Título: Explotación de repositorios OAI a través de la armonización de vocabularios controlados

Autor: Diego Ferreyra
Correo electrónico: diego@r020.com.ar
Institución: Universidad de Buenos Aires, Facultad de Filosofía y Letras, Departamento de Bibliotecología y Ciencia de la Información

Abstract:

Este trabajo presenta una experiencia orientada al análisis de la viabilidad técnica y metodológica del marco de armonización de vocabularios controlados utilizado para el desarrollo de servicios en línea basados en la explotación cooperativa de contenidos. Para el desarrollo de la misma se han articulado diferentes tecnologías, unas orientadas a la portabilidad de metadatos y otras a la de vocabularios controlados. Para realizar el mencionado análisis se desarrolló e implementó un sitio web denominado Paperlandia en el que se agregaron diversas fuentes de metadatos bajo una misma infraestructura de servicios.

Presentación

Este trabajo presenta una experiencia orientada al análisis de la viabilidad técnica y metodológica del marco de armonización de vocabularios controlados utilizado para el desarrollo de servicios en línea basados en la explotación cooperativa de contenidos. Para el desarrollo de la misma se han articulado diferentes tecnologías, unas orientadas a la portabilidad de metadatos y otras a la de vocabularios controlados. Para realizar el mencionado análisis se desarrolló e implementó un sitio web denominado Paperlandia en el que se agregaron diversas fuentes de metadatos bajo una misma infraestructura de servicios.

En primer lugar se adoptaron fuentes OAI-PMH como vía de acceso sistemática a fuentes de metadatos, luego ser identificaron los vocabularios utilizados en los repositorios, como ser el OAI-PMH, luego se analizaron los vocabularios utilizados por cada uno de los repositorios y se desarrolló un vocabulario controlado específico para el repositorio de agregación. Se estableció un mapeo entre los términos de los vocabularios empleados en cada repositorio y el vocabulario del repositorio de agregación. Finalmente, se implementó la solución tecnológica para la exposición y consulta de los metadatos en línea.

La explotación cooperativa de contenidos basada en repositorios OAI-PMH

Desde hace ya más de una década y con el objetivo de desarrollar modelos y metodologías para la difusión y explotación cooperativa de recursos documentales, surgió el protocolo para cosecha y exposición de metadatos OAI-PMH. Puntualmente, esta iniciativa tuvo su origen en 1999 en una reunión en Santa Fe (Nuevo México, USA) orientada a aumentar su impacto y visibilidad de los eprints. Esta iniciativa se desarrolló y formalizó una arquitectura compuesta por un oferente de metadatos (servidor OAI-PMH)y un cliente que, utilizando una conjunto simple y acotado de comandos, puede acceder masivamente a metadatos estructurados a través de XML y utilizando Dublin Core no cualificado.

Este esquema, basado en el alto grado de interoperabilidad que permite la utilización del lenguaje de marcado (en este caso XML) para la estructuración de datos, la identificación de servicios y recursos a través de URIs y la semántica prácticamente trivial del esquema Dublin Core no cualificado, ha permitido la consolidación y su adopción como estrategia abierta para la exposición y difusión de recursos y modelos federados de explotación cooperativa, ya sea en comunidades especializadas, iniciativas gubernamentales, instituciones y como una capacidad estándar de un gran

número de plataformas de gestión de recursos basadas en metadatos.

El modelo de explotación cooperativa de contenidos en base a las capacidades del OAI-PMH contempla, tal como se ha indicado anteriormente, la existencia de servidores y clientes; la articulación de estos dos roles ha generado un escenario en el que es posible identificar los siguientes actores en términos de tipos de repositorios:

- Repositorios agregadores: servicios basados en el cosechado de otros repositorios.
- Repositorios disciplinares: repositorios especializados según actividades, tópicos, disciplinas, etc.
- Repositorios institucionales: repositorios dedicados a garantizar la disponibilidad de recursos en base a producción o políticas de contenidos específicas de una institución.
- Repositorios gubernamentales: repositorios dedicados a garantizar la disponibilidad de recursos en base a lineamientos gubernamentales.

