Un experimento de creación de biblioteca digital con *Greenstone*

Por **Piedad Garrido Picazo** y **Jesús Tramullas Saz**

Piedad Garrido Picazo, profesora del Depto. de Informática e Ingeniería de Sistemas, Univ. de Zaragoza y licenciada en documentación. Sus líneas de investigación se centran en bases de datos xml y topic maps. Pertenece al Grupo de Investigación en Información Digital Infodig y a la Red Temática de Investigación sobre Documentación Digital.

Resumen: *Este trabajo revisa y evalúa las prestaciones que ofrece la herramienta de software libre Greenstone para la creación de colecciones de documentos digitales. Analiza las diferentes interfaces disponibles, así como los problemas de procesamiento de información detectados.*

Palabras clave: *Bibliotecas digitales, Desarrollo de colecciones, Evaluación de interfaces, Greenstone.*

Title: An experience in digital library creation using *Greenstone*

Abstract: *This paper reviews and evaluates the features of Greenstone, a free software tool, for creating and developing digital document collections. An analysis is given of the different interfaces that are available and problems related to information processing are identified.*

Keywords: *Digital libraries, Collection development, Interface evaluation, Greenstone.*

Garrido Picazo, Piedad; Tramullas Saz, Jesús. *"Un experimento de creación de biblioteca digital con Greenstone". En: El profesional de la información, 2004, marzo-abril v. 13, n. 2, pp. 84-92.*

Jesús Tramullas Saz, profesor de documentación automatizada, Univ. de Zaragoza. Su investigación se centra en bibliotecas digitales y en el diseño y evaluación de servicios de información digital. Pertenece al Grupo de Investigación en Información Digital Infodig y a la Red Temática de Investigación sobre Documentación Digital.
http://tramullas.com

1. Planteamiento

La creación y desarrollo de bibliotecas digitales es una de las tareas que están llevando a cabo las bibliotecas en los últimos años. Independientemente del enfoque desde el cual se aborden, la complejidad de las mismas, así como los requerimientos técnicos, delimitan sobremanera la capacidad de crear colecciones digitales. La disponibilidad de softwares que faciliten este tipo de servicios es una de las necesidades más acuciantes para conseguir extender los servicios y la información disponible a todos los usuarios. En este contexto, la mayoría de las plataformas para bibliotecas digitales son propietarias, y son contadas las herramientas que pueden conseguirse e instalarse libremente (Dspace, Ghanesa, etc.) y casi únicas las que requieren una plataforma informática fácilmente instalable (Greenstone). Greenstone es la herramienta de software libre más desarrollada e instalada a nivel mundial para la creación de bibliotecas digitales. Se encuentra reconocida e incluida en el programa de información y documentación de la Unesco a tal fin, la cual desarrolla un programa de promoción y formación, a escala global, de este programa. La revisión de la bibliografía sobre esta aplicación permite apreciar la creciente importancia de la misma, así como el interés que genera tanto en la comunidad científica como en la profesional.
http://www.greenstone.org

La planificación, el trabajo y las complicaciones que plantea la construcción y desarrollo de una colección de documentos digitales, que por supuesto superan largamente los problemas relacionados con la mera digitalización de fondos, obligan a probar directamente las posibilidades, y a testear las situaciones que puedan producirse en un contexto real de actividad. En estas circunstancias, el presente trabajo analiza los procedimientos técnicos necesarios para construir una colección con *Greenstone*, así como los problemas téc-

Artículo recibido el 12-12-03
Aceptación definitiva: 28-02-04

nicos y de interfaz encontrados. No se abordan en este análisis las cuestiones referidas a la recuperación y acceso a la información dentro de la misma. La colección *Jbidi*, construida y analizada para este experimento, se encuentra, además, disponible en línea.
http://imhotep.unizar.es/greenstone

2. Servicios básicos de las bibliotecas digitales

Independientemente de la variedad de conceptos y definiciones que subyacen al metatérmino "biblioteca digital", pueden rastrearse en casi todas ellas un conjunto de elementos que, si bien no llegan a completar una definición consensuada, sí que permiten, por contra, establecer un panorama de componentes que en principio deberían encontrarse presentes para poder calificar a un servicio de información como una biblioteca digital (**Tramullas**, 2002). Los elementos básicos de una arquitectura y de un servicio de biblioteca digital deberían responder a:

—Colección: creación y gestión de colecciones de documentos digitales, locales o distribuidos, independientemente de su formato.

—Servicios de valor añadido: el mero acceso no debería considerarse un servicio, pues éstos serían aquellos productos o facilidades creados para poner en valor el contenido de la colección, adecuados a las necesidades y a los requerimientos de sus usuarios.

—Personalización: capacidad para que el usuario pueda establecer su propia organización y selección de los elementos de la colección.

—Ciclo de vida de la información: la colección digital puede tener unas fases diferentes en sus diversas etapas que las que ofrece una colección tradicional en soporte físico.

