



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA DE POSTGRADO

CATALOGACIÓN Y BÚSQUEDA SEMÁNTICA EN UN SITIO WEB

JUAN MANUEL BARRIOS NÚÑEZ

MIEMBROS DE LA COMISIÓN EVALUADORA

SR. CLAUDIO GUTIÉRREZ, PROFESOR GUÍA

SR. CARLOS HURTADO, PROFESOR INTEGRANTE

SR. ALEJANDRO BASSI, PROFESOR INTEGRANTE

SRA. ANDREA RODRÍGUEZ, PROFESOR INVITADO

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

SANTIAGO, CHILE

MAYO 2006

Índice general

1. Introducción	1
2. Marco Teórico	4
2.1. Web Semántica	4
2.2. Catálogos	5
2.3. RDF	6
2.4. Ontologías	7
2.5. Dublin Core	8
3. Motivación y trabajo relacionado	10
3.1. Características de la Web	10
3.2. Buscadores sintácticos e intranets	11
3.3. Buscadores sintácticos con metadatos	13
3.4. Directorios Web	14
3.5. Metadatos Universales	15
3.6. Trabajo Relacionado	15
4. Catalogación de un sitio web	17
4.1. Roles de usuario	19
4.2. Servidor de catalogación	20

4.3. Cliente de catalogación	21
4.4. Esquemas de metadatos	21
4.5. Mantenimiento de metadatos	24
4.6. Visualizar metadatos	25
4.7. Búsqueda de recursos	25
4.8. Ordenamiento de resultados	27
4.9. Instalación y puesta en marcha	29
5. Sistema Catalogo	31
5.1. Servidor de catalogación	32
5.1.1. Módulo de Administración	33
5.1.2. Módulo de Catalogación	36
5.1.3. Módulo de Búsquedas	41
5.2. Cliente de catalogación	49
5.2.1. Ver metadatos	50
5.2.2. Modificar metadatos	50
6. Caso de estudio	52
6.1. Procedimiento	52
6.2. Esquema de metadatos	54
6.3. Catalogación inicial	56
6.4. Instalación del sistema	56
6.5. Mantenimiento de metadatos	57
7. Discusión y Conclusiones	59
7.1. Conclusiones	59
7.2. Recomendaciones	61

7.2.1. Creación de ontologías	61
7.2.2. Catalogación de páginas	62
7.2.3. Búsqueda de recursos	63
7.3. Trabajo futuro	63
7.3.1. Investigación del modelo de catalogación	63
7.3.2. Desarrollo del sistema Catalogo	64
A. Código Fuente de Ejemplo	69
A.1. Sincronizador.xml	69

Resumen

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. Para aumentar la comprensión de los computadores, los humanos deben extraer la información relevante de cada documento y mantenerla como datos agregados o *metadatos*.

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. Junto con el éxito en el crecimiento de la Web se ha verificado además un aumento del número de sitios de intranets, lo que ha generado una necesidad de motores de búsqueda con características especiales para éstos. Sin embargo, se han realizado estudios que señalan el pobre desempeño de las búsquedas en sitios de intranet y la alta frustración de los usuarios con este tipo de búsquedas.

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.

Se desarrolló una implementación de este modelo, en software libre, llamada sistema *Catalogo*. Se realizó además un caso de estudio con la instalación de este sistema para un sitio web corporativo, el cual consistió en crear un esquema de metadatos, efectuar una catalogación automática y manual del sitio, para luego realizar búsquedas de pruebas y evaluar las características y los resultados de este modelo.

Al utilizar el sistema de catalogación se verificó que la cantidad de resultados encontrados es menor que la cantidad que se puede encontrar con un buscador sintáctico, siendo el primero normalmente el conjunto de los resultados más relevantes para la búsqueda realizada. El sistema permite realizar diferentes tipos de búsquedas: búsquedas por texto, siendo posible restringir el contexto de las palabras, búsquedas de recursos asociados con algún objeto o que referencien a algún otro 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 modificar cada sitio ya publicado.

Capítulo 1

Introducción

El éxito en el crecimiento de la Web ha transformado a Internet en la fuente de conocimientos más grande de la historia. En la actualidad la cantidad de páginas indexadas por Google supera los 8 mil millones y la *World Wide Web Consortium* (W3C) estima que el número de usuarios para el año 2005 será de más de mil millones.

Los servicios de búsqueda públicos se han convertido en la principal puerta de entrada a esta vasta gama de documentos, permitiendo buscar y acceder a información que se encuentra disponible en la Web en la forma de textos escritos en lenguaje natural. 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 [Ame04].

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*, 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 [CHHL99, Ste99]. 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 [Key05].

Por otra parte, la Web Semántica, propuesta por Tim Berners-Lee el año 2001, 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 [BLHL01]. Esto permite a las computadoras aumentar la “comprensión” de la información acercándolo hacia un nivel más humano, transformando la Web desde una red de “conexiones” hacia una red de “conocimiento” [HHL03].

En la actualidad, para que un sitio web cuente con una opción de búsqueda puede optar por usar el servicio de algún buscador público, en el caso que el sitio sea accesible desde Internet, o puede instalar alguno de los motores de búsqueda existentes. En caso de elegir un motor de búsqueda, la selección debe depender de las necesidades particulares de cada organización, de la información publicada y de las búsquedas a realizar, entre otros aspectos [Ste99].

Un informe de Jakob Nielsen del año 2002 [Nie02] señala que para ambos casos, servicios públicos o motor propio, las búsquedas tienen un pobre desempeño al ejecutarse sobre sitios de intranet. Los principales problemas encontrados fueron que los resultados no son priorizados correctamente y que la información en el despliegue de resultados no es suficientemente explicativa para que un usuario encuentre lo buscado. Esto se debe a que los motores de búsqueda muestran problemas cuando grandes grupos de páginas contienen similares títulos, resúmenes y/o contenidos (algo común en páginas de un mismo sitio que tratan sobre un mismo tema), y cuando la referencia entre páginas está dada principalmente por motivos de navegación y estructura del sitio.

El presente trabajo utiliza conceptos de la Web Semántica con el objetivo de conseguir mejores resultados que los que se obtienen actualmente al utilizar búsquedas tradicionales en un sitio de intranet. Para esto se propone realizar una catalogación semi-automática del sitio, donde se crea un conjunto de metadatos sobre los contenidos de las páginas y recursos disponibles en el sitio, los que son organizados formando un catálogo. Los metadatos en el catálogo deben cumplir con una o más ontologías que han sido definidas previamente por un administrador. Un grupo de usuarios catalogadores crean y mantienen los metadatos en el catálogo, a través de los cuales los usuarios finales pueden realizar búsquedas de los recursos existentes en el sitio según diferentes criterios.

