

Proyecto: OAC: Open Archives Catalog

¿?.¿?: OAI-PMH. Estado del arte

año 2004

DELi

**[Inés Jacob, Joseba Abaitua, JosuKa Díaz,
Fernando Quintana, Garikoitz Echebarria]**

Este documento pretende ser una introducción a la OAI (Open Archives Initiative) y más concretamente al protocolo OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Se analizan sus orígenes, su evolución y se describe su arquitectura y conceptos técnicos fundamentales. Finalmente se describe en qué situación se encuentra actualmente el protocolo y cuáles son sus perspectivas de futuro.

1. OPEN ARCHIVES INITIATIVE. EL ORIGEN DEL PROTOCOLO OAI-PMH

Algunas disciplinas comenzaron, a principios de la década de los 90, a crear archivos o repositorios de documentos electrónicos (eprints) para conseguir una rápida comunicación de sus publicaciones. De esta forma los resultados de investigaciones en estas áreas eran conocidos más rápidamente que hasta entonces.

Sin duda el más conocido de estos archivos es arXiv.org, creado por Paul Ginsparg en Los Alamos (USA) para el área de la Física. Entre otros también podemos mencionar CogPrints en el área de la Psicología, NCSTRL en el campo de la Informática o REPEC en el campo de la Economía.

Se permite acceder a estos repositorios mediante interfaces web con ayuda de búsquedas. Cada uno de estos repositorios tiene un interfaz web distinto por lo que un usuario interesado en varios de ellos debe aprender tantos interfaces como repositorios quiere acceder.

En julio de 1999 Paul Ginsparg, Rick Luce y Herbert Von de Sompel convocaron a expertos en archivos de eprints para una reunión en Santa Fe (Nuevo México, Estados Unidos) en octubre de ese mismo año. El objetivo de esta reunión era desarrollar y promover estándares de interoperabilidad entre los archivos de eprints. Con esta misión fue creada la OAI en dicha reunión y decidieron establecer una solución minimalista para que estos estándares fueran mayoritariamente adoptados entre la comunidad de proveedores de eprints. Por

ello la recopilación de metadatos (metadata harvesting) fue la solución de interoperabilidad elegida. A su vez, fueron tomados los primeros acuerdos técnicos y organizativos, conocidos como la Convención de Santa Fe. Entre dichos acuerdos técnicos se incluyeron consensos sobre un protocolo de recopilación de metadatos basado en el protocolo Dienst (utilizado por el archivo NCSTRL), un estándar de metadatos común (Open Archives Metadata Set) y un esquema de identificación uniforme.

Los resultados de Santa Fe fueron publicados en febrero de 2000 y pronto comenzaron a surgir comunidades que vieron que esta iniciativa podía cubrir algunas de sus necesidades. Los bibliotecarios y museólogos se mostraron muy interesados en descubrir formas de publicar en Internet partes de las colecciones de bibliotecas y museos. Por otra parte pronto surgieron voces que ponían en duda esta iniciativa debido a la falta de una estructura organizacional.

Rápidamente atajaron este último imprevisto. En agosto de 2000, la DLF (Digital Library Federation) y el CNI (Coalition of Networked Information) anunciaron que ofrecerían el soporte de su organización a la iniciativa. En ese momento se crearon dos comités, uno de gestión y otro técnico, que se encargarán de la coordinación de la iniciativa. Además, se revisaron las decisiones tomadas en Santa Fe para ampliar el objeto de trabajo más allá de los eprints e incluir disciplinas que no tuvieran este tipo de documentación. Los aspectos técnicos que sólo eran aplicables a los eprints también fueron reconsiderados.

En enero de 2001 se publicaron las especificaciones revisadas y se publicó la versión 1.0 del protocolo OAI-PMH. Desde ese momento comenzó la implementación del protocolo y empezaron a aparecer las primeras instituciones que lo utilizaron para publicar sus metadatos en Internet. La adopción fue lenta pero progresiva y en junio de 2002 el comité técnico elaboró la versión 2.0 del protocolo teniendo en cuenta los problemas y ambigüedades que se fueron descubriendo durante la implementación de la versión 1.0.

2. EL PROTOCOLO OAI-PMH

Una de las prioridades a la hora de decantarse por un modelo de interoperabilidad era que fuera lo más sencillo posible para conseguir una amplia adopción de la iniciativa. Como ya se ha mencionado, el elegido fue el modelo de recolección de metadatos frente al modelo de búsqueda distribuida.

El modelo de recolección de metadatos (metadata harvesting) permite a los proveedores de documentos electrónicos exponer sus metadatos a través de una interfaz, con el objetivo de que la misma pueda ser utilizada como base para desarrollar servicios de valor añadido.