3. *Greenstone* es software para bibliotecas digitales

Es una aplicación para la construcción y explotación de bibliotecas digitales, creada y mantenida por *New Zealand Digital Library Project*, en la *University of Waikato*, en un proyecto coordinado por **Ian Witten**. Tiene como núcleo el motor de indización y recuperación de información textual *MG* (**Witten**; **Moffat**; **Bell**, 1999), que ha sido integrado con un entorno web, sobre el que se dispone la interfaz básica de usuario. Dicho motor crea una representación vectorial de los documentos textuales. La aplicación está formada por diferentes macros, programadas en *Perl*, encargadas del tratamiento y recuperación de la información textual, y por un conjunto de plugins que actúan como filtros de importación para diferentes formatos de documentos digitales. Como sistema de gestión de bases de datos para soporte a los procesos, se utiliza *GDBM* (*GNU Database Manager*). *Greenstone* ofrece la posibilidad de exportar colecciones a soporte cd, mediante la instalación del paquete *Export*. Se están desarrollando, por otra parte, aplicaciones complementarias como *Phind* que extienden sus capacidades. Para aquellos casos en los que sea necesario habilitar interfaces de búsqueda con múltiples campos, se sustituye el motor *MG* por *MGPP/MG+*. Actualmente se encuentran disponibles versiones para plataformas *Linux/Unix, Windows* y *Apple Macintosh*. Se distribuye bajo la licencia *GNU General Public License*.

«La mayoría de las plataformas para bibliotecas digitales son propietarias y son pocos los programas que pueden conseguirse e instalarse libremente (Dspace, Ganesha, Greenstone)»

El funcionamiento básico de *Greenstone* es sencillo (**Witten**; **Bainbridge**, 2003): se construye una colección, importando documentos, para luego poder acceder al contenido de los mismos mediante diferentes criterios. Admite documentos digitales en numerosos formatos, como texto ascii, *Word*, *Adobe Acrobat*, *Excel*, *PowerPoint*, xml, html (tanto locales como remotos), *Refer* y *BibText*, *PostScript*, correo electrónico y recientemente *CDS-ISIS*. Toda la información textual es procesada y transformada en ficheros en formato xml con caracteres *Unicode*, *UTF-8*. Como implementa el protocolo *Corba*, puede deducirse que pueda trabajar tanto con colecciones locales como distribuidas. Está prevista la posibilidad de que se puedan crear e integrar nuevos plugins para otros formatos en el sistema. La aplicación también está preparada para trabajar con metadatos, tanto detectados automáticamente como asignados de forma manual, así como para poder modificar y ajustar las interfaces de usuario final de recuperación y consulta de información.

El proceso estándar responde a las siguientes fases:

—Creación de la colección: definición de su nombre, descripción de su contenido y responsable de su gestión.

—Selección de los documentos: especificación de aquellos que se van a incluir en la colección y de la ubicación de los mismos.

—Configuración de la colección: parámetros básicos de tratamiento, indización y presentación de la colección. Definición de la interfaz de usuario. *Greens-*

tone ofrece una configuración estándar, que puede ser modificada.

—Construcción de la colección: de acuerdo con los parámetros establecidos en la fase anterior, los documentos son importados, procesados por los plugins correspondientes, transformados en xml, e indizados. Se crean los índices textuales y las estructuras de organización y acceso, llamadas clasificadores.

—Consulta de la colección: procesos de localización y acceso a los documentos, utilizando la interfaz disponible, mediante búsqueda en índices textuales o utilización de los clasificadores.

4. La evaluación de bibliotecas digitales

La evaluación de herramientas de gestión documental, así como de sus interfaces de usuario, especialmente en el campo de los opacs, las prestaciones de los mismos, las capacidades de los lenguajes de recuperación, y otros aspectos, es un clásico en la bibliografía especializada sobre biblioteconomía y documentación desde los años 80. El desarrollo de las bibliotecas digitales en la segunda mitad de la década de 1990 ha traído consigo la cada vez mayor importancia prestada a los procesos de evaluación de las herramientas y las colecciones, especialmente desde la perspectiva del usuario final. Prueba de lo anterior es la celebración de sesiones específicas dedicadas a la evaluación en las *Joint conference ACM/IEEE on digital libraries* (*JCDL*) y en las *European conference on digital libraries* (*ECDL*). La investigación ha mostrado un creciente interés por desarrollar métodos y técnicas de evaluación de bibliotecas digitales, que han atendido especialmente a tres áreas específicas, como han indicado **Fuhr,** et al., (2002):

—Desarrollo de métricas: creación de métodos y técnicas para evaluar, de manera mensurable, las prestaciones de los sistemas software y hardware para bibliotecas digitales. En este campo se presta particular interés a los sistemas de indización y recuperación de la información. Destaca especialmente el grupo de trabajo establecido en *Ercim Delos network on excellence on digital libraries* que en 2002 dedicó un seminario específico a la cuestión (**Borgman**, 2002).

—Análisis de colecciones: estudio y análisis de los procesos de formación y organización de las colecciones digitales que ha dado especial importancia a los problemas relacionados con la calidad de los procedimientos de digitalización de textos e imágenes.

—Usuarios: estudio de los comportamientos y pautas de uso de las bibliotecas digitales, por parte de los usuarios finales (**Bollen**; **Luce**, 2002) con especial énfasis en las técnicas de diseño centrado en el usuario y usabilidad. Los trabajos sobre la aplicación del diseño centrado en el usuario, en el contexto de las bibliotecas digitales, inciden en estos aspectos (**Tramullas**, en prensa).

