EXPERIENCIA EN EL HARVESTING DE DOCUMENTOS OAI EN EL PROYECTO SEDICI

DE GIUSTI, M. R.[1]
MARMONTI, E.
VILA, M. M.
LIRA, A.
SOBRADO, A.

Universidad Nacional de La Plata
E-mail: emarmonti@sedici.unlp.edu.ar

RESUMEN

En este artículo se presentan las estrategias y particularidades encontradas en el proceso de "harvesting" de información académica realizado sobre diecisiete (17) repositorios de información. Esta iniciativa, llevada a cabo en la Universidad Nacional de La Plata, tiene como objetivo el brindar material de valor académico complementario al de propia producción institucional ya residente en la Biblioteca Digital del Portal SeDiCI. En el trabajo se tratan temáticas tales como: dificultades que el protocolo debería considerar para las posibilidades tecnológicas y de conectividad de nuestros países, dificultades en lo referente a la uniformidad de la información que se encuentra en los diferentes repositorios, tanto como los hallazgos y el potencial aprovechamiento máximo que el protocolo puede brindar para el descubrimiento de nuevas fuentes de información.

Palabras Clave: Protocolo Archivos Abiertos; Metadatos; Convención Dublin Core; Recursos de información; Libre Acceso.

---

1. Investigador Comisión de Investigaciones Científicas de la Provincia de Buenos Aires - CIC y Directora del Proyecto de Enlace de Bibliotecas (PrEBi) y del Servicio de Difusión de la Creación Intelectual (SeDiCI) de la Universidad Nacional de La Plata, Argentina. Dirección de consulta: marisadg@ing.unlp.edu.ar

INTRODUCCIÓN

El Servicio de Difusión de la Creación Intelectual (SeDiCI) forma parte del Proyecto de Enlace de Bibliotecas (PrEBi) perteneciente a la Universidad Nacional de La Plata (UNLP). Su principal objetivo es el de difundir por medios tecnológicos los trabajos académicos creados por la comunidad de docentes, alumnos e investigadores de la mencionada casa de estudios.

SeDiCI contempla una gran diversidad de tipos documentales, cada uno de los cuales posee una convención de carga (entrada de metadatos) propia. La tipología documental abarca desde Tesis (de grado y de postgrado), artículos de Publicaciones Periódicas, documentos multimediales (con sonidos e imágenes), libros electrónicos, proyectos de investigación, etcétera.

SeDiCI trabaja de acuerdo a las pautas de las Bibliotecas Digitales modernas, pone a disposición los recursos digitales que se producen en la comunidad académica de la Institución, tanto como aquellas que potencialmente podrían llegar a querer utilizar quienes realizan las consultas [AGENJO 2005]; bajo esta idea es que se han desarrollado diferentes estrategias para maximizar la cantidad de documentos obtenidos intentando, a la vez, minimizar el esfuerzo de procesamiento y conexión que implica esta tarea: la estrategia de "harvesting" da origen el presente trabajo.

Existen pocos "Service Provide" que oferten una cantidad importante de documentos obtenidos como resultado de la *cosecha*. El mejor ejemplo es OAIster [Oaister] , que posee información de casi de 6 millones de documentos. SeDiCI aspira a administrar en forma inteligente, volúmenes comparables de referencias con un procesamiento automático que permita agregar descriptores adicionales para la interpretación/catalogación/búsqueda de la información referencial. Actualmente, SeDiCI maneja alrededor de 1.100.000 documentos de casi 20 repositorios de todo el mundo.

En la propuesta inicial del proyecto SeDiCI se incluía la de integrar los repositorios de información propia a portales como el *Latinoamerican Open Archives Portal* bajo el rol de Data provider [Latam].

CARACTERIZACIÓN DEL PROBLEMA. HARVESTING DESDE LATINOAMÉRICA?

Open Archives es un estándar, incluso como vía de indexación de sitios cuya información documental reside en bases de datos, los cuales no pueden ser indexados de manera directa por parte de metabuscadores como Google o Yahoo [AJENJO, 2005] [Google].

