

Catalogación y búsqueda semántica en un sitio web

Juan Barrios N. - Claudio Gutiérrez

Universidad de Chile, Departamento de Ciencias de la Computación,
Blanco Encalada 2120, Santiago, Chile
jbarrios@dcc.uchile.cl - cguetierr@dcc.uchile.cl

Resumen

La Web Semántica es una propuesta de la W3C que permite automatizar el procesamiento semántico de la información en la Web actual. Una de las aplicaciones que más se potencia con este enfoque es la *catalogación*, es decir, el proceso de creación de información agregada a nivel semántico. Este trabajo propone un modelo para la catalogación semi-automática de un sitio web a partir de la creación de un conjunto de metadatos sobre los contenidos de un sitio. A partir de esto crea un catálogo y ofrece al usuario distintos buscadores sobre estos conceptos semánticos. Este enfoque mejora los resultados de los buscadores sintácticos en el ámbito de intranets, donde las técnicas de recuperación de información clásica no han mostrado los éxitos que tienen en Internet global. Este artículo reporta el modelo, su implementación, un caso de estudio, y su comparación con buscadores sintácticos en un sitio.

Palabras claves: Catálogos, Metadatos, Web Semántica, Intranet, Búsqueda Semántica

Abstract

The Semantic Web is a proposal of the W3C to allow automatic processing of semantic information in the current Web. One of the applications which benefits more with this approach is *cataloguing*, that is, the process of creation of aggregate information at a semantic level. This work proposes a model for semi-automatic cataloguing of a Web site based on the creation of a set of metadata about the site contents. It builds a catalog and offers the user different search procedures based on semantic concepts. This approach improves results of syntactic search engines in the scope of intranets, where classical information retrieval techniques are far from having the success they enjoy in the global Internet. This paper reports a model, its design and implementation, tests it against a study case, and compares it with syntactic search engines in a single Web site.

Keywords: Catalog, Metadata, Semantic Web, Intranet, Semantic Search

1. Introducción

La Web Semántica es una extensión de la Web tradicional donde a la información publicada en lenguaje natural se le agrega un significado estructurado, con el objetivo de permitir que el contenido de un documento pueda ser procesado y entendido por una computadora [3]. Para aumentar la comprensión de los computadores, los humanos deben extraer la información relevante de cada documento y mantenerla como datos agregados o *metadatos*. Una de las aplicaciones más interesantes que ha potenciado este enfoque es la *catalogación*, debido a las técnicas proporcionadas para la creación de información agregada a nivel semántico. En particular, hoy tenemos las herramientas conceptuales para enfrentar la tarea de catalogación de páginas web. Los Directorios Web -como el *Open Directory Project*- son una demostración de las posibilidades que proporciona la catalogación dentro de la Web, como reunir sitios relacionados o realizar búsquedas restringidas a ámbitos temáticos.

La catalogación ha sido utilizada con anterioridad a la existencia de la Web, particularmente en las bibliotecas, donde se extrae información de cada libro por bibliotecarios especializados creando una ficha correspondiente bajo un formato y reglas definidas, formando un *catálogo*. Sin embargo, aún cuando este proceso sea una técnica antigua y esté estandarizada desde los años '60 en formatos como el MARC (Machine Readable Catalogue Format) [6], los documentos digitales de la Web no pueden catalogarse en el sentido

tradicional y estricto de una biblioteca [11]. Esto se debe a que existen características específicas de la información electrónica que hacen que un registro de metadatos de un documento electrónico (un texto, un sonido, una imagen digital, un programa, etc.) difiera de los registros catalográficos tradicionales de la información tangible (libros, revistas, etc.). Cualquier forma de catalogación debe tomar en cuenta, además, la naturaleza de la Web y sus principales dificultades para esta área: *Masividad, Dinamismo y Distribución* [2, 7].

Por otra parte, los buscadores sintácticos de Internet se han transformado en la principal puerta de acceso a la gran cantidad de información disponible en línea. Los usuarios incluso han comenzado a utilizarlos como “buscadores de respuestas”, visualizando a la Web como un solo gran recurso que proporciona información sin importar de donde provenga ésta [14]. Diversos estudios de satisfacción demuestran la buena evaluación que tienen los buscadores de Internet por parte de los usuarios, por ejemplo, en un estudio de la *American Customer Satisfaction Index* publicado en agosto del 2004, la satisfacción de los usuarios con los motores de búsquedas alcanza 80 puntos de un máximo de 100 [1].

Junto con el éxito en el crecimiento de Internet se ha verificado un aumento del número de sitios de intranets, es decir, de sitios web intra-organizacionales separados de Internet por firewalls o proxies, lo que ha generado una necesidad de motores de búsqueda con características especiales para intranets. Sin embargo, aún cuando la búsqueda de páginas en Internet ha recibido gran atención académica y comercial, se ha realizado poca investigación sobre formas específicas para realizar búsquedas dentro de un sitio web [5, 16]. Un estudio de *Keynote Systems* publicado en enero del 2005 además de reafirmar la supremacía de Google en las búsquedas de Internet, hace notar la alta frustración de los usuarios (un 22%) con las búsquedas locales en un sitio [9].

Un informe de Jakob Nielsen del año 2002 [13] señala que las búsquedas en sitios de intranet tienen un pobre desempeño, independiente de si son implementadas con un motor de búsqueda propio o utilizando alguno de los servicios públicos. Los problemas encontrados se deben principalmente a que el conjunto de resultados no es priorizado correctamente y la información en el despliegue de resultados no es suficientemente explicativa para que un usuario encuentre lo buscado. Las causas principales para estas fallas son:

- Grandes grupos de páginas normalmente contienen idénticos títulos y resúmenes aún cuando la información contenida sea diferente, lo que afecta el despliegue de información encontrada en su detalle.
- Los links entre páginas no existen debido a la importancia de la información de una página, sino que principalmente por motivos de navegación y estructura del sitio, lo que afecta los algoritmos de ranqueo de páginas.
- El contexto en el cual una página existe y las relaciones con otras páginas no puede ser vista a través de la visualización estándar de resultados [5].