En un trabajo de referencia, **Saracevic** (2000) estableció que, dada la complejidad del espacio de información digital al que identificamos como biblioteca digital, los procesos de diseño y evaluación debían utilizar como pautas las nociones de constructos (¿qué es una biblioteca digital?), contextos (¿qué es evaluar una biblioteca digital?), criterios (¿qué evaluar en el contexto?), medidas (¿qué medidas aplicar a los criterios?) y métodos (¿cómo evaluar?). **Marchionini** (2000), en su evaluación de la biblioteca digital *Perseus* estableció, en primer lugar, los objetivos y el contexto de la biblioteca, para definir áreas principales de interés, sobre las cuales llevar a cabo la evaluación. La toma de datos se realizó mediante técnicas de observación y análisis. Se definieron cinco áreas principales correspondientes a infraestructura física y conceptual, ventaja mecánica, incrementos, desarrollo de comunidad y cambio sistemático. **Tramullas** (2003), para un estudio de bibliotecas digitales en universidades españolas, definió el modelo de análisis *Cabdu*, que evaluaba atendiendo a seis áreas: identificación y contextualización, organización y desarrollo, contenidos, colecciones y digitalización, servicios, infraestructura y percepción del usuario.

«Greenstone es el software libre más desarrollado e instalado a nivel mundial para la creación de bibliotecas digitales»

El trabajo que se plantea a continuación se encuentra enmarcado en las actividades de evaluación definidas en los párrafos anteriores. Establece como contexto de actuación la creación de una biblioteca digital de documentos ofimáticos mediante la aplicación *Greenstone*, para evaluar dentro del mismo las siguientes cuestiones:

—La capacidad técnica de tratamiento de los documentos textuales.

—La capacidad y prestaciones disponibles para la organización de la colección.

—La extracción de metadatos para su uso en organización y recuperación.

—Las interfaces de usuario final disponibles para la creación de colecciones.

—La generación de interfaces de usuario final para el acceso y consulta de la colección.

5. Desarrollo del experimento

Establecidas las premisas básicas en apartados anteriores, el objetivo principal de la investigación ha sido determinar las prestaciones de *Greenstone* para la creación, desarrollo y mantenimiento de bibliotecas digitales. Se trata de un trabajo novedoso, no realizado con anterioridad sobre esta herramienta, que tiene una utilización creciente, como se ha demostrado en párrafos anteriores. Los problemas que plantea la interfaz de desarrollo de colecciones son lo suficientemente importantes como para dedicar un trabajo, que sirva como elemento de valoración y crítica ante la aplicación. Este tipo de evaluaciones debe tener un componente técnico ineludible. Basta revisar los contenidos de las diferentes *JCDL* o *ECDL*, los foros más prestigiosos a nivel mundial sobre la cuestión, para comprobar la validez del enfoque.

«Greenstone admite documentos digitales en numerosos formatos, como texto ascii, Word, Adobe Acrobat, PostScript, Excel, PowerPoint, xml, html (tanto locales como remotos), Refer y BibText, correo electrónico, y recientemente CDS-ISIS. Toda la información textual es tranformada en Unicode»

Ya se han explicado la arquitectura y componentes de *Greenstone*, así como las capacidades básicas de este software. Sin embargo, y como ya se ha indicado, es necesario ajustar la configuración y prestaciones de la aplicación en y para cada caso específico. En este sentido, el experimento se planteó desde el punto de vista de la creación de una colección de documentos ofimáticos comunes, como los formatos *Adobe Acrobat* (pdf) y *Microsoft Word* (doc), con capacidad de búsqueda a texto completo. La versión de *Greenstone* utilizada fue la 2.40 —desde el 8 de diciembre de 2003 se encuentra disponible la versión 2.41—, tanto en modo local, como a través de servidor web (en este caso *Apache 1.33*); el servidor instalado utilizaba *Windows 2000 Professional SP4*. Con la finalidad de comparar las prestaciones de *Greenstone*, esta colección se creó y desarrolló utilizando las tres interfaces de usuario disponibles:

```
indexes         document:text document:Title document:Source
defaultindex    document:text

plugin          ZIPPlug
plugin          GAPlug
plugin          TEXTPlug
plugin          HTMLPlug
plugin          EMAILPlug
plugin          PDFPlug
plugin          RTFPlug
plugin          WordPlug
plugin          PSPlug
plugin          ArcPlug
plugin          RecPlug


classify        AZList -metadata Title
classify        AZList -metadata Source
```

Figura 1. Fichero collect.cfg predefinido

—*Collector*, mediante un cliente web.

—La interfaz de línea de órdenes o comandos.

—*GLI* (*Greenstone Librarian Interface*).

En los tres casos se han llevado a cabo dos procesos de creación diferentes:

—En el primero, al que denominaremos directo, los documentos son tratados directamente, dejando a *Greenstone* las tareas de extracción de sus metadatos, así como la creación de estructuras de organización y acceso a la colección.

—Un segundo proceso, al que denominaremos mediado, donde los metadatos de los documentos han sido facilitados en ficheros xml por los autores, y se han creado estructuras específicas de organización y acceso a la colección.

El panel de test se ha compuesto con 30 documentos, en los formatos arriba indicados, correspondientes a las *Actas de las Jornadas de bibliotecas digitales Jbidi* en sus ediciones de 2000, 2001, 2002 y 2003. La estructura formal de los documentos corresponde a la establecida por *Springer* para sus publicaciones.

