Herramientas para obtener, mapear y filtrar recursos académicos desde repositorios digitales

Autores

Lira, Ariel
PREBI-SEDICI, Universidad Nacional de La Plata y CESGI, Comisión de Investigaciones Científicas
Soloaga Ignacio
PREBI-SEDICI, Universidad Nacional de La Plata
Villarreal, Gonzalo Luján
PREBI-SEDICI, Universidad Nacional de La Plata y CESGI, Comisión de Investigaciones Científicas
De Giusti, Marisa Raquel
PREBI-SEDICI, Universidad Nacional de La Plata y CESGI, Comisión de Investigaciones Científicas

Resumen

Uno de los principales objetivos de los repositorios digitales es el brindar mecanismos de interoperabilidad, a fin de permitir la recuperación de sus registros por otros sistemas y de ofrecer a sus usuarios información y servicios a partir de registros obtenidos desde otros repositorios. El intercambio de registros entre repositorios digitales brinda un mecanismo para agilizar el poblamiento cruzado, mediante el cual un repositorio puede obtener recursos de otro repositorio para corroborar datos cruzados, completar información faltante o desactualizada, e incluso para incorporar nuevos registros a su acervo propio. Esto resulta de particular interés para repositorios institucionales con investigadores de múltiple dependencia, que quizás realizan el autoarchivo de su producción en el repositorio de una de las instituciones en las que realizan sus actividades, o en muchos casos en repositorios temáticos utilizados tradicionalmente en cada área, como ser arXiv, REPEC o PubMed Central.

Existen múltiples formas de obtener documentos en masa de distintos repositorios, y una de ellas es mediante el protocolo OAI-PMH, que permite la transmisión de registros de metadatos bajo el esquema Dublin Core. Haciendo uso de las interfaces que expone cada repositorio, se puede realizar una cosecha de los metadatos deseados. Si bien la obtención de los documentos a importar a un repositorio es crucial para el proceso de importación en masa, existen una serie de etapas subsiguientes a ésta que son de suma importancia y que a su vez presentan ciertos obstáculos que deben ser solucionados para garantizar la integridad y calidad de los datos. Estas etapas pueden agruparse, además de la obtención de los documentos, en transformación de los metadatos y detección de registros duplicados.

Se describe el proceso general de obtención de documentos e importación de los mismos a un repositorio, así como también las herramientas desarrolladas para dar soporte a las distintas etapas del proceso.

Abstract: One of the main objectives of digital repositories is to provide interoperability mechanisms in order to allow the recovery of their records by other systems and to offer their users information and services from records obtained from other repositories. The exchange of records between digital repositories provides a mechanism to streamline cross-population, whereby a repository can obtain resources from another repository to corroborate cross-data, complete missing or outdated information, and even to incorporate new records into its own collection. This is of particular interest for institutional repositories with multi-dependency researchers, who may self-archive their production in the repository of one of the institutions in which they carry out their activities, or in many cases in thematic repositories traditionally used in each area, such as arXiv, REPEC or PubMed Central.

There are multiple ways to obtain mass documents from different repositories, and one of them is through the OAI-PMH protocol, which allows the transmission of record metadata under the Dublin Core scheme. Using the interfaces that each repository exposes, you can harvest the desired metadata. While obtaining the documents to be imported into a repository is crucial for the mass import process, there are a series of steps that are consequential to it, which are of the utmost importance and which in turn present certain obstacles that must be solved to guarantee the data integrity and quality. These stages can be grouped, in addition to obtaining the documents, in transformation of the metadata and detection of duplicate records.

The general process of obtaining documents and importing them into a repository is described below, as well as the tools developed to support the different stages of the process.

