Software libre para gestión de recursos de información digital

Jesús Tramullas
Depto. Ciencias de la Documentación, Univ. de Zaragoza
http://tramullas.com

1. La gestión de recursos de información

La Sociedad de la Información ha puesto al alcance de los ciudadanos una cantidad y variedad ingente de recursos de información, que aquellos pueden utilizar para resolver todo tipo de necesidades, desde cuestiones de salud a vacaciones, pasando por educación, economía... A estos recursos de información acceden los usuarios a través de interfaces web, ya que forman parte de servicios de información, tanto públicos como privados, disponibles a través de Internet.

La creación, desarrollo y mantenimiento de portales de servicios de información es una disciplina que ha disfrutado de un notable auge en el último lustro, merced a la preocupación creciente de las administraciones públicas y de las organizaciones privadas por orientar su presencia en Internet a satisfacer las necesidades de los usuarios, antes que a otros objetivos (o al menos, así debería ser). Los dos valores fundamentales que ofrecen los portales a sus usuarios radican en la posibilidad de acceder a información necesaria para tomar decisiones, y en la capacidad de interactuar para efectuar transacciones, generalmente basadas precisamente en la información disponible. Para el objetivo de este texto, la interacción no resulta importante, y el lector interesado puede acudir al abundante cuerpo de conocimiento recogido en la bibliografía especializada, y que puede consultarse mediante la *Human Computer Interaction Bibliography* (http://www.hcibib.org).

Sí que resulta clave, para nuestro objetivo, abordar la gestión de información, en concreto la gestión de recursos de información, como componente nuclear de los servicios de la Sociedad de la Información. La gestión de los recursos de información es una disciplina integrada en las Ciencias de la Información y la Documentación (*Information Science*), y está íntimamente relacionada con disciplinas como la informática, la gestión y la planificación de proyectos y programas o los estudios de usuarios. Las organizaciones desarrollan sus actividades mediante la toma de decisiones, fundamentada en la información disponible. Este tipo de actividad estratégica ya fue identificada y formalizada desde finales de la década de 1970, y ha sido objeto de abundante bibliografía, que culminó en la década de 1990 con el auge de la gestión del conocimiento. La tendencia actual hacia la integración de información en las organizaciones (la tan de moda *Bussiness Intelligence*) no es sino el desarrollo lógico del largo proceso de la puesta en valor de la información, tanto propia como ajena, en todos los planos de actividad.

Es en este contexto donde cabe situar la gestión de recursos de información como una disciplina altamente especializada. Los recursos de información tienen un valor estratégico para todas las organizaciones, y su adecuada gestión es necesaria para alcanzar los objetivos fijados. Al igual que cualquier otro producto, los recursos de información deben ser cuidadosamente planificados y diseñados, deben gestionarse rigurosamente, deben estar sujetos a las correspondientes auditorías (de información, se entiende), y muestran procesos complejos de ciclo de vida.

Dada la variedad de recursos de información que pueden existir, no existen herramientas informáticas directamente aplicables. Si se trata de recursos de información muy especializados, como por ejemplo, un archivo de documentos digitales, o un base de datos documental, pueden encontrarse en el mercado soluciones propietarias específicas, normalmente con un coste muy elevado. Sin embargo, gran parte de los recursos de información digital actuales están especialmente construidos, y tienen peculiaridades y características específicas, lo cual ha

favorecido que se desarrollen utilizando herramientas para la gestión de información, las más conocidos de las cuales son los sistemas de gestión de contenidos, o CMS, como por ejemplo el software propietario Documentum. es en este campo de actividad donde se van a reseñar las principales herramientas de software libre que se pueden utilizar, y se están utilizando, para la gestión de recursos de información.

2. Colecciones de documentos: sistemas de etiquetado de metadatos

La característica principal de los recursos de información es la disponibilidad de un conjunto de documentos, que forman una colección documental. Un recurso de información puede contener una o varias colección de documentos, en diferentes formatos, no sólo de texto, sino que en la actualidad han evolucionado para incluir materiales gráficos (estáticos y/o dinámicos), bases de datos, y colecciones de enlaces a otros recursos de información disponibles a través del web.