El modelo se compone de un cliente de catalogación y de un servidor de catalogación. El cliente de catalogación es una herramienta que permite ver y asignar metadatos a páginas web mientras se visita el sitio. El servidor de catalogación es un servicio que se encarga de la persistencia de la información del catálogo y de ejecutar acciones sobre éste. Permite definir los esquemas de metadatos, actualizar los metadatos en el catálogo, monitorear el estado de la

información contenida y proveer diferentes tipos de visualizaciones y búsquedas de recursos.

Una implementación de este modelo es el sistema *Catalogo* el cual se compone de un conjunto de aplicaciones web desarrolladas utilizando tecnología Java sobre un servidor web con una base de datos relacional y de un plug-in para el navegador Mozilla para apoyar la navegación usando una barra lateral. Se desarrolló además un caso de estudio con la instalación de este sistema para un sitio web corporativo, el cual consistió en crear el esquema de metadatos, efectuar una catalogación automática y manual del sitio, para luego realizar búsquedas de pruebas y evaluar las características y los resultados de este modelo.

El presente trabajo se estructura de la siguiente forma: en el capítulo 2 se resumen los temas generales sobre los cuales se enmarca este trabajo. En el capítulo 3 se describen las características y problemas relacionados a Internet, los metadatos y los buscadores tradicionales que motivan el presente trabajo. En el capítulo 4 se detalla el modelo de catalogación, sus componentes y características como perfiles de usuario, búsquedas de recursos y ranqueo de resultados. En el capítulo 5 se presenta el sistema *Catalogo*, sus propiedades y su implementación. En el capítulo 6 se describe un caso de estudio con el uso del sistema para un sitio web en particular, su instalación, las ontologías utilizadas, el ingreso de metadatos y el tiempo de implantación asociado. En el capítulo 7 se discuten los resultados obtenidos y se hace una comparación de este modelo con los motores de búsqueda tradicionales, para finalizar con recomendaciones generales y trabajo futuro por realizar.

Capítulo 2

Marco Teórico

El presente trabajo utiliza conceptos de la Web Semántica para aplicarlos en búsquedas de intranet. Por esto, los temas involucrados son los relacionados con metadatos, RDF, catalogación bibliográfica y ontologías, los cuales son tratados a continuación.

2.1. Web Semántica

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 automáticamente por una computadora. Es dar un espacio a las computadoras para que puedan entender la información escrita por humanos en lenguaje natural en Internet. En este contexto para Tim Berners-Lee, el creador de la Web Semántica y actual director de la W3C, la palabra *semántico* se debe entender como *procesable por máquinas*.

Para aumentar la comprensión de los computadores, los humanos deben extraer la información relevante de cada documento, en especial los más difíciles de automatizar como contexto, resumen, campo temático, etc., y mantenerla como datos agregados o *metadatos*.

Existen tres elementos básicos para lograr el objetivo de la Web Semántica: XML como el medio estructurado para el registro e intercambio de la información, RDF como el modelo para expresar información de un objeto y sus relaciones con otros objetos, y Ontologías para definir la estructura de la información registrada permitiendo realizar acciones como búsquedas o inferencias.

2.2. Catálogos

Catalogar corresponde al proceso de crear un registro sustituto o *metadato* para grupos de información como libros, vídeos, discos, sitios web, etc. [Atk02]. 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 se 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 [MÃ©02].

Los catálogos han sido utilizados con anterioridad a la existencia de Internet, particularmente en las bibliotecas, donde se extrae información de cada libro por bibliotecarios especializados creando una ficha correspondiente con un formato y reglas definidas. Las fichas luego son organizadas según diferentes criterios formando catálogos donde se efectúan las búsquedas de libros. Las reglas y los formatos de los catálogos fueron estandarizadas para facilitar el intercambio entre bibliotecas en el MARC (Machine Readable Catalogue Format) originado en los años '60 [DH97].

Sin embargo, aún cuando la catalogación se realice desde hace décadas en bibliotecas, los documentos digitales de la Web no deben catalogarse en sentido tradicional y estricto como si se tratase de un tipo de documento más que integra la colección de la biblioteca [MÃ©02]. 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.). Las diferencias se ponen de manifiesto en cuanto a:

- Localización. Un registro de catálogo hace referencia a una localización física asociada en la biblioteca. Un registro de metadatos hará referencia a localizaciones remotas

donde se requerirá contemplar los detalles del modo de acceso (HTTP, FTP, etc.) y sus restricciones (passwords, etc.). Es frecuente además que el documento se ubique en varias localizaciones de Internet.

- Formato de los documentos. Un documento electrónico puede existir en diferentes formatos (HTML, PDF, PS, etc.). En general, los esquemas de metadatos permiten representar las distintas versiones de un mismo documento en sólo un registro, mientras que en los catálogos tradicionales se describe este hecho como distintas ediciones de una misma publicación donde cada una de éstas tiene asociado un registro diferente.
- Falta de estabilidad. Los datos en Internet muchas veces tienen poca estabilidad: los archivos se mueven, desaparecen o los autores de las páginas cambian y actualizan los contenidos de éstas, sin variar su URI. Con esto un documento consultado en dos momentos distintos puede haber variado en forma y contenido de manera sustancial.
- Nivel de detalle (Granularidad). En los documentos tradicionales el catalogador debe decidir el nivel de análisis (por ejemplo, una descripción como libro o por cada uno de sus capítulos) así como los campos que se van a incluir (autor, título, materias, etc.). Los registros MARC permiten relacionar niveles de análisis inferiores con el análisis de ítem completo. Con los recursos electrónicos se debe tomar una decisión similar, es decir, se debe decidir cuál es el objeto de información mínimo susceptible de asignarle metadatos, ya sea cada página, sus imágenes, cada archivo, el sitio completo, etc. Sin embargo, los formatos de metadatos deben llegar más allá que describir distintos niveles jerárquicos de las páginas, sino que deben ser flexibles para representar tanto un ítem singular (como una página web o una imagen) como un ítem colectivo (una colección de páginas que forman un sitio o un conjunto de imágenes que forman una colección)¹.

2.3. RDF

Resource Description Framework (RDF) es un modelo para expresar afirmaciones del tipo: un *recurso* tiene una *propiedad* con un cierto *valor*. Por tanto, sus sentencias son triplas de la forma sujeto-predicado-objeto, donde, por ejemplo, *sujeto* puede ser una persona, una página web, etc.; *predicado* puede ser la relación “es autor de”, “es hermano de”, etc.; y *objeto*

¹Particularmente, para este aspecto Dublin Core ha introducido la utilización de calificadores para el elemento *Relation* los cuales son: *isVersionOf*, *hasVersion*, *isReplacedBy*, *Replaces*, *isRequiredBy*, *Requires*, *isPartOf*, *hasPart*, *isReferencedBy*, *References*, *isFormatOf* y *hasFormat*.

puede ser un libro, otra persona, etc. Cada uno de los tres elementos de una sentencia se representa únicamente mediante un identificador único, conocido como URI. En particular, el identificador único de una página web corresponde normalmente a su ubicación en Internet dado por su URL.