Estos problemas dificultan la utilización de las intranets y hacen perder tiempo a los usuarios en búsquedas ineficaces. Una mejora en las intranets, en su diseño, forma de navegación y búsquedas podría disminuir estas pérdidas de tiempo y dinero en hasta un 43% [13].

El presente trabajo utiliza conceptos de la Web Semántica con el objetivo de obtener mejores resultados que los que se obtienen actualmente al utilizar búsquedas tradicionales en un sitio de intranet. Para esto, se propone un modelo para la catalogación semi-automática del sitio, creando un conjunto de metadatos sobre los contenidos de los recursos disponibles en un sitio, los que son organizados formando un *catálogo*. Los visitantes del sitio web realizan diferentes tipos de consultas en el *servidor de catalogación* para encontrar el o los recursos del sitio que responden a sus necesidades. Los metadatos son creados según un esquema particular del sitio y son ingresados, revisados y mantenidos por un grupo de usuarios catalogadores que agregan a su navegador una herramienta especializada llamada el *cliente de catalogación*.

Se desarrolló una implementación de este modelo llamada *Sistema Catálogo*¹. Éste se compone de un conjunto de aplicaciones web para búsqueda de recursos, mantención de metadatos y definición de esquemas de metadatos, y de un plugin para navegadores basados en Mozilla. Se realizó además un caso de estudio con la instalación de este sistema para un sitio web corporativo. Para esto se creó un esquema de metadatos particular al sitio, se efectuó una catalogación automática y manual, y luego se realizaron búsquedas de pruebas para evaluar las características y los resultados de este modelo.

Al utilizar el sistema de catalogación para realizar búsquedas se verificó que la cantidad de resultados encontrados es menor que la cantidad que se puede encontrar con un buscador sintáctico, siendo el primero

¹<http://putu.dcc.uchile.cl/catalogo/>

normalmente el conjunto de los resultados más relevantes para la búsqueda realizada. El sistema permite realizar búsquedas similares a las de un buscador tradicional, siendo posible además restringir el contexto de las palabras. Se pueden realizar también búsquedas de recursos asociados con alguna instancia de clase (por ejemplo, con alguna persona en particular) o de páginas que referencien a cierto recurso.

Este modelo de catalogación permite, además, utilizar metadatos en la web actual sin necesidad de modificar las páginas web publicadas, lo que permite comenzar a utilizar la Web Semántica sin necesidad de intervenir cada sitio ya existente. Es importante notar también que el catálogo es extensible y puede tener diferentes usos aparte de realizar búsquedas de recursos disponibles en el sitio.

A continuación se presenta el contexto en el cual se desarrolló este trabajo y las investigaciones que lo han influenciado. En la sección 2 se explican detalles generales sobre la catalogación y se presenta el modelo de catalogación propuesto. En la sección 3 se presenta el software desarrollado con la implementación del modelo de catalogación. La sección 4 muestra un caso de estudio con la instalación del sistema para un sitio corporativo. Finalmente la sección 5 muestra los resultados de la catalogación aplicada a un sitio web y se presenta una comparación entre ésta y un buscador tradicional en un sitio web.

1.1. Trabajo relacionado

Este trabajo se enmarca dentro de los proyectos realizados por el Grupo Metadatos de la Universidad de Chile para avanzar hacia la Web Semántica y está basado en la presentación de Tesis de Magíster del año 2003 *Catalogación semántica de sitios web* [12]. Entre los proyectos realizados por el Grupo Metadatos, el presente trabajo se relaciona con *DepMark* [10] al utilizar parte de su ontología para la instalación del sistema de catalogación en el sitio web del Departamento de Ciencias de la Computación de la Universidad de Chile (DCC).

El problema de los buscadores sintácticos relacionado con el despliegue de resultados es tratado por el sistema *Cha-Cha* [5] de la Universidad de California. Este sistema propone un cambio en el diseño de la visualización de los resultados de búsqueda para permitir ver el contexto temático de cada página. Sin embargo, este contexto es deducido según los directorios contenidos dentro de la URL que tiene asignada cada página y no como información que ha sido agregada por humanos, como lo propone la Web Semántica.

El proyecto *Annotea* [19, 8] de la W3C tiene por objetivo permitir la creación y publicación de comentarios sobre documentos web utilizando un esquema basado en RDF y XML. Estas anotaciones son acumuladas en servidores centrales y son ingresadas y visualizadas por programas clientes creados o adaptados para ello. Un programa cliente de este proyecto es *Annozilla*², el cual es un plugin para el navegador Mozilla que permite ingresar anotaciones para las páginas web que se estén visitando. El presente trabajo utilizó el código fuente de Annozilla como ejemplo para la implementación del cliente de catalogación.

Los Directorios Web intentan lograr una catalogación global de Internet clasificando todos los sitios disponibles en la Web a través de un árbol temático universal. La clasificación es realizada y mantenida por personas -ya sea voluntarios o contratados- que asignan manualmente cada sitio en uno o más grupos dentro del árbol. Los directorios más utilizados en la actualidad son *Open Directory Project*, *Yahoo! Directory* y *LookSmart*³. Sin embargo, los Directorios Web no han tenido impacto como una forma para efectuar búsquedas en Internet a través de un catálogo, sino que han tenido mayor efectividad como apoyo a los buscadores sintácticos, presentando temas y sitios relacionados a las palabras buscadas [4].

El enfoque más utilizado actualmente en Internet para usar metadatos es agregar etiquetas <META>, también llamadas *meta-etiquetas*, a un documento HTML, con información relevante del texto como palabras claves, título, descripción, tiempo de actualización, etc. Las meta-etiquetas tienen el objetivo de guiar a los buscadores sintácticos en la indexación, búsqueda de resultados y medición de relevancia. Sin embargo, principalmente debido al mal uso dado a este tipo de información extra, los mayores motores de búsqueda han disminuido, y algunos eliminado, el soporte para las meta-etiquetas, con lo cual cada vez pierden mayor importancia, al menos en el ámbito de la búsqueda tradicional [17]. Por esta razón y a la infactibilidad de modificar la web existente para agregar meta-etiquetas, es que el presente trabajo apoya el uso de catálogos sobre sitios web como opción para el uso de metadatos en Internet.