El rol de Data Provider consiste en mapear (trasladar) el conjunto de metadatos usado en forma local, por la plataforma de bibliotecas digitales, hacia la convención Dublin Core; esta operación genera problemas y hasta inconsistencias en la información obtenida desde los "harvesters".

El rol de Service Provider significa un desafío ya que implica recolectar información académica desde distintos repositorios, almacenando la información de los mismos en motores de bases de datos o indexadores que admitan una buena performance y un costo de mantenimiento/optimización adecuado. El rol de Service Provider es la cara visible al usuario final, y de acuerdo a estudios internacionales hay una relación 1:5 entre la cantidad de Service y Data Providers, tal número muestra que representa una innovación en cuanto a los servicios que debe prestar una biblioteca digital y especialmente, una biblioteca digital temática.

Las velocidades de acceso a Internet de las instituciones académicas en Argentina, sumado a la falta de equipos de características adecuadas, dificulta la tarea de un Service Provider académico en el contexto económico actual. En el caso de SeDiCI se dispone de un servidor dedicado a la recolección y aún así persiste una carga

sobre la red, debido a que el proceso está recolectando información de manera constante.

Una característica importante es el manejo de documentos eliminados [*OpenArchives*]; cada repositorio puede tomar una de las siguientes políticas: no mostrar nunca documentos borrados, mostrarlos siempre, o durante un cierto intervalo de tiempo. Cuando el harvester se encuentra con un registro que dice haber sido borrado, no lo agrega a la Base de Datos para no generar inconsistencias, vale aclarar que la cantidad de registros borrados, en la mayoría de los repositorios, es muy pequeña. Es posible realizar un mantenimiento de la Base de Datos periódicamente, pidiéndole al repositorio los documentos cosechados previamente para corroborar que en el repositorio remoto no existan documentos eliminados, y en el caso de existir, eliminar todos los documentos que fueron borrados desde la "cosecha" anterior. Este punto resulta de gran importancia, dado que parte de las responsabilidades de un Service Provider consisten en no presentar al usuario información errónea , inexistente o no pertinente.

En el protocolo OAI/PMH [OpenArchives], cuando la respuesta a una solicitud es muy numerosa en cantidad de documentos, se suele subdividir la misma en varias partes, adicionando a cada parte una "etiqueta de continuación" o "resumption token" única, excepto a la primera. Con cada fracción de la respuesta devuelta al cliente, se adjunta el resumption token correspondiente a la siguiente fracción, lo cual permite ir recuperando secuencialmente todos los resultados. El resumption token puede poseer un tiempo de validez, que depende del servidor, si el tiempo de validez expiró, no se puede utilizar para recuperar datos.

En la implementación del algoritmo, además de las potenciales desconexiones inducidas por problemas de conectividad locales (o agotamiento del repositorio), persiste un problema un problema que se considera es un impedimento serio. El protocolo no especifica un orden obligatorio de recuperación de los registros, si bien esto se cumplimenta habitualmente, ciertos repositorios no devuelven las entradas de metadatos bajo un orden específico en la recuperación. Cuando se

realiza un pedido al Data Provider los datos no vuelven ordenados por un criterio en particular (como sería esperable, dc:identifier), y si la cantidad de registros es grande se utilizan resumption tokens para recuperar los registros en forma parcial; si la conexión con el servidor se interrumpe, y no es posible retomarla rapidamente, el resumption token pierde su validez y no es posible continuar con el proceso de recuperación de los registros que restan a partir de un punto medio, debiendo reiniciar la cosecha.

Otro obstáculo frecuente consiste en el tratamiento de los caracteres especiales; por ejemplo los caracteres acentuados, o símbolos especiales. Tanto las Bases de Datos como las piezas de *software* que actúan como intérpretes del lenguaje de marcado XML, manejan un conjunto limitado de caracteres especiales, debido a esto suelen producirse errores en el proceso de intepretación de los resultados devueltos.

Un desafío importante para el proceso de harvesting de un Service Provider consiste en el uso de los Sets [OpenArchives]. Un repositorio admite una ilimitada cantidad de Sets, y un documento puede, potencialmente, pertenecer a todos ellos, lo que implica un procesamiento adicional sobre los documentos obtenidos, ya que al usuario final del Service Provider debe tener la posibilidad de aprovechar esta división conceptual, temática u organizativa que el repositorio de origen ha contemplado para sus documentos.

