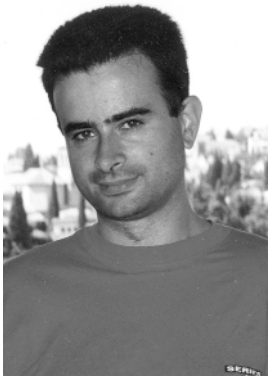
Open archives initiative. Protocol for metadata harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo

Artículo

Por **José Manuel Barrueco** e **Imma Subirats Coll**



José Manuel Barrueco, licenciado en documentación (Universidad Politécnica de Valencia). Actualmente trabaja en la Biblioteca de Ciències Socials de la Universitat de València. Ha publicado varios trabajos sobre bibliotecas digitales y revistas electrónicas. Mantiene servicios como DoIS (Documents in Information Science) o WoPEc (Working Papers in Economics).
http://dois.mimas.ac.uk
http://netec.mcc.ac.uk/WoPEc

***Resumen**: Se describe el protocolo OAI-PMH (Open Archives Initiative–Protocol for Metadata Harvesting) utilizado para la transmisión de metadatos en internet. Se analiza el contexto en el que nació, las comunidades de depósitos de documentos científicos y cómo se ha desarrollando y extendido su alcance a cualquier material en formato electrónico. Se describe brevemente su arquitectura basada en el modelo cliente–servidor donde los primeros, llamados archivos, ponen a disposición del público metadatos en formato Dublin Core para que puedan ser recuperados por los segundos. La comunicación se realiza mediante el protocolo http. Las respuestas están codificadas en xml. Finalmente se hace una revisión de las principales instituciones que lo han implementado, los servicios que se han basado en él y se presenta una serie de herramientas que facilitan la creación de archivos abiertos.*

***Palabras Clave**: Bibliotecas digitales, Protocolos de internet, Metadatos, Xml, Eprints, Open archives initiative-Protocol for metadata harvesting, OAI-PMH.*

Imma Subirats Coll, licenciada en documentación (Universitat de Barcelona), actualmente trabaja en el Departamento de Política Territorial i Obres Públiques de la Generalitat de Catalunya como documentalista. Ha publicado varios trabajos sobre bibliotecas digitales y gestión de documentación electrónica. Mantiene servicios como DoIS (Documents in Information Science) y participa en E-LIS (Eprints in Library and Information Science).
http://dois.mimas.ac.uk
http://eprints.rclis.org/

Title: Open archives initiative. Protocol for metadata harvesting (OAI-PMH): description, functions and applications of a protocol

***Abstract:** The authors describe the OAI-PMH protocol (Open Archives Initiative Protocol for Metadata Harvesting) used for the interchange of metadata on the internet. Analysed are questions such as the context in which the protocol was born —communities of scientific documents repositories— and how it has evolved and extended its reach to include any material in electronic format. Its architecture, based on the client/server model, is briefly described. Clients, which are called "archives", place Dublin Core metadata in the public domain in order for servers or service providers to retrieve this metadata and to build added value services for end users. Communication between archives and services takes place using the http protocol. Questions and replies are coded in xml. The authors conclude with a review of the main institutions that have implemented the protocol, user services based on it, and software tools that facilitate the creation of open archives.*

***Keywords:** Digital libraries, Internet protocols, Metadata, Xml, E-prints, Open archives initiative-Protocol for metadata harvesting, OAI-PMH.*

***Barrueco, José Manuel; Subirats Coll, Imma. "**OAI-PMH: protocolo para la transmisión de contenidos en internet". En: El profesional de la información, 2003, marzo-abril, v. 12, n. 2, pp. 99-106.*

0. Introducción

El trabajo que presentamos en este número monográfico tiene como objetivo divulgar entre los profesionales de nuestro país un nuevo protocolo para la transmisión de contenidos en internet denominado *OAI-PMH* (*Open Archives Initiative–Protocol for Metadata Harvesting*).

http://www.openarchives.org

Si bien es de aparición reciente, los primeros trabajos para su desarrollo se iniciaron en 1999 y el inte-

Artículo recibido el: 25-01-03
Aceptación definitiva: 07-03-03

rés que ha despertado entre la comunidad de bibliotecarios de todo el mundo ha sido muy grande. Esta muestra de atención viene probada por la gran cantidad de jornadas celebradas en distintos países, por el número de artículos publicados en revistas especializadas, por la importancia de las instituciones que lo han apoyado desde el primer momento y por los numerosos proyectos de investigación financiados en el último año tanto por la Unión Europea como por la *National Science Foundation* de los EUA.

