3ra Conferencia Internacional de Biblioteca Digital y Educación a Distancia

Expandiendo las Fronteras del Conocimiento

Mérida – Venezuela • 12 y 13 de noviembre 2009

PLATAFORMA DE RECOLECCIÓN EN FUENTES HETEROGÉNEAS DE LA WEB Y SU APLICACIÓN A LA ORGANIZACIÓN DE UN REPOSITORIO SEMÁNTICO EN SEDICI: PRELIMINARES

Marisa Raquel De Giusti, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y Proyecto de Enlace de Bibliotecas, marisa.degiusti@sedici.unlp.edu.ar

Ariel Sobrado, Proyecto de Enlace de Bibliotecas, asobrado@sedici.unlp.edu.ar

Agustín Vosou, Proyecto de Enlace de Bibliotecas, agustinvosou@sedici.unlp.edu.ar

Gonzalo Luján Villarreal, Consejo Nacional de Investigaciones Técnicas y Científicas y Proyecto de Enlace de Bibliotecas gonetil@sedici.unlp.edu.ar

.

3ra Conferencia Internacional de Biblioteca Digital y Educación a Distancia

Expandiendo las Fronteras del Conocimiento

Mérida – Venezuela • 12 y 13 de noviembre 2009

PLATAFORMA DE RECOLECCIÓN EN FUENTES HETEROGÉNEAS DE LA WEB Y SU APLICACIÓN A LA ORGANIZACIÓN DE UN REPOSITORIO SEMÁNTICO EN SEDICI: PRELIMINARES

Marisa Raquel De Giusti, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y Proyecto de Enlace de Bibliotecas, marisa.degiusti@sedici.unlp.edu.ar

Ariel Sobrado, Proyecto de Enlace de Bibliotecas, asobrado@sedici.unlp.edu.ar

Agustín Vosou, Proyecto de Enlace de Bibliotecas, agustinvosou@sedici.unlp.edu.ar

Gonzalo Luján Villarreal, Consejo Nacional de Investigaciones Técnicas y Científicas y Proyecto de Enlace de Bibliotecas gonetil@sedici.unlp.edu.ar

Resumen

Se presenta una plataforma de recolección destinada a relacionar y unificar información disponible en distintos lugares de la Web _que siguen diferentes convenciones_ para crear un repositorio temático que puedan navegar los usuarios. La plataforma será usada en el Servicio de Difusión de la Creación Intelectual (SeDiCI) y utiliza de manera combinada ontologías y tesauros para brindar información mejor clasificada.

Actualmente, la información está diseminada en recursos de la Web y los motores de búsqueda tradicionales le devuelven al usuario listas rankeadas sin proveer ninguna relación semántica entre documentos. Los usuarios pasan gran cantidad de tiempo para vincular unos documentos con otros y saber cuáles atacan el dominio completo del problema; recién al localizar las semejanzas y las diferencias entre fragmentos de información éstas se trasladan a su trabajo y sirven para la creación de nuevo conocimiento.

La plataforma propuesta separa los módulos de funcionamiento de los diferentes dominios de interés (temas) para permitir su utilización en distintas áreas de conocimiento. El desarrollo incluye dos agentes que recorren las URLs almacenadas en una base de datos (uno responsable de poblar una ontología y otro de obtener URLs relacionadas), un módulo capaz de reconocer las páginas marcadas, interpretar las etiquetas y proveer las reglas para extraer la información y guardarla en un fichero RDF; tras esta etapa se aplica una homogeneización y la información así transformada se clasifica en función de una ontología de dominio.

La plataforma vuelve más eficientes los procesos de extracción automática y búsqueda de información en fuentes heterogéneas que representan los mismos conceptos siguiendo distintas convenciones.

Palabras Clave: SeDiCI, repositorio semántico, ontologías y tesauros.

.