Evidentemente, cualquier recopilación de documentos o de enlaces no forma, por sí misma, una colección. Para que sea merecedora de tal consideración, una colección de documentos debe haber sido objeto de un conjunto de técnicas y tratamientos que analicen, describan y faciliten la búsqueda y el acceso a los documentos por parte de los usuarios. El contenido informativo de los documentos debe ser analizado, descrito y representado según unas normas o estándares, de forma que se disponga de un medio de acceso al documento. El nombre genérico que se da a estas descripciones es el de "conjunto de metadatos". El conjunto más utilizado en el entorno bibliotecario es el formato MARC, cuya trayectoria se remonta a la década de 1960. Sobre este trabajo de descripción de valor añadido, que debe caracterizar a cualquier colección de documentos merecedora de tal nombre, debe estructurarse la organización de la colección, que se utiliza, en los entornos digitales, para ofrecer al usuario una primera forma de acceso, generalmente mediante una navegación de tipo jerárquico. La segunda forma de acceso que se ofrece al usuario adopta la forma de motor de búsqueda, que permite al usuario buscar sobre las descripciones y metadatos anteriormente indicados, o sobre el contenido textual completo de los documentos.

Una cuestión clave en la gestión de recursos de información la representan los esquemas y sistemas de etiquetado de la información. En el contexto en el cual se sitúa este texto, este tipo de esquemas resulta ser tan importante como la existencia de herramientas de software libre capaces de trabajar con ellos. Afortunadamente, y gracias al trabajo y al soporte de organizaciones como el W3C, los sistemas de etiquetado se han desarrollado de manera abierta, y se han convertido en estándares de libre acceso, al alcance de todos los desarrolladores y usuarios. Es necesario que las herramientas libres para gestión de recursos de información sean capaces de trabajar con este tipo de estándares.

Aunque los esquemas existentes son numerosos, cabe destacar entre todos ellos *Dublin Core,* MODS (*Metadata Object Description Schema*) y RDF (*Resource Description Framework*). En primer lugar, todos estos esquemas soportan su formulación dentro de XML. En segundo lugar, todos ellos están pensados para ofrecer soporte a la descripción de documentos digitales de cualquier tipo. Si bien *Dublin Core* y MODS están pensados para ser utilizados directamente, RDF, en realidad, establece un marco general, dentro del cual el desarrollador puede establecer su propio esquema de descripción. Por último, debe citarse SKOS (*Simple Knowledge Organization System*), cuya finalidad es precisamente describir, en XML, los lenguajes y esquemas de clasificación utilizados en la descripción del contenido informativo de los documentos, como pueden ser clasificaciones, tesauros, ontologías...

- *Dublin Core*, o DC: es un conjunto de quince metadatos que permiten la descripción de cualquier recurso de información digital, atendiendo a aspectos de autoría y responsabilidad, descripción y contenido informativo-documental. La URL de referencia es

http://www.dublincore.org.

- *Metadata Object Description Schema*, o MODS: es un conjunto de metadatos que tiene el mismo objetivo que DC, pero que busca superar los problemas y limitaciones detectados en la aplicación de DC. Para ello, toma como punto de partida el formato MARC, del que selecciona diferentes elementos. La organización responsable del desarrollo de MODS en la Library of Congress estadounidense. La URL de referencia es http://www.loc.goc/standards/mods/
- *Resource Description Framework*, o RDF: es un sistema de etiquetado que pretende actuar como marco general, dentro del cual formular y describir cualquier tipo de recursos de informacion digital, según las necesidades que puedan surgir. La URL de referencia es http:///www.w3c.org/RDF, ya que el W3C se encarga del mantenimiento de la norma. Como ejemplo de herramientas para desarrolladores en RDF, véase el Redland RDF Application Framework (http://librdf.org)

Fig.1. Redland RDF Application Framework

Un elemento clave relacionado con los conjuntos de metadatos, es el protocolo utilizado para poder intercambiar información entre diferentes herramientas. En este sentido, cabe citar, en primer lugar, Z39.50, que sirve para poder consultar catálogos de bibliotecas, e integrar el resultado de los mismos, si es necesario. En estos momentos se está redactando una nueva formulación del mismo, más avanzada y que ofrece más prestaciones, a la que se ha denominado ZING. En segundo lugar debe comentarse el protocolo OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), permite la consulta de colecciones entre repositorios de forma remota, a través del conjunto de metadatos Dublin Core, codificados en UTF. En el siguiente apartado se puede observar que las herramientas para repositorios de documentos incorporan el protocolo OAI, lo que da idea de su importancia. La URL de referencia es http://www.openarchives.org/OAI/openarchivesprotocol.html