5.1. Proceso directo

Consistió en alimentar directamente a *Greenstone* con los documentos originales, sin complementos, para observar los errores que se pudieran producir en su fase de captura y tratamiento, y analizar las estructuras de organización y acceso que ofrece a partir de los datos obtenidos de los mismos. En primer lugar, se procedió a crear la colección directamente desde la interfaz *Collector* (la documentación básica para ello se encuentra recogida en la *Greenstone user's guide*). En la misma fue necesario indicar individualmente todos y cada uno de los documentos ya que, a pesar de que se puede introducir un directorio (lo que teóricamente supone la captura de todos los documentos) en realidad el proceso, por razones no aclaradas, falló en dos de las tres ocasiones en las que se llevó a cabo. No se modificó el fichero de configuración *collect.cfg* que *Greenstone* incluye por definición (figura 1) y se procedió

```
WARNING: No plugin could process \HASH8321.dir\doc.xml
*** creating auxiliary files
test collection built successfully
installing the test collection
build: Failed to install collection to C:\program files\gsdl\collect\test
Collection will remain at C:\program files\gsdl\tmp\tbuild2\test
```

Figura 2. Información de errores en Collector

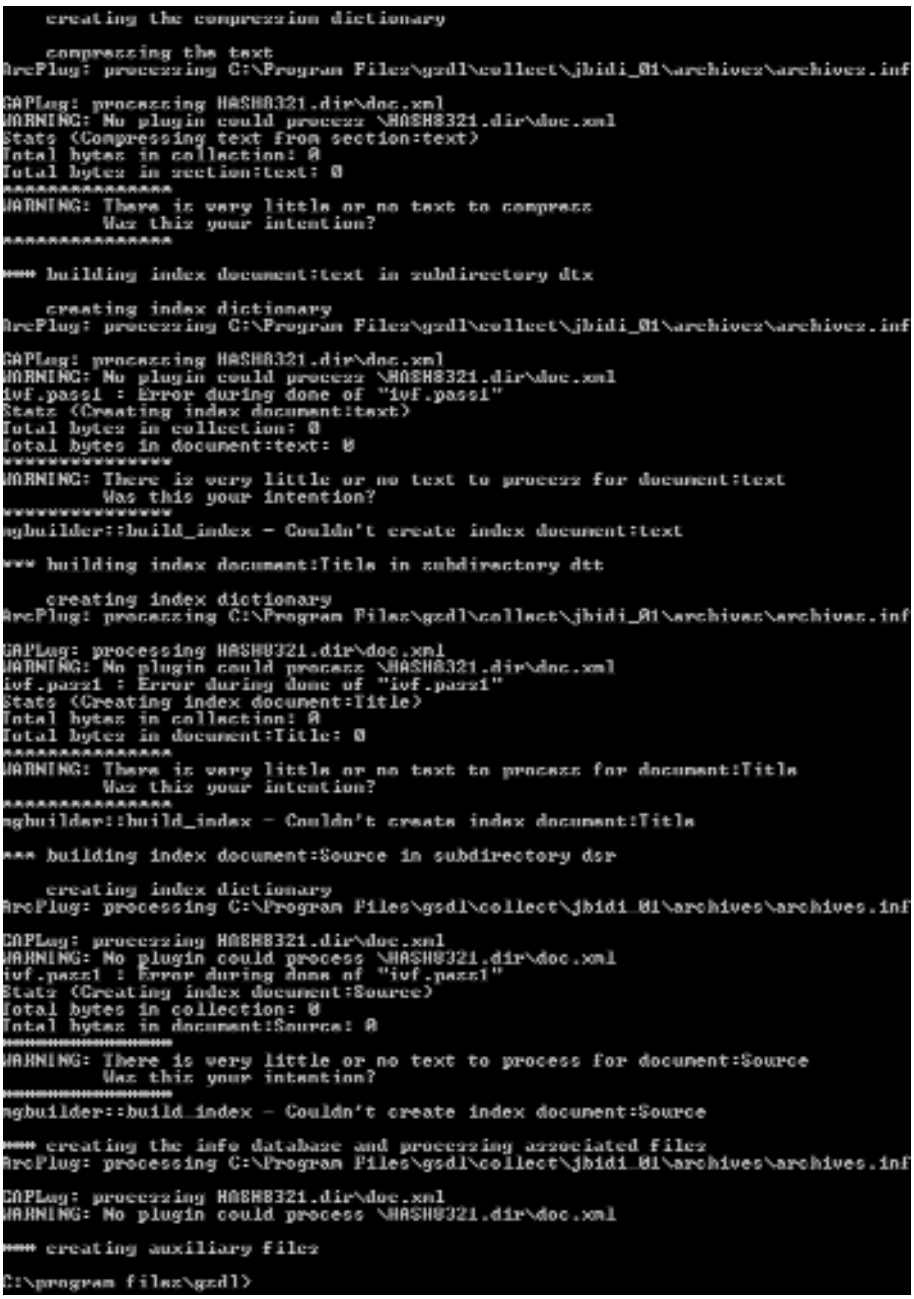

Figura 3. Información detallada en la interfaz de línea de órdenes

datos que ha conseguido detectar. En este proceso sólo se revelaron como metadatos los nombres de fichero, lo que resulta a todas luces insuficiente.