Las sentencias de RDF son representadas mediante grafos dirigidos, donde el sujeto tiene un arco hacia el objeto mediante el predicado. Un grupo de sentencias forma un grafo RDF, el cual contiene todas las relaciones existentes entre los elementos involucrados en las sentencias.

Una de las formas de representación de RDF es a través de XML, donde cada sentencia corresponde a una etiqueta `<Description>`, el sujeto es el atributo `about` y el predicado son elementos anidados. Por ejemplo:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://www.dcc.uchile.cl/web/channel.html">
    <dc:title>Departamento de Ciencias de la Computación</dc:title>
  </rdf:Description>
</rdf:RDF>
```

La sentencia dice que el sujeto `http://www.dcc.uchile.cl/web/channel.html` tiene un título (como lo define el campo `title` de la ontología de URI `http://purl.org/dc/elements/1.1/`) cuyo valor es *Departamento de Ciencias de la Computación*. XML es el encargado de resolver los asuntos de codificación de caracteres (en el ejemplo corresponde a UTF-8) y de representación de referencias a ontologías (a través de namespaces).

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.

2.4. Ontologías

Para describir recursos y lograr una interoperabilidad y reutilización del conocimiento entre diferentes sistemas es necesario especificar inicialmente un vocabulario común que defina las características que serán representadas por las sentencias de conocimiento. Esta especificación de un vocabulario para un dominio común es llamada una *Ontología*, el cual es un término importado desde la filosofía donde se refiere a la teoría sobre la naturaleza

de la existencia. Para la representación del conocimiento, una ontología corresponde a una descripción explícita y formal de los conceptos para un dominio de discusión [NM01].

Para crear una ontología se deben definir los conceptos que se usarán, es decir las clases que la componen señalando posibles jerarquías de superclases y subclases, por ejemplo: Persona, Curso, Alumno, etc. Cada clase debe definir sus propiedades, es decir las diferentes características que la describen y que son de interés para el dominio de la ontología, por ejemplo: nombre, edad, estatura, etc. Las propiedades pueden tener restricciones sobre el conjunto de sus posibles valores, declarando que pueden contener textos, números, fechas, un valor de un conjunto predeterminado, etc.

Con una ontología creada se puede crear después una base de conocimientos, que corresponde a la información ingresada que cumple con la ontología. Para esto se deben crear las instancias individuales de cada clase ingresando sus propiedades correspondientes. Por ejemplo para una clase Persona se ingresan sus instancias que corresponderán a las personas que existirán en la base de conocimientos ingresando su nombre, edad y restantes propiedades definidas. Se pueden definir que el valor de una propiedad corresponda a una referencia a otra instancia de la misma o distinta clase. Así se puede ingresar información que relacione instancias, como por ejemplo se puede ingresar que una persona está inscrita en cierto curso o que tiene un parentesco con otra persona.

Las ontologías pueden ser definidas utilizando RDF, a través del lenguaje *RDF Schema*, o utilizando un lenguaje especializado para crear ontologías llamado *OWL*.

2.5. Dublin Core

Entre los modelos de metadatos usados comúnmente en Internet se ha destacado el *Dublin Metadata Core Element Set*, el cual es una lista básica de elementos para el registro de metadatos, que se diseñó inicialmente para ayudar a los motores de búsquedas a encontrar y recuperar páginas web y que ha ido evolucionando hacia un formato de registro para el intercambio y recuperación de información digital. Este modelo es el producto del primer encuentro sobre metadatos organizado en forma conjunta por NCSA (*National Center for Supercomputing Applications*) y OCLC (*On Line Library Computer Center*) en Dublin (Ohio, USA) en marzo de 1995. Desde entonces el interés por este modelo ha ido creciendo y se ha desarrollado a través de sucesivos encuentros o talleres que han servido para impulsarlo.

El objetivo de Dublin Core es definir un conjunto simple de elementos de dato para

que los autores y publicadores de documentos de Internet puedan crear sus propios registros sin gran entrenamiento previo. El enfoque de Dublin Core es tener un nivel de control bibliográfico intermedio entre un acercamiento estructurado y detallado como MARC y uno simple y automatizable como los utilizados por los buscadores de Internet. Además permite que cada autor tenga la capacidad de aumentar este conjunto con elementos más especializados en caso de necesitarlo.

El conjunto básico de Dublin Core contiene quince elementos y se puede dividir en tres grupos:

- Elementos relacionados con el contenido del documento. A este grupo pertenecen: título, materia, descripción, fuente, lengua, relación y cobertura/alcance.
- Elementos relacionados con la propiedad intelectual del documento. Son las etiquetas relativas al autor/creador, editor, otros autores/colaboradores y derechos de autor.
- Elementos relacionados con la temporalidad del documento. Estos elementos son: fecha, tipo de recurso, formato e identificador del recurso.

Además de estos elementos básicos, Dublin Core define otro conjunto de elementos que corresponde al refinamiento de éstos, y que permite describir cada documento con mayor precisión, conocidos como *calificadores*. Se espera que los autores utilicen estos calificadores e incluso creen sus propios elementos para otros dominios particulares.

Capítulo 3

Motivación y trabajo relacionado

A continuación se describen las características y problemas relacionados con Internet, con los metadatos y con los buscadores tradicionales que motivan el presente trabajo. En particular, se plantean cinco temas existentes en la actualidad además de los proyectos relacionados que influyeron en las decisiones tomadas para llevar a cabo este trabajo.

3.1. Características de la Web

Cualquier medio de catalogación debe tomar en cuenta la naturaleza de la web y sus principales dificultades: masividad, dinamismo y distribución:

- *Masividad*: La cantidad de información disponible en Internet es tan vasta y creciente que conocer su tamaño con precisión es imposible. Incluso, las estimaciones y predicciones hechas sobre el tamaño de Internet para la actualidad son muy menores al número de páginas indexadas por Google¹. Aún cuando los metadatos permiten efectuar índices sobre información seleccionada, normalmente de menor tamaño que la fuente misma, lograr crear un índice o catálogo completo de Internet es -siendo optimista- muy difícil.
- *Dinamismo*: Debido a la naturaleza electrónica de los documentos la información en Internet tiene poca estabilidad: los archivos se mueven o desaparecen, la información puede ser revisada y actualizada e incluso modificada completamente. En particular, se deben diferenciar las páginas que son inherentemente inestables (como páginas que