Palabras clave

Repositorios digitales; metadatos; interoperabilidad;

Introducción

La incorporación en masa de nuevos documentos a un repositorio digital puede ser un proceso complejo, potenciado en aquellos casos en que la cantidad de procesamientos a realizar sobre los datos a ser incorporados es considerable. Las etapas que componen al proceso varían según la necesidad de cada repositorio, y puede resultar en la ejecución de múltiples tareas con distintos propósitos y objetivos (De Giusti, Marisa Raquel; 2011). Debido a esto, es importante la definición de un flujo de procesos que sea tan automatizado como sea posible, con tareas controladas y resultados definidos, para asegurar

la calidad de los metadatos a importar a lo largo de todo el proceso, desde la obtención de los mismos hasta la importación de los registros al repositorio.

Existen distintos métodos utilizados en cada etapa de obtención, transformación y filtrado de los datos, así como también herramientas para dar soporte a las mismas.

Cosecha de documentos a través de OAI-PMH

El primer paso en el flujo de trabajo es recuperar documentos de diversos repositorios que contienen publicaciones institucionales haciendo uso de las interfaces OAI-PMH (Open Archives Initiative, 2015) que los mismos exponen. Este protocolo permite la cosecha de colecciones de documentos definidas dentro del mismo repositorio, y muchas veces éstas contienen cantidades considerables de documentos que dificulta la exportación, y por eso el protocolo resuelve el problema entregando las colecciones solicitadas de partes.

Una cosecha puede constar de 100 o más partes dependiendo el tamaño de la colección, y realizar la misma manualmente resulta en un proceso muy lento y repetitivo. Para solucionar este problema y aportar nuevas funcionalidades surge el desarrollo del Harvester.

Esta herramienta es utilizada para realizar cosechas de documentos sobre distintos repositorios a través del protocolo OAI-PMH, manejando de manera automática la descarga total de los documentos, y pudiendo programar cosechas para que sean ejecutadas regularmente.

Si bien la versión actual del Harvester soporta únicamente cosechas mediante el protocolo OAI-PMH, puede extenderse en un futuro para soportar la descarga de documentos mediante otros protocolos, como por ejemplo ResourceSync (ResourceSync Framework Specification) o interfaces específicas de proveedores de servicios como por ejemplo Scopus (Elsevier Scopus APIs) o Unpaywall (REST API). La herramienta actualmente se compone de tres módulos principales, detallados a continuación.

(1) Alta y planificación de cosechas.

Módulo encargado del alta de repositorios y definición y scheduling de cosechas. Es el encargado de comunicar al cosechador indicado (OAI-PMH) la ejecución de una cosecha en particular.

(2) Recuperadores / cosechadores

Módulo encargado de realizar una cosecha sobre un repositorio mediante el protocolo indicado. Actualmente ya existe un Cliente OAI separado como servicio, el cual puede realizar cosechas, manejar los problemas que puedan surgir en este proceso, y volcar los datos a disco. Además se puede consultar por el estado de una cosecha en particular en cualquier momento. En un futuro, podrían existir distintos cosechadores para diferentes protocolos, por ej. ResourceSync u OAI-ORE (ORE Specifications and User Guides, 2014).

(3) Almacenadores / Indexadores

Módulo encargado de indexar o guardar en una base de datos el resultado obtenido de una cosecha. En principio trabajamos con Solr, pero podría extenderse a bases de datos como Elasticsearch, eXist, entre otras. Para el servicio encargado de indexar a Solr en particular, existe una librería SolrJ que facilita la comunicación para aplicaciones Java.

Como resultado de una cosecha OAI-PMH se obtiene un directorio con múltiples archivos XML conteniendo los metadatos de cada documento perteneciente a la colección objetivo, bajo el esquema de metadatos Dublin Core.

Caso de uso - Cosecha OAI-PMH

El repositorio arXiv guarda pre-publicaciones de archivos científicos en el campo de las matemáticas, física, ciencias de la computación y biología cuantitativa y expone mediante su interfaz OAI-PMH los metadatos de un millón seiscientos mil documentos. Debido al gran volumen de registros y a la calidad de los metadatos, se vuelve un repositorio de gran interés para ser cosechado.