En segundo lugar, se llevó a cabo todo el proceso a través de la interfaz que ofrece la línea de órdenes o comandos. En este caso es necesario abrir una sesión de consola *DOS* y activar el intérprete *Perl* para formular las instrucciones y los parámetros de las mismas en la secuencia establecida (la documentación básica para ello se encuentra recogida en la *Greenstone developer's guide*). De esta manera fue posible procesar directamente todo el contenido del directorio. De los treinta documentos del panel de prueba fueron rechazados los mismos diez que lo fueron en la prueba con la interfaz *Collector*. Así pues, al poder disponer de los mensajes de error fue posible conocer rápidamente cuáles no habían sido aceptados y la causa de ello. De la misma forma, la importación del contenido y su tratamiento y transformación en documentos xml es referenciada y explicada claramente, lo que facilita obtener gran cantidad de información sobre los procesos desarrollados por *Greenstone* (figura 3). Una vez creada y activada la colección, la consulta de la misma reveló que se repetían los mismos índices para búsqueda textual, y los mismos clasificadores insuficientes obtenidos con el proceso anterior.

al proceso de tratamiento del contenido de los documentos.

Se observó que diez de los documentos propuestos no eran procesados, sin apenas mayor información de los errores producidos, lo que dificultaba el estudio de las causas del fallo (figura 2). Una vez disponible la colección, las estructuras de organización y acceso, a las que denomina clasificadores, tampoco resultaron especialmente adecuadas. La interfaz estándar para colecciones incluye la posibilidad de buscar a través de un índice de texto completo, y crea clasificadores utilizando los meta-

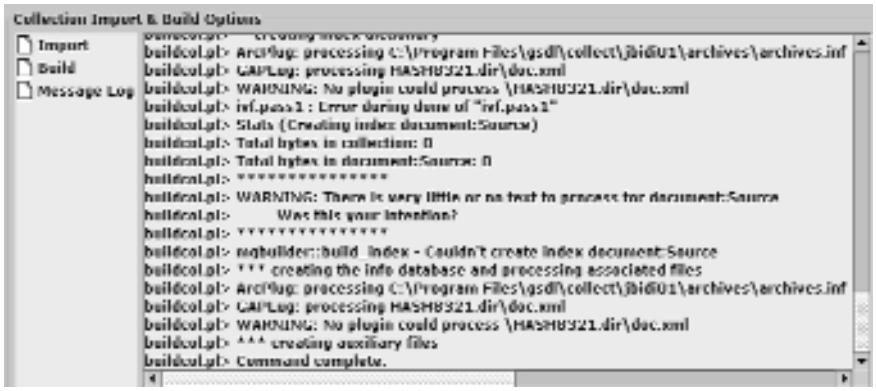

Figura 4. Información de tallada en GLI

Por último, se llevó a cabo el proceso utilizando la nueva interfaz *GLI*, una aplicación de entorno gráfico, desarrollada en *Java*, que se incorporó a *Greenstone* en la versión 2.40. Este entorno facilita todo el proceso de creación directa de colecciones mediante la guía de los procesos necesarios, así como la posibilidad de configurar toda la colección de forma más intuitiva. Por el momento esta aplicación no se ha visto acompañada de una documentación adecuada, lo que dificulta su correcta utilización. En este experimento la aplicación falló en tres ocasiones en la creación de la colección bloqueando los procesos debido a la utilización de una versión inadecuada de *Java Runtime Environment* (1.3), la cual fue sustituida por la 1.4.1, con lo que la aplicación funcionó correctamente. *GLI* aceptó el procesamiento del directorio, e informó adecuadamente del proceso de creación de la colección con igual detalle que el ofrecido por la interfaz de línea de órdenes (figura 4). Al igual que en los dos casos anteriores, fueron rechazados los mismos diez documentos. La consulta de la colección a través de *Greenstone* reveló que los clasificadores obtenidos fueron los mismos que en los dos procesos anteriores.

«El desarrollo de todo el potencial que ofrece Greenstone sólo es posible mediante la consulta detallada de la información contenida en la Greenstone developer's guide»

La comparación entre los procesos directos a través de las tres interfaces permitió obtener las siguientes conclusiones:

—El proceso de creación, captura e importación es el mismo, independientemente de la interfaz utilizada, lo que explica la reiteración de errores. En realidad, tanto *Collector* como *GLI* lo que hacen es trasladar las instrucciones del usuario a la interfaz de línea de órdenes.

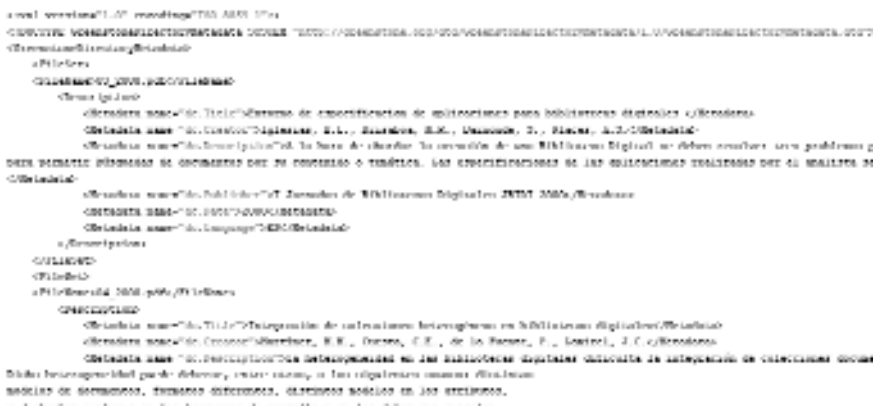

Figura 5. Metadatos en xml según la dtd de Greenstone para Dublin Core

—*Collector* no informa adecuadamente de los problemas que se pueden producir en la captura e importación de documentos, es rígida y ofrece poca ayuda al usuario.