¹Google en la actualidad muestra en su página principal una cantidad de más de 8 mil millones de páginas indexadas.

presentan titulares), de las semi-estables (donde la información principal no cambia, pero sí su entorno como títulos, barras, etc.), de las estables (donde su modificación es poco frecuente como páginas estáticas, archivos binarios, etc.). Por esto, para que los metadatos no pierdan su utilidad con el tiempo, se debe proporcionar, en lo posible, un mecanismo para mantener coherente cada metadato con su documento, ya sea informando el estado o permitiendo actualizarlo.

- *Distribución*: Por ser de acceso igualitario y de crecimiento descentralizado, en la web no existen estándares para el vocabulario utilizado ni estándares de calidad ni control previo a efectuar una publicación. En este aspecto, los metadatos permiten asignar una estructura más rígida y estandarizada a la información publicada.

3.2. Buscadores sintácticos e intranets

Los motores de búsqueda tradicionales, aún cuando han mostrado avances sorprendentes, poseen falencias en el paradigma utilizado que quedan de manifiesto en la cantidad de resultados no útiles obtenidos y en el tiempo utilizado por un usuario para búsquedas de información [DFKO01].

En general, un usuario no deseará conocer las páginas donde aparecen ciertas palabras (a menos que desee hacer estadísticas) sino que desea encontrar una página que contenga cierta información que busca, quedando en sus manos hacer la relación entre lo que se desea buscar y las palabras claves que permiten encontrarlo. Esto puede ser muy dificultoso cuando el usuario no tiene conocimientos previos sobre el tema o cuando existen muchos sinónimos o diferentes significados de una palabra, y depende de las habilidades que tiene cada uno para hacer estas relaciones contenido-palabras claves.

A pesar de esto, 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 [Nie04]. 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 [Ame04].

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 [CHHL99, Ste99]. 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 [Key05].

Actualmente, para proveer de búsquedas específicas a un sitio, los publicadores deben decidir entre dos opciones:

- Usar un servicio de algún buscador público, el cual restringe su dominio de búsqueda a uno particular para el sitio. Con estos, el sitio puede dar un servicio de búsqueda local rápidamente y sin grandes necesidades de mantención o administración y con un costo reducido o incluso gratis. Sin embargo, este enfoque no puede utilizarse cuando la intranet tiene acceso privado (donde no puede acceder un robot externo), cuando se desee hacer búsquedas muy detalladas (en general no se indexan cada una de las páginas existentes), o cuando no se desee tener publicidad externa (para el caso de servicios gratuitos) [Sea01a].
- Instalar un motor de búsqueda particular, lo que permite tener un control mayor sobre parámetros de búsqueda y de resultados, pero hay que tener presente aspectos como instalación, mantención, recursos utilizados, precio, operabilidad, plataformas disponibles, etc. que deben ser evaluados bajo diferentes criterios [Sea01b, Ste99].

Un informe de Jakob Nielsen del año 2002 [Nie02] señala que las búsquedas en sitios de intranet tienen un pobre desempeño, independientemente de ser 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 pueden ser apreciadas a través de la visualización estándar de resultados [CHHL99].

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 % [Nie02].

3.3. Buscadores sintácticos con metadatos

Los metadatos presentan una opción de mejora a los buscadores sintácticos al permitir agregar a cada documento información estructurada. El enfoque más utilizado actualmente en Internet 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, este enfoque ha presentado algunos problemas para su desarrollo:

- Mentir sobre el contenido de un sitio, agregando grandes cantidades de metadatos falsos, con el objeto de mejorar la probabilidad de aparición dentro de los primeros resultados ha minado drásticamente las esperanzas sobre este enfoque. Incluso han debido intervenir las leyes de derecho de autor por acciones legales interpuestas por empresas dueñas de marcas registradas que son utilizadas usualmente como palabras claves sin la debida autorización [Sul02c].
- Las meta-etiquetas podrían ser creadas junto con cada nueva página a publicar, pero difícilmente podrían ser agregadas a todas las páginas ya existentes dado el tamaño actual de la web y la variedad de sus autores, por lo cual los beneficios que se podrían llegar a obtener tendrían efecto menor sobre las búsquedas actuales.
- Crear metadatos con el único fin de guiar buscadores sintácticos es, sin lugar a dudas, subutilizar las capacidades de los metadatos para los costos y dificultades que conlleva

crearlos. Si se emprende una tarea de asignar metadatos en la web debe ser posible hacer un mayor uso que sólo guiar búsquedas sintácticas.

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 [Sul02a].

3.4. Directorios Web

Un Directorio Web es una clasificación de los sitios existentes en Internet utilizando un árbol temático universal. La clasificación es realizada y mantenida por personas -ya sean voluntarios o contratados- los 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*².

Existen dos formas de obtener información desde un directorio: navegando sus categorías o buscando en sus contenidos. La navegación es efectuada a través de una visualización de la jerarquía de clasificación donde se pueden ver los sitios asignados a cada nodo. Las búsquedas son de palabras o frases dentro de los registros del directorio, donde el universo de búsqueda puede ser restringido a una categoría y sus descendientes.

Un Directorio Web por catalogar todos los sitios de Internet sin mediar más división que un árbol temático, presenta problemas en la cantidad de registros a clasificar por nodo versus la cantidad de nodos del árbol. No se puede clasificar demasiada cantidad de información en un nodo puesto que éste perdería sentido como buen discriminador, pero al tener un árbol demasiado grande (con muchos nodos) se hace difícil la navegación a través de él³.

Los Directorios Web han tenido mayor efectividad como apoyo a los buscadores sintácticos, para presentar temas y sitios relacionados como parte de sus resultados, más que como una nueva forma de efectuar búsquedas en Internet a través de la navegación de categorías clasificadas manualmente [CV01].

²<http://www.dmoz.org>, <http://dir.yahoo.com> y <http://www.looksmart.com>, respectivamente.