Paradójicamente el interés entre los investigadores y profesionales de nuestro país parece escaso cuando no nulo. En las últimas jornadas técnicas sobre el tema celebradas en Ginebra y Lisboa el pasado año, la asistencia de participantes españoles fue simbólica. Las últimas jornadas sobre bibliotecas digitales, *Jbidi 2002*, no incluyeron ninguna presentación sobre el mismo. En el estudio sobre la actividad relacionada con *Oai* (*Open Archives Initiative*) en Europa (**Dobratz**, 2002) no aparece ninguna mención a nuestro país, mientras que sí aparecen otras naciones de nuestro entorno como Francia, Italia o Portugal.

http://library.cern.ch/Announcement.htm

http://www.oaforum.org/workshops/lisb_invitation.php

http://mariachi.dsic.upv.es/jbidi/jbidi2002/

Con objeto de paliar esta deficiencia en un asunto que consideramos de vital interés para la in-

Para saber más

Van de Sompel, Herbert; **Lagoze, Carl**. "The Santa Fe convention of the Open Archives Initiative". En: *D-Lib*, 2000, febrero, v. 6, n. 2.

La bibliografía básica sobre el tema está encabezada obligatoriamente por los dos autores que constituyen el motor de la iniciativa, **Van de Sompel** y **Lagoze**. Este artículo, publicado muy poco después de la reunión de Santa Fe, está escrito para dar a conocer al mundo el desarrollo de la reunión y el nacimiento de *OAI*. Es importante notar la importancia que se da a los archivos de eprints en este primer momento.

http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Van de Sompel, Herbert; **Lagoze, Carl**. "The Open Archives Initiative: building a low-barrier interoperability framework". En: *Jcdl*, 2001.

En esta comunicación los autores hacen un repaso de lo que supuso el primer año de existencia de *OAI*. Es un documento más técnico que el anterior y donde, tras una reseña histórica, se entra a describir el marco técnico para la implementación de un archivo abierto: se tratan cuestiones de metadatos, registros y una descripción exhaustiva del protocolo.

http://www.openarchives.org/documents/oai.pdf

Van de Sompel, Herbert; **Lagoze, Carl** (ed.) "The Open Archives Initiative Protocol for Metadata Harvesting". Consultado el: 10-01-03.

Este documento es la especificación técnica del *PMH*. Todo lo que hay que saber para implementar un archivo o un proveedor de servicios. Contiene numerosos ejemplos.

http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Van de Sompel, Herbert; **Lagoze, Carl**. "Notes from the interoperability front: a progress report on the Open Archives Initiative". En: *Ecdl*, 2002.

Otro informe más sobre el avance de la iniciativa. En éste, el más reciente, se hace un repaso de proyectos y programas de investigación financiados tanto en EUA como en Europa relacionados con *OAI*. Así mismo se describen ejemplos de archivos y servicios y se hace una presentación de la versión 2 del protocolo.

http://www.openarchives.org/documents/ecdl-oai.pdf

Harnad, Stevan. "Free at last: the future of peer-reviewed journals". En: *D-Lib*, 1999, v. 5, n. 12.

Finalmente un artículo de **Harnad**, donde expone sus teorías sobre el futuro de las revistas científicas en competencia con los eprints. Este autor ha estado en el corazón de la iniciativa desde la reunión de Santa Fe. Frente a los anteriores autores que compondrían la parte técnica de la misma, las implicaciones que busca **Harnad** en *OAI* son más culturales y de transformación de los procesos de comunicación científica.

http://www.dlib.org/dlib/december99/12harnad.html

vestigación en bibliotecas digitales, presentamos este trabajo con un carácter eminentemente divulgativo. La importancia de *OAI-PMH* se puede resumir en una frase: está llamado a ser a las bibliotecas digitales lo que http es hoy al web.

«OAI-PMH está llamado a ser a las bibliotecas digitales lo que http es hoy al web»

El resto del trabajo se estructura de la siguiente forma: la sección 1 da una visión general del protocolo abordando cuestiones como cuáles son sus objetivos, cómo funciona o qué instituciones lo respaldan. Con posterioridad se hace una breve historia de su evolución desde su nacimiento en 1999 hasta la última versión publicada en junio de 2002. El punto 3 entra, sin profundizar demasiado, en las cuestiones técnicas del protocolo, mientras que el 4 describe las principales instituciones que lo han adoptado y los servicios que se han creado utilizando las funcionalidades que aporta. Finalmente la sección 5 detalla una serie de herramientas que nos pueden ayudar a la hora de implementarlo.