—La aplicación falla en la identificación de metadatos útiles para el usuario final, y en la creación de estructuras de organización de la colección, o clasificadores, cuando estos metadatos no han sido objeto de una preparación y de un tratamiento previo.

—La presentación de las listas de resultados, y de los documentos específicos dificulta la identificación de los mismos, y la consulta de su contenido.

—Los fallos en diez documentos se deben a los plugins o conectores utilizados por *Greenstone* para importar los diferentes tipos de fichero. Este tipo de errores ya han sido reconocidos y analizados en sus listas de correo y son debidos a problemas internos de los plugins utilizados para los documentos pdf y doc, los cuales no son producto del propio proyecto, sino que han sido desarrollados por terceros como software libre. La única solución ofrecida para este tipo de inconvenientes es retirar los documentos fallidos de la colección.

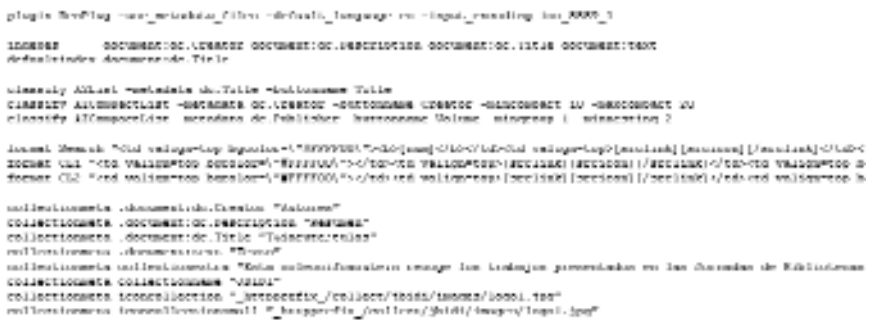

Figura 6. Fichero collect.cfg modificado

5.2. Proceso mediado

Tras los resultados obtenidos en el proceso directo, se llevó a cabo la construcción de la colección en lo que se denominó proceso mediado, en cuanto se asociaron metadatos a los documentos y se configuró una estructura de organización y acceso derivada de los mismos. En este segundo procedimiento ya no fueron procesados los diez documentos erróneos, ya que los resultados no se iban a ver afectados por su presencia o ausencia, dado que esta fase se enfocaba a evaluar la posibilidad de organizar adecuadamente la colección. En primer lugar, se creó un documento xml que incorporaba un conjunto seleccionado de metadatos para cada uno de los documentos, según la norma Dublin Core (figura 5), ya que es uno de los conjuntos de metadatos reconocidos por *Greenstone* (la dtd correspondiente se encuentra disponible en la url indicada a continuación).
http://greenstone.org/dtd/GreenstoneDirectoryMetadata/1.0/GreenstoneDirectoryMetadata.dtd

En segundo lugar, se crearon en el fichero de configuración los clasificadores necesarios para ofrecer diferentes formas de organización y acceso a los documentos de la colección, así como índices de texto específicos para los metadatos seleccionados (la documentación básica para ambas tareas se recoge en la *Greenstone developer's guide*). La integración de ambos elementos varió según la interfaz utilizada. Por último, se definió un esquema de presentación especial para los listados de resultados.

En primer lugar, la creación de la colección se llevó a cabo mediante *Collector*. En este caso ya se había creado previamente el fichero *metadata.xml* que contuviera todos los metadatos asociados a los documentos. Tras la captura de documentos, en la fase de configuración fue necesario modificar manualmente el fichero para añadir las líneas correspondientes a la utilización de metadatos, la creación de índices de texto específico y la definición de clasificadores (figura 6), así como la presentación de resultados. Evidentemente, estas tareas requieren que el usuario tenga los conocimientos técnicos precisos para ello, ya que no existe ningún tipo de ayuda en esta interfaz. Tras un primer proceso de importación, la consulta de la colección permitió apreciar que la estructura de organización disponible no era la prevista. Dada la falta de información, se optó por llevar a cabo el proceso mediante la interfaz de línea de órdenes para detectar el error, que era debido a un fallo en el fichero *metadata.xml*. Una vez corregido, y tras repetir el proceso, la colección se creó correctamente tanto en organización (figura 7) como en la disponibilidad de varios índices de texto contra los que ejecutar las consultas (figura 8). También la presentación de resultados facilitaba la identificación y presentación de documentos (figura 9).

Figura 7. Organización del acceso a la colección, por búsqueda, autores, títulos y volúmenes

Figura 8. Índices textuales disponibles para formular expresiones de búsqueda

Los problemas detectados obligaron a crear la colección mediante la utilización de la interfaz de línea de órdenes. Se utilizó el mismo fichero *metadata.xml* que en el proceso anterior. En este caso, fue necesario editar previamente el *collect.cfg* para añadir manualmente las mismas líneas de configuración de índices, clasificadores y presentación que las introducidas en *Collector*. En este caso, se detectó que el documento xml no estaba bien formado en un carácter. Tras su corrección, la creación de la colección se llevó a cabo sin