³El árbol de Open Directory Project actualmente contiene 590 mil categorías, con una anchura mayor a 40 categorías en algunos casos y con una profundidad mayor a 8 en algunas ramas.

3.5. Metadatos Universales

Lograr una asignación universal de metadatos de la web en forma minuciosa y confiable es una tarea que, a menos que el mundo y la sociedad actual cambiase, muy difícilmente podrá ser lograda [Doc01]. Por ejemplo, la primera dificultad es lograr definir la ontología que define el mundo con la cual se registrará toda la información importante de un recurso. La tarea es prácticamente imposible, al menos planteada de esa forma, ya que no existe sólo un punto de vista para definir el mundo. Incluso, aparecen más dificultades cuando se pueden obtener ventajas sobre los competidores al no cumplir con la fidelidad de los metadatos, como ya sucede cuando los buscadores sintácticos utilizan metadatos (ver sección 3.3).

En general, estas dificultades provienen de intentar catalogar el universo bajo una sola métrica y por diferentes personas que no necesariamente tendrán la intención de colaborar desinteresadamente por el bien común [Doc01].

Sin embargo, se pueden obtener buenos resultados cuando el objetivo de catalogación es reducido a un subconjunto de tamaño tratable y cuando los metadatos pertenecen a un mismo grupo-organización donde no debiese existir mayor competencia interna.

3.6. 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* [Nun03]. Entre los proyectos realizados por el Grupo Metadatos, el presente trabajo se relaciona con *DepMark* [MG02] 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* [CHHL99] 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* [Wor02, KK01] 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*⁴, el cual es un plug-in 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.

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 búsqueda en sitios de intranet. En este aspecto el presente 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*⁵ 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.

En el siguiente capítulo se presenta el modelo de catalogación propuesto, junto con sus características principales y formas de uso.

⁴<http://annozilla.mozdev.org>

⁵<http://metabrowser.spirit.net.au/>

Capítulo 4

Catalogación de un sitio web

El presente trabajo tiene por objetivo la creación de un sistema para la catalogación de sitios web utilizando técnicas de Web Semántica, el cual permita obtener mejores resultados que los que se obtienen actualmente al utilizar búsquedas tradicionales de recursos existentes en una intranet. Para esto se propone realizar una catalogación semi-automática del sitio, donde se crea un conjunto de metadatos sobre los contenidos de las páginas y recursos disponibles en el sitio, los que son organizados formando un catálogo.

Se entenderá por sitio web a un subconjunto de Internet cuyas páginas pertenecen a una misma entidad u organización (por ejemplo, un sitio corporativo o un sitio de intranet), la cual es la interesada en la creación de un catálogo de sus páginas y en la calidad de los datos que contenga. Por tanto la definición exacta de sitio web será una decisión que debe ser tomada por esta organización y normalmente estará dada por la unidad temática de las páginas. Así, se puede decidir crear un catálogo para todas las páginas bajo cierto dominio de Internet, para un conjunto de dominios de Internet o para un grupo de páginas que pertenecen a un portal.

Al existir el interés de una organización por la catalogación de un sitio y ser responsable de los datos en éste, se evitan las luchas por aparecer en el primer lugar de los resultados de búsquedas y se evitan los problemas derivados de intentar catalogaciones universales.

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

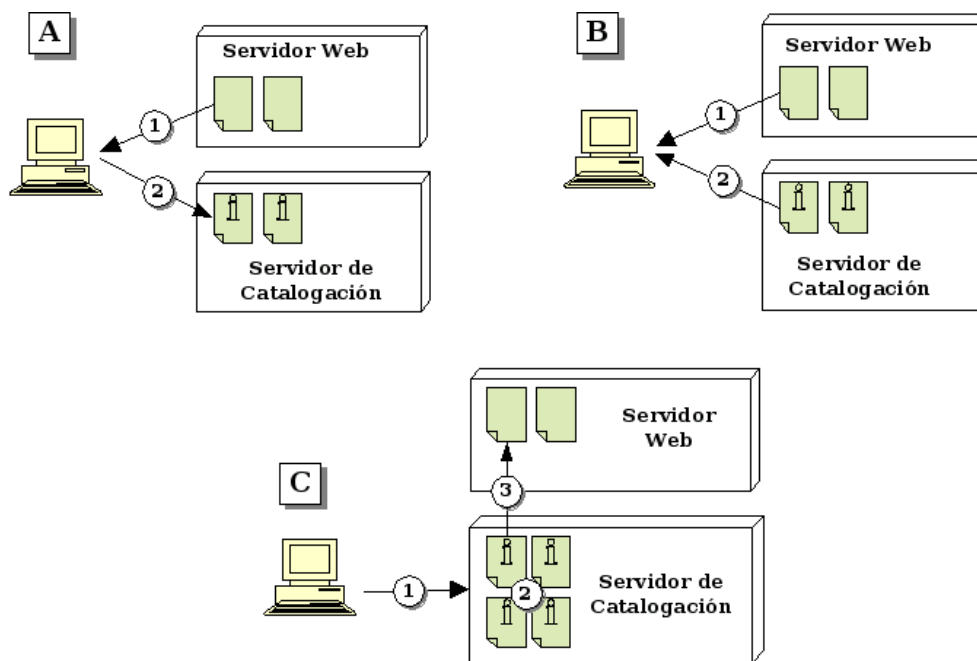


Figura 4.1: Acciones del sistema de catalogación. **A.** *Agregar metadatos.* 1. Ver Página. 2. Asignar metadatos a URL. **B.** *Apoyar la navegación.* 1. Ver Página. 2. Ver metadatos de URL. **C.** *Buscar recursos.* 1. Ingresar condiciones de búsqueda. 2. Búsqueda entre metadatos. 3. Redirigir hacia URL de los metadatos encontrados.

tipos de buscadores y navegadores de metadatos. El cliente de catalogación corresponde a una herramienta que reside en el computador del usuario como un plug-in 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 4.1 resume los tres posibles usos que permite el modelo de catalogación:

- Agregar metadatos (figura 4.1-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.
- Apoyar la navegación (figura 4.1-B), que corresponde a utilizar el cliente de catalogación para obtener los metadatos en el catálogo de cierta página que se esté visualizando.
- Buscar recursos (figura 4.1-C), que corresponde a realizar búsquedas o navegaciones en el catálogo para encontrar páginas con la información requerida.

A continuación se procederá a detallar el sistema de catalogación, sus características

y sus componentes.

4.1. 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. Sus tareas son:

- Definir el tamaño o los límites del sitio a catalogar.
- Definir quiénes serán los usuarios Catalogadores del sistema. Por ejemplo, pueden ser los autores de las páginas publicadas en el sitio, un grupo de personas especializadas en catalogación o un grupo abierto al público.
- Señalar los tipos de búsquedas especiales recurrentes que debe ser implementados en el sitio, en el caso que no se desee utilizar los buscadores genéricos.

Administrador Corresponde al usuario experto en el sistema de catalogación. Sus tareas 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 instancias para las clases definidas, la que puede ser manual a través de la herramienta de ingreso de instancias, o automática creando un *script* de creación de instancias.
- Realizar una carga inicial de metadatos de las páginas. Debe generar un *script* que ingrese masivamente metadatos iniciales para la mayor cantidad de páginas posibles.
- Implementar el o los buscadores especiales que haya definido el Dueño del sitio, utilizando como base los buscadores genéricos.

- Una vez que el sistema se encuentra en funcionamiento, monitorear el sistema a través de indicadores generales, revisando la calidad de los metadatos.

Catalogadores Corresponde al grupo de usuarios encargados de mantener los metadatos en el catálogo. Estos usuarios deben tener conocimientos básicos sobre catalogación de páginas, la herramienta de mantención de metadatos y el esquema de metadatos usado en el sitio. Sus tareas son:

- Crear y revisar manualmente los metadatos existentes en el sistema, utilizando la herramienta desarrollada para esto.
- Crear y revisar manualmente las instancias de las clases existentes en el sistema, utilizando la herramienta desarrollada para esto.
- Recibir y procesar las notificaciones de metadatos erróneos en una página.