Una desventaja actual de los sets es la falta de un estándar internacional para su especificación. Los sets tienen tres atributos esenciales: un nombre, un identificador y una descripción. La descripción es posiblemente el campo más útil para el cliente final, ya que le permite ver de qué tratan los documentos pertenecientes al set en cuestión. Pero en general, este campo esta desaprovechado, ya sea tanto por su falta de uso por parte de los Data Provider como por su desorganización o no uso de estándares en el volcado de la información que contiene.

PROBLEMAS EN EL MAPEO DE METADATOS.

Muchas arquitecturas de bibliotecas digitales administran una convención propia o basada en algún estándar para el almacenamiento de los objetos digitales que administran. Convenciones de este tipo pueden nombrarse ETD-BR, MTS, MARC21, METS, etc.

El protocolo OAI-PMH admite la recolección de metadatos indicando cual es la convención que debe ser usada; debido a la existencia de diferentes convenciones, se unifica la colecta en la convención Dublin Core [DublinCore]. SeDiCI actualmente sólo colecta metadatos en formato Dublin Core, aunque es posible que en un futuro cercano también coseche Marc21 o algún formato de metadatos mas completo.

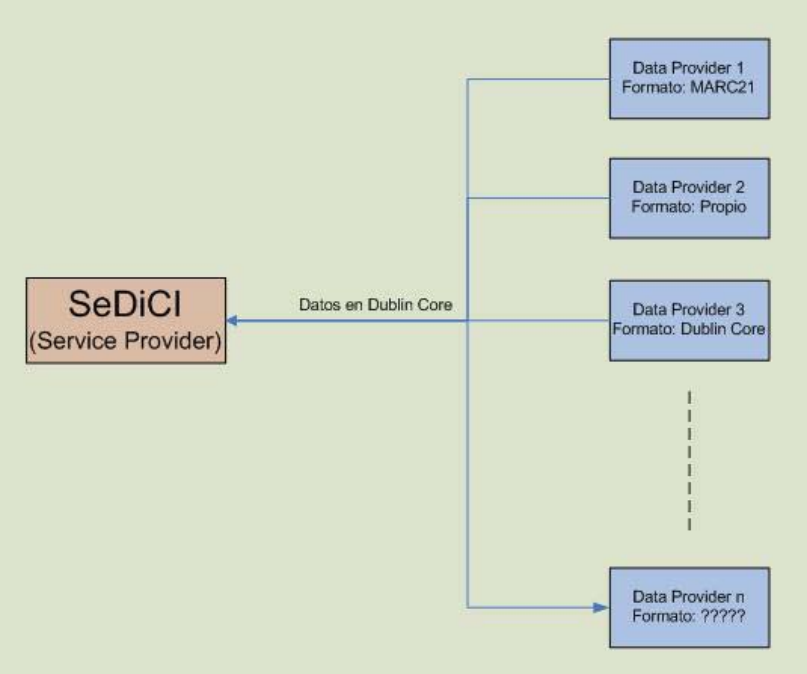

Figura - Flujo considerado de metadatos hacia el Service Provider.

El *software* que efectúa la provisión de datos debe realizar un mapeo entre la información que administra originalmente y la convención Dublin Core, tal mapeo no siempre resulta consistente. Esto significa que el harvesting de un repositorio devuelve, en un tag determinado de Dublin Core, información apropiada de acuerdo a su definición de dicho tag, no existiendo una definición Standard para todos los repositorios. Esto hace que la calidad de la información del Service Provider se vea afectada por la cantidad de repositorios que no realizan apropiadamente dichos mapeos, sumado a que no siempre se respetan normativas en relación a la carga de información textual; una tesis de postgrado puede aparecer como "Postgraduated Thesis" o simplemente como "Thesis", sin determinar el grado de la misma.

Este último problema introduce complicaciones a la hora de intentar realizar procesamientos o mediciones sobre el conjunto de datos obtenidos.

CARACTERIZACIÓN DEL PROCESO DE HARVESTING EN SEDICI.