PLATFORM FOR COLLECTION FROM HETEROGENEOUS WEB SOURCES AND ITS APPLICATION TO A SEMANTIC REPOSITORY ORGANIZATION AT SEDICI: PRELIMINARIES

Marisa Raquel De Giusti, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y Proyecto de Enlace de Bibliotecas, marisa.degiusti@sedici.unlp.edu.ar
Ariel Sobrado, Proyecto de Enlace de Bibliotecas, asobrado@sedici.unlp.edu.ar
Agustín Vosou, Proyecto de Enlace de Bibliotecas, agustinvosou@sedici.unlp.edu.ar
Gonzalo Luján Villarreal, Consejo Nacional de Investigaciones Técnicas y Científicas y Proyecto de Enlace de Bibliotecas gonetil@sedici.unlp.edu.ar

Abstract

Presentation of a web collection platform designed to relate and unify information available on different standard web sources with a view to creating a user-browseable thematic repository. The platform will be used at the Servicio de Difusión de la Creación Intelectual (SeDiCI) [Intellectual Creation Diffusion Service] combined with ontologies and thesaurus to provide improved data sorting.

Data is currently spread on web resources and traditional search engines return ranked lists with no semantic relation among documents. Users have to spend a great deal of time relating documents and trying to figure out which ones fully address the issue domain. It is only after locating similarities and differences that information fragments are applied to the user's work, enabling knowledge creation.

The proposed platform sorts out the different theme domain functioning modules to allow their use in various knowledge areas. Development includes two agents that searches data base stored URLs, one is capable of identifying bookmarked pages, interpreting labels and providing rules for extracting information and storing it in a RDF data file; on the other hand, the other agent is in charge of getting related URLs from the given one. After this stage, homogenization is applied and transformed information is sorted out according to domain ontologies.

The platform allows for more efficient automatic extraction processes and information search among heterogeneous sources that represent the same concepts using different standards.

Keywords: SeDiCI, semantic repository, ontology and thesaurus.

.

Plataforma de recolección en fuentes heterogéneas de la WEB y su aplicación a la organización de un repositorio semántico en SeDiCI: Preliminares

Marisa R. De Giusti, Ariel Sobrado, Agustín Vosou y Gonzalo L. Villarreal

I. Introducción

El Servicio de Difusión de la Creación Intelectual (SeDiCI: http://sedici.unlp.edu.ar) fue creado, inicialmente, para exponer la creación de las distintas unidades académicas de la UNLP como vía de socialización de conocimientos. SeDiCI oferta sus contenidos siguiendo el protocolo de la Iniciativa Open Archives (OAI) y a la vez recolecta información académica externa libre bajo este protocolo. Un objetivo constante del servicio es brindar a los usuarios información cada vez mayor y más pertinente. Para lograr este objetivo se ha pensado en cómo obtener de la web información libre de distintas temáticas de interés chequeando las fuentes y estructurando adecuadamente esta información en la biblioteca digital para permitir a los usuarios búsquedas más adecuadas.

La información volcada en Internet es normalmente buscada por los usuarios en buscadores de tipo general: Google, Yahoo!, etc. El problema que se presenta es que estos buscadores realizan búsquedas por palabras clave en lugar de una búsqueda por conceptos: de este modo, se pierden relaciones de importancia y con ello la información obtenida es diferente aún en el caso de que las palabras utilizadas para la búsqueda sean sinónimos (Abian, 2003).

Una de las alternativas actuales de la web es la denominada web semántica (W3C Semantic Web, 2009 y Wikipedia, 2009), una iniciativa del W3C liderada por Tim Berners Lee, (Berners-Lee, 1999). Sin embargo, en este caso, dada la naturaleza de la biblioteca digital, la propuesta no está dirigida a trabajar generando y compartiendo ontologías (más allá de que las consecuencias de este trabajo lleven a este logro), sino que se pretende hacer un uso combinado de ontologías (Gruber, 1992) y de los tesauros (Wikipedia, 2009) que actualmente utiliza el servicio para incrementar así la eficiencia de los procesos de obtención automática de información de fuentes heterogéneas (la web, determinados repositorios, etc.) por parte de SeDiCI y la búsqueda por parte de los usuarios. Cuando se habla de fuentes heterogéneas, como advertimos en el párrafo previo, y aunque estemos buscando información en el mismo dominio de interés, las distintas fuentes utilizan distintas convenciones para representar los mismos conceptos. Con este fin es que se propone la creación de una plataforma capaz de extraer datos de páginas de distintos portales marcadas semánticamente y realizar la unificación del formato, almacenando la información en un repositorio ontológico sobre el que los usuarios puedan buscar ya no con palabras claves sino semánticamente.