La presente experiencia está centrada en el caso de los repositorios agregadores

Definición del problema

El marco OAI-PMH ha permitido la articulación de distintos contextos, comunidades y contenidos a través de un protocolo sencillo y un esquema de representación de metadatos simple. Esto se ha logrado básicamente a través de normalización provista por el protocolo mismo y sus esquemas de estructuración y representación de metadatos asociados. Sin embargo, poco se ha avanzado en los procesos de normalización o armonización basados en los contenidos de los metadatos.

Esto conlleva una serie de problemas que se acentúan al momento de consolidar diversas fuentes de metadatos en un único repositorio y bajo una misma infraestructura de servicios como es el caso de los repositorios agregadores. A continuación se mencionan algunos de los inconvenientes relevados:

- Variabilidad en la utilización del esquema Dublin Core: considérese que en un esquema como el Dublin Core, “todo es optativo, todo es extensible, todo es modificable”[1], esto comporta un alto grado de variabilidad, aun en contextos de alta normalización local, ya que se trata en realidad de la integración de más de un repositorio provenientes de diferentes comunidades y lenguas que aún disponiendo de políticas de normalización para el llenado de metadatos, no tienen por qué ser homogéneas.

- Reproducción de problemas de normalización de origen: cada repositorio implica asimismo diferentes grados de cumplimiento de las mencionadas políticas de llenado de metadatos. Es decir, más allá de que pudieran adoptar tal o cual pauta normativa o tal o cual vocabulario controlado, cada repositorio cumple y controla estas pautas de diferente forma. Los metadatos pueden ser llenados con diversos criterios, políticas y por distintas personas en diferentes momentos en un mismo repositorio.

- Dificultad para implementar modelos locales: los metadatos obtenidos en los distintos repositorios rara vez responden a los modelos de dominio locales o a las prácticas lingüísticas del repositorio agregador. En este sentido, el repositorio agregador representa un modelo de mediación emergente a partir de la agregación de los diferentes modelos de mediación implementados por cada uno de los repositorios agregados.

A partir de las problemáticas antes mencionadas, surgió la posibilidad de explorar una vía de trabajo

1 :Lagoze, C., Lynch, C. A., and Daniel, R. (1996). The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata. Cornell Computer Science Technical Report TR96-1593.

a través de la utilización de las herramientas provistas por el marco de armonización de vocabularios controlados aplicados, en este caso, para el mapeo de los diferentes vocabularios y estándares de valores utilizados por los repositorios. Se consideró que si acaso es posible utilizar crosswalk para resolver la portabilidad entre esquemas de metadatos, la armonización de vocabularios podría considerarse y utilizarse un modelo de solución para lograr portabilidad de contenidos entre diferentes comunidades de interpretación.

Para el desarrollo de la experiencia se resolvió seguir los siguientes pasos:

- Selección y determinación de repositorios a cosechar.
- Consolidación del vocabulario local del agregador
- Establecimiento de los mapeos terminológicos entre los vocabularios de los repositorios y el vocabulario del repositorio de agregación.

A continuación se describen brevemente los pasos realizados.

Selección y determinación de repositorios a cosechar

Inicialmente, se consideró natural en el contexto de un proyecto orientado a la interoperabilidad entre vocabularios controlados utilizar el campo Subject del esquema Dublin Core como fuente para el análisis y mapeo de términos. En base a esta consideración y con el interés de evaluar el modelo de solución en escenarios variados, se intentó identificar repositorios que utilizaran diferentes tipos de vocabularios entre sí y diferentes idiomas.

Se seleccionaron tres repositorios disciplinares en el campo de la bibliotecología y las ciencias de la información:

– Digital Library of Information Science and Technology (DLIST): inicia sus actividades en el año 2002 y se propone establecer un archivo abierto dedicado a las ciencias de la información. Es mantenido por la School of Information Resources and Library Science and Learning Technologies Center de la University of Arizona. Utiliza para el llenado del campo Subject del esquema Dublin Core un vocabulario en inglés que se correspondería con lo que en el contexto de la recomendación ANSI/NISO Z39.19-2005 se considera una **lista de términos**, ya que no dispone de relaciones de ningún tipo entre los términos (ni jerárquicas, ni asociativas ni de equivalencia).