Público General Corresponde al usuario que visita el sitio y utiliza el sistema para buscar recursos en él. No tiene conocimiento previo del sistema. Sus posibles 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.
- Notificar posibles datos erróneos o imprecisos existentes en el catálogo. Estos reportes los reciben los usuarios catalogadores para procesarlos.

4.2. Servidor de catalogación

El servidor de catalogación se encarga de la persistencia de la información del catálogo y de publicar aplicaciones para que los diferentes tipos de usuarios puedan hacer uso del catálogo. Proporciona al menos tres funcionalidades:

- Administración. Permite que un usuario Administrador defina la o las ontologías en el sistema, es decir, la estructura que se utilizará para el registro de metadatos. Permite

además definir los valores de configuración del sistema y puede presentar indicadores sobre el uso del sistema y la calidad de los metadatos.

- **Catalogación.** Permite que un usuario Catalogador ingrese y mantenga los metadatos de los recursos catalogados y de los objetos para las ontologías definidas.
- **Búsquedas.** Permite que el público general que visita un sitio pueda utilizar alguno de los diferentes tipos de buscadores de recursos presentados o pueda usar alguno de los visualizadores de información existente en el catálogo.

El servidor de catalogación proporciona un servicio independiente y externo a un servidor web, por lo que puede residir en el mismo servidor o en un servidor independiente de las páginas web catalogadas.

4.3. Cliente de catalogación

El cliente de catalogación corresponde a la herramienta que reside en el computador del cliente como un plug-in para un navegador. Permite efectuar acciones sin necesidad de acceder directamente al servidor de catalogación y manteniendo preferencias particulares del usuario. Las acciones que puede realizar un usuario dependen del perfil al que pertenezca:

- **Catalogadores.** Permite revisar los metadatos asignados a las páginas que esté visitando. En el caso que encuentre que los metadatos son incorrectos, antiguos o inexistentes puede actualizarlos o ingresar unos nuevos.
- **Público general.** Permite visualizar los metadatos pertenecientes a la página que se esté visitando. Puede permitir además invocar búsquedas sobre el servidor de catalogación las cuales pueden contener parámetros predefinidos por el usuario, como restricciones sobre ciertos campos o cantidad máxima de resultados.

4.4. Esquemas de metadatos

Para poder ingresar metadatos y realizar búsquedas en el catálogo, es necesario que el usuario administrador defina previamente el esquema de metadatos a utilizar. Un esquema de metadatos corresponde a la definición de un conjunto de campos de diferentes tipos, organizados en forma de árbol, con el objetivo de registrar y organizar los metadatos

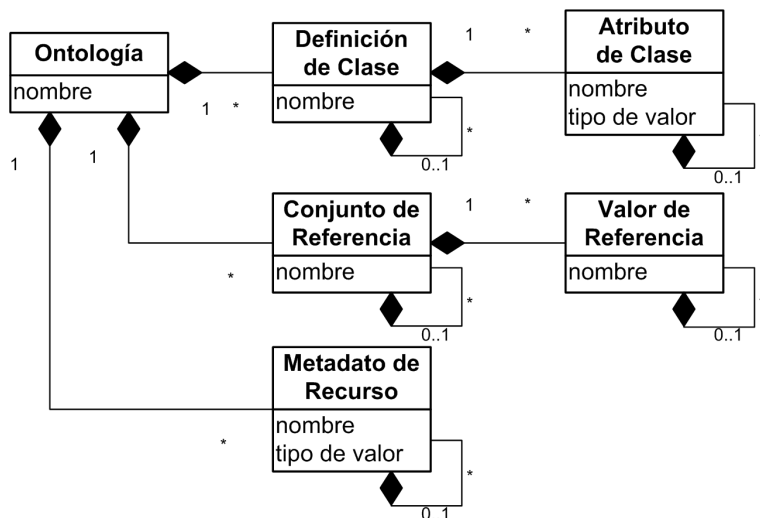


Figura 4.2: Modelo que cumple cada esquema de metadatos definido en el servidor de catalogación.

en el catálogo. Los seis posibles tipos de campos en un esquema de metadatos y sus relaciones entre sí son descritos en la figura 4.2:

- **Ontología.** Es el campo raíz del esquema de metadatos, permite agrupar los campos en una sola unidad temática. Existen tantos esquemas independientes como campos tipo Ontología se hayan definido. Puede contener tres tipos de hijos: Definición de Clase, Conjunto de Referencia y Metadato de Recurso.
- **Definición de Clase.** Permite definir una clase o una subclase dentro de una Ontología para crear una jerarquía de clases. Por ejemplo, una ontología podría contener un campo tipo Definición de Clase de nombre “Persona”, el cual a su vez puede contener los campos tipo Definición de Clase “Alumno”, “Funcionario” y “Académico”. Una vez creado el esquema, es tarea del usuario catalogador administrar las instancias correspondientes para las clases definidas en cada ontología.
- **Atributo de Clase.** Permite definir un atributo para una clase definida en una ontología. Por ejemplo, el campo tipo Definición de Clase de nombre “Persona” podría contener como hijo los campos tipo Atributo de Clase “Nombre” y “Fecha de Nacimiento”, la clase “Alumno” podría contener un campo tipo Atributo de Clase “Carrera que cursa”, la clase “Funcionario” un atributo “Cargo”, la clase “Académico” un atributo “Área de investigación”, etc.
- **Metadato de Recurso.** Permite definir cada aspecto que puede ser extraído de los recursos existentes en catálogo, es decir, corresponde a los metadatos que serán creados

por los usuarios catalogadores para catalogar los recursos. Cada campo puede ser hijo de otro campo tipo Metadato de Recurso para definir una especificación del campo, según el sentido dado para la refinación de elementos de los calificadores de Dublin Core. Por ejemplo, una ontología podría definir para la catalogación de recursos los campos “Título”, “Resumen” y “Autor” y podría definir el campo “Título Alternativo” que correspondería una refinación del campo “Título”.

- **Conjunto de Referencia.** Permite definir un conjunto de valores que pueden ser asignados a un campo tipo Atributo de Clase o Metadato de Recurso. Por ejemplo, una ontología podría contener un campo tipo Conjunto de Referencia de nombre “Áreas de investigación” donde se registren jerárquicamente distintas áreas de investigación que luego pueden ser asignadas como valor al atributo “Área de investigación” de la clase “Académico”.
- **Valor de Referencia.** Permite definir cada posible valor dentro de un Conjunto de Referencia. Cada campo debe ser hijo de un campo tipo Conjunto de Referencia o de un campo tipo Valor de Referencia en el caso de definir una especificación de éste. Por ejemplo, para el Conjunto de Referencia “Áreas investigativas” se pueden definir campos de tipo Valor de Referencia con los nombres “Algoritmos”, “Inteligencia Artificial”, etc. y campos de especificación como “Sistemas Expertos”, “Redes Neuronales”, etc.

Cuando se define un campo tipo Atributo de Clase o Metadato de Recurso se debe definir además el tipo del valor asignable al momento de la catalogación. Existen cinco posibles tipos para el valor que puede ser asignado a estos campos:

- **Texto.** Es un campo cuyo valor es un texto sin restricciones. Por ejemplo, un atributo “Nombre” o un metadato “Título” tendrían un tipo de valor Texto.
- **Fecha.** Es un campo cuyo valor corresponde a una fecha seleccionable de un calendario. Por ejemplo, un atributo “Fecha de nacimiento” o un metadato “Fecha de creación” tendrían un tipo de valor Fecha.
- **URL.** Es un campo cuyo valor representa una referencia hacia un recurso web. Por ejemplo, un atributo “Página personal” o un metadato “Información relacionada” podrían tener un tipo de valor URL donde su valor sería la URL del recurso asociado (el cual puede ser parte o no del catálogo).
