Datos climatológicos Estación Mucujún período 2000-2008
- Datos climatológicos Estación Santa Rosa período 1974-2001
- Datos climatológicos Estación San Juan período 1995-2001
- Datos climatológicos Estación Pico Espejo-MARS período 2004-2007
- Datos climatológicos Estación La Hechicera 2000-2004
- Datos climatológicos Estación La Glorieta1 período 2005-2008
- Datos climatológicos Estación La Glorieta2 período 2008
- Datos climatológicos Estación La Chiquinquirá período 2005-2008

- Datos de evaluación de la enfermedad Sigatoka Negra en una plantación de plátano Hartón (*Musa AAB cv. Hartón*)

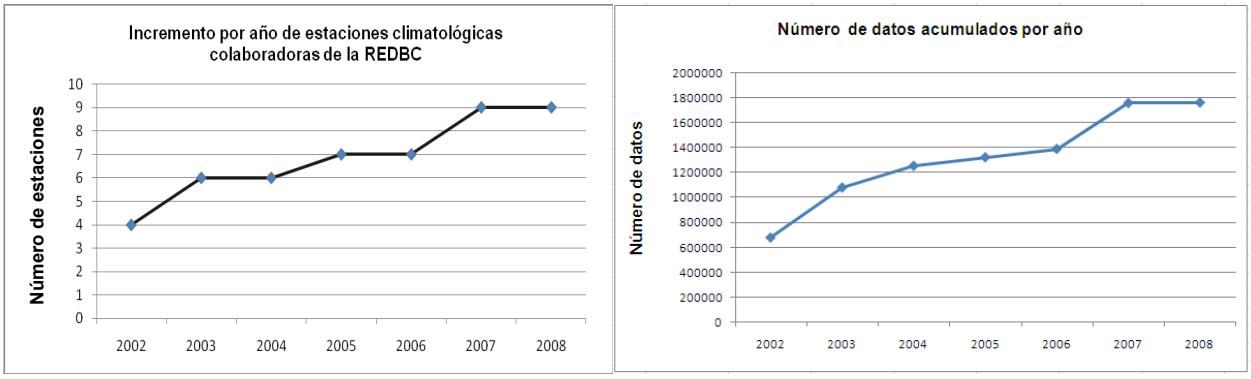

Figura 5: Estadísticas de crecimiento en la preservación y catalogación de datos. Es claro el incremento del número de estaciones y registros: cuatro estaciones y 679.094 registros en 2002 pasamos a nueve estaciones 1.755.986 para finales del 2007.

5.5 Datos libres de uso creciente e impacto fitosanitario

Actualmente, la RedBC proporciona al usuario, a través de Internet, una descripción completa de once colecciones de datos, de tal modo que estas puedan ser descubiertas, transferidas y usadas en el contexto apropiado por otros usuarios potenciales. En la Figura 5 cuadrante I, se puede apreciar el ritmo de crecimiento del número de estaciones que se han ido sumando a la RedBC. Se comenzó en el 2002 con cuatro estaciones, dos en la zona Sur del Lago de Maracaibo y dos en la meseta de Mérida, y hoy en el 2008 se dispone de nueve. Si bien el número de estaciones es limitado, el éxito de este enfoque voluntario, de esta prueba de concepto, se corrobora también con el incremento de la cantidad de datos que se han almacenado: de 679.094 registros en 2002 pasamos a 1.755.986 para finales del 2007 (ver Figura 5 cuadrante II).

Recientemente (Freitez, 2007) un nuevo conjunto de datos se ha venido a correlacionar con los datos climáticos. El cultivo de musáceas (plátanos y bananos), en Venezuela, es una de las actividades económicas más importantes en la región del Sur del Lago de Maracaibo. Más de la mitad del total de superficie que en el país se destina a este tipo de cultivo se encuentra concentrada en esta región. En la última década, un hongo (la Sigatoka Negra o *Mycosphaerella Fijiensis Morelet*) ha impactado la producción de rubro disminuyendo la calidad del producto y aumentando los costos por el uso periódico de agroquímicos, los cuales, además contaminan significativamente aguas y suelos. Por lo tanto, el asociar la incidencia de la Sigatoka a variables climáticas es de gran interés ya que focaliza el uso de agroquímicos cuando la severidad de la patología lo requiera. Por ello, los investigadores del INIA, CIPLAT y ULA colocaron una serie

de "caza esporas" las cuales registran la densidad del inóculo en el ambiente. Con estos nuevos datos se ha comenzado a intentar correlacionar variables bioclimáticas. Los resultados preliminares muestran que en algunos casos se logra predecir la intensidad de la enfermedad, con y sin discriminación por temporada climática (lluvia y sequía) con correlaciones hasta de 79,22 %. Para la temporada de sequía (octubre-marzo) se destacan como variables correlacionadas con la densidad del inóculo en el ambiente: la humedad relativa, precipitación, velocidad del viento y energía solar, mientras que en el caso de la temporada lluviosa (abril-septiembre), las variables mejor correlacionadas son la temperatura del aire, evapotraspiración, humedad relativa y precipitación.