Existe una gran variedad de software de catalogación para la implementación de bibliotecas digitales, sin embargo el presente trabajo no tiene por objetivo final la catalogación formal de recursos en Internet, sino que uno más pragmático como es la mejora de búsquedas en sitios de intranet. En este aspecto el presente

²<http://annozilla.mozdev.org>

³<http://www.dmoz.org>, <http://dir.yahoo.com> y <http://www.looksmart.com>, respectivamente.

trabajo se aleja -en principio- de las bibliotecas digitales y sus proyectos relacionados y de los softwares de representación del conocimiento.

La compañía australiana *Metabrowser Systems*⁴ ha desarrollado dos productos comerciales para la creación de depósitos de metadatos. El primero es un navegador de Internet basado en MS IE que permite ver las meta-etiquetas de las páginas visitadas y crear nuevos metadatos según diferentes esquemas de metadatos, entre ellos Dublin Core. El segundo es un software repositorio de metadatos que ha sido liberado recientemente como software de prueba. Está basado en tecnología *.NET* y permite contener y administrar los metadatos ingresados. Metabrowser utiliza una arquitectura similar a la propuesta por el presente trabajo, aunque este último la utiliza con el objetivo de mejorar las búsquedas en una intranet proporcionando el software libre necesario para lograrlo.

1.2. Contribuciones

Las principales contribuciones de este trabajo son las siguientes:

- Presentar y detallar una opción para mejorar los resultados que se obtienen actualmente al utilizar búsquedas sintácticas en un sitio de intranet que utiliza técnicas de Web Semántica.
- Proporcionar la implementación del software libre que permite mantener catálogos de recursos en la Web y de una herramienta de catalogación en la forma de plugin de navegador.
- Realizar una instalación piloto en un sitio particular, a partir de la cual se realizaron pruebas y obtuvieron conclusiones sobre el modelo propuesto, comparándolo con un buscador tradicional.

2. Catalogación de un sitio web

Catalogar corresponde al proceso de crear un registro sustituto o *metadato* para grupos de información como libros, vídeos, discos, sitios web, etc. [2]. El conjunto de registros conforma un *catálogo* y cumple con tres funciones básicas:

- Conocer qué recursos hay disponibles.
- Conocer dónde se encuentra cada uno de estos recursos.
- Reunir recursos relacionados.

Para el caso particular de documentos electrónicos, el proceso de creación de metadatos se puede definir como la actividad que consiste en extraer y añadir información sobre documentos publicados en aras de su posterior recuperación o para incrementar su utilidad. Se trata, por tanto, de un *arte* de índole técnico, ya que requiere cierta destreza y conocimiento de lenguajes de marcado, formato en que está realizada la publicación electrónica, así como del estándar de metadatos aplicable a tal efecto y del sistema de búsqueda utilizado [11]. El responsable de la creación de los metadatos puede ser el autor del documento mismo, en el caso de bibliotecas digitales tenderán a ser personas especialistas en técnicas de catalogación, y en el caso de sistemas de información especializados los catalogadores deben manejar además los detalles de los conceptos específicos involucrados.

Para el caso de este trabajo, la catalogación es realizada por un conjunto de usuarios que deben tener conocimientos básicos en la herramienta de catalogación y del esquema de metadatos usado. Los autores de los documentos pueden ser -y normalmente los son- usuarios catalogadores del sistema, donde pueden ingresar o actualizar los metadatos de los documentos existentes en el sitio.

Uno de los modelos más utilizados en la actualidad para la creación de la información agregada de un documento es el *Resource Description Framework* (RDF), que permite expresar afirmaciones del tipo: un *recurso* tiene una *propiedad* con un cierto *valor*. Por tanto, sus sentencias son tripletas de la forma sujeto-predicado-objeto, donde *sujeto* puede ser, por ejemplo, una persona, una página web, etc.; *predicado* puede ser la relación “es autor de”, “es hermano de”, etc.; y *objeto* puede ser un libro, otra persona, etc. Las sentencias de RDF son representadas mediante grafos dirigidos, donde el sujeto tiene un arco hacia

⁴<http://metabrowser.spirit.net.au/>

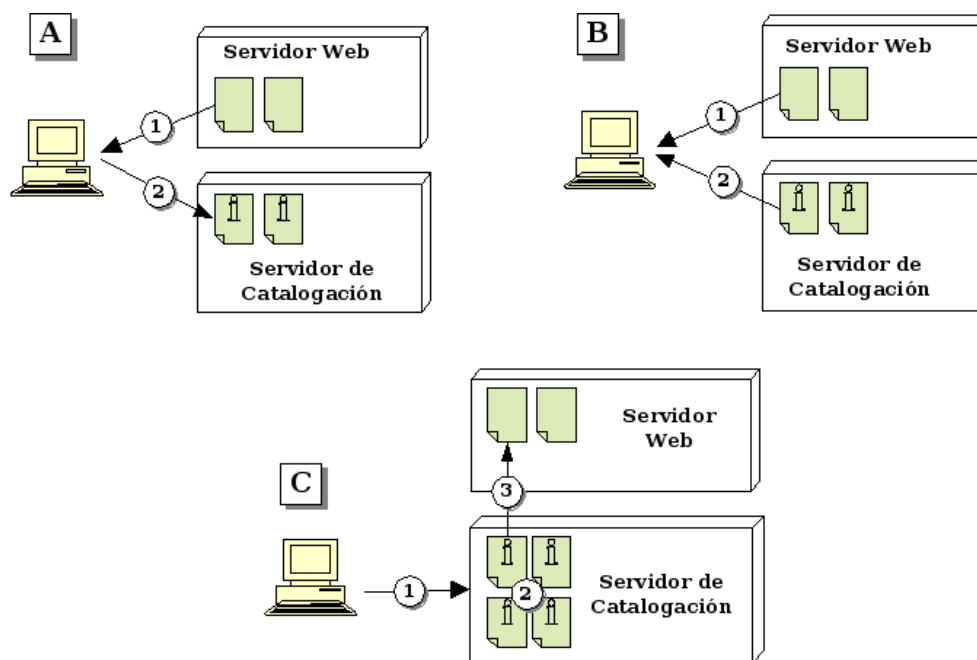


Figura 1: Acciones del sistema de catalogación. **A.** *Agregar metadatos.* 1. Ver Página. 2. Asignar metadatos a URL. **B.** *Apoyar la navegación.* 1. Ver Página. 2. Ver metadatos de URL. **C.** *Buscar recursos.* 1. Ingresar condiciones de búsqueda. 2. Búsqueda entre metadatos. 3. Redirigir hacia URL de los metadatos encontrados.

el objeto mediante el predicado. Un grupo de sentencias forman un grafo RDF, el cual contiene todas las relaciones existentes entre los elementos involucrados en las sentencias [18].