3. Las herramientas de software libre para gestión de recursos de información

La disponibilidad de herramientas libres capaces de gestionar recursos de información digital no es elevada, si se compara con otros ámbitos, como la programación, la multimedia, o el desarrollo de portales. Como puede deducirse de los anteriores apartados, será el uso que se haga de las herramientas disponibles lo que diferencie a un servicio de información de otro que no lo sea, además de la ausencia o presencia de servicios y características básicos. Como puede imaginarse, es perfectamente posible desarrollar y operar un servicio de información digital utilizando Zope, por ejemplo, o gestionar bases de datos XML con Exist. Sin embargo, para la recopilación, evidentemente no exhaustiva, que sigue a estas líneas se ha optado por seleccionar a aquéllos que han sido desarrollados pensado en dar soporte a las funcionalidades necesarias para las colecciones de documentos o de otros tipos de colecciones de recursos de información.

Herramientas para repositorios de documentos digitales:

- CDSware: soporte del afamado CERN Document Server, y desarrollado por el propio CERN, su misión es actuar como plataforma para colecciones de documentos, bibliotecas digitales, pre-prints, etc. Como estándar de descripción de contenidos utiliza MARC21. Sigue el modelo LAMP, y necesita Python. La documentación de referencia incluye todos los módulos necesarios para disponer una instalación de CDSware completa (por ejemplo, es capaz de realizar gráficas de los trabajos citados, pero para ello necesita GNUPlot). La integración de documentos ofimáticos se lleva a cabo usando parsers de terceros. Un servidor CDSware es capaz de soportar gran cantidad de colecciones, y puede ofrecer

portales diferentes para cada una de ellas. Integra un potente motor de búsqueda, y permite la personalización de servicios para usuarios registrados. Además puede exportar la información a formatos HTML, XML, MARC, e incorpora OAI. La URL de referencia es http://cdsware.cern.ch

- Fedora: es una aplicación para desarrollo de repositorios de documentos, creada por Cornell University y la biblioteca de la University of Virginia. ha sido desarrollado en Java, por lo que se necesita el JSDK de Sun. El concepto básico de Fedora es trabajr con objetos digitales, que poseen un ciclo de vida y unas relaciones entre ellos, a lo largo de un período de tiempo. Ofrece una arquitectura completamente modular, y ofrece servicios web. Además, se están desarrollando aplicaciones en otros lenguajes para aumentar las prestaciones del repositorio. Como estándar de metadatos usa Dublin Core, y puede aplicar RDF para determinar las relaciones entre los diferentes contenidos. También cumple el estándar OAI. La URL de referencia es http://www.fedora.info
- DSpace: es un desarrollo conjunto del MIT y Hewlett-Packard, y se trata de un repositorio de información digital, pensado para almacenar, indizar el contenido y difundir los trabajos de investigación de una organización, lo que explica su creciente difusión en repositorios de documentos de universidades. DSpace incorpora Dublin Core y el protocolo OAI. Está escrito en Java, y necesita el JSDK de Sun, Apache Ant, Postgres u Oracle y Jakarta Tomcat. En DSpace es posible crear grupo de usuarios, definir niveles de seguridad, crear flujos de trabajo, personalizar la presentación... La recuperación de información, como en otras plataformas, se realiza mediante Lucene. DSpace incorpora también la posibilidad de empaquetar sus contenidos en formato METS (*Metadata Encoding and Transmission Standard*), lo que facilita el intercambio de colecciones con otras herramientas, por ejempolo, Greenstone. La URL de referencia es http://dspace.org
- E-Prints: el objetivo de esta herramienta es crear un repositorio de documentos de libe acceso, y sirve para dar soporte a repositorios, archivos de e-prints y a revistas digitales. ha sido desarrollado en el marco del projecto Open Access, por la School of Electronics and Computer Science, University of Southampton. E-Prints necesita un servidor LAMP, ya que está escrito en Perl, y parsers de terceros para procesar documentos ofimáticos. Una carencia (o ventaja) es que no incorpora motor de búsqueda propio. Está especialmente pensado para interactuar con otros servidores gracias a su potente integración OAI. Permite crear RSS de las novedades de contenido. La URL de referencia es http://www.eprints.org/software/