.

Marisa R. De Giusti, Ariel Sobrado, Agustín Vosou y Gonzalo L. Villarreal

II. Descripción de la plataforma

En esta sección se describen los componentes de la plataforma observables en la Figura 1.

Figura 1: Esquema de la plataforma

Como puede observarse en la figura la plataforma puede dividirse en dos grandes sistemas: sistema de recolección y sistema de búsqueda. El sistema de recolección, tiene como tareas principales las de recorrer las páginas web, detectar aquellas marcadas, extraer el contenido de las mismas y organizarlo atendiendo a las ontologías definidas para la temática y a los tesauros seleccionados. El sistema de búsqueda, sobre el cual no nos centraremos en esta primera

.

Marisa R. De Giusti, Ariel Sobrado, Agustín Vosou y Gonzalo L. Villarreal

presentación, estará encargado básicamente de ayudar al usuario en la realización de búsquedas inteligentes (semánticas) sobre el repositorio (semántico también) que ha sido poblado por el sistema de recolección. El desarrollo del módulo de búsquedas aún no ha sido iniciado, pues hemos determinado que resulta prioritario avanzar primero con el sistema de recolección a fin de obtener una buena cantidad de información para luego poder ser buscada. En ambos módulos las ontologías definidas y los tesauros tienen un rol importante. En esta primera etapa los esfuerzos están destinados a los componentes del sistema de recolección, cuyo desarrollo ha avanzado considerablemente aunque de momento, y por estar en una fase preliminar de pruebas y mejoras, no se ha puesto accesible para el público en general.

Sistema de recolección

El sistema de recolección está compuesto por los siguientes elementos (que serán descritos en detalle más adelante): 1) una base de datos que contiene las urls a visitar; 2) un agente (Robot 1) que toma urls no visitadas de la base de datos e invoca al Web crawler; 3) un Web crawler capaz de obtener urls embebidas dentro de otra; 4) un agente (Robot 2) que procesa urls y puebla la ontología; 5) un procesador GRDDL que permite aplicar transformaciones a un documento (X)HTML; 6) un tesauro con los términos homogeneizados; y por último 7) una ontología que utilizaremos para representar los datos obtenidos.
Dichos componentes están separados en dos módulos, de búsqueda y de procesamiento, los cuales se pueden ejecutar paralelamente sin interrumpir la ejecución del otro.