6 Conclusiones y Recomendaciones

Sin duda, este lustro de operación del Repositorio de Datos Libres de la RedBC ha mostrado su viabilidad. Además, su mantenimiento en el tiempo y su creciente uso lo respaldan. Una serie de encuestas informales a usuarios que se comunican con nuestro personal de soporte, muestra su amplia variedad. Son estudiantes de educación media, estudiantes universitarios, investigadores, profesores, tesistas de pre y postgrado de distintas instituciones nacionales y extranjeras. En el éxito de este tipo de proyectos juegan un papel primordial los esquemas culturales y costumbres institucionales, por lo tanto es preciso tener claro que el convencer a los investigadores e instituciones generadoras de datos para que los resguarden en un repositorio y los compartan, es un proceso muy lento. Pero, si por estas dificultades, no se toma la iniciativa, nunca se logrará que se acelere el crecimiento de la producción científica, intelectual y por ende el desarrollo económico de nuestros países. El mayor de los logros de esta red es la cooperación y en ello ha sido clave el adiestramiento técnico y la visión cooperativa en la creación del conocimiento. Varios talleres y cursillos de adiestramiento en el manejo de datos han surgido y se seguirán desarrollando dentro de la Red. Igualmente, la consultoría y asesoría en el mantenimiento de estaciones es casi permanente. Aunque existen no pocas dificultades para convencer algunos investigadores y grupos de investigación de la importancia para compartir los datos y colocarlos con acceso libre, el fácil acceso de los datos en la Web y su perpetuación, así como también el reconocimiento recibido de otras instituciones que los utilizan, ha estimulado el deseo de pequeños grupos a contribuir con sus datos. Es así como el éxito de la experiencia de la RedBC se puede medir por las estadísticas de incorporación de estaciones y de preservación de datos.

Como hemos mencionado arriba, no han sido pocas las dificultades que se han enfrentado y se seguirán enfrentando en este camino de preservar el patrimonio de los datos. Entre las mayores dificultades encontradas se pueden mencionar:

- Escaso conocimiento de la importancia de la preservación y posibilidades de uso secundario de los datos.
- Poca disposición para compartir los datos. La mayor parte de los investigadores suponen que los datos son de su propiedad.
- Poca receptividad para aportar metadatos que documenten las colecciones de datos.
- Bajo interés por parte de los dueños de los datos en publicar sus colecciones de datos (retardo en envío de datos-pérdida de información).

- Información incorrecta sobre la calidad y cantidad de información que dicen poseer algunas instituciones (entorpece la investigación).
- Algunas instituciones no disponen de conexión a Internet.
- Costos de los instrumentos de captura de datos.
- No existe la figura de "gerente local de información" en el caso de las estaciones.

La RedBC proporciona al usuario, a través de Internet, de una descripción completa de las colecciones de datos de tal modo, que estas puedan ser descubiertas, transferidas y usadas en el contexto apropiado por otros usuarios potenciales. La creación de este repositorio de datos científicos ha hecho posible que éste funcione como un proyecto que le permite a una parte de la comunidad científica venezolana empezar a catalogar y sistematizar sus colecciones de datos ambientales y, más importante aún, saber donde buscar al realizar investigaciones posteriores.

Si esta experiencia sirve de estímulo a otras instituciones y se logra dar un efecto multiplicador, nuestro país saldrá beneficiado y fortalecido al disponer de redes de colaboración en todos los estados que inclusive pudieran interconectarse para ofrecer un mejor servicio. Estas conformarían redes de manejo de conocimientos basadas en datos y metadatos (intención de KNB). Un primer paso en esta replicación de esfuerzos en la preservación, lo ha constituido la iniciativa de la UNESUR en seguir el ejemplo de la RedBC y conformar también un repositorio para el almacenamiento, resguardo y difusión de sus datos bajo el asesoramiento de CeCalCULA.

Agradecimientos

Este proyecto ha sido financiado por el programa de la Agenda Plátano del CDCHT de la Universidad de Los Andes bajo el número CVI PIC AGM01099. Igualmente los autores agradecen a Fundacite Mérida por el apoyo financiero brindado. Adicionalmente uno de nosotros (LAN) agradece también el financiamiento del Fondo Nacional de Investigaciones Científicas y Tecnológicas bajo los proyectos S1-2000000820 y F-2002000426 y la hospitalidad del Grupo de Relatividad y Gravitación de la Universidad de las Islas Baleares en España, que permitió la finalización de la redacción de este trabajo.

References

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

Arzberger, P., Schroeder, A., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Uhlir, P., y Wouters, P. (2004a). *Promoting access to public research data for scientific, economic, and social development*. Data Science Journal, **3**:135–152.