En tareas de catalogación de páginas web, RDF permite describir los contenidos mediante sentencias que contienen al recurso web como sujeto y los predicados corresponden a cada uno de los aspectos a registrar, los que son definidos mediante *ontologías*. Una ontología es la especificación de un vocabulario para un dominio común, es decir, es un modelo para el registro de información, que puede ser definido utilizando RDF a través del lenguaje *RDF Schema*, o utilizando un lenguaje especializado para crear ontologías llamado *OWL*.

Entre los modelos comúnmente usados para registrar datos en Internet se ha destacado el *Dublin Metadata Core Element Set*⁵. Dublin Core es una lista básica de quince elementos diseñada para que los autores y publicadores de documentos de Internet puedan crear sus propios registros sin gran entrenamiento previo.

Para este trabajo, se plantea la creación de un esquema de metadatos específico al sitio en catalogación diseñado pensando en las búsquedas que se desee mejorar, o en su defecto, la utilización de un esquema de metadatos genérico basado en Dublin Core.

2.1. Modelo de catalogación

El modelo propuesto para el sistema de catalogación está compuesto de dos elementos: **el servidor de catalogación** y **el cliente de catalogación**. El servidor de catalogación se encarga de la persistencia de la información del catálogo y de publicar aplicaciones para que los diferentes usuarios puedan hacer uso del catálogo. Permite definir el esquema de metadatos a utilizar, crear y mantener metadatos de los recursos catalogados y presenta diferentes tipos de buscadores y navegadores de metadatos. El cliente de catalogación corresponde a una herramienta que reside en el computador del usuario como un plugin para un navegador. Permite visualizar los metadatos asignados a cada página que se esté visitando, y en el caso que el usuario tenga los permisos necesarios, permite agregar y modificar metadatos en el catálogo.

La figura 1 resume los tres posibles usos que permite el modelo de catalogación:

- Agregar metadatos (figura 1-A), que corresponde a crear un registro para cierta página y agregarlo al catálogo. El ingreso del metadato se puede realizar ya sea utilizando el cliente de catalogación o una aplicación adecuada en el servidor de catalogación.

⁵<http://dublincore.org/>

- Apoyar la navegación (figura 1-B), que corresponde a obtener los metadatos en el catálogo de cierta página que se esté visualizando, utilizando el cliente de catalogación.
- Buscar recursos (figura 1-C), que corresponde a realizar búsquedas o navegaciones en el catálogo para encontrar páginas con la información requerida.

2.2. Roles de usuario

En el sistema de catalogación se diferencian cuatro diferentes roles involucrados, cada uno con diferentes tareas y responsabilidades:

Interesado o Dueño del sitio Corresponde a la persona u organización que desea tener un catálogo para mejorar las búsquedas en su sitio. Las labores principales que realiza en el sistema son: definir los límites del sitio a catalogar; asignar los usuarios Catalogadores del sistema; y proponer posibles necesidades de interfaz gráfica en los buscadores de recursos.

Administrador Corresponde al usuario experto en el sistema de catalogación. Sus principales acciones en el sistema son: estudiar el sitio a catalogar, su nivel de estructuración, su ámbito temático y sus características particulares; definir el o los esquemas de metadatos a utilizar en el sistema, es decir, debe decidir los metadatos a capturar de una página, las definiciones de clase que contendrá el esquema y definir los valores de referencia a utilizar; realizar una carga inicial de metadatos de las páginas creando un *script* que ingrese masivamente metadatos para la mayor cantidad de páginas posibles; y monitorear el sistema a través de indicadores generales con el objeto de revisar la calidad de los metadatos una vez que el sistema esté en funcionamiento.

Catalogador Corresponde al usuario encargado de mantener los metadatos en el catálogo. Normalmente corresponde a una gran cantidad de personas que deben tener conocimiento básico sobre catalogación de páginas, la herramienta de mantención de metadatos e instancias, y el esquema de metadatos usado en el sitio. Sus labores principales en el sistema son: crear, revisar y mantener los metadatos existentes en el sistema; crear, revisar y mantener las instancias de las clases existentes en el sistema; y recibir y procesar las notificaciones de metadatos erróneos en una página.

Público General o Visitantes Corresponde al usuario que visita el sitio y utiliza el sistema para buscar recursos en él. No tiene conocimiento previo del sistema. Sus acciones son: realizar búsquedas de recursos utilizando alguno de los diferentes buscadores que provee el sistema; navegar el sitio y, en el caso de contar con el cliente de catalogación, puede obtener los metadatos de una página como guía para su navegación; y notificar posibles datos erróneos o imprecisos existentes en el catálogo.

2.3. Esquema de metadatos

Para poder ingresar metadatos dentro del catálogo, es necesario definir previamente el o los esquemas de metadatos que se utilizarán. La definición del esquema de metadatos está representado como un conjunto de tipos de campo organizados en forma de árbol donde cada uno debe tener un elemento padre. Existen cinco posibles tipos de campo (ver figura 2):

- **Ontología.** Es el campo raíz de un esquema de metadatos, permite agrupar los campos en una sola unidad temática. Existen tantos esquemas de metadatos como campos tipo Ontología se hayan definido. Cada Campo de Información hijo corresponde a cada uno de los tipos de metadatos que se capturarán de cada recurso del sitio.
- **Definición de Clase.** Permite declarar una clase dentro de la ontología. Debe ser hijo de un campo Ontología o, en el caso que se defina una subclase, otro campo Definición de Clase. Cada Campo de Información hijo corresponde a cada uno de los atributos de la clase que define.
- **Conjunto de Referencia.** Permite declarar un grupo de valores que corresponden a las posibles opciones de un campo de tipo Elección de Referencia. Debe ser hijo de un campo Ontología o, en el caso que se defina un subgrupo, otro Grupo de Referencia. Cada campo tipo Valor de Referencia hijo corresponde a cada una de las opciones posibles de elección.