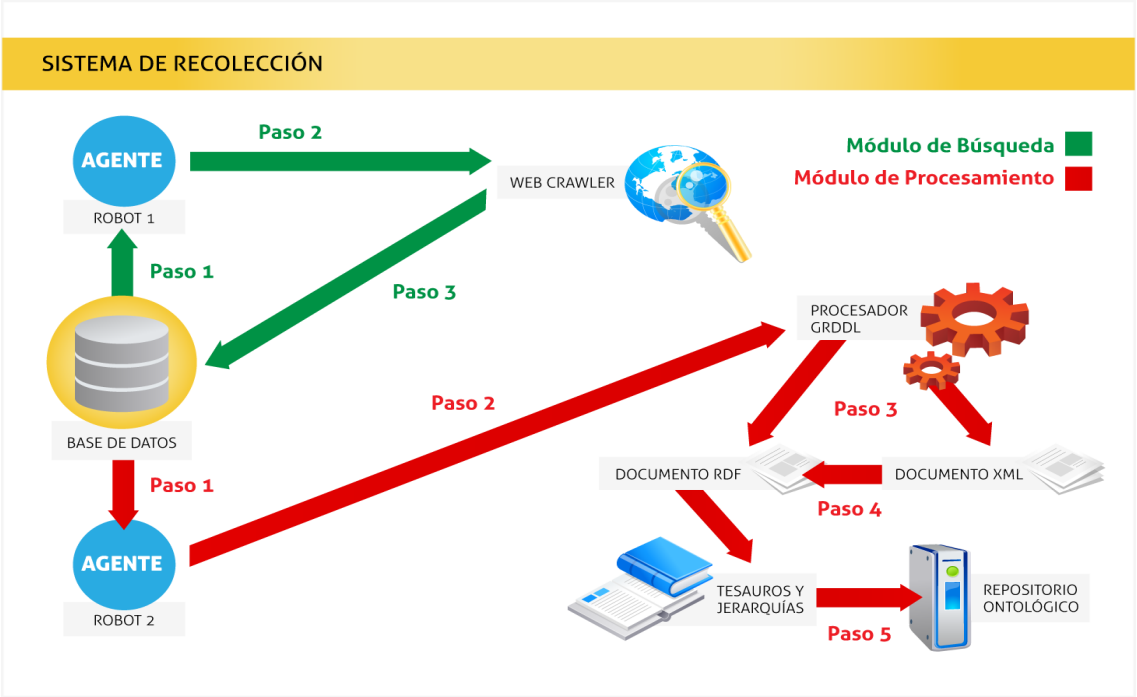

Figura 2: Sistema de recolección

.

Marisa R. De Giusti, Ariel Sobrado, Agustín Vosou y Gonzalo L. Villarreal

Pasos de ejecución del módulo de búsqueda:

1. El Robot 1 toma las URLs no visitadas de la Base de Datos.
2. El Robot 1 envía al Web crawler la URL obtenida. También la marca como visitada.
3. El Web crawler verifica dicha URL, busca las URLs embebidas dentro del código HTML (identificadas dentro del tag <a>) y las agrega a la Base de Datos.

Pasos de ejecución del módulo de procesamiento:

1. El Robot 2 toma de la Base de Datos una URL que ya haya sido visitada, pero no haya sido procesada. Esa dirección le es pasada al procesador GRDDL.
2. El procesador GRDDL aplica ciertas transformaciones XSLT (mediante hojas de transformaciones) para obtener un documento XML o RDF.
3. El Robot 2 marca la URL como procesada.
4. En caso de ser necesario transforma el fichero XML a RDF.
5. Se buscan los términos en el tesauro definido y se crean instancias de la ontología para poblar el repositorio.

El sistema de recolección se encarga de las siguientes tareas: a través de un agente (Genesereth, 1994) recorre una lista de páginas web que envía a un Web crawler (Wikipedia, 2009 y SUN, 2009), encargado de recolectar otros enlaces embebidos en la página en cuestión e incorporarlos a la lista previa. Un agente llamado robot 2 toma una URL y la envía a un procesador del tipo GRDDL ("Gleaning Resource Descriptions from Dialects of Languages") (W3C, 2008) que aplica una transformación, en nuestro caso mediante XSLT, de un documento XHTML o XML a XML (W3C, 1999). Este documento textual sólo contiene ahora las tuplas de interés, lo que permitirá a la aplicación extraer de forma automática información de páginas web estructuradas para integrarla en un repositorio.

El procesador GRDDL detecta en las páginas los microformatos de interés y con la indicación de la dirección de la hoja de transformación (XSL) que usamos para capturar los microformatos (Wikipedia, 2009 y Microformats, 2005) y la URL del lugar de extracción, devuelve un XML con la información extraída por la hoja de transformación. El documento XML es a posteriori transformado a RDF (Wikipedia, 2009) para que los datos sean homogeneizados, clasificados y se pase a poblar la ontología. La homogeneización se realiza mediante conversiones (uso de sinónimos) y también traducciones para tener un idioma único. La clasificación de la información se hace mediante una ontología de dominio; durante esta etapa se crean instancias con atributos y relaciones a partir del RDF: dichas instancias deben cumplir con su pertenencia a la clase raíz de la ontología. Además se busca en el RDF las relaciones de dominio marcadas por la ontología.

.