1. ¿Qué es *OAI-PMH*?

En primer lugar hemos de matizar varios puntos respecto a su nombre. El término "archivo" refleja los orígenes de la iniciativa en el seno de las comunidades de eprints donde es sinónimo de depósito de documentos científicos a texto completo. No tiene nada que ver aquí con el concepto tradicional de archivo con connotaciones de preservación y conservación. Se utiliza por lo tanto con un sentido mucho más amplio, como un depósito para almacenar cualquier tipo de información. El término "abierto" se refiere al punto de vista de la arquitectura del sistema. Se trata de definir interfaces que faciliten la disponibilidad de contenidos procedentes de una variedad de proveedores. Apertura tampoco significa gratuidad o acceso ilimitado a dicha información.

La *OAI* se creó con la misión de desarrollar y promover estándares de interoperabilidad para facilitar la difusión eficiente de contenidos en internet. Surgió como un esfuerzo para mejorar el acceso a archivos de publicaciones electrónicas (eprints), en definitiva, para incrementar la disponibilidad de las publicaciones científicas. Los trabajos iniciales se centraron en el desarrollo de marcos de compatibilidad para la federación de archivos de eprints. Pronto pareció evidente que dichos marcos (permitir el intercambio de múltiples formatos bibliográficos entre distintas máquinas utilizando un protocolo común) tenían aplicaciones más allá de esta comunidad. Por ello se adoptó un objetivo mucho más amplio: abrir el acceso a un rango de materiales digitales.

Por lo tanto, la *OAI* no es solamente un proyecto centrado en publicaciones científicas, sino en la comunicación de metadatos sobre cualquier material almacenado en soporte electrónico. No hay nada en el protocolo que impida a los implementadores transmitir el contenido propiamente dicho de esos materiales. No obstante éste no es el objeto principal de *OAI-PMH*.

Los metadatos a transmitir vía *OAI-PMH* deberán codificarse en *Dublin Core* sin calificar con objeto de minimizar los problemas derivados de las conversiones entre múltiples formatos. Aunque se está investigando la creación de servicios tales como una interfaz de búsqueda a través de formatos heterogéneos de metadatos, una solución menos complicada, y por lo tanto más fácil de realizar, es requerir a los implementadores convertir sus datos a un formato común. Los 15 elementos de *Dublin Core* han evolucionado a lo largo de los pasados años como el estándar de facto para los metadatos simples y multidisciplinares.

¿Qué relación existe con otros protocolos como el Z39.50? El marco diseñado por *OAI* es intencionalmente simple con el propósito de proporcionar una mínima complicación para las instituciones que deseen ponerlo en práctica. Los protocolos como el Z39.50 tienen una funcionalidad más completa, por ejemplo, tratan cuestiones como el manejo de sesiones, gestión de conjuntos de resultados y permiten la especificación de predicados para filtrar los resultados obtenidos. Sin embargo, esto acarrea un incremento en la complejidad de la implementación y, en consecuencia, de los costes. Por lo tanto no se trata de reemplazar otras iniciativas, sino de desarrollar una alternativa que sea fácil de llevar a cabo y de usar para propósitos diferentes de los que ya tratan los sistemas de interoperabilidad existentes. El futuro juzgará si esta barrera mínima es realista y funcional.

La *OAI* no define o prescribe ningún esquema para la gestión de derechos. Los temas relacionados con restricciones en el acceso y gestión de la propiedad intelectual son responsabilidad de los proveedores de datos. En este sentido la iniciativa deja de lado los aspectos culturales y sociales que pueda acarrear la implementación del protocolo y se centra exclusivamente en cuestiones técnicas. No obstante ya hemos mencionado que uno de los apoyos más fuertes que sostienen *OAI* son las comunidades de eprints. Surgieron con el objetivo de establecer canales alternativos a los tradicionales editores comerciales para la distribución de información científica. Si bien hoy nadie discute la necesidad de la existencia de tales intermediarios en el sistema de publicación de la ciencia, se prevé que en el

futuro cambien los papeles que desempeñan. De las dos funciones básicas que llevan a cabo los editores, la de establecer un control de calidad de los contenidos continuará siendo crucial, pero la distribución de los trabajos en beneficio de los autores y la comunidad científica es la que está más en entredicho.