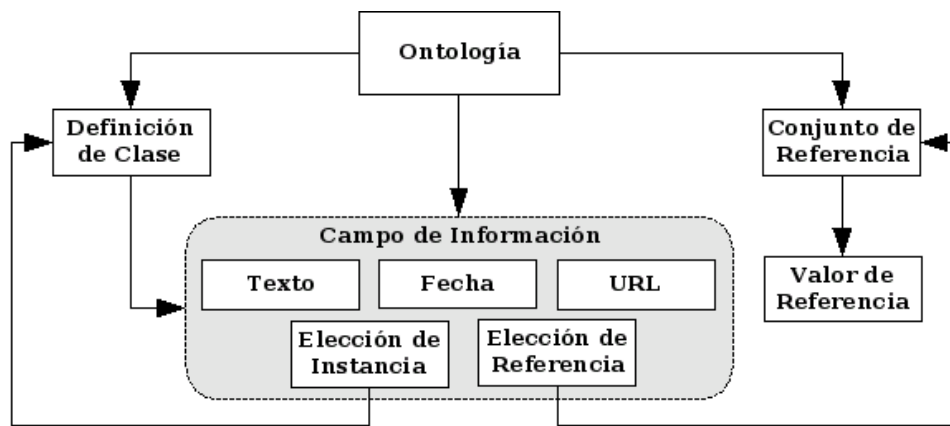


Figura 2: Estructura que cumple cada esquema de metadatos definido en el servidor de catalogación.

- **Valor de Referencia.** Es un campo que representa una opción dentro de un Conjunto de Referencia. Debe ser hijo de un Grupo de Referencia o, en el caso que se defina una especificación, otro Valor de Referencia.
- **Campo de Información.** Es un campo que permite que un usuario catalogador ingrese información sobre cierto elemento. Esta información corresponderá a metadatos de un recurso del sitio en el caso que el campo sea hijo de una ontología, o a atributos de una instancia en el caso que el campo sea hijo de una Definición de Clase. Cada campo de Información puede ser hijo de otro campo de Información para señalar una especificación del campo. Este tipo de campo debe ser uno de los siguientes cinco tipos según el formato aceptable para su valor:
 - *Texto.* Es un campo cuyo valor es un texto sin restricciones.
 - *Fecha.* Es un campo cuyo valor corresponde a una fecha seleccionable de un calendario.
 - *URL.* Es un campo cuyo valor representa una página web. En el caso que la página referenciada exista dentro del servidor esta última aumentará su relevancia base.
 - *Elección de Instancia.* Es un campo cuyo valor debe ser alguna de las instancias creadas para una Definición de Clase definida.
 - *Elección de Referencia.* Es un campo cuyo valor debe ser uno de los Valores de Referencia dentro de un Conjunto de Referencia definido.

2.4. Instalación y puesta en marcha

Para poder contar con el sistema de catalogación en un sitio web se debe realizar una serie de tareas cada una con diferentes responsables:

1. El usuario Interesado debe decidir el tamaño del sitio a catalogar.
2. El usuario Administrador debe estudiar el sitio, la información que contiene y su nivel de estructuración para decidir los recursos que deben ser catalogados. El nivel de estructuración muestra además como se puede realizar la carga inicial de datos, lo que permite estimar los beneficios que se pueden lograr y el tiempo requerido.
3. El usuario Administrador debe definir e ingresar en el servidor de catalogación el o los esquemas de metadatos a utilizar. Para esto se debe decidir los metadatos a ingresar por cada página y las clases y valores de referencia a crear.
4. El usuario Administrador debe ingresar al catálogo las instancias conocidas de antemano para las clases definidas en el esquema de metadatos.



Figura 3: Buscador de recursos catalogados. En el ejemplo se buscan las páginas que posean algún metadato que contenga las palabras “clase” y “auxiliar” y que pertenezcan al curso “CC10A - Computación I”.

5. El usuario Administrador debe hacer una carga inicial de metadatos de las páginas del sitio, que contengan la mayor cantidad de datos del esquema. En lo posible que contengan el título de la página, las referencias entre páginas y las asociaciones con las instancias ya creadas.
6. El usuario Interesado debe decidir quienes cumplirán la labor de usuarios Catalogadores del sistema.
7. Los usuarios Catalogadores deben realizar la catalogación manual de páginas, verificando el marcado automático e ingresando nuevas páginas y metadatos al catálogo.
8. El usuario Administrador debe implementar buscadores particulares al sitio y/o modificar los buscadores genéricos para asemejarse al diseño gráfico del sitio.
9. El usuario Administrador debe hacer ajustes sobre los puntos de ranqueo de los campos de metadatos y de las referencias entre páginas, realizando búsquedas de prueba hasta verificar resultados satisfactorios en el ranqueo de resultados.

Una vez que el sistema está en funcionamiento, los visitantes del sitio realizan búsquedas de recursos utilizando las aplicaciones correspondientes. En ese momento se inicia el proceso de mantención de metadatos, que comprende las siguientes tareas:

1. Los usuarios Visitantes del sitio, en el caso de encontrar anomalías en los metadatos existentes en el catálogo, notifican los posibles problemas de datos en el sistema.
2. Los usuarios Catalogadores deben realizar mantención periódica de los metadatos en el sistema, agregando metadatos para nuevas páginas, actualizando metadatos para páginas que hayan sido modificadas, o eliminando metadatos de páginas borradas del sitio. Reciben y procesan además las notificaciones de metadatos erróneos recibidas.
3. El usuario Administrador monitorea la calidad de los metadatos del sistema a través de indicadores proporcionados por el servidor de catalogación.