Marisa R. De Giusti, Ariel Sobrado, Agustín Vosou y Gonzalo L. Villarreal

El paso siguiente es el chequeo de que las instancias satisfagan todas las restricciones impuestas por su clase a través de un módulo de razonamiento encargado del chequeo sobre las instancias con las cuales se ha poblado la ontología. Si la instancia en cuestión pertenece a la clase se le agregan los atributos heredables. Terminada esta etapa se realiza el almacenamiento en un repositorio semántico accesible vía web.

III. Estudio de caso

La elección del caso estuvo sometida a la limitación de la existencia de páginas marcadas en la web actual y requirió hacer cambios sobre la marcha a fin de probar la plataforma de manera fehaciente en lo relativo al sistema de recolección. Como primer ejercicio trabajamos directamente con la biblioteca digital SeDiCI (que operaría como una web heterogénea); para poder hacerlo se la debió modificar para que comience a utilizar el microformato Dublin Core (Dublin Core Metadata Initiative DCMI, 2003 y Dublin Core Metadata Initiative DCMI, 2009) en la representación de los registros (Méndez, 2008) sobre el repositorio. El ejemplo de búsqueda elegido consistió en la búsqueda de material que en el campo de descriptores contuviera: “Física del estado sólido”.

IV. Definición de la Ontología

Nuestra ontología SediciON volcada a Protégé (Stanford Center for Biomedical Informatics Research, 2009) posee una clase denominada MATERIAL de la cual es subclase el tipo de documentos que estamos analizando. Nuestra ontología hace reúso de la Dublin-Core Ontology http://protege.stanford.edu/plugins/owl/dc/protege-dc.owl.

1. Nivel sintáctico: Para la representación de la ontología SediciON se ha utilizado el lenguaje OWL-DL, recomendación del W3C (W3C, 2004). Las características de OWL-DL como lenguaje aseguran la interoperabilidad con otros sistemas y formatos. Otras características de OWL-DL, como su capacidad para la inferencia en sistemas de organización conceptual basados en jerarquías, se utilizarán para proporcionar más funcionalidad al sistema y describir de una forma más rica los recursos involucrados en él.

2. Nivel semántico: Se ha utilizado una ontología de alto nivel para la organización general de repositorios académicos y para garantizar su interoperabilidad con otros sistemas similares. No sólo es necesario el uso de un formato sintáctico de representación estándar, sino también asegurar la compatibilidad semántica futura con otras extensiones, en caso de que sea necesario añadir información sobre nuevos recursos relacionados con la biblioteca digital SeDiCI; de ahí el uso de una ontología estándar de alto nivel.

.

Marisa R. De Giusti, Ariel Sobrado, Agustín Vosou y Gonzalo L. Villarreal

Las entidades reutilizadas de la Ontología DC aparecen referenciadas muy brevemente arriba, pero se remite al lector a su descripción original en caso de que necesite una aclaración de su significado. La ontología modela el sistema de 15 definiciones semánticas descriptivas de Dublin Core.

V. Ejemplo de extracción

En nuestro ejemplo extraeremos los datos de documentos marcados de la biblioteca digital SeDiCI. Durante el proceso de adquisición el módulo de recolección procesa únicamente las páginas con microformatos Dublin Core. El fichero RDF creado contiene los datos de los trabajos referidos a la temática de interés, es decir su título, su autor y una serie de descriptores. Con los datos RDF el módulo de población de la ontología crea las instancias de los archivos. En nuestro caso, antes de almacenarlos (en un futuro próximo) podríamos buscar entre los valores del atributo Description todos aquellos valores que se correspondan con los términos alternativos de algún tesauro distinto del que opera en SeDiCI.
El módulo de población podría también aplicar transformaciones de idioma.
La información y los datos sobre los diferentes recursos se van disponiendo en un repositorio ontológico de SeDiCI: FisSol. Las entidades de esta base de datos se deben mapear a instancias RDF. Las instancias se organizan de acuerdo al modelo conceptual de la ontología SediciON.