– E-prints in Librarianship, Information Science (E-lis): Surge en base a dos iniciativas RePEc y DoIS en el año 2003. Dispone de una red de editores de 45 países de todo el mundo que participan voluntariamente. Utiliza para el llenado del campo Subject del esquema Dublin Core un vocabulario en inglés denominado JITA Classification System of Library and Information Science. Dicho vocabulario se correspondería con lo que en el contexto de la recomendación ANSI/NISO Z39.19-2005 se considera una **taxonomía**, ya que dispone de relaciones jerárquicas entre términos, pero no dispone de relaciones asociativas o de equivalencia.

– Temaria: Es sostenido por el Grup de recerca Organització i Recuperació de Continguts Digitals de la Facultat de Biblioteconomia i Documentació (Universitat de Barcelona) desde el año 2005 y se dedica a describir y ofrecer acceso a artículos de revistas científicas españolas de Información y Documentación. Utiliza para el llenado del campo Subject del esquema Dublin Core el Tesauro de Biblioteconomía y Documentación elaborado por el CINDOC en español. Dicho vocabulario se correspondería con lo que en el contexto de la recomendación ANSI/NISO Z39.19-2005 se considera un **tesauro**, ya que dispone de relaciones jerárquicas entre términos, relaciones asociativas y de equivalencia. El vocabulario cuenta con equivalencias al catalán, inglés y francés.

Esta selección de repositorios permitió evaluar el modelo de solución utilizando fuentes de datos de diferentes idiomas y que utilizan diferentes tipos de vocabularios para el llenado del campo Subject (DC:subject). Tal como se observa, se trata de lograr una infraestructura de servicios que consolide un acceso a los recursos a partir de un vocabulario propio en base a relaciones de equivalencia con tres tipos de vocabularios distintos: una lista de términos en inglés, una taxonomía en inglés y un tesauro en español.

Para el cosechado de los repositorios y la consolidación del repositorio local de agregación, se utilizó la herramienta de cosechado de metadatos PKP Open Archives Harvester desarrollada por el Public Knowledge Project. El mencionado repositorio local de agregación, y el proyecto en general, fue denominado Paperlandia para su publicación en la WWW.

El vocabulario local del agregador

Para la consolidación del vocabulario local del repositorio agregador se optó por desarrollar un vocabulario basado en una traducción al español del JITA Classification System of Library and Information Science. Al vocabulario resultante se le agregaron relaciones asociativas y equivalencias.

Para la elaboración y gestión del vocabulario se utilizó la herramienta de gestión de vocabularios controlados TemaTres.

Mapeo entre los vocabularios de los repositorios y el vocabulario del repositorio de agregación Paperlandia

La formalización de las relaciones de mapeo entre los vocabularios comprendió los siguientes niveles de armonización:

- Armonización terminológica: orientada a eliminar inconsistencias terminológicas entre vocabularios cooperantes como ser variaciones en la puntuación, reglas sintácticas, nombres propios y regionales. Por ejemplo:
  - Diseño, desarrollo, implementación y mantenimiento
    - IK. Design, development, implementation and maintenance
    - IK. Design, development, implementation and maintenance.

- Referencias cruzadas: este tipo de armonización se basa en la formalización de relaciones de equivalencia entre vocabularios cooperantes. Se utilizaron las siguientes formas de equivalencia semántica entre conceptos:
  1. Equivalencia exacta: cuando dos conceptos tienen el mismo alcance y significado.
  2. Equivalencia parcial: cuando el alcance de un concepto en un vocabulario se corresponde completamente con el de un concepto de otro vocabulario, pero esto no ocurre a la inversa.
  3. Equivalencia inexacta: cuando el alcance de dos conceptos se solapan entre sí, sin llegar corresponderse por completo.

  No se utilizó la relación de no-equivalencia, referida al caso de no existencia en el vocabulario de destino de un concepto que contenga el alcance del concepto del vocabulario fuente, ya que en términos funcionales, ese tipo de relación resultó expresada por la nulidad

o no existencia de relación de relación entre dos conceptos. De esta manera, para el desarrollo de esta experiencia se consideró que las relaciones de equivalencia entre conceptos de diferentes vocabularios, debían ser expresados siempre de manera positiva, considerando la no existencia de relación como una no-equivalencia.

- Términos comunes: este tipo de armonización se basa en la utilización de términos idénticos para conceptos idénticos. Fue posible utilizar este tipo sólo con el *Tesauro de Biblioteconomía y Documentación* elaborado por el CINDOC (utilizado por Temaria) ya que es el único en español.