3. Sistema Catálogo

Sistema Catálogo es el sistema de catalogación desarrollado en el marco de este trabajo con el objetivo de demostrar las capacidades del modelo presentado en obtener mejores resultados que los buscadores sintácticos en un sitio web. Se compone del servidor de catalogación y cliente de catalogación descritos a continuación.

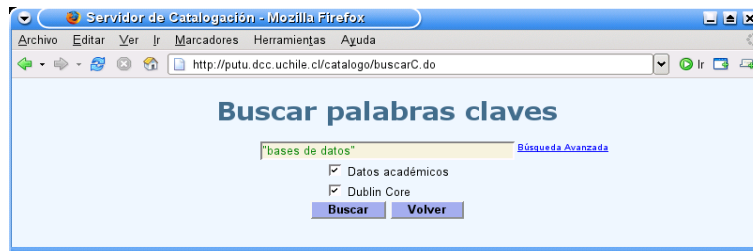


Figura 4: Buscador de recursos por palabras claves. En el ejemplo se buscan las páginas que posean algún metadato que contenga la frase “bases de datos”.

3.1. Servidor de catalogación

Es un conjunto de aplicaciones desarrolladas utilizando tecnología Java sobre un servidor web para las implementaciones y una base de datos relacional para la persistencia de los esquemas y metadatos. Contiene tres aplicaciones web: Búsqueda, Catalogación y Administración, las que pueden ser accedidas por un usuario tipo Visitante, Catalogador y Administrador, respectivamente. Los requerimientos técnicos para instalar el servidor de catalogación son los siguientes:

- J2SDK 1.4.
- Jakarta Tomcat 5.0.
- PostgreSQL 7.4.

La figura 3 muestra un buscador de recursos catalogados. El ingreso de los parámetros se realiza definiendo primero la ontología o el campo específico sobre el cual se desea realizar la búsqueda y luego el valor a buscar. Según el tipo del campo seleccionado se despliega la forma de ingresar el valor correspondiente: una lista de instancias para un campo tipo Elección de Instancia, una lista de Valores de Referencia para un campo tipo Elección de Referencia, un calendario para un campo tipo Fecha o un cuadro de texto para un campo tipo Texto Libre o URL. En el ejemplo de la figura se realiza una búsqueda de las páginas cuyo conjunto de metadatos cumpla con dos condiciones simultáneamente: que posea un metadato clasificado bajo la ontología *Datos Académicos* cuyo texto contenga las palabras “clase” y “auxiliar”, y que posea un metadato para el campo *Pertenece al Curso* cuyo valor sea una referencia a la instancia “CC10A - Computación I” de la clase Curso.

La figura 4 muestra un buscador de recursos basado sólo en palabras claves. Su interfaz es similar a un buscador sintáctico tradicional, permite ingresar un texto a buscar entre los metadatos definidos para las páginas catalogadas, restringiendo el universo de búsqueda a los metadatos asignados bajo una o más ontologías. En el ejemplo de la figura se realiza una búsqueda de las páginas que contengan en alguno de sus metadatos la frase “bases de datos”. En el caso de la búsqueda avanzada, se pueden realizar múltiples búsquedas de palabras restringiendo cada una a los metadatos asignados bajo una ontología o algún campo específico de ésta.

3.2. Cliente de catalogación

Es una herramienta desarrollada utilizando tecnología de Mozilla (XUL y JavaScript), que puede ser agregada como barra lateral del navegador. El cliente está empaquetado en un archivo XPI lo que permite que sea instalado automáticamente dentro del navegador. Una vez instalado el cliente se debe agregar una referencia en la barra lateral del navegador a la dirección `chrome://cliente/content/panel.xul`. Los requerimientos técnicos para instalar el cliente es utilizar un navegador Mozilla o Mozilla-Firefox 1.0 o mayor, que tenga habilitada la capacidad de instalar software.

El plugin permite dos tareas: Ver metadatos y Modificar metadatos, las que pueden ser accedidas por un usuario tipo Visitante y Catalogador, respectivamente.

La figura 5 muestra el navegador con la barra lateral del cliente de catalogación. Al presionar el botón *Refresh*, se realiza una consulta al servidor de catalogación por los metadatos correspondientes a la URL de la página que se encuentra en la ventana central del navegador, los que son desplegados en forma de árbol.

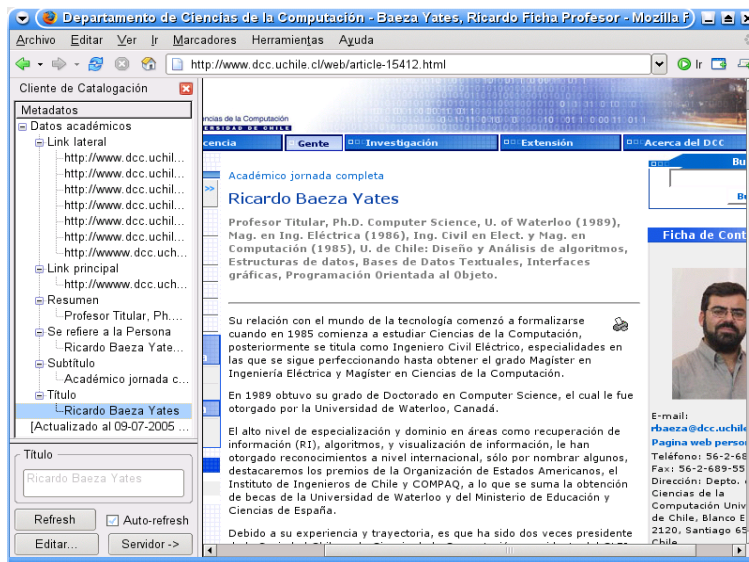


Figura 5: Visualización de los metadatos de un recurso catalogado utilizando el cliente de catalogación.

Al marcar el recuadro *auto-refresh* se habilita el modo automático, donde cada 10 segundos se realiza la acción del botón *Refresh*.