Para la recolección de la información desde los distintos repositorios, se ha construido un proceso en el lenguaje de programación Java, reaprovechando una librería diseñada por la organización OCLC (OAIHarvester2) [OCLC]. El equipo de programación ha realizado ajustes en la librería de base y ha generado diversos elementos de parametrización los cuales permiten ajustar el comportamiento del proceso de harvesting propiamente dicho.

Parte de los desafíos planteados para mejorar los aspectos de eficiencia consiste en paralelizar el proceso de harvesting; Se puede pensar que si dicho proceso es paralelizable, la obtención de resultados será mucho más rápida, sobre todo para los nuevos repositorios o aquellos que tienen un buen nivel de actualización.

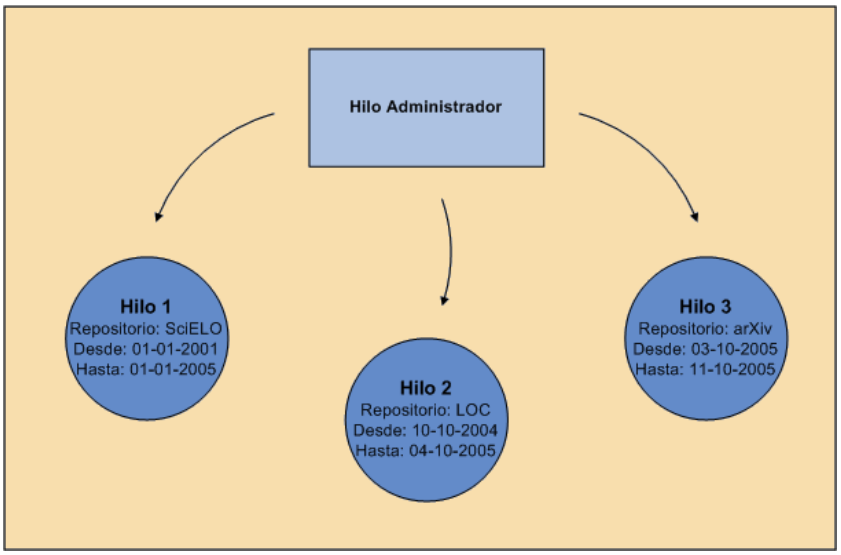

Figura - Ejemplo del Harvester ejecutando 3 hilos en paralelo.

El harvester de SeDiCI, consiste en varios hilos que funcionan de forma concurrente, donde cada hilo cosecha los datos de un repositorio entre dos fechas. También se utiliza un hilo de ejecución Administrador, cuyo objetivo es sincronizar los hilos anteriormente mencionados, manteniendo estadísticas globales sobre el proceso de harvesting (basándose en las estadísticas parciales de cada hilo).

El hilo Administrador también es el encargado de iniciar los cosechadores individuales, para ello lee información sobre los repositorios de la base de datos, y con los datos de configuración calcula las fechas entre las cuales deben hacerse el harvesting y finalmente crea instancias de los hilos con los parámetros adecuados. A medida que dichos hilos van finalizando, recoge las estadísticas sobre los datos recolectados por el hilo y actualiza la base de datos.

Al iniciar el proceso de Harvesting se crea un archivo de texto plano que registra los eventos que se van sucediendo, indicando fecha y hora, número de hilo y una descripción del evento. Este archivo permite conocer los repositorios contra los que se conecta cada uno de los hilos, los parámetros de la conexión, la cantidad de

documentos procesados, recolectados y borrados y por último, el momento de la finalización de las conexiones.