Se asume que una solución de este tipo tiene limitaciones de funcionalidad. Existen otros estándares de interoperabilidad, por ejemplo el Z39.50 (basado en el modelo de búsqueda distribuida), que recogen algunos aspectos de una forma más completa. Sin embargo, las estrategias de interoperabilidad generalmente aumentan en dificultad de implementación con un incremento de funcionalidad. Es por esto que la elección del modelo de recolección de metadatos responde a la necesidad de dar una solución de interoperabilidad con bajo coste de implementación.

Este enfoque hace que identifiquemos claramente dos roles: Los “proveedores de datos” y los “proveedores de servicios”. A los primeros se les suele llamar “clientes” y son los que proporcionan los metadatos que hacen referencia a los recursos publicados y los segundo son los “servidores” que son los que recolectan los metadatos de los proveedores de datos con el objetivo de darles algún valor añadido y presentarlos a los usuarios finales. Es posible que un participante tome ambos roles, ofreciendo tanto metadatos como servicios.

2.1 METADATOS

Técnicamente existían dos necesidades básicas a la hora de abordar el asunto de los metadatos. Por una parte debería garantizarse la interoperabilidad y por otra la

extensibilidad o especificidad para cada comunidad. Estos aspectos han sido, y en menor medida siguen siendo, muy discutidos en la comunidad.

Para satisfacer la necesidad de interoperabilidad en los metadatos se decidió que todos los proveedores de datos proveyeran sus metadatos en un formato común. Este formato es Dublin Core sin cualificar. Se barajaron otras alternativas para garantizar la interoperabilidad pero obligar a utilizar un formato común pareció la solución más prudente.

La decisión de elegir Dublin Core sin cualificar como formato común fue muy meditada. De hecho, en un principio en Santa Fe definieron un formato de metadatos (OAMS) pensando solamente en las comunidades de eprints. Sin embargo la ampliación de objetivos de la OAI forzó a reconsiderar esta decisión y elegir un formato de metadatos que fuera adecuado para distintos dominios.

Por otro lado, se anima a las distintas comunidades a que, además de en DC, expongan sus metadatos en formatos específicos para cada comunidad. El único requisito a cumplir es que los registros en dichos formatos estén estructurados como documentos XML y que tengan su XML schema correspondiente para validarlo. A pesar de todo, son muy pocas las disciplinas que ofrecen los metadatos en un formato propio.

2.2 RECURSOS, ITEMS Y REGISTROS

Evidentemente un proveedor de datos publicará los metadatos de sus recursos electrónicos. Estos recursos son los referenciados por los metadatos. Un item es la representación de dicho recurso y contiene un identificador único. Por último, cada item puede estar representado en varios formatos de metadatos. Existirán tantos registros asociados a un item como formatos de metadatos sean utilizados para expresar un item.

Como hemos mencionado, cada item debe tener un identificador único. Se llegó a un acuerdo en cuanto a la identificación. Debe tener tres componentes:

- a) oai - Una cadena de caracteres fija.

- b) <IDRepositorio> - Un identificador del repositorio.
 c) <IDItem> - Un identificador del item dentro del repositorio.

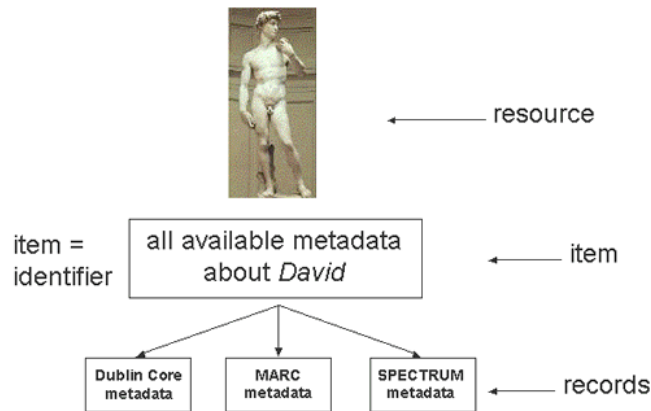


Figura 1

Por lo tanto, un ejemplo válido podría ser oai:arXiv:doc000001

Un registro tiene tres partes: header, metadata y about. La header contiene información común a todos los registros (independientemente del formato de metadatos del registro) como son el identificador (único) y el datestamp, que indica la fecha de creación, borrado o último cambio de los metadatos del registro. En la parte metadata contiene metadatos en un formato y en la parte about es flexible y puede ser utilizada para indicar algo sobre los metadatos (típicamente es utilizado para indicar información sobre los derechos sobre los metadatos).

```
<record>
  <header>
    <identifier>oai:sigir:ws3</identifier>
    <datestamp>2001-08-13</datestamp>
  </header>
  <metadata>
    <dc>
      <title>OAI Workshop at SIGIR</title>
      <creator>Hussein Suleman</creator>
      <language>English</language>
    </dc>
  </metadata>
  <about>
    <metadataID>oai:sigir:ws3md</metadataID>
  </about>
</record>
```