4. Caso de estudio

Como prueba del sistema, se realizó la catalogación del sitio web del Departamento de Ciencias de la Computación de la Universidad de Chile (DCC), cuya URL es <http://www.dcc.uchile.cl>.

Se utilizó un esquema de metadatos particular al sitio, el cual es el resultado del estudio del conjunto de páginas a catalogar y de la ontología utilizada por el proyecto *DepMark* [10]. Contiene cuatro campos de tipo Texto: *Título*, *Subtítulo*, *Resumen* y *Sección*; dos campos de tipo URL: *Link principal* y *Link lateral*; y tres campos de tipo Elección de Instancia: *Pertenece a la Carrera*, *Pertenece al Curso* y *Se refiere a la Persona* que referencian a instancias de las clases *Carrera*, *Curso* y *Persona*, respectivamente. Además del esquema particular al sitio, se ingresó en el sistema el esquema de datos definidos por Dublin Core para sus elementos básicos y sus calificadores.

Para realizar la carga automática primero se hizo una copia local del sitio web publicado. Se recolectaron 15 MB en 835 archivos, de las cuales sólo se seleccionaron 353 páginas para ser catalogadas con un espacio total de 1 MB. Las páginas restantes no fueron catalogadas por ser páginas de enlace, que sólo presentan resúmenes de otras o son versiones imprimibles.

Luego se procedió a implementar un conjunto de *scripts* para procesar el texto del HTML utilizando expresiones regulares. Se capturaron los campos de tipo Texto y URL, y se insertaron en la base de datos del sistema de catalogación. La creación de instancias fue manual y correspondió a un total de 99 instancias: 4 carreras, 51 cursos y 44 personas. Después se realizaron, también manualmente, las asociaciones entre instancias y páginas a través de los campos de tipo Elección de Instancia.

El conjunto de metadatos ingresados finalmente, sumando metadatos de páginas y atributos de instancia,

	Tarea	t
1.	Definición del conjunto a catalogar y estudio de la estructura del sitio.	2 días
2.	Definición el esquema de metadatos.	2 días
3.	Ingreso de instancias en forma manual.	1 día
4.	Carga inicial de metadatos.	2 días
5.	Marcado manual del sitio.	7 días
6.	Pruebas de búsquedas y ajuste de los puntos de ranqueo del esquema.	2 días
	<i>Total</i>	<i>16 días</i>

Cuadro 1: Resumen del tiempo para la puesta en marcha del sistema.

Aspecto	Catalogación de un sitio	Buscador sintáctico de un sitio
Costos	Mayor cantidad de trabajo y tiempo para lograr catalogar un sitio. Necesidad de un experto en el sistema de catalogación.	Baja cantidad de tiempo y conocimientos necesarios para tener el sistema en funcionamiento.
Mantenición	Requiere de un usuario administrador para monitorear el estado del sistema y de usuarios catalogadores para verificar y actualizar los metadatos.	Requiere de un usuario administrador para monitorear el estado del sistema.
Resultados	Menor cantidad de resultados encontrados, pero los encontrados son de mayor relevancia para la búsqueda.	Al encontrar todas las páginas donde se encuentra cierta palabra, normalmente los resultados son una gran cantidad de páginas muy similares.
Formas de búsqueda	Diferentes tipos de buscadores que pueden ser usados según la cantidad de información que se tenga sobre lo buscado. En el caso de tener pocos conocimientos se puede intentar una navegación del catálogo.	Interfaz simple de búsqueda. Poca utilización de las búsquedas avanzadas. Difícil de utilizar en el caso de tener poco conocimiento en el área buscada.
Contexto de resultados	Permite conocer el contexto de cada página, independiente de la forma de navegación.	No existe forma de conocer el contexto de una página.
Recursos indexados	Se puede agregar al catálogo todo tipo de documento, incluido cualquier archivo binario.	Se pueden indexar archivos de texto y archivos binarios que puedan ser transformados automáticamente en texto.
Uso de metadatos	Permite hacer uso de metadatos en la web sin necesidad de modificar la web existente.	Para hacer uso de metadatos requiere de la modificación de las páginas web ya publicadas para agregar las meta-etiquetas correspondientes.

Cuadro 2: Comparación entre el sistema de catalogación y los buscadores sintácticos de un sitio.

fue un total de 2440 los que ocuparon un espacio aproximado de 104 KB (tamaño en bytes de los textos de todos los metadatos). Por lo cual, los metadatos correspondieron a aproximadamente un 10% del tamaño de las páginas catalogadas y a un 0,7% del tamaño total del sitio.

El cuadro 1 resume el tiempo tomado para que el sistema de catalogación haya quedado disponible para realizar búsquedas al público. El esfuerzo realizado fue de aproximadamente tres semanas por una sola persona. Se verificó que la tarea que toma mayor tiempo es la marcación manual del sitio. Una de las razones que incidió en esto fue que la carga inicial de metadatos no incluyó asociaciones con instancias las cuales debieron ser enlazadas en forma manual.

El sistema se encuentra actualmente disponible para uso público en la dirección web <http://putu.dcc.uchile.cl/catalogo/>.

5. Conclusiones

Al utilizar el sistema de catalogación para realizar búsquedas se verifica que la cantidad de resultados encontrados es menor que la cantidad que se puede encontrar con un buscador sintáctico, correspondiendo normalmente al conjunto de los resultados más relevantes para la búsqueda realizada.

Al desarrollar el buscador de metadatos quedó de manifiesto que en un catálogo no puede existir sólo una interfaz de búsqueda, sino que debe permitir múltiples y variadas formas para realizar consultas. Es de esperarse que en un principio el tipo de buscador más utilizado sea el presentado en la figura 4 por su mayor similitud con un buscador sintáctico. Sin embargo, una vez que los usuarios adquieren conocimiento sobre el catálogo y sus capacidades, se hace más factible utilizar alguno de los navegadores o buscadores proporcionados que hacen mayor uso del potencial de los metadatos.