- **Elección de Referencia.** Es un campo cuyo valor corresponde a un campo tipo Valor de Referencia que pertenece a un Conjunto de Referencia determinado. Por ejemplo, el

atributo “Área de investigación” de la clase “Académico” podría tener un tipo de valor Elección de Referencia seleccionando un valor del Conjunto de Referencia “Áreas de investigación”.

- **Elección de Instancia.** Es un campo cuyo valor corresponde una instancia creada para una Definición de Clase determinada. Por ejemplo, un metadato “Autor” podría tener un tipo de valor Elección de Instancia seleccionando como valor una instancia de la clase “Persona” o alguna de sus subclases.

Un ejemplo más detallado de un esquema de metadatos puede ser visto en el esquema creado para la instalación del sistema en el caso de estudio (capítulo 6).

4.5. Mantención de metadatos

El usuario tipo Catalogador realiza el ingreso, modificación y eliminación de metadatos en el catálogo. Existen dos posibles formas de realizar la mantención de datos, a través del cliente de catalogación y a través del servidor de catalogación:

Cliente de catalogación El usuario Catalogador ingresa al sitio web con su navegador y cuando desee agregar, modificar o eliminar metadatos asociados a una página que esté visitando utiliza el cliente de catalogación previamente instalado en su navegador. El cliente de catalogación solicita los metadatos de la página al servidor de catalogación y los muestra al usuario para que realice modificaciones sobre éstos.

En el caso que el metadato a ingresar o modificar sea del tipo Elección de Referencia o Elección de Instancia el cliente consulta al servidor los valores posibles y los muestra al usuario para su elección. Una vez concluida la modificación el cliente envía los datos al servidor para la actualización de los metadatos de la página.

Servidor de catalogación Cada recurso dentro del catálogo es identificado según la URL con que está publicado en el sitio web. El usuario Catalogador debe ingresar a la aplicación de mantención de metadatos e ingresar la URL de la página que desee agregar o actualizar. El servidor muestra los metadatos existentes para la URL dada y permite al usuario agregar, modificar o borrar metadatos.

El servidor de catalogación puede tener implementado un algoritmo particular al sitio

catalogado que permita obtener automáticamente un conjunto de metadatos a partir del texto de una página web publicada. Estos *scripts* permiten realizar una carga inicial automática de metadatos a partir de los cuales comenzar a catalogar el sitio web. Una vez que el sistema de catalogación se encuentre en funcionamiento estos scripts permiten realizar una *sincronización* entre los metadatos automatizables y la página web a la que pertenecen. Normalmente la sincronización corresponderá a los metadatos que contienen extractos textuales de la página como el título, subtítulo o las URL de referencias a otras páginas evitando que éstos sean mantenidos por los usuarios catalogadores.

4.6. Visualizar metadatos

Un usuario visitante puede visualizar los metadatos asociados a una página a través del cliente de catalogación. Para esto debe instalar esta herramienta en su navegador y utilizarla para consultar por los metadatos de cada página. El cliente podría además monitorear la navegación del usuario y consultar periódicamente por los metadatos de cada página que se encuentre visitando.

Los datos recuperados pueden servir de información extra sobre el contenido, como autor, fecha o área temática, o puede servir de guía para la navegación al presentar el contexto en el cual existe la página, como páginas que la referencian y páginas referenciadas.

4.7. Búsqueda de recursos

Las búsquedas de recursos corresponden a realizar consultas sobre el catálogo para recuperar un conjunto de páginas que cumplan una o más condiciones. Las búsquedas pueden ser divididas en tres grupos según la forma en la cual se definen las restricciones:

Búsquedas por palabras claves El usuario ingresa primero el texto a buscar y luego especifica el conjunto de campos del esquema sobre el cual desea realizar la búsqueda. Este tipo de buscador sería similar a un buscador sintáctico tradicional donde se puede restringir su dominio de búsqueda a cierto conjunto de metadatos. Cada valor de metadato dentro del dominio de búsqueda debe ser convertido automáticamente a un texto según el tipo de valor del campo al que pertenece:

- *Texto*. Se utiliza el valor del metadato.

- *Fecha*. Se convierte según el formato de fecha definido en la configuración local del servidor (*dd-mm-yyyy*, *yyyy-mm-dd*, etc.).
- *URL*. Se utiliza el valor del metadato.
- *Elección de Referencia*. Se utiliza el nombre del campo Valor de Referencia seleccionado.
- *Elección de Instancia*. Se utiliza el nombre de la instancia seleccionada. Se define el nombre de una instancia como la concatenación de la conversión a texto de todos sus atributos sin incluir los atributos tipo Elección de Instancia.

Búsquedas por valor de metadato El usuario especifica primero el campo del esquema para realizar la búsqueda y luego la aplicación despliega una forma de ingreso acorde con el tipo de valor del campo seleccionado:

- *Texto*. Despliega un campo de texto. Realiza una búsqueda similar a una búsqueda por palabras claves.
- *Fecha*. Despliega un calendario para seleccionar una fecha. Al desplegar un calendario se evita que el usuario deba conocer el formato que se utiliza para el registro de fechas (*dd-mm-yyyy*, *yyyy-mm-dd*, etc.).
- *URL*. Despliega un campo de texto y verifica que el valor ingresado tenga una forma válida. En este caso la lista de resultados corresponde a todas las páginas catalogadas que contienen una referencia a la URL ingresada. Esto permite conocer el contexto al cual pertenece cierto recurso (no necesariamente igual al de navegación) y permite encontrar información relacionada con éste, lo que es especialmente útil para el caso de archivos binarios.
- *Elección de Referencia*. Se despliega una selección de los valores de referencia disponibles para el Conjunto de Referencia que defina el campo.
- *Elección de Instancia*. Se despliega un listado de las instancias disponibles para realizar la búsqueda. Este listado de instancias, con el correr del tiempo, puede llegar a ser muy extenso. En ese caso es mejor que el sistema permita hacer una búsqueda anidada según un valor en alguno de los atributos de la instancia.

Navegación jerárquica del catálogo El usuario puede recorrer la jerarquía de clases del esquema de metadatos y al seleccionar una clase se listan las instancias que pertenecen a ella. Luego, el usuario elige una instancia a partir de la cual se listan todos los recursos del catálogo que contienen algún metadato con ésta como valor. Esta navegación jerárquica permite crear un índice temático de páginas donde se muestran las páginas asociadas a cierto tema en particular de una forma similar a los Directorios Web [MÃ©02, p. 236].