VI. Conclusiones y Trabajos futuros

Se han presentado los primeros lineamientos de una plataforma de búsqueda semántica en fuentes heterogéneas que combina el uso de ontologías y tesauros**,** que dará mayor pertinencia a la búsqueda de los usuarios de SeDiCI. Para el grupo de trabajo de PrEBi esta es una primer experiencia que ha servido para conocer el ámbito de trabajo y su aplicabilidad a la biblioteca digital según un objetivo prioritario**:** brindar mejor y más estructurada información. El sistema de recolección deberá analizarse en cuanto a su eficiencia: el uso de dos robots, las herramientas y librerías seleccionadas, especialmente pensando que tras una primera etapa de semantización de la biblioteca, ésta deberá completarse con nuevos contenidos desprendidos de la WEB**.** En este sentido, la búsqueda de páginas marcadas, tal cual la aplicación actual**,** puede resultar una limitación y será necesario pensar en otras técnicas de extracción. Finalmente**,** será necesario pensar en cómo manejar otras relaciones definidas en tesauros y ontologías más complejas.

.

El sistema de búsqueda deberá realizarse por completo pues la idea de esta plataforma es que los usuarios accedan a los repositorios semánticos creados y puedan realizar búsquedas más pertinentes, para lo cual el módulo de búsquedas deberá permitirles un acceso vía web que brinde una pantalla de búsqueda donde puedan buscar por conceptos, seleccionar atributos y elegir restricciones para que finalmente se les devuelva una lista de resultados que muestren los atributos seleccionados que cumplen con las restricciones establecidas.

VII. Referencias y citas bibliográficas.

Abian, M.A. “El futuro de la web. Xml,rdf/rdfs, ontologías y la web semántica”. http://www.javahispano.org/contenidos/es/el_futuro_de_la_web/

W3C. (2009). W3C Semantic Web Activity. [En línea]. http://www.w3.org/2001/sw/grddl-wg/td/grddl-tests#spaces-in-rel/. [2009, Julio 10].

Wikipedia.(2009).WebSemántica.[En línea]. http://es.wikipedia.org/wiki/Web_semántica. [2009,Julio 10].

Berners-Lee, T. y Fischetti, M. (1999). “Weaving the Web: The original Design and Ultimate Destiny of the World Wide Web by its Inventor. San Francisco:Harper.

Wikipedia.(2009).Web Crawler.[En linea]. http://en.wikipedia.org/wiki/Web_crawler [2009,Julio 10].

Sun.(2009). Writing a Web Crawler in the Java Programming Language. [En linea]. http://java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler/ [2009,Julio 10].

Gruber, T. R.(1992) “What is an Ontology?”.[En linea] http://www-ksl.stanford.edu/kst/what-is-an-ontology.html [2009 , Julio 5].

W3C. (1999). XSL Transformations (XSLT) Version 1.0 [En linea]. http://www.w3.org/TR/xslt. (2009,Junio 5).

Wikipedia. (2009). Tesauro. [En linea]. http://es.wikipedia.org/wiki/Tesauro. (2009,Julio 7).

Wikipedia. (2009). Resource Description Framework. [En linea]. http://es.wikipedia.org/wiki/Resource_Description_Framework.

.

Wikipedia.(2009). Microformato. [En linea]. http://es.wikipedia.org/wiki/Microformato [2009, Julio 9].

Microformats.(2005). Microformats.[En linea]. http://microformats.org/. [2009, Julio 6].

Dublin Core Metadata Initiative (DCMI).(2003).Expressing Dublin Core in HTML/XHTML meta and link elements[En linea]. http://dublincore.org/documents/dcq-html/. [2009 , julio 8].

Dublin Core Metadata Initiative(DCMI).(2009). Dublin Core Metadata Initiative [En linea] http://dublincore.org. [2009 , julio 8].

Mendez, E.. “DCMF:DC y microformatos, a good marriage”. International Conference on Dublin Core and Metadata Applicationes, 22-26 September 2008. [En línea]. http://dc2008.de/wp-content/uploads/2008/09/dc2008_mendezetal.pdf. . [2009, Junio 6].

Stanford Center for Biomedical Informatics Research.(2009). Welcome to protégé. [En linea] http://protege.stanford.edu/. [2009,julio 7].

W3C.(2004). Web Ontology Language.[Em linea]. http://www.w3.org/2004/OWL/. [2009, junio 30].

.