Fig. 2. E-LIS, un repositorio desarrollado sobre E-Prints

Herramientas para bibliotecas digitales:

- Greenstone: es una aplicación para la construcción y explotación de bibliotecas digitales, creada , desarrollada y mantenida por New Zealand Digital Library Project, en la University of Waikato. Esta aplicación tiene como núcleo el motor de indización y recuperación de información textual MG/MG++, aunque las últimas versiones incorporan también Lucene. La aplicación está formada por diferentes macros, programados en Perl, encargados 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. Es capaz de procesar e incorporar a las colecciones documentos en numerosos formatos, incluyendo los ofimáticos más comunes. 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. Greenstone también lleva incorporado un servidor OAI. La URL de referencia es http://www.greenstone.org

Fig. 3. Greenstone

Herramientas para directorios de recursos de información digital:

- Scout Portal Toolkit: esta herramienta, desarrollada por el Internet Scout Project de la Universidad de Wisconsin-Madison, permite crear un portal de acceso a los recursos de información digital, de todo tipo, que pueda necesitar una organización. Pone especial énfasis en las tareas de administración y control de metadatos, ya que usa Dublin Core (aunque puede ajustarse a las necesidades del usuario). Incorpora métodos de valoración de calidad de los recursos por parte de los usuarios, y otras prestaciones que lo hacen especialmente interesante. Está escrito en PHP, y usa MySQL como base de datos. La URL de referencia es http://scout.wisc.edu/Projects/SPT/
- MyLibrary: es un desarrollo de Eric Lease Morgan, bibliotecario de la University of Notre Dame, y una firma reconocida en el campo del software libre para servicios de información digital. Pone el énfasis en la utilización de clasificaciones para organizar los recursos disponibles, pero no usa, por ejemplo, Dublin Core para la descripción de recursos. Suele ofrecer esquemas de navegación jerárquica. Está escrito en Perl, y requiere MySQL, Postgres u otro RDBM. La URL de referencia es http://dewey.library.nd.edu/mylibrary/
- iVia: es una herramienta para crear portales basados en colecciones de recursos de información digital. Desarrollado por iVia Project, de la University of California (Riverside), sigue el clásico esquema LAMP. En realidad, es una integración de diferentes paquetes, algunos de los cuales pueden instalarse por separado. La aplicación es una combinación de C++ y Java. Tampoco sigue el esquema Dublin Core, pero ofrece gran potencial para la clasificación de recursos y la creación de clasificadores. La URL de referencia es http://ivia.ucr.edu
- ROADS: herramienta en Perl, supuso un importante hito en el desarrollo de directorios temáticos especializados de recursos web. Desde 2004 ya no se desarrolla. La URL de referencia es http://roads.sourceforge.net

Fig. 4. Scout Portal Toolkit

Herramientas para automatización de bibliotecas:

Este tipo de herramientas están pensados para dar soporte a las tareas de gestión de información (catalogación, mantenimiento de autoridades, recuperación de información mediante OPAC...) y a las tareas administrativas (préstamos, reservas, publicaciones seriadas, control de usuarios, administración...) que se llevan a cabo en los servicios y unidades de información que pertenecen al tipo biblioteca. Es el ámbito en el cual más soluciones y herramientas se han desarrollado. La mayor parte de ellas responde al modelo LAMP, y han sido escritas en Perl o PHP. Pueden ofrecen mayores o menores prestaciones, pero básicamente es necesario que den soporte al formato MARC, que permitan la gestión bibliográfica del catálogo, y que incorporen un OPAC avanzado. Bastantes de estas herramientas también ofrecen servidores Z39.50, y posibilidades de importar y exportar información bibliográfica a otros formatos. Sin pretender ser exhaustivo, el siguiente listado contiene los más implantados:

- OpenBiblio: http://obiblio.sourceforge.net
- Koha: http://www.koha.org
- Emilda: http://www.emilda.org
- PHPMyLibrary: www.phpmylibrary.org
- Gnuteca: http://www.gnuteca.org.br
- PMB: http://www.sigb.net