Se realizó una cosecha de todos los documentos expuestos por arXiv y dio como resultado un directorio con más de 2 gigabytes de archivos XML. Debido a la cantidad de solicitudes que se deben realizar para descargar todo el conjunto de documentos, arXiv impone una limitación de solicitud cada cierto tiempo, indicando un lapso de espera para realizar la siguiente petición.

La herramienta Harvester permitió programar una cosecha de todo el repositorio manejando automáticamente todas las respuestas del repositorio, dejando pasar el tiempo indicado por el mismo, entre solicitud y solicitud, y realizando la cosecha total en el orden de 10 horas (dependiendo la velocidad de la red utilizada). Esta prueba sirvió para corroborar la potencia y flexibilidad del Harvester, permitiendo cosechar repositorios de gran tamaño sin inconvenientes.

Mapeo y transformación de metadatos

Cada repositorio decide bajo qué esquema de metadatos guardar sus documentos, y si bien existen algunos esquemas más adoptados que otros (Dublin Core, MODS, etc.), cada repositorio decide qué elementos de dichos esquemas utilizar, cómo combinar metadatos entre distintos equemas y cuáles son los metadatos que se priorizan sobre otros. Es por ello que, más allá del esquema utilizado por la fuente desde donde se obtuvieron los registros, por lo general se requiere aplicar uno o más procesos de transformación o mapeo entre esquemas, tanto para facilitar su procesamiento como también para permitir su importación en el repositorio destino.

Para facilitar esta transformación, se desarrolló una herramienta que permite realizar mapeos entre esquemas de metadatos, a partir de archivos de

configuración personalizables, y que además permite definir para cada metadato una serie de filtros predefinidos para ser aplicados sobre los mismos, como por ejemplo eliminación de caracteres especiales, espacios en blanco, transformación a mayúsculas y/o minúsculas, etc.

Muchas veces un repositorio está interesado en incorporar de cada documento obtenido sólo un subconjunto de sus metadatos, por ejemplo en aquellos casos en los que se busca el enriquecimiento de metadatos, debiendo filtrar y tal vez cruzar a los mismos. También existen casos en los que los metadatos ofrecidos por las interfaces OAI-PMH de algunos repositorios no son suficientes para el estándar de importación al repositorio y los mismos deben completarse recuperando información de diversas fuentes externas al mismo, haciendo uso de identificadores persistentes como ser DOI, Scopus ID, etc.

La herramienta de mapeo de metadatos permite trabajar en conjunto con la herramienta de detección de documentos duplicados detallada a continuación, haciendo el pasaje al esquema genérico que entiende la misma.

Detección de documentos duplicados

Uno de los riesgos más grandes que se corre al importar grandes cantidades de documentos de forma masiva a un repositorio, es incluir en la importación documentos ya presentes en el repositorio. Cuando las publicaciones son subidas una a una, la detección del mismo documento dentro de un repositorio es bastante sencilla y puede realizarse manualmente. Esto no sucede cuando las cantidades de registros a subir al repositorio son muy grandes, no sería viable buscar uno a uno los documentos de forma manual.

Para solucionar esto, se desarrolló una herramienta que detecta documentos duplicados a partir de dos listados de recursos. Existen múltiples condiciones en los datos que dificultan de gran manera esta tarea, como se detalla a continuación.

La dificultad principal presente en la detección de recursos iguales en el dominio de los repositorios académicos y científicos es la desambiguación de valores. La estructura bajo la cual se cargan registros en distintos repositorios depende de los tesauros adoptados, el esquema de metadatos, y en gran parte, del factor humano. De esta manera y a modo de ejemplo, el artículo titulado ‘Políticas territoriales y construcción del paisaje cultural’ y subtitulado ‘Caso Región Gran La Plata’ puede estar cargado de esta manera en un repositorio; y puede figurar como un artículo titulado ‘Politicas territoriales y construccion del paisaje cultural - Caso region gran La Plata’ en otro. En este último, el registro no distingue entre título y subtítulo, hay palabras que no llevan las tildes correspondientes y existen diferencias en la presencia de mayúsculas y minúsculas.