La *OAI* ha obtenido financiación en EUA de la *National Science Foundation*. De la gestión administrativa y técnica se encargan dos comités que están coordinados por **Herbert Van de Sompel** y **Carl Lagoze**, ambos de la *Cornell University*.

2. Un poco de historia, de la *Convención de Santa Fe* a la *OAI*

Los orígenes de *OAI* radican en un creciente interés en la búsqueda de alternativas a los modelos tradicionales de comunicación científica. En algunas disciplinas, principalmente en ciencias, comenzaron a surgir los llamados archivos o repositorios de documentos electrónicos para la rápida comunicación de resultados de investigación. Esos documentos se han llamado eprints de forma genérica. Este nuevo concepto agrupa tanto aquellos que no han pasado por un proceso de certificación o peer review (preprints) como los que sí lo han pasado, o postprints (artículos, libros, etc.). El más conocido de estos archivos es sin duda *arXiv.org* creado por **Paul Ginsparg** en Los Alamos (EUA) para el área de física.

En octubre de 1999 se organizó una reunión en Santa Fe (Nuevo México, EUA) con la idea de que la interoperabilidad de estos archivos de eprints era clave para aumentar su impacto entre la comunidad académica. Con ella se podrían federar varios archivos, intercambiar registros o realizar búsquedas en disciplinas relacionadas al mismo tiempo. Los participantes en esta reunión fueron especialistas en bibliotecas digitales, así como representantes de los principales archivos existentes:

—*ArXiv.org*. Considerado como el primer ejemplo de archivo de eprints y fundado en 1991. Aunque comenzó como archivo de prepublicaciones ha evolucionado para incluir también artículos publicados en revistas tradicionales. Igualmente comenzó centrado en física de altas energías pero ha incorporado otras disciplinas relacionadas como las matemáticas, informática, etc.

http://arxiv.org

—*CogPrints*. Es un proyecto de la *University of Southampton* en el Reino Unido. Es una exportación del modelo de *arXiv.org* al campo de la psicología y disciplinas relacionadas.

http://cogprints.soton.ac.uk

—*Ncstrl* (*Networked Computer Science Technical Reference Library*). Es una colección de informes y documentos en informática. Está basado en una arquitectura distribuida en la que los documentos son almacenados en archivos distribuidos y son hechos disponibles a través de servicios que se comunican utilizando el protocolo *Dienst* (*Distributed interactive extensible network server for techreports*) de la *Cornell University*.
http://www.ncstrl.org

—*Ndltd* (*Networked Digital Library of Theses and Dissertations*). Su objetivo es construir una biblioteca digital de tesis en formato electrónico cuyos autores sean estudiantes de las instituciones miembros.
http://www.ndtl.org

—*RePEc* (*Research Papers in Economics*). También está basado en un modelo distribuido y proporciona a los autores la opción de remitir sus documentos de trabajo a un archivo local de su propia institución o, si no existe uno, al *EconWPA*, un archivo mantenido por la *Washington University at Saint Louis* siguiendo el modelo de *arXiv.org*. Todos los archivos utilizan el denominado *Protocolo de Guildford* que garantiza la compatibilidad entre los archivos y los servicios a los usuarios finales.
http://repec.org
http://econwpa.wustl.edu

La interoperabilidad de los archivos tiene varias facetas como son por ejemplo sistemas de identificación comunes, formatos de metadatos, modelos de documentos o protocolos. Los participantes establecieron que una solución minimalista era imprescindible si se quería alcanzar una amplia adopción entre la comunidad de proveedores de eprints. La solución adoptada fue la recolección de metadatos (*metadata harvesting*). Esto permite a los proveedores de eprints exponer sus metadatos a través de una interfaz, con el objeto de que la misma pueda ser utilizada como la base para el desarrollo de servicios de valor añadido.

El resultado de la reunión fue un conjunto de acuerdos técnicos y organizativos conocidos como la *Convención de Santa Fe*. Los aspectos técnicos incluían 3 puntos fundamentales: un formato para los metadatos, un protocolo basado en el antiguo *Dients* y un sistema de identificación.