Arzberger, P., Schroeder, P., Beaulieu, A., y Bowker, G. (2004b). *Science and government: An international framework to promote access to data*. Science, **303**:1777-1778.

Barros, E., Laender, A., Gonçalves, M., y Cota, R. (2007). *Transitioning from the ecological fieldwork to an online repository: a digital library solution and evaluation*. Int J Digit Libr **7**:109-112.

Borgman, C. (2006). *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.

Borgman, C., Wallis, J., y Enyedy, N. (2007). *Little science confronts the data deluge: habitat ecology, embedded sensor networks and digital libraries*. Int J Digit Libr, **7**:17–30.

Clark, R. (2007). *Database protection in europe—recent developments and modest proposal*. Data Science Journal, **6**:OD12–OD20.

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

Collins, L., Martínez, M., Mane, K., Powell, J., Kieffer, C., Simas, T., Heckethorn, S., Varjabedian, K., Blake, M., y Luce, R. (2007). *Collaborative escience libraries*. Int J Digit Libr, **7**:31–33.

Dávila, J. A., Núñez, L., Sandia, B., Silva, J. G., y Torrens, R. (2006a). *www.saber.ula.ve: un ejemplo de repositorio institucional universitario*. Interciencia, **31**(1):29–37.

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

De Roure, D. y Frey, J. (2007). *Three perspectives on collaborative knowledge acquisition in e-science*. En Workshop on Semantic Web for Collaborative Knowledge Acquisition. Twentieth International Joint Conference on Artificial Intelligence, Hyderabad, India, Enero 2007.

Dozier, J., Alexander, S., Courain, M., Dutton, J. A., Emery, W., Gritton, B., Jenne, R., Kurth, W., Lide, D., Richard, B. K., y Warnow-Blewett, J. (1995). *Preserving scientific data on our physical universe: A new strategy for archiving the nation's scientific information resources*, Reporte Técnico, National Research Council, Inglaterra.

Foster, I. (2005). *Service-oriented science*. Science, **308**:814–817.

Freitez, J.A. (2007). *Desarrollo de un Modelo Predictivo del Brote de Sigatoka Negra para las Plantaciones de Plátano al Sur del Lago de Maracaibo* Tesis de Maestría, Ingeniería de Sistemas, Facultad de Ingeniería, Universidad de Los Andes, Mérida, Venezuela.

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

Hey, T. y Trefethen, A. (2003a). *The data deluge: An e-science perspective*. En Grid Computing: Making the Global Infrastructure a Reality, Berman, F., Fox, G., y Hey, T., editores, John Wiley & Sons Ltd, New York, pag 809–824.

Hey, T. y Trefethen, A. E. (2003b). *e-science and its implications*. Phil. Trans. R. Soc. Lond. A, **361**:1809–1825.

Hey, T. y Trefethen, A. E. (2005). *Cyberinfrastructure for e-science*. Science, **308**:817–821.

Karasti, H., Baker, K., y Halkola, E. (2006). *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.

Lessig, L. (2004). *Free Culture*. The Penguin Press, New York edition.

Lord, P. y Macdonald, A. (2003). *Data curation for e-science in the uk: an audit to establish requirements for future curation and provision*. Reporte Técnico, The JISC Committee for the Support of Research, The Digital Archiving Consultancy Limited.

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities and relationships* Reporte Técnico, UKOLN Bath Inglaterra.

Michener, W., Brunt, J., Helly, J., Kirchner, T., y Stafford, S. (1997). *Nongeospatial metadata for the ecological sciences*. Ecological Applications, **7**(1):330–342.

Murray-Rust, P. (2008). *Open data in science*. precedings.nature.com.

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

Pilat, D. y Fukasaku, Y. (2007). *OECD principles and guidelines for access to research data from public funding*. Data Science Journal, **6**(Open Data Issue):OD4–OD11.

Sabourin, M. y Dumouchel, B. (2007). *Canadian national consultation on access to scientific research data*. Data Science Journal, 6(Open Data Issue):OD26–OD35.

Simberloff, D., Barish, B. C., Droegemeier, K. K., Etter, D. M., Fedoroff, N. V., Ford, K. M., Lanzerotti, L. J., Leshner, A., Lubchenco, J., Rossmann, M. G., y White Jr., J. A. (2005). *Long-lived digital data collections enabling research and education in the 21st century*. Reporte Técnico NSB-05-40, National Science Foundation.

Torrens, R. (2003). *Desarrollo de sistemas de información bio-climática*. Tesis de Maestría, Ingeniería de Sistemas, Facultad de Ingeniería, Universidad de Los Andes, Mérida, Venezuela.

Uhlir, P. y Schröder, P. (2007). *Open data for global science*. Data Science Journal **6**(Open Data Issue):OD36-OD53.

Xu, G. (2007). *Open access to scientific data: Promoting science and innovation*. Data Science Journal, *6*(Open Data Issue):OD21–OD25.