Asimismo, detectar que dos autores son el mismo es muy dificultoso, principalmente por la forma en que los mismos son cargados en un registro. Por ejemplo, la autora ‘García, María Ana’ puede aparecer como ‘García, María

A.', 'García, M. Ana' ó 'García, M. A.', en distintos repositorios. Aunque con distintas funciones de aproximación entre strings, estos casos pueden detectarse, incluso aún existe un problema que excede a la similitud entre valores, y es que pueden existir dos autoras llamadas 'García, María Ana', que sin presencia de un identificador único no es posible diferenciar entre ambas. Este problema es atacado con el identificador ORCID, pero en este caso no presenta una solución debido a que el porcentaje de autores cargados con ORCID oscila entre porcentajes sumamente bajos y/o nulos.

Para un humano, la detección de patrones y similitudes entre distintos valores no es un gran problema, pero sí lo es para las computadoras. Teniendo esto presente, se pensó un modelo de aplicación dónde, mediante el uso de un sistema de reglas, con pesos relativos definidos, se establece un porcentaje de similitud o coincidencia entre recursos de dos repositorios. Cada regla del sistema evalúa aspectos específicos de un documento, analizando uno o varios metadatos en conjunto, y existen distintas reglas dependiendo los tipos de documentos que vayan a ser comparados.

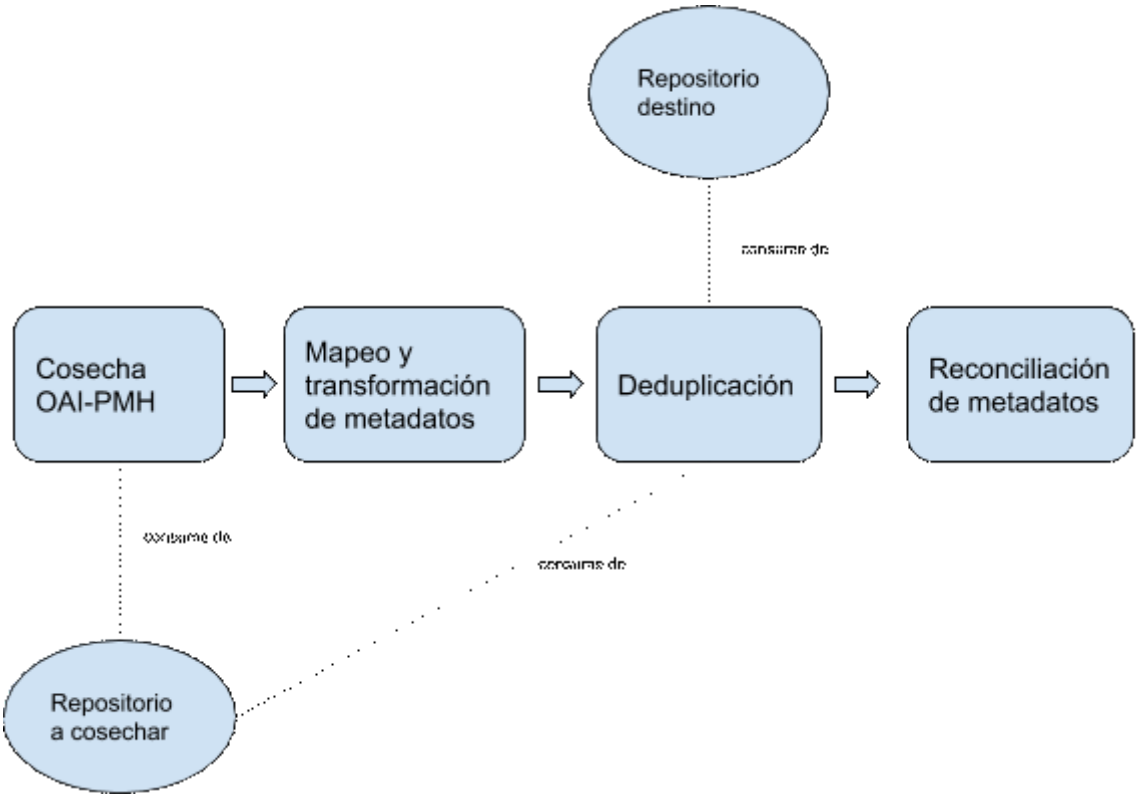
En base a la información generada por la herramienta de deduplicación pueden seleccionarse, utilizando los porcentajes de confianza de cada detección y definiendo umbrales de aceptación y rechazo, los documentos a ser importados a un repositorio.