Tras hacer públicos los resultados de la reunión en febrero de 2000, quedó claro que había un interés en esta iniciativa más allá de las comunidades de eprints. En principio, bibliotecarios y museólogos se mostraron interesados en descubrir formas de hacer visibles a los motores de búsqueda en internet partes de las colecciones de bibliotecas y museos. Estas necesidades se expresaron en una serie de reuniones celebradas en

el contexto de las principales jornadas sobre bibliotecas digitales celebradas tanto en EUA como en Europa. Respondiendo a este amplio interés se procedió a la reconsideración de las decisiones tomadas en Santa Fe. Así es decidió ampliar el objeto de trabajo más allá de los eprints para incluir disciplinas que no tuvieran este tipo de documentación con lo que los aspectos técnicos aplicables exclusivamente a eprints fueron reconsiderados. Además, la credibilidad del esfuerzo era incierta debido a la falta de una estructura organizativa. Los profesionales son lógicamente reacios a adoptar estándares cuando los responsables de promoción y mantenimiento de los mismos son cuestionables.

El último punto, credibilidad, fue el primero en tratarse y así en agosto de 2000, la *Digital Library Federation* y la *Coalition of Networked Information* de los EUA anunciaron que ofrecerían el soporte de su organización a la iniciativa. A partir de este momento comenzaron a funcionar dos comités, uno de gestión y otro técnico, que se encargarían de su coordinación.

Las especificaciones revisadas fueron hechas públicas en enero de 2001 con la publicación del *Open Archives Initiative–Protocol for Metadata Harvesting* (*OAI-PMH*) versión 1.0. La intención era que este protocolo, con mínimas modificaciones, permaneciera estable al menos durante un año, mientras las distintas comunidades lo probaban y experimentaban con él.

Pronto aparecieron las primeras instituciones que lo utilizaron para poner en internet sus metadatos. El *OAI-PMH* es una tecnología que sigue lo que **Shapiro** y **Varian** (**Shapiro**, 1999) denominan efectos de red: la adopción inicial es lenta y progresiva pero la respuesta positiva a la misma aumenta de forma muy rápida la tasa de adopción. Esto se ha cumplido en los dos años que lleva funcionado el protocolo. Ya son más de cien las instituciones que han creado archivos abiertos, el número de servicios basados en la utilización de la información almacenada en los anteriores no ha parado de crecer tanto en número como en calidad de los valores añadidos que ofrecen. En estos momentos hay registrados en el servidor de *OAI* unos doce servicios. Igualmente ha aparecido toda una serie de herramientas de software destinadas a facilitar la creación y mantenimiento de archivos.

Han sido muchos los proyectos de investigación que se han concedido durante el pasado año para estudiar la aplicación del protocolo y temas relacionados. Así en EUA está por ejemplo la *Metadata Harvesting Initiative* de la *Fundación Mellon* en el seno de la que se han financiado cuatro proyectos por valor de 1,5 millones de US$ con objeto de crear servicios basados en *OAI-PMH*. La *National Science Digital Library*, un proyecto de la *National Science Foundation* tiene como objeto la creación de lo que será la mayor biblioteca digital hasta el momento. Ha adoptado el protocolo como base para la comunicación de metadatos entre los participantes. En Europa se han financiado proyectos por parte de la UE como por ejemplo el *Open Archives Forum* cuyo objeto es la creación de una comunidad de interés en *OAI* por medio de la organización de jornadas y actividades de soporte a la implementación de archivos y servicios.

Inmediatamente después de la difusión de la versión 1 comenzó el trabajo del comité técnico para tratar los problemas de definición o funcionalidad que se fueran descubriendo. Ese trabajo desembocó en la elaboración de la versión 2 del protocolo anunciada en junio de 2002. Los principales cambios que se introdujeron fueron relacionados con la clarificación de ambigüedades o mejores medios para expresar las funcionalidades existentes. Es decir, no se introdujeron cambios sustanciales.

«La iniciativa deja de lado los aspectos culturales y sociales que pueda acarrear la implementación del protocolo y se centra exclusivamente en cuestiones técnicas»

Entre los planes para el futuro está la creación de una versión *Soap* (*Simple Object Access Protocol*) del protocolo. Se espera que ésta se convierta en una parte integral del trabajo en bibliotecas digitales. De hecho, en los pasados meses se ha pasado de hablar fundamentalmente del protocolo mismo a comentar proyectos en los que se usa, y después a analizar proyectos sin ni siquiera mencionarlo. Otro área de interés son los formatos de metadatos, básicamente determinar si cumple su función el sistema utilizado actualmente (*Dublin Core* no calificado). También se estudiará la utilidad del protocolo más allá de la descripción de recursos, en cuestiones como certificación, estadísticas de uso, datos sobre citas, etc. Un aspecto que merece especial atención es una vuelta a la misión original de *OAI*, los eprints. Para ello se trabajará en un perfil de *OAI-PMH* para este tipo de documentos.

http://www.w3.org/TR/SOAP

3. El protocolo *OAI-PMH*

Los participantes en Santa Fe tomaron una decisión clave en cuanto a la arquitectura del protocolo. Adoptaron un modelo que rechazaba la búsqueda distribuida (como hace Z39.50) a favor de simplemente tener servidores proporcionando metadatos, sujetos sólo a criterios de alcance bastante simples, tales como

por ejemplo todos los registros añadidos o cambiados desde una fecha específica.