Fig. 5. Koha

Motores de búsqueda:

- Lucene: es el motor de búsqueda textual desarrollado por la Apache Foundation. En realidad, es una librería escrita en Java, sobre la cual pueden crearse motores de búsqueda, o bien ser integrada en todo tipo de aplicaciones. Trabaja con ficheros XML, lo que hace necesario que todo tipo de documentos diferente sea previamente transformado. Evidentemente, ello requiere un proceso de análisis e instalación más complejo, por al necesidad de integrar librerías o parser externos. Para web, recientemente se ha presentado Nutch, especialmente diseñado para este entorno, y que permite instalar rápidamente un motor de búsqueda web, incorporando parser para HTML, crawler, gráficos para links, etc. La URL de referencia es http://lucene.apache.org
- Xapian: al igual que Lucene, es una librería que actúa como motor probabilístico para recuperación de información. Está escrito en C++, y ofrece binfings para otros 14 lenguajes, incluyendo PHP, Perl, Python... ofrece una versión diseñada para motor de búsqueda web, llamada Omega. La URL de referencia es http://www.xapian.org
- ht://Dig: es una herramienta pensada para crear motores de búsqueda de alcance limitado, como un dominio o una organización. Sólo trabaja con ficheros HML, aunque es capaz de seguir la estructura recursiva de los enlaces. La URL de referencia es http://www.htdig.org
- Swish-e (Simple Web Indexing System for Humans – Enhanced): este motor es un veterano del desarrollo de motores, tanto para intranets cono para motores de búsqueda en Internet. Derivado del primer Swish, escrito en 1994 por Kevin Hughes, es un herramienta potente y rápida, sobre la cual, a su vez, se han desarrollado interfaces y aplicaciones específicas. Es capaz de indizar el contenido textual de ficheros XML, HTML, Microsoft Office, pdf, correo electrónico, etc. Está escrito en Perl, y utiliza como parser XML la librería libxml2 de Gnome. La URL de referencia es http://swish-e.org

Fig 6. Lucene

La categorización previa no cubre todas las posibilidades disponibles para el desarrollo de servicios de información digital. Por ejemplo, una buena plataforma para la gestión de documentos parece ofrecerla Alfresco, que acaba de lanzar su primera versión (http://www.alfrescosoftware.com). KnowledgeTree es presentado como una solución corporativa para la gestión documental (http://www.ktdms.com). Si tiene que editar una revista científica, la mejor solución es Open Journal System, OJS (http://pkp.sfu.ca), desarrollado dentro del Public Knowledge Project. Para desarrollar catálogos colectivos de bibliotecas, contra servidores Z39.50, la solución es MOCCAM (http://server4.hosting.cri74.org/ccy2/). Para sencillos directorios temáticos, Potnia puede ser un ejemplo (http://potnia.sourceforge.net).

Referencias

Barragán, C. (2005). “Programari lliure: introducció i estat de la qüestió per als professionals de la informació i la documentació.” Lorente, M. (coord.) *Anuario Bibliodoc*. Barcelona: COBDC, p. 59-76.

Chawner, B. (2005). *Open Source Software and Libraries Bibliography*. URL : http://www.vuw.ac.nz/staff/brenda_chawner/biblio.html [consultado 7-10-2005]

FreeBiblio.info. L'actuallité du logiciel libre et gratuit pour bibliothéques. URL: http://www.freebibilio.info [consultado 3-10-2005]

Grup de treball de programari lliure per als professionals de la información. URL:

http://www.soft-libre.net [consultado 2-5-2005]

Grupo español de usuarios de Greenstone. URL: http://greenstone.docunautica.com [consultado 2-5-2005]

Lease, E.L. (2004). *Open source software in libraries: A workshop*. URL: http://www.infomotions.com/musings/ossnlibraries-workshop/ [consultado 9-9-2005]

Open Source Systems for Libraries. URL: http://www.oss4lib.org [consultado 16-10-2005]

Sturman, R. (2004). E*l programario de código abierto para la gestión integrada de la biblioteca: ¿un nuevo recurso?*. URL: http://www.soft-libre-net/docs/trad_spa.htm [consultado 29-6-2005]

Witten, I.H., Bainbridge, D. (2002). *How to Build a Digital Library*. San Francisco: Morgan Kaufmann.