Reconciliación de metadatos y mapeo al esquema esperado

La herramienta de deduplicación realiza un número de comparaciones sobre algunos metadatos de cada registro en un formato genérico, lo que implica que una vez obtenido el resultado de la deduplicación, y seleccionados los documentos a importar, se debe recuperar el total de metadatos obtenido mediante OAI-PMH. Para esto, se deben cruzar ambos conjuntos de datos, aquel que dio como resultado el deduplicador, y aquel que contiene el conjunto total de metadatos.

Como se comentó en la segunda etapa, los repositorios definen el esquema bajo el cual se guardan los metadatos de cada documento a importar, y muchas veces el formato Dublin Core que ofrece OAI-PMH no es suficiente. Es entonces cuando la herramienta de mapeo y transformación vuelve a ser de gran utilidad. Haciendo uso de la misma se define un nuevo archivo de configuración indicando la correspondencia de cada metadato dublin core con cada metadato esperado por el repositorio.

Gráfico ilustrativo del proceso



Trabajo futuro

Queda pendiente la extensión de la funcionalidad de cada herramienta y la definición de nuevos servicios para proveer mayor flexibilidad a la hora de filtrar, transformar y enriquecer los datos obtenidos.

Actualmente la versión del cosechador OAI-PMH funciona como una aplicación monolítica que dificulta la escalabilidad y adaptabilidad de la misma. Para solucionar este problema y para permitir la cosecha mediante otros protocolos más allá de OAI-PMH, se propone un modelo de arquitectura orientada a servicios, interconectados entre sí. Las razones que conllevan a esta decisión son principalmente la necesidad de:

- Mejorar la escalabilidad del sistema.
- Flexibilizar la solución de errores acotando el dominio del problema.
- Soporte de múltiples tecnologías (en caso de ser necesario).
- Reuso y combinación de servicios particulares en distintos contextos.

Otros trabajos pendientes a realizar es el agregar nuevos filtros a la herramienta de mapeo de metadatos para proveer mayor flexibilidad a la hora de establecer el procesamiento sobre los datos, y explorar el campo de machine learning aplicado a la detección de documentos duplicados, evaluar opciones y comparar resultados con el modelo utilizado actualmente.

Bibliografía

- The Open Archives Initiative Protocol for Metadata Harvesting (2015). Recuperado de http://www.openarchives.org/OAI/openarchivesprotocol.html .
- ResourceSync Framework Specification (ANSI/NISO Z39.99-2017), (2 February 2017). Recuperado de http://www.openarchives.org/rs/1.1/resourcesync
- Elsevier Scopus APIs, Elsevier Developers. Recuperado de https://dev.elsevier.com/sc_apis.html
- Unpaywall, REST API. Recuperado de https://unpaywall.org/products/api
- De Giusti, Marisa Raquel; Lira, Ariel Jorge y Oviedo, Néstor Fabián (Mayo, 2011). Extract, transform and load architecture for metadata collection. VI Simposio Internacional de Bibliotecas Digitales. Pontificia Universidad Católica de Río Grande Do Sul, Porto Alegre.
- ORE Specifications and User Guides - Table of Contents (2014). Recuperado de http://www.openarchives.org/ore/1.0/toc.