- Asociación: se trata de la articulación consistente entre un vocabulario especializado en el contexto de la estructura jerárquica de uno vocabulario más general. Este tipo de armonización fue el predominante en el caso del *Tesauro de Biblioteconomía y Documentación* (Temaria), ya que ofrecía un nivel de especialización y granularidad muy superior al del vocabulario final a ser utilizado en Paperlandia, contando el primero con más de 1000 términos y el segundo con cerca de 150 términos. Por ejemplo:
    - Catalogación, control bibliográfico
    - Catalogación
    - Descripción bibliográfica
    - Directrices
    - Normas
    - Normas internacionales

La armonización resultó ser un proceso apropiado para la resolución de problemáticas generales de la normalización y control de autoridades, como ser el uso concurrente de diferentes políticas de metadatos en un mismo repositorio, etc.

Para el proceso de armonización fueron utilizadas las funcionalidades orientadas para la declaración de relaciones entre vocabularios de la herramienta de gestión de vocabularios controlados TemaTres. A continuación se presenta un gráfico ilustrativo del esquema de mapeo.

Además de las herramientas antes mencionadas (PKP Open Archives Harvester y TemaTres), fueron desarrolladas dos herramientas de consultas y presentación de contenidos, a saber:

1. Herramienta para la detección de términos: orientada a detectar términos utilizados en un repositorio que no dispongan de un mapeo en el vocabulario del repositorio agregador (Paperlandia).

2. Herramienta para presentación de contenidos: orientada a desarrollar las opciones de navegación, la presentación de metadatos y el esquema de búsqueda temático que permite acceder a los metadatos agregados en base al vocabulario del repositorio agregador (Paperlandia) y no según los vocabularios o términos de los repositorios de origen.

A partir de las primeras pruebas realizadas, se evaluó la factibilidad y conveniencia de extender la misma metodología a otros campos previstos en el esquema Dublin Core e informados por los servicios OAI-PMH. Puntualmente fueron tratados con la misma metodología los campos de *idioma* (language) que indica el o los idiomas utilizado en el recurso descripto y *tipo de recurso* (type) que da cuenta del tipo, clase o género a la que pertenece el recurso descripto. Estos campos o categorías de descripción fueron seleccionados considerando que su llenado en general se basa en listas de

términos relativamente estables sin llegar a ofrecer una normalización homogénea en todos los casos.

El procesamiento de estos dos campos permitió el desarrollo e implementación de un esquema de navegación facetado utilizando como navegación principal el vocabulario temático y los vocabularios de idioma y tipo de recurso como opciones de filtrado.

Conclusiones

A partir de la experiencia y la implementación concreta del repositorio de agregación bajo el nombre de Paperlandia, fue posible desarrollar algunos aprendizajes metodológicos y evaluar el alcance y factibilidad operativa de los procesos de mapeo terminológico utilizados para la resolución de infraestructuras de servicios basadas en la explotación cooperativa de contenidos. A continuación se describen brevemente los resultados alcanzados:

- **Autonomía local e interoperabilidad global:** el mapeo terminológico permitió construir una solución para la normalización de fuentes autónomas de datos en el contexto de redes federadas de cooperación, manteniendo condiciones de autonomía de cada uno de los nodos (repositorios) y el nodo agregador (Paperlandia) a partir de las condiciones de interoperabilidad garantizadas por el marco OAI-PMH.

- **Metadatos y normalización**: los desarrollos realizados para comparar los contenidos de un campo específico en el contexto de un esquema de metadatos (en este caso DublinCore) y los términos disponibles en un vocabulario determinado resultaron una herramienta útil y fundamental a la hora de detectar de variaciones, frecuencias, regularidades y coincidencias en el contexto de cada repositorio en particular y en el contexto del repositorio agregador. Eventualmente, esta misma metodología podría utilizarse como vía de solución para la detección de problemáticas de normalización y control de autoridades.

- **Interoperabilidad de contenidos**: fue posible explorar una concepción amplia y compleja de la noción de interoperabilidad más allá de la portabilidad de datos y estructuras. Una noción capaz contemplar los modelos de mediación de cada comunidad de interpretación que ofrece una marco metodológico para la apropiación de los contenidos según concepciones locales.