No vamos a entrar aquí en una descripción técnica del protocolo, pero básicamente utiliza transacciones http para emitir preguntas y obtener respuestas entre un servidor o archivo y un cliente o servicio recolector de metadatos. El segundo puede pedir al primero que le envíe metadatos según determinados criterios como por ejemplo la fecha de creación de los datos. En respuesta, el primero devuelve un conjunto de registros en formato xml, incluyendo identificadores (urls por ejemplo) de los objetos descritos en cada registro.

Las peticiones se emiten utilizando los métodos *get* o *post* del protocolo http y constan de una lista de opciones con la forma de pares del tipo *clave=valor*. Existen seis peticiones que un cliente puede realizar a un servidor:

—*GetRecord*. Recupera un registro concreto. Necesita dos argumentos: identificador del registro pedido y especificación del formato bibliográfico en que se debe devolver.

—*Identify*. Utilizado para recuperar información sobre el servidor: nombre, versión del protocolo que utiliza, dirección del administrador, etc.

—*ListIdentifiers*. En lugar de los registros completos recupera sólo sus encabezamientos. Permite argumentos como el rango de fechas entre los que queremos recuperar los datos.

—*ListRecords*. Igual que el anterior pero recupera los registros completos.

—*ListSets*. Hace posible la recuperación de un conjunto de registros, los cuales son creados opcionalmente por el servidor para facilitar una recuperación selectiva de los registros. Sería una clasificación de los contenidos según diferentes entradas. Un cliente puede pedir que se recuperen sólo los registros pertenecientes a una determinada clase. Los conjuntos pueden ser simples listas o estructuras jerárquicas.

—*ListMetadataFormats*. Devuelve la lista de formatos bibliográficos que utiliza el servidor.

El protocolo soporta múltiples formatos para expresar los metadatos. No obstante requiere que todos los servidores ofrezcan los registros utilizando *Dublin Core* no calificado, codificado en xml. Además de este formato, cada servidor es libre de ofrecer los registros en otro/s adicionales (marc, por ejemplo). Un cliente puede pedir que los registros se le sirvan en cualquiera de los soportados por el servidor. La idea subyacente aquí es que en el futuro las diferentes comunidades que utilicen el protocolo definan sus propios formatos que sean más ricos y más precisos que *Dublin Core*. Por ejemplo, la comunidad de archivos de eprints está trabajando en un formato denominado *AMF* (*Academic Metadata Format*) que sea capaz de describir todos los elementos que intervienen en el proceso de comunicación científica: documentos, autores, instituciones y canales de distribución de documentos.
http://amf.openlib.org/doc/ebisu.html

Las respuestas del servidor estarán formateadas según el protocolo http con los encabezamientos adecuados. Serán documentos xml correctos que se podrán validar contra el esquema definido en el protocolo y disponible en la dirección:
http://www.openarchives.org/OAI/2.0/

Aspectos que no trata el protocolo son por ejemplo cuestiones de gestión o autorización para el acceso de los clientes. El servidor deberá recurrir a métodos externos si desea limitar los clientes a los que sirva información. En relación con este punto está la utilización que los clientes hagan de los datos pues también queda fuera del protocolo. Finalmente, tampoco trata el tema de cómo los clientes pueden localizar aquellos servidores que contengan los datos que necesitan.

4. Proveedores de datos y de servicios

De la sección anterior se desprende que la arquitectura de *OAI-PMH* se basa en clientes y servidores. Los primeros son los archivos que proporcionan la información, y los segundos son los recolectores o servicios que toman los datos, con el objetivo de incorporarles algún valor añadido y presentarlos a los usuarios finales.

Desde enero de 2001 se ha mantenido un registro de todos los archivos que han implementado el protocolo. Al no ser algo obligatorio, se supone que son muchos los archivos que existen y no se han dado de alta. Desde esa fecha, el incremento del número de archivos ha sido constante llegando hasta 45 en la actualidad (sólo los que han adoptado la versión 2). Entre ellos tenemos como más destacables: *arXiv.org* junto con el resto de iniciativas que mencionamos en el punto 2; *Cern*, que recoge informes y prepublicaciones en el área de física, o *Citebase*, que proporciona datos sobre citas recibidas por los eprints almacenados en varios archivos.