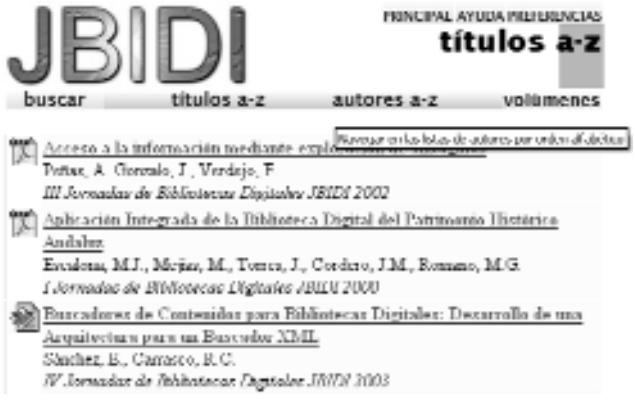

Figura 9. Listado de documentos por título

problemas y pudo repetirse, a su vez, también mediante *Collector*. Por último, se procedió a crearla utilizando *GLI*; en este caso, y a través de su diálogo *Enrich*, permite que el usuario seleccione el tipo de conjunto de metadatos a utilizar (se seleccionó Dublin Core) así como la edición interactiva de los datos de cada documento (figura 10). *GLI* crea automáticamente el fichero *metadata.xml* e introduce el parámetro oportuno en *collect.cfg*. También la creación de índices, clasificadores y esquema de presentación es posible directamente desde la interfaz gráfica.

Sin embargo, el proceso de creación informaba continuamente de problemas en el fichero *metadata.xml*, lo que impedía la correcta disponibilidad de la colección. Tras varias pruebas, se detectó que el error se debía a la identificación del conjunto de caracteres utilizado en el fichero *metadata.xml* como *UTF-8*, cuando en realidad era *ISO-8859-1*. *Greenstone* trabaja internamente con el primero, que no ofrece problemas en el tratamiento de caracteres acentuados y especiales, siendo capaz de detectar y ajustar otros conjuntos de caracteres, incluyendo chino o cirílico. Sin embargo, *GLI* acepta cualquier conjunto de caracteres, pero no hace la traslación a *UTF-8* en el fichero *metadata.xml*. Esto hace que se incluyan en el mismo caracteres extraños que producen errores de documentos xml mal formados. No fue posible encontrar ninguna forma de configurar *GLI* para definir el fichero *metadata.xml* como *ISO-8859-1*. Tampoco admitió correctamente estos caracteres codificados como *UTF-8*. Como último recurso, se probó introducir un conjunto plano de caracteres, sin acentos ni otras letras especiales, con lo que *GLI* funcionó correctamente. Por último, y a pesar de su facilidad de uso, la falta de una documentación adecuada impide aprovechar todo su potencial, siendo de nuevo imprescindible la consulta de la *Greenstone developer's guide*.

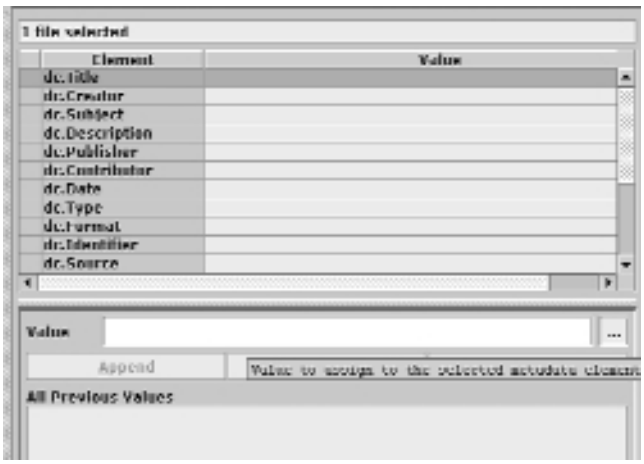

Figura 10. Introducción de metadatos Dublin Core con GLI

La comparación entre los procesos mediados permitió obtener las siguientes conclusiones:

—A pesar de su pretendida sencillez, es necesario recurrir constantemente a la *Greenstone developer's guide* para la creación de clasificadores, la definición de índices y la modificación de la interfaz de usuario final.

—El proceso de *Collector* y de la interfaz de línea de órdenes es el mismo. En cambio, *GLI* da soporte directo a la creación del fichero *metadata.xml* y a la creación de clasificadores, índices y presentaciones que incorpora automáticamente al *collect.cfg*.

—*GLI* es la mejor interfaz para la creación de colecciones complejas, dada la facilidad y rapidez para el diseño y la producción, pero no había implementado en la versión 2.40a el soporte multilingüe necesario. Por el momento, es necesario introducir los caracteres directamente en codificación *UTF*.

—Se debe disponer previamente de un analizador de documentos xml para comprobar que el fichero metadata.xml está formado correctamente antes de proceder a su utilización.

5. Conclusiones

El experimento llevado a cabo demuestra que *Greenstone* es una buena herramienta de soporte a una biblioteca o archivo digital cuando se utilizan sus prestaciones más allá de la mera recopilación e importación de documentos. En este caso, sólo se han usado documentos en dos formatos, por tratarse de los más comunes en los entornos ofimáticos y de publicación actuales, pero la integración de otros (html, xml, etc.) parece plantear similares problemas de detección automática de metadatos y de creación de índices textuales y de clasificadores. Es casi ineludible la creación de ficheros que contengan los metadatos asociados a cada documento, a no ser que puedan ser detectados los necesarios, lo que evidentemente es más sencillo en xml o en otros formatos altamente estructurados, como mensajes de correo o referencias bibliográficas, pero resulta complejo en ficheros como los utilizados en este experimento.