Este archivo de "log" se usa asimismo para registrar errores de las conexiones (desconexiones repentinas y fin de archivo inesperados). Gracias a la utilización de este archivo log, se puede conocer una vez terminado el proceso de harvesting que servidores respondieron y cuales no (conexiones rechazadas), que conexiones fueron suspendidas y el motivo de la suspensión, y principalmente la cantidad de documentos que se procesaron para cada repositorio, conociendo la cantidad de registros nuevos y la cantidad de registros ya existentes en la Base de Datos, para así mantener la información estadística sobre el proceso realizado.

```
Thu Jun 30 20:34:20 ART 2005 -> Comienzo del Harvesting: Thu Jun 30 20:34:20 ART 2005
Thu Jun 30 20:34:20 ART 2005 -> Se deben procesar 20 repositorios, mediante 3 threads

Thu Jun 30 20:34:20 ART 2005 -> Thread 0: Inicio del Harvesting sobre el Repositorio Library of Congress Open Archive
Initiative Repository 1
Thu Jun 30 20:34:20 ART 2005 -> Thread 0:
http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&metadataPrefix=oai_dc&from=2003-01-01&until=2005-06-30
Thu Jun 30 20:34:20 ART 2005 -> Thread 1: Inicio del Harvesting sobre el Repositorio NDLTD Union Catalog
Thu Jun 30 20:34:20 ART 2005 -> Thread 2: Inicio del Harvesting sobre el Repositorio ARCHIVE OF EUROPEAN INTEGRATION
Thu Jun 30 20:34:32 ART 2005 -> Thread 1:
http://alcme.oclc.org/ndltd/servlet/OAIHandler?verb=ListRecords&metadataPrefix=oai_dc&from=2003-01-01&until=2005-06-30
Thu Jun 30 20:35:05 ART 2005 -> Thread 2:
http://aei.pitt.edu/perl/oai2?verb=ListRecords&metadataPrefix=oai_dc&from=2003-01-01&until=2005-06-30
Thu Jun 30 20:35:23 ART 2005 -> Thread 0: http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&resumptionToken=7eoo
Thu Jun 30 21:00:37 ART 2005 -> Thread 2: http://aei.pitt.edu/perl/oai2?verb=ListRecords&resumptionToken=0/7385991/oai_dc
Thu Jun 30 21:00:38 ART 2005 -> Thread 1:
http://alcme.oclc.org/ndltd/servlet/OAIHandler?verb=ListRecords&resumptionToken=1120174472846:6300:194561:oai_dc
...
Thu Jun 30 21:00:39 ART 2005 -> Thread 2: Se han procesado 2056 registros.
Thu Jun 30 21:00:39 ART 2005 -> Thread 2: Ha finalizado el Harvesting sobre el Repositorio ARCHIVE OF EUROPEAN
       INTEGRATION (59)
Thu Jun 30 21:00:39 ART 2005 -> Thread 3: Inicio del Harvesting sobre el repositorio arXiv
Thu Jun 30 21:01:18 ART 2005 -> Thread 0: http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&resumptionToken=dRDt
Thu Jun 30 21:01:34 ART 2005 -> Thread 3:
http://arXiv.org/oai2?verb=ListRecords&metadataPrefix=oai_dc&from=2003-01-01&until=2005-06-30
Thu Jun 30 21:01:52 ART 2005 -> Thread 1:
http://alcme.oclc.org/ndltd/servlet/OAIHandler?verb=ListRecords&resumptionToken=1120174472846:6600:194561:oai_dc
Thu Jun 30 21:02:28 ART 2005 -> Thread 0: http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&resumptionToken=HnIU
```

Figura - Fracción del archivo de log generado usando el cual es posible analizar los verbos, su respuesta, los hilos de ejecución paralelos y los resultados de las cosechas.

El listado de los conjuntos de cada repositorio es actualizado automáticamente por un proceso especial que se ejecuta cada 15 días. Este proceso recorre el listado de repositorios disponibles y solicita a cada uno de ellos el listado de sets que posee. Al mantener esta información actualizada, se logra un proceso de harvesting más eficiente y actualizado, accediendo a los nuevos conjuntos de cada repositorio y evitando el acceso a conjuntos inexistentes.

Otro de los procesamientos que el harvester realiza, es la obtención de repositorios "friends" o relacionados con los que actualmente se encuentra explorando. Esta funcionalidad permite obtener nuevos repositorios, los cuales deberán ser aprobados manualmente por el encargado desde la interfaz de administración central antes de ser incluidos en los próximos procesos de harvesting. Al no ser el sitio de Open Archives el único referente acerca de un listado de repositorios en línea, esta funcionalidad ayuda a descubrir y proponer repositorios potencialmente valiosos, que son referidos por algún repositorio de nuestro interés.