Habría que destacar también los archivos que se están abriendo en el área de biblioteconomía y documentación. En estos momentos hay tres disponibles, aunque solamente uno de ellos está registrado:

—*@rchiveSIC* es un proyecto de colaboración entre varias instituciones francesas (universidades y centros de investigación como el *Cnrs*). En estos momentos almacena unos 80 documentos, la mayor parte de

ellos en francés. Incluye materiales de áreas relacionadas como museología.
http://archivesic.ccsd.cnrs.fr/

—*Dlist* (*Digital Library of Information Science and Technology*). Es un archivo creado por la *School of Information Resources and Library Science* y *Arizona Health Sciences Library* (*University of Arizona*). Almacena más de 100 documentos. Su objetivo es recoger todo tipo de textos científicos en documentación pero con dos áreas temáticas de mayor énfasis: materiales educativos y bibliometría. Solamente aceptan documentos en inglés.
http://dlist.sir.arizona.edu/

—*E-LIS* (*Eprints in Library and Information Science*). Es el proyecto más reciente dado que aún no se ha hecho público. Es un esfuerzo internacional para crear un archivo multinacional y multilingüe de documentos científicos en las áreas de biblioteconomía y documentación. Ha sido financiado parcialmente por el *Ministerio de Educación y Cultura* español.
http://eprints.rclis.org

Igualmente se mantiene un registro de los servicios creados utilizando los datos proporcionados por los anteriores. Tampoco aquí es obligatorio el registro, por lo que es imposible saber cuántos servicios existen realmente en la actualidad. Sin duda son muchos más que los 12 que aparecen en el registro oficial. Como más destacables:

—*ARC*. Es un servicio experimental creado con el objetivo de investigar temas relacionados con la recolección de metadatos siguiendo el protocolo *OAI-PMH* y cómo hacerlos disponibles a los usuarios. Más que un servicio en sí mismo es un software que podría ser utilizado por instituciones que quieran crear sus propios servicios. El código fuente está disponible en la Red de forma gratuita. Ha sido desarrollado por el *Digital Library Research Group* de la *Old Dominion University*.
http://arc.cs.odu.edu/

—*OAIster.* Es un proyecto financiado por la *Fundación Mellon* con el objetivo de crear una amplia colección de recursos digitales gratuitos útiles, que previamente eran de muy difícil acceso, y ponerla al alcance de cualquier usuario de la forma más sencilla posible. Es decir, trata de sacar a la luz colecciones que antes eran invisibles. Todos los recursos tienen el texto completo disponible en la Red de forma que siempre se pueda llegar a los contenidos. Recoge datos de todos los archivos conocidos. En total 122 archivos con más de un millón de registros.
http://oaister.umdl.umich.edu/

—*Perseus*. Es una biblioteca digital especializada en humanidades. Su servicio *OAI* también recupera datos de todos los servicios conocidos. Está financiada por la *National Science Foundation* en los EUA.
http://www.perseus.tufts.edu/

—*Cyclades*. Es un proyecto sufragado por la Unión Europea. Su objetivo no está directamente relacionado con OAI ya que es crear un marco de colaboración entre los investigadores de los centros que participan en el proyecto. Intenta fomentar la colaboración entre los mismos, emitir recomendaciones y crear servicios personalizados. Otros servicios también financiados por la UE son ICite y Torii, ambos dentro del proyecto Tools for Innovative Publishing in Science.
http://www.ercim.org/cyclades
http://icite.sissa.it/
http://tips.sissa.it/

En resumen, los proveedores están proliferando y cada vez están proporcionando servicios más sofisticados. Se puede decir que existe un mercado donde los servicios pueden competir, por ejemplo existen hasta 10 interfaces diferentes a los datos proporcionados por *arXiv.org*, cada uno de ellos con unas características diferenciadas.

5. ¿Cómo crear un archivo abierto?

Como hemos visto en el punto 3, *OAI-PMH* solamente es una interfaz sumamente sencilla para acceder a la información bibliográfica disponible en un archivo o repositorio. Por lo tanto, cualquiera puede realizar una implementación del mismo para poner a disposición de la comunidad internet los datos que hasta ahora estaban escondidos en bases de datos o catálogos. En este sentido bastaría con disponer de un servidor web y un programa cgi (en *Perl* o *PHP*) que recibiera las peticiones *OAI-PMH*, interrogara nuestra base de datos y devolviera la respuesta.

