Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 439-440 recibido 20-05-2005; aceptado 01-06-2005

OAC-onto: Open Archive Cataloger, ontologías y metadatos

Inés Jacob, Joseba Abaitua, JosuKa Díaz, Fernando Quintana, Jon Fernández, Txus Sánchez, Garikoitz Echevarría
Grupo DELi
Universidad de Deusto
Ap. 1 - 48080 Bilbao
ines@eside.deusto.es

Josu Azpillaga
CodeSyntax
Azitaingo Industrialdea 3K
20600 Eibar (Gipuzkoa)
jazpillaga@codesyntax.com

Resumen: El proyecto OAC (Open Archives Cataloger) tiene como objetivo mejorar la gestión de metadatos documentales para grandes corpora de textos multilingües y permitir su almacenamiento en repositorios distribuidos. Con este fin, se ha realizado una implementación del protocolo OAI-PMH para el servidor web Zope, que incluye el repertorio completo de funciones: proveedor de datos, recolector de metadatos, pasarela estática y proveedor de servicios.
Palabras clave: OAI, interoperabilidad y formatos de metadatos, gestión de corpora multilingüe, diseminación de metadatos.

Abstract: OAC (Open Archives Cataloger) is a project aimed at improving metadata management for large corpora of multilingual texts, allowing document storage in distributed repositories. The main technical contribution is the implementation of the OAI-PMH protocol for the Zope web server, including the full repertory of functions: data provider, metadata harvester, static repository gateway, and service provider.
Keywords: OAI, interoperability and metadata formats, multilingual corpora management, metadata disemination.

1. Datos del proyecto

El proyecto *OAC-onto: Open Archives Cataloger, ontologías y metadatos* está siendo desarrollado por el grupo DELi de la Universidad de Deusto (Bilbao) y la empresa CodeSyntax (Eibar, Gipuzkoa). Su directora es Inés Jacob (DELi), comenzó en octubre de 2004 y su duración es de 15 meses. Está financiado por el Dpto. de Industria, Comercio y Turismo del Gobierno Vasco, dentro del Programa SAIOTEK (S-OD04UD04), y es continuación del proyecto *OAC: Open Archives Cataloger*, desarrollado entre octubre de 2003 y diciembre de 2004 por los participantes bajo el mismo programa (S-OD03UD09).

2. Objetivos

La motivación inicial del proyecto es mejorar la gestión de metadatos documentales para grandes corpora de textos multilingües, partiendo de la experiencia en sistemas como SARE-Bi (Díaz et al., 2003), y teniendo como objetivo posibilitar la distribución e intercambio de los metadatos y/o textos.

Se ha elegido el estándar de interoperabilidad OAI (*Open Archives Initiative*, http://www.openarchives.org/) de Lagoze y de Sompel (2001), dada su actual aceptación, como referente para la arquitectura del sistema, y el servidor web Zope (http://www.zope.org/) para su implementación.

OAI facilita que la diseminación de metadatos por parte de proveedores de contenidos sea aprovechada por otras aplicaciones recolectoras de dichos metadatos, como buscadores y catalogadores de información. No obstante, OAC posibilita asimismo la distribución de los propios contenidos, si ello resulta preciso para las aplicaciones que se definan, en consonancia con prácticas similares de OAI (de Sompel, Young, y Hickey, 2003), en particular la de Los Álamos National Laboratory (Jerez et al., 2004).

OAC implementa los módulos de diseminación de metadatos mediante la tecnología del servidor web Zope, a partir del producto *ZOpenArchives* creado por la compañía francesa Pentila (http://www.pentila.com/), al que se añaden varios componentes:

- pasarela estática,
- indización automática de la base de registros,
- funciones primitivas de consulta, y
- la infraestructura que soportará el desarrollo de la interfaz de usuario de la aplicación.

Un aspecto crucial, motivado por la heterogeneidad de fuentes y de recursos a los que se planea aplicar los resultados del proyecto, ha sido la modificación del recolector (*zOAIHarvester*) para que admitiera metadatos codificados con cualquier esquema XML, lo cual además de su interés propio, es la base para la distribución de contenidos con OAI. Se ha incluido la capacidad para almacenar dichos metadatos (*zOAIRecord*) tanto en el agregador como en el proveedor de servicios. El problema se ha resuelto mediante un superconjunto o unión semántica de metadatos (es decir, sin repetición de elementos con la misma información), denominado *lenguaje neutro de metadatos*, similar al concepto de *lingua franca* de Chan (2005).

En la actualidad, además de los 15 elementos de Dublin Core (DC, http://dublincore.org/) sin calificar (que aseguran la interoperabilidad básica), están incluidos DC calificado, BibTeX, TEI (http://www.tei-c.org/) y MARCXML (http://www.loc.gov/standards/marcxml/). Se han programado funciones de conversión entre el lenguaje neutro y cada uno de los demás (DC, BibTeX, MARCXML), en ambos sentidos. Un recolector usa la función de conversión en sentido directo para transformar los metadatos obtenidos (en MARCXML, por ejemplo) al lenguaje neutro. Un proveedor de servicios puede usar la conversión en sentido inverso (lenguaje neutro a MARCXML) para las aplicaciones, por ejemplo para que la interfaz de estas puedan adaptarse a un formato de metadatos concreto, si así se desea, para un usuario con experiencia en el mismo.

La utilización del servidor Zope como soporte del gestor documental se ha revelado muy adecuada gracias a su arquitectura de base de datos de objetos, que permite entre otras cosas la incorporación de sencillos mecanismos como la *diseminación selectiva* o la *diseminación por consulta*, que aportan flexibilidad al uso del protocolo OAI-PMH sin comprometer en ningún caso el cumplimiento del estándar.

En la actualidad, estamos experimentando el uso de ontologías (OWL para BibTeX y DC) con vistas a mejorar el aprovechamiento de los recursos y facilitar su descubrimiento. Todo el desarrollo se ha realizado bajo postulados de software libre.

3. Agradecimientos

Agradecemos especialmente a Gari Araolaza, Eneko Astigarraga y Luistxo Fernández (CodeSyntax) su apoyo constante al proyecto. También deseamos agradecer a Badihardugu, Gaztelupeko Hotsak, Gerediaga Elkartea, Ibinagabeitia Proiektua, Inguma, Lanbide Ekimena y Megadenda la colaboración prestada como agentes de contenidos en euskera, y en particular, la aportación de partes significativas de sus bases de datos documentales para las pruebas del proyecto.

Bibliografía

Chan, Lois Mai. 2005. Metadata Interoperability. A Study of Methodology. En *The 3rd China-US Library Conference*, Shanghai (China), marzo.

Díaz, JosuKa, Joseba Abaitua, Inés Jacob, Fernando Quintana, y Garikoitz Araolaza. 2003. Metadata for multilingual content management. En *Translating and the Computer 25. Conference Proceedings*, páginas 151–170, Londres, noviembre.

de Sompel, Herbert Van, Jeff Young, y Thom Hickey. 2003. Using the OAI-PMH ... Differently. *D-Lib Magazine*, 9(7/8).

Jerez, Henry, Xiaoming Liu, Patrick Hochstenbach, y Herbert Van de Sompel. 2004. The Multi-faceted Use of the OAI-PMH in the LANL Repository. En *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, páginas 11–20, Tuscon, AZ (EE.UU.), junio.

Lagoze, Carl y Herbert Van de Sompel. 2001. The Open Archives Initiative: Building a low-barrier interoperability framework. En *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, páginas 54–62, Roanoke, VA (EE.UU.), junio.