Se trata de una herramienta que demanda un usuario final con conocimientos avanzados para poder aprovechar todo su potencial. En este sentido cabe destacar la reciente incorporación de *GLI* a la distribución estándar de *Greenstone*. Su aceptación y difusión dentro del programa de comunicación e información de la *Unesco*, el creciente número de colecciones disponibles, soportadas por *Greenstone*, la utilización del protocolo *Corba*, la disponibilidad de pasarelas OAI y Z39.50, así como la aparición reciente un plugin para procesar e incorporar bases de datos de *CDS/ISIS*, indican que nos encontramos ante una plataforma que lleva camino de convertirse casi en un estándar.

Cabe plantear en este caso la aceptación y la integración de esta herramienta en el contexto bibliotecario y documental hispanoparlante. Evidentemente, es un software que puede resolver perfectamente los problemas básicos de tratamiento y recuperación de información textual que puedan plantearse en una unidad de información de tamaño pequeño o medio. Los requerimientos técnicos mínimos tienen un coste perfectamente asumible y la curva básica de aprendizaje es rápida. Para el usuario de consulta, el proceso de aprendizaje del comportamiento de la interfaz es inmediato. En las pruebas realizadas la aplicación no ha dejado de responder correctamente a las peticiones que se formulaban en ninguna ocasión, ni ha mostrado problemas o errores en los procesos básicos de recuperación de información. Baste revisar los enlaces disponibles al respecto para poder apreciar las posibilidades de organización y tratamiento, incluyendo colecciones con material gráfico.

http://www.greenstone.org/cgi-bin/library?e=p-en-home-utfZz-8&a=p&p=examples

Por el momento, las experiencias conocidas con *Greenstone* en España son escasas, limitándose al ámbito investigador y de formación de las universidades. La mayor difusión de esta herramienta va a permitir, en plazo breve, que unidades de información de todo tipo desarrollen proyectos e iniciativas de biblioteca digital que hasta ahora estaban limitadas a entidades y empresas que dispusiesen de grandes presupuestos.

Bibliografía

Bainbridge, B.; Buchanan, G.; McPherson, J.; Jones, S.; Mahoui, A.; Witten, I. H. "Greenstone: a platform for distributed digital library applications". En: *Fifth European conference on research and advanced technology for digital libraries*, 2001, Springer, pp. 137-148.

Bainbridge, B.; McKay, D.; Witten, I. H. *Greenstone digital library developer's guide*. Dept of Computer Science, Univ. of Waikato, 2003.

Bollen, J.; Luce, R. "Evaluation of digital library impact and user communities by analysis of usage patterns". En: *D-Lib magazine*, June 2002. Consultado en: 12-01-04.
http://www.dlib.org/dlib/june02/bollen/06bollen.html

Borgman, C. L. "What are digital libraries? Competing visions". En: *Information processing and management*, 1999, n. 35, pp. 227-243.

Borgman, C. L.; Solvberg, G.; Kóvacs, L. (eds.) *Fourth Delos workshop: evaluation of digital libraries: testbeds, measurements, and metrics*. Budapest: Hungarian Academy of Sciences, 2002.

Fuhr, N., [et al.]. "Digital libraries: a generic classification and evaluation scheme". En: *2001 European conference on digital libraries*, lecture notes on computer science, 2001, Springer, pp. 187 y ss.

Marchionini, G. "Evaluating digital libraries. A longitudinal and multifaceted view". En: *Library trends*, 2000, v. 49, n. 2, pp. 304-333.

Rowlands, I.; Bawden, D. "Digital libraries: a conceptual framework". En: *Libri*, 1999, n. 49, pp. 192-202.

Saracevic, T. "Digital library evolution: toward an evaluation of concepts". En: *Library trends*, 2000, v. 49, n. 2, pp. 350-369.

Tramullas Saz, J. "Propuestas de concepto y definición de la biblioteca digital". En: *Actas de las III Jornadas de bibliotecas digitales Jbidi 2002*, pp. 11-20.

Tramullas Saz, J. "Análisis preliminar de bibliotecas digitales en las universidades españolas". En: *Actas de las 8as Jornadas españolas de documentación, Documat*, 2003, pp. 95-106.

Tramullas Saz, J. "El diseño centrado en el usuario en la biblioteca digital". En: *X Jornadas nacionales de información y documentación en ciencias de la salud*, Málaga, 2003 (en prensa).

Witten, I. H.; Bainbridge, D. *How to build a digital library*. San Francisco: Morgan Kaufmann, 2003.

Witten, I. H.; Bainbridge, B.; Boddie, S. J. "Power to the people: end-user building of digital library collections". En: *First Acm/Iee-CS Joint conference on digital libraries*, Acm, 2001, pp. 94-103.

Witten, I. H., Bodie, S. *Greenstone digital library user's guide*. Dept of Computer Science, Univ. of Waikato, 2003.

Witten, I. H.; Moffat, A.; Bell, T. C. *Managing gigabytes: compressing and indexing documents and images*. San Francisco: Morgan Kaufmann, 1999.

Piedad Garrido Picazo, *Depto. Informática e Ingeniería de Sistemas, Univ. de Zaragoza.*
piedad@unizar.es

Jesús Tramullas Saz, *Depto. Ciencias de la Documentación, Univ. de Zaragoza.*
http://tramullas.com
tramullas@unizar.es