En cuanto a la plataforma subyacente al harvester, se usa un servidor con un procesador Intel Pentium 4 de 2 Ghz, con 1 GB de memoria RAM y un disco de 120 GB. El Sistema Operativo residente está basado en la versión 9 de la distribución Red Hat de Linux, con algunas modificaciones propias.

ALMACENAMIENTO Y LÍNEAS DE TRABAJO

Uno de los desafíos más importantes que SeDiCI enfrenta en estos momentos consiste en la selección de un motor adecuado para el almacenamiento y recuperación de la información cosechada. Se está actualmente realizando evaluaciones de performance, recall y pertinencia sobre el repositorio obtenido usando diversas tecnologías.

Desde el perfil de Data provider, el desafío consiste en adaptar la construcción de *sofware* a las tendencias actuales [Jstor]. Dicha adaptación permitirá obtener una

solución única para el acceso de usuarios reales (humanos) como de servicios de indexación sobre el mismo repositorio. Muchas de las soluciones, tal como se expresa en [Jstor] no son pre-construídas y deben ser adaptadas a los requerimientos que SeDiCI propone.

Una de las líneas de trabajo planteadas por la Dirección del proyecto consiste en la obtención de información significante a partir de la cosecha. Debe ser posible automapear los descriptores y la clasificación de la información con el Tesuaro, o las herramientas taxonómicas propias de SeDiCI, ejemplo de esta modalidad podemos encontrar en el proyecto de Biblioteca Electrónica Universia [Universia] sobre el tesauro de UNESCO. Esto permitirá abstraerse de las diferencias de criterio usadas por los diferentes catalogadores de los distintos repositorios y las diferencias terminológicas en la clasificación documental, las cuales representan un obstáculo para el acceso a la información obtenida.

CONCLUSIONES

El protocolo OAI/PMH es una herramienta de formidable valor para la obtención de documentación técnico académica. Diversos autores citan la problemática existente en relación con los estándares y el desafío que implica la coexistencia de diversas convenciones de metadatos sobre los distintos repositorios [Anderson]. Superar estos desafíos implica la adhesión a “buenas prácticas” en el uso de los metadatos, a un trabajo cooperativo y coordinado entre los responsables de procesos técnicos de recursos digitales.

Por el lado tecnológico, es importante que los administradores de los repositorios que implementan el servicio de Data Provider OAI-PMH consideren todas las realidades acerca del acceso potencial por parte de proveedores de servicio latinoamericanos. Facilitar y estandarizar el comportamiento de los repositorios,

redundará en forma directa sobre la democratización del acceso a los recursos de información por parte de las comunidades académicas de estos países.

REFERÊNCIAS BIBLIOGRÁFICAS

[Agenjo 2005] Agenjo, Xavier, "Recursos Digitales: Un reto para las Bibliotecas Nacionales". Jornada sobre Bibliotecas Nacionales. Biblioteca Valenciana. 2005. http://bv.gva.es/documentos/Ponencias/Agenjo.pdf

[Anderson] Anderson, Ian and Ross, Seamus. Discovering Good Practice: Metadata and the NINCH Guide. In Proceedings 3rd. Open Archives Forum Workshop, Berlin. 2003. http://eprints.rclis.org/archive/00001158/Principio del formulario

[Open Archives] Open Archives Metadata Initiative Protocol. http://www.openarchives.org

[Latam] LatinAmerican Open Archives Portal. http://www1.lanic.utexas.edu/project/laoap/indexesp.html

[DublinCore] Dublin Core Metadata Initiative. http://www.dublincore.org

[OCLC] OAIHarvester2. OCLC. http://www.oclc.org/research/*software*/oai/harvester2.htm

[Oaister] OAISter. Michigan University. http://oaister.umdl.umich.edu/o/oaister/

[Google] Google admite la inclusión de Open Archives como medio de indexación de *Site*maps móviles o dinámicos. https://www.google.es/*Web* masters/*Site*maps/docs/es/other.html

[Jstor] Krot, Michael and Yakimischak, David. Building the JSTOR OAI-PMH service: a technical case study in best practices. E-Prints in Library and Information Science. 2004. http://eprints.rclis.org/archive/00000999/

[Universia] Biblioteca Virtual de Objetos de Aprendizaje Universia. <http://biblioteca.universia.net>.