2.3 ESPECIFICACIÓN DEL OAI-PMH

El protocolo se compone de seis tipos de peticiones o verbos. Estas peticiones se realizan 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. Mediante estas peticiones el proveedor de servicios pide al proveedor de datos sus metadatos. Se puede realizar esta recolección de metadatos de forma selectiva, en base a rangos de fechas (para recolectar solamente los registros modificados en un rango de fechas) y en base a sets (los metadatos pueden estar clasificados en grupos). Como respuesta, el proveedor de datos devuelve un conjunto de registros en formato XML. En caso de que haya algún tipo de error en la petición (falta de algún argumento, errores léxicos,...) el proveedor de datos devolverá la excepción adecuada.

A continuación vamos a detenernos en una descripción más técnica del protocolo. Las peticiones OAI tiene la siguiente estructura:

- URL base. El host y puerto del servidor que actúa como repositorio de datos. Se le puede añadir un path.
- Argumentos. Son una lista de pares clave-valor. Como mínimo, cada petición OAI tiene un par clave-valor que especifica el tipo de petición que se quiere realizar.

Un ejemplo realizando una petición de tipo GetRecord (describiremos más adelante cada una de ellas) al host `http://edoc.hu-berlin.de/OIA-2.0` sería `http://archive.org?verb=GetRecord&identifier=oai:HUBerlin:3000819&metadataPrefix=oai_dc`

Los seis tipos de peticiones que un proveedor de servicios puede realizar a un proveedor de datos son las siguientes:

- Identify. Se utiliza para obtener información sobre el repositorio: nombre, versión del protocolo que utiliza, contacto con el administrador,... No tiene ningún argumento adicional.

- ListMetadataFormats. Devuelve la lista de formatos bibliográficos que utiliza el servidor.
- ListSets. Devuelve la estructura de un repositorio.
- ListIdentifiers. Recupera solamente los encabezados de los registros. Es obligatorio añadir el argumento metadataPrefix para indicar en qué formato queremos que nos devuelva los encabezados. Permite, opcionalmente, los argumentos from y until para indicar un rango de fechas (no hay por qué usar los dos a la vez) y el argumento set para indicar los registros de qué conjunto queremos obtener.
- ListRecords Recupera registros de un repositorio. Los parámetros posibles son los mismos que para el verbo ListIdentifiers
- GetRecord. Devuelve un registro de un repositorio. Tiene dos parámetros que son obligatorios: metadataPrefix e identifier (el identificador único del registro a recuperar).

Tres de estos tipos de petición pueden devolver largas listas de entradas. La especificación de OAI-PMH soporta particionamiento de las respuestas. La decisión de implementar o no el particionamiento está en el proveedor de datos pero los proveedores de servicios deben soportarlo obligatoriamente.

Una respuesta particionada incluye la lista incompleta de registros, un resumption token (que se utilizará para pedir el siguiente grupo de registros) y una fecha de expiración del resumption token. Para realizar una petición pidiendo el siguiente grupo de registros solamente deberemos indicar el verbo (tipo de petición) y el argumento resumptionToken con el valor obtenido en la anterior respuesta. Cuando se reciba el último grupo de registros el resumptionToken estará vacío.



Figura 2

2.4 REPOSITORIOS ESTÁTICOS

Aunque ya de por sí OAI-PMH está pensado para que no ponga muchas trabas a quien quiera implementarlo, su implementación no es trivial y el tamaño de la colección de metadatos puede no justificar la investigación e implementación del protocolo.

Para hacer aún más sencilla la exposición de los metadatos la OAI creó los repositorios estáticos. Están pensados para colecciones relativamente pequeñas de metadatos (de 1 a 50.000 registros). De esta forma, quién quiera publicar sus metadatos tiene que crear un fichero XML que contenga información sobre el repositorio y sus metadatos y colocarlo en un servidor web. Evidentemente, existe un XML Schema que debe cumplir dicho fichero XML.

Un repositorio estático es accesible via OAI-PMH mediante la intermediación de una pasarela de repositorio estático (Static Repository Gateway) que, utilizando los registros de metadatos y la información sobre el repositorio que alberga el fichero XML, es capaz de responder a los seis tipos de peticiones OAI-PMH.