Por otro lado, la iniciativa *OAI* nace del movimiento de eprints, cuyo objetivo es poner a disposición del público documentos en formato electrónico vía repositorios. Para facilitar esta tarea ha aparecido una serie de programas que permiten a cualquier institución (universidad o centro de investigación) crear su propio archivo al tiempo que hacerlo compatible con *OAI-PMH*. Ejemplos de algunos de estos programas son:

a. *Eprints*. Es el más popular de todos ya que está siendo utilizado en más de 30 instituciones. Es un software desarrollado en el seno del *Open Citation Project* dirigido por **Stevan Harnad** en la *Universidad de Southampton* (UK). Está diseñado con el objetivo de ser fácil, rápido de instalación y gratuito. Se distribuye bajo la licencia *GNU*, lo cual significa que el código fuente es accesible y modificable por cualquier programador, con la condición de que las modificaciones se hagan también accesibles públicamente. *Eprints* puede funcionar en cualquier ordenador con sistema operativo *Linux*. Sus principales características son:
http://www.eprints.org

—Facilidad de instalación y configuración. Es un objetivo que no se ha alcanzado aún. Si bien el proceso está automatizado en gran parte se necesitan conocimientos técnicos para llevarlo a cabo. Es difícil que en el estado actual pueda ser instalado por investigadores sin asistencia de administradores de sistemas.

—Se pueden almacenar documentos en cualquier formato, así como almacenar un mismo documento en varios formatos. La carga de ficheros se realiza mediante una interfaz web muy sencilla.

—Permite utilizar cualquier formato para almacenar la información bibliográfica sobre los documentos.

—Hace posible que los usuarios se registren como lectores o como autores para obtener un mayor aprovechamiento de sus funciones.

b. *Dspace*. Es el más reciente de los programas ya que se anunció en noviembre de 2002. Está desarrollado por la empresa *HP* y las bibliotecas del *MIT*. También es un software con las fuentes disponibles públicamente (open source) cuyo objetivo es permitir a una organización almacenar, describir y gestionar documentos electrónicos, distribuirlos a través del web mediante un sistema de búsqueda y recuperación de la información y finalmente proporcionar un sistema para el almacenamiento a largo plazo de los documentos. Está pensado para funcionar en varias plataformas y soporta la versión 2 de *OAI-PMH*.
http://www.dspace.org

c. *CDSware*. Su primera versión se hizo pública en agosto de 2002. Está desarrollado, mantenido y utilizado por el *Cern Document Server* de Ginebra. Su objeto es permitir a una institución crear su propio servidor de eprints, catálogos de sus fondos o un sistema documental a través del web. Es compatible con *OAI-PMH*. Lo más destacado es que utiliza el formato marc 21 para almacenar los registros bibliográficos. Igual que en los casos anteriores, es un software gratuito distribuido bajo la licencia *GPL* (*General Public License*). En la biblioteca del *Cern* se utiliza para gestionar más de 350 colecciones formadas por más de 565.000 registros con unos 220.000 de ellos representando documentos a texto completo. El incremento se sitúa en torno a los 1.000 semanales.
http://cdsware.cern.ch

—*VT ETD-db*. Creado en el *Virginia Polytechnic Institute* y la *Virginia State University* (EUA). Como en los casos anteriores se trata de un software para crear depósitos de documentos y está siendo usado por la *Université Catholique de Louvain*. Trata de proporcionar una interfaz para que los usuarios puedan introducir y gestionar información bibliográfica relativa a colecciones de tesis en formato electrónico.
http://scholar.lib.vt.edu/ETD-db/

6. Bibliografía

Shapiro, Carl; Varian, Hal R. *Information rules: a strategic guide to the network economy*. Boston: Harvard Business School Press, 1999.

Dobratz, Susanne; Matthaei, Birgit. *Overview of European Open Archives Activities on OAI by the Open Archives Forum*. Consultado el: 03-03-03.
http://www.oaforum.org/otherfiles/lisb_overview.ppt

José Manuel Barrueco, *biblioteca de Ciències Socials. Universitat de València.*
jose.Barrueco@uv.es

Imma Subirats Coll, *biblioteca del Dept. de Política Territorial i Obres Públiques. Generalitat de Catalunya.*
immasubirats@myrealbox.com