- **Muchos metatados, poca normalización**: a partir de la implementación real de Paperlandia, fue posible evaluar la viabilidad real del marco de armonización de vocabularios como esquema de solución para repositorios caracterizados por la alta variabilidad y el bajo nivel de normalización de sus metadatos en escenarios de cooperación abiertos pero delimitados.

- **Universalidad vs. Apropiación**: en el marco de armonización de vocabularios en particular y el modelo de vocabularios multilingües en general es posible identificar una impronta orientada al logro de una comunicación transparente entre lenguas. En el caso de la presente experiencia, el modelo de trabajo estuvo más bien orientado al desarrollo y experimentación de una metodología de apropiación basada en la utilización de los vocabularios controlados como modelos de mediación entre lenguas pero también entre comunidades de

interpretación. En este sentido, podría decirse que la experiencia opera sobre la base de la formalización y visibilidad del modelo de mediación y no sobre su transparencia o invisibilidad.

Considerando un escenario de alta disponibilidad y exposición de metadatos a través de servicios de cosechado u otros medios, y en virtud de algunas de las conclusiones alcanzadas en esta experiencia, es posible identificar un rol diferencial para los vocabularios controlados: el de operar como herramientas de representación y apropiación de recursos.

Con el interés de revisar estos aprendizajes y extender la base de casos, se prevé llevar adelante una experiencia similar pero orientada a repositorios de objetos de aprendizaje, un campo que comparte algunas de las problemáticas tratadas en el presente trabajo.

Bibliografia

Bowker, G. C., & Star, S. L. (1999). Sorting things out Classification and its consequences. Inside technology. Cambridge, Mass: MIT Press.

European Committee for standardization (2005). CWA 15453:2005. Harmonisation of vocabularies for eLearning. Recuperado octubre 20, 2009, de : http://fire.eun.org/cwa15453-00-2005-Nov.pdf. Fecha de consulta: 10/10/2008

Heath, B. P., McArthur, D. J., McClelland, M. K., and Vetter, R. J. (2005). Metadata lessons from the iLumina digital library. Commun. ACM 48, 7 (Jul. 2005), 68-74.
International Federation of Library Associations and Institutions. (2009). Guidelines for multilingual thesauri: Working group on guidelines for multilingual thesauri IFLA classification and indexing section. The Hague: International Federation of Library Associations and Institutions. Recuperado octubre 20, 2009, de : http://www.ifap.ru/library/book411.pdf

Lagoze, C., Lynch, C. A., and Daniel, R. (1996). The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata. Cornell Computer Science Technical Report TR96-1593.

National Information Standards Organization (U.S.). (2005). Guidelines for the construction, format, and management of monolingual controlled vocabulary. National information standards series. Bethesda, Md: NISO Press. Recuperado octubre 20, 2009, de : http://www.niso.org/kst/reports/standards/kfile_download?&pt=RkGKiXzW643YeUaYUqZ1BFwDhIG4-24RJbcZBWg8uE4vWdpZsJDs4RjLz0t90_d5_ymGsj_IKVa86hjP37r_hONsJghRDv2N-zj4TZCh8Dp01rZbmK3O-8vcVjh4hezP

Rajapakse, R.K, Mushens, B, Johnson, C. (2007). The use of keyphrases for selectin metadata form taxonomies. Creating Collaborative Advantage Through Knowledge and Innovation (Series on Innovation and Knowledge Management) (Series on Innovation and Knowledge Management) (pp. 329-345). Sassari: World Scientific Publishing Company.

Shreeves, Sarah L.; Knutson, Ellen M.; Stvilia, Besiki; Palmer, Carole L.; Twidale, Michael B.; Cole, Timothy W. (2005) Is 'Quality' Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections. En H.A. Thompson (Ed.) Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, Abril 7-10 2005, Minneapolis, MN. Chicago, IL: Association of College and Research Libraries. p. 223-237. Recuperado octubre 20, 2009, de : http://www.ideals.illinois.edu/handle/2142/145

Sitios y herramientas web:

1. Paperlandia: http://paperlandia.r020.com.ar
   - Vocabularios empleados: http://paperlandia.r020.com.ar/vocabularios/

2. PKP Open Archives Harvester : http://pkp.sfu.ca/?q=harvester

3. Tematres gestión de vocabularios controlados: http://tematres.r020.com.ar/