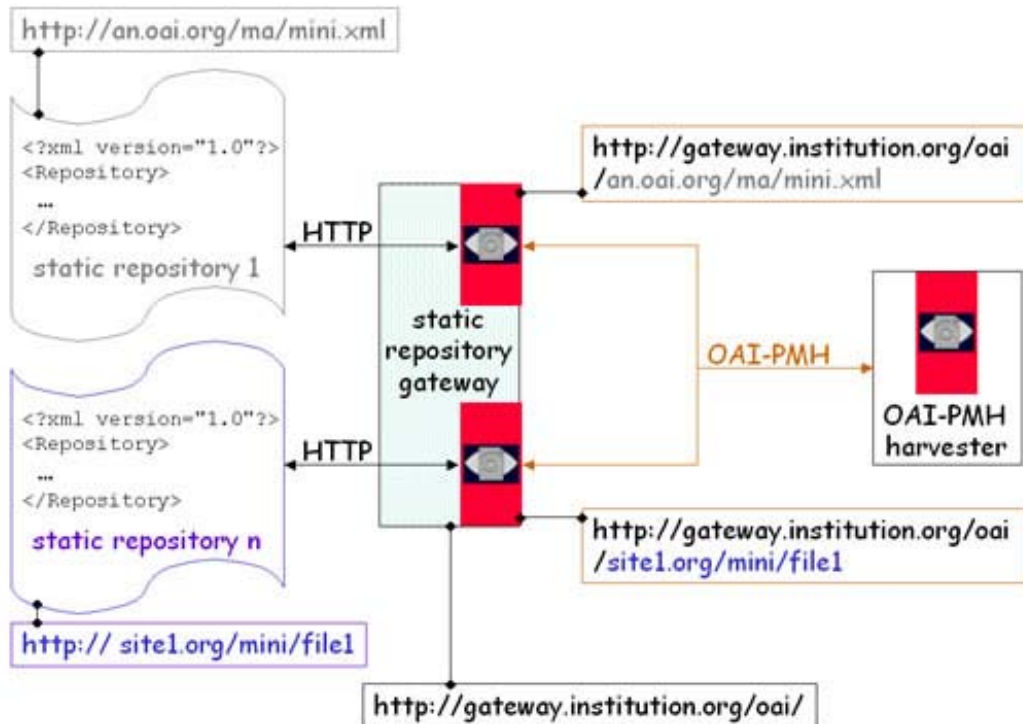


Figura 3

Cada repositorio estático es accesible por una URL que se crea uniendo la dirección de la pasarela de repositorio estático y la dirección en la que está alojado en fichero XML. Evidentemente, una pasarela de repositorio estático puede ser utilizada por varios repositorios estáticos.

Evidentemente, esta técnica para publicar los metadatos exige un esfuerzo técnico mucho menor y menos tiempo ya que basta con crear el fichero XML con los metadatos y asociarte a una pasarela de repositorio estático.

3. MIRANDO AL FUTURO

Actualmente OAI-PMH es un protocolo estable. Desde que se publicó la versión 2.0 del protocolo éste no ha sufrido cambios y no hay planes para una versión 3.0. El núcleo del protocolo no será extendido y, como mucho, podrían surgir versiones 2.x a medio plazo en las que no habría cambios funcionales sino tipográficos.

En cualquier caso, existen algunos aspectos que aún esperan ser mejorados. Los proveedores de datos, que mayormente publican sus datos en DC, están a la espera de mejores metadatos. La propiedad intelectual está siendo también un tema muy debatido. Se espera que en el futuro las guías de implementación del protocolo especifiquen mecanismos para expresar los derechos de autor en OAI-PMH.

4. REFERENCIAS

En la actualidad existen dos referencias fundamentales para la comunidad OAI: la propia Open Archives Initiative a través de su página web (<http://www.openarchivesinitiative.com>) y el proyecto Open Archives Forum.

La Open Archives Initiative publica en su página web los documentos más importantes de la OAI, noticias, FAQs y permite subscribirte a listas de distribución sobre la OAI.

El proyecto Open Archives Forum tenía como objetivo crear una comunidad de interés en la OAI. Para ello desarrollaron una página web en la que ofrecen servicios muy interesantes a la comunidad. La principal aportación a la comunidad es un excelente tutorial introductorio a OAI-PMH. Además, tiene listados de organizaciones, proyectos y software relacionados con la iniciativa y te permite registrar tus proyectos o desarrollos para que la comunidad los conozca. Por último, ofrece también información (programas, presentaciones,...) de los workshops realizados.

Desde estas dos referencias se puede acceder a muchas otras referencias sobre la OAI.