Desarrollar una interfaz genérica para los diferentes buscadores es un problema que no pudo ser solucionado satisfactoriamente. Se intentó realizar una interfaz más amigable incluso disminuyendo la potencia del buscador (permitiendo ingresar búsquedas anidadas en un solo nivel cuando el motor permite múltiples niveles), sin embargo el problema aún está abierto para mejores soluciones genéricas. Se puede afrontar este problema implementando buscadores especializados para cada sitio en particular que se adecuen al diseño de éste y a sus esquemas de metadatos.

Una característica importante del sistema de catalogación es que no es necesario modificar las páginas

web existentes para hacer uso de él. Esto permite la utilización de metadatos para mejorar las búsquedas sin necesidad de modificar un sitio ya existente.

La información contenida en el catálogo además es extensible para otros posibles usos independientes del buscador semántico. Por tanto, el catálogo es un recurso que tiene gran potencial y sirve de base para desarrollar futuras ideas y proyectos relacionados con la Web Semántica, como por ejemplo realizar mapas de sitio o implementar búsquedas inter-sitios.

A modo de resumen, el cuadro 2 presenta una comparación entre el modelo de catalogación propuesto y los buscadores sintácticos de un sitio, según diferentes aspectos.

6. Trabajo futuro

- Estudiar los problemas asociados a la unicidad de una página según su URL. En particular estudiar los casos de páginas dinámicas, páginas por defecto y mirrors de sitios.
- Estudiar el problema de la interfaz de usuarios para un buscador semántico. Debe ser lo suficientemente poderosa para permitir el ingreso de consultas complejas, pero debe permitir que un usuario promedio sin conocimiento previos pueda hacer uso de ella.
- Estudiar formas de integración entre catálogos, y por consiguiente, estudiar la integración entre diferentes esquemas de metadatos.
- Estudiar el uso del sistema de catalogación como depósito de referencias de Internet. Se puede utilizar el mismo software de catalogación con el objeto de organizar en forma colaborativa un conjunto muy grande de referencias a páginas de Internet. Se ingresan datos para cada página de interés según cierto esquema de metadatos y luego se realizan consultas sobre estos metadatos para encontrar los links relevantes entre el conjunto de referencias.
- Estudiar el *perfilamiento* en la web. Al poder realizar búsquedas restringiendo los ámbitos de interés a ciertos temas en particular (por ejemplo, artes) y cada página ser catalogada según una persona con el perfil del área (por ejemplo, un artista), se pueden realizar búsquedas y navegar la web bajo cierta forma de ver la información, es decir bajo un perfil particular.

Agradecimientos

Los autores agradecen financiamiento al Proyecto FONDECYT 1030810, “Metadatos para describir y consultar la Web Oculta”. Claudio Gutiérrez agradece también al Nucleo Milenio, Centro de Investigación de la Web, P04-067-F, Mideplan.

Referencias

- [1] American Customer Satisfaction Index. *Second Quarter Scores: Manufacturing/Durable Goods & E-Business: Search Engines*, Agosto 2004. <http://www.theacsi.org>.
- [2] Susan Atkey. *Issues in Cataloguing the Web*. School of Library, Archival and Information Studies/UBC, Diciembre 2002.
- [3] Tim Berners-Lee, James Hendler, and Ora Lassila. *The Semantic Web*. Scientific American, Inc, Mayo 2001.
- [4] Fidel CACHEDA and Angel Viña. *Understanding how people use search engines: a statistical analysis for e-Business*. Proceedings of the e-Business and e-Work Conference and Exhibition (e-2001), Venice, Italy, Octubre 2001.
- [5] Michael Chen, Marti Hearst, Jason Hong, and James Lin. *Cha-Cha: A System for Organizing Intranet Search Results*. Proceedings of the 2nd USENIX Symposium on Internet Technologies and SYSTEMS (USITS), Octubre 1999.
- [6] Lorcan Dempsey and Rachel Heery. *A review of metadata: a survey of current resource description formats*. UKOLN Metadata Group, Marzo 1997.
- [7] Jeff Hefflin, James Hendler, and Sean Luke. *SHOE: A Blueprint for the Semantic Web*. Data and Knowledge Engineering. Spinning the Semantic Web. MIT Press, Cambridge, Marzo 2003.
- [8] José Kahan and Marja-Riita Koivunen. *Annotea: An Open RDF Infrastructure for Shared Web Annotations*. World Wide Web Consortium, Mayo 2001.
- [9] Keynote Systems. *Yahoo! Search and MSN Search Close the Gap with Google*. Press Release 05-01-13, Enero 2005. <http://www.keynote.com>.
- [10] Ernesto Krsulovic Morales and Claudio Gutiérrez. *Building Yearbooks with RDF*. Centro de Investigación de la Web. Departamento de Ciencias de la Computación. Universidad de Chile, Diciembre 2002.

- [11] Eva M^a Méndez. *Metadatos y recuperación de información: Estándares, problemas y aplicabilidad en bibliotecas digitales*. Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid. Ediciones Trea. ISBN: 84-9704-055-4, Junio 2002.
- [12] Juan Manuel Barrios N. *Presentación Tema de Tesis para Magister en Ciencias, mención Computación: Catalogación semántica de sitios web*. Departamento de Ciencias de la Computación. Universidad de Chile, Diciembre 2003.
- [13] Jakob Nielsen. *Intranet Usability: The Trillion-Dollar Question*. Useit.com Alertbox, Noviembre 2002.
- [14] Jakob Nielsen. *When Search Engines Become Answer Engines*. Useit.com Alertbox, Agosto 2004.
- [15] Natalya Fridman Noy and Deborah L. McGuinness. *Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*. Stanford Knowledge Systems Laboratory, Marzo 2001.
- [16] Dick Stenmark. *A Methodology for Intranet Search Engine Evaluation*. In Käkölä, T. (ed.), Proceedings of IRIS22, August 7-10, Department of CS/IS, University of Jyväskylä, Finland, Agosto 1999.
- [17] Danny Sullivan. *Death Of A Meta Tag*. Search Engine Watch, Octubre 2002.
- [18] World Wide Web Consortium. *Resource Description Framework Model and Syntax Specification. W3C Recommendation*, Febrero 1999.
- [19] World Wide Web Consortium. *Annotea Protocols. W3C Draft*, Diciembre 2002.