Cabe señalar que el cliente de catalogación podría implementar uno o varios de estos buscadores de recursos, el cual podría contener preferencias particulares al usuario. Por ejemplo, el usuario podría preferir siempre los metadatos asociados a cierto esquema o las páginas que contengan cierto valor entre sus metadatos.

4.8. Ordenamiento de resultados

Todas las búsquedas de recursos deben retornar el conjunto de recursos que cumplen con las condiciones de búsqueda especificadas por el usuario, ordenados según el nivel de relevancia de cada uno para la búsqueda en particular.

Se propone como algoritmo de ordenamiento de resultados uno basado en puntos de ranqueo, según el cual se aplica un ordenamiento de mayor a menor según el puntaje conseguido por cada recurso para cada búsqueda. Es decir, si r es un recurso en el catálogo y b las condiciones de una búsqueda, el ordenamiento de los resultados se debe realizar según el puntaje de ranqueo $P(r, b)$ calculado para cada recurso encontrado según las condiciones de la búsqueda en particular.

Para calcular este valor, cada campo del esquema de metadatos debe tener asignado un valor de ranqueo, que debe ser ingresado por el usuario administrador en el momento de la definición del esquema. Sea C un campo del esquema de metadatos y M un metadato en el catálogo, se define:

- $tipo(C)$ como el tipo de valor del campo C . Los posibles valores para $tipo(C)$ son: *texto*, *fecha*, *URL*, *elección referencia* o *elección instancia*.
- $puntaje(C)$ como el valor de ranqueo asignado al campo C . Es un valor definido por el administrador del sistema al crear cada campo del esquema de metadatos.
- $recurso(M)$ como el recurso del catálogo al que pertenece el metadato M .

- $campo(M)$ como el campo del esquema de metadatos al que está asociado el metadato M .
- $valor(M)$ como el valor asignado al metadato M . El valor asignable a un metadato depende de su tipo de valor, es decir, depende de $tipo(campo(M))$.
- $id(r)$ como el identificador del recurso r en el catálogo, el cual corresponde a la URL con la cual se puede encontrar en la web.
- M_r como el conjunto de metadatos que pertenecen al recurso r , es decir:

$$M_r = \{M \in Catalogo \mid recurso(M) = r\}$$

Cada recurso contiene un puntaje de ranqueo estático S_r el cual es calculado como la suma de los puntos de ranqueo de los campos tipo *URL* cuyo valor es igual al identificador del recurso. Así, un recurso puede recibir referencias de distinta valoración según la calidad o importancia de ésta, según los puntos de ranqueo del campo que contiene la referencia. Es decir:

$$S_r = \sum_{M \in R_r} puntaje(campo(M)) \text{ , donde:}$$

$$R_r = \{M \in \{Catalogo \setminus M_r\} \mid tipo(campo(M)) = URL \wedge valor(M) = id(r)\}$$

Por otra parte, a cada recurso encontrado en una búsqueda se le asigna un puntaje dinámico $D_{r,b}$, calculado como la suma de los valores de los campos cuyos metadatos cumplen con las condiciones de la búsqueda. Así, si un valor buscado se encuentra en varios metadatos de una misma página el puntaje dinámico corresponderá a la suma de los puntos de ranqueo de los campos a los que corresponden esos metadatos. Es decir:

$$D_{r,b} = \sum_{M \in B_{r,b}} puntaje(campo(M)) \text{ , donde:}$$

$$B_{r,b} = \{M \in M_r \mid valor(M) \text{ cumple con } b\}$$

Finalmente, los recursos son ordenados según la cantidad total de puntos alcanzados, que corresponde a la suma de los puntos de ranqueo estáticos del recurso más los puntos de ranqueo dinámicos asociados a la búsqueda particular. Es decir:

$$P(r, b) = S_r + D_{r,b}$$

Como el valor S_r no depende de una búsqueda en particular si no que de los metadatos existentes en el catálogo, puede ser precalculado para cada recurso catalogado. Además S_r puede ser mantenido en el tiempo, actualizando su valor cada vez que un usuario ingrese o modifique un metadato que pertenezca a un campo de tipo *URL* y que contenga o haya contenido una referencia a r .

4.9. 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 Dueño del sitio debe decidir el tamaño del sitio a catalogar, basado en la información publicada que desee utilizar como universo de las búsquedas.
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 los esquemas de metadatos a utilizar. Para esto debe decidir los metadatos a ingresar para 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.
5. El usuario Administrador debe hacer una carga inicial de metadatos para las páginas del sitio, incluyendo la mayor cantidad de campos del esquema. En lo posible debe contener el título de la página, las referencias entre páginas y las asociaciones con las instancias ya creadas.
6. El usuario Dueño del sitio 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 ordenamiento 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. Programa y realiza sincronizaciones periódicas de los metadatos automatizables.

En el siguiente capítulo se presenta una implementación del sistema de catalogación y luego su implantación de prueba para un sitio de Internet.

Capítulo 5

Sistema Catalogo

Catalogo es el sistema de catalogación desarrollado en el marco de este trabajo con el objetivo de demostrar las capacidades del modelo presentado para realizar búsquedas de recursos en un sitio web.

Está compuesto de un servidor de catalogación el cual contiene y administra los metadatos, y un cliente de catalogación el cual permite acceder al servidor de catalogación. Contiene tres posibles tipos de usuario: administrador, catalogador y visitante. La forma de uso del software es la siguiente:

- El usuario administrador ingresa al servidor de catalogación para definir los esquemas de metadatos a registrar de las páginas.
- El usuario administrador realiza una carga automática con metadatos iniciales de las páginas.
- Los usuarios catalogadores ingresan y mantienen metadatos referentes a cada recurso del sitio, utilizando la aplicación web correspondiente o el cliente provisto por el sistema.
- Los usuarios visitantes que deseen hacer búsquedas de recursos del sitio pueden acceder a la aplicación web de búsquedas y utilizar alguno de los buscadores ahí presentados.
- El usuario administrador monitorea el sistema a través de los indicadores de calidad de metadatos.

Catalogo es software libre (con licencia GPL) y fue aceptado para pertenecer a los proyectos del portal *SourceForge.net*. A través de la página <http://catalogo.sourceforge.net> se pue-



Figura 5.1: Página principal módulo de administración.

de obtener documentación del proyecto, documentación del modelo de catalogación, bajar versiones con binarios y código fuente para contribuir al desarrollo del software a través de la comunidad del software libre.

A continuación se describe el servidor de catalogación y el cliente de catalogación. Para la descripción se utilizarán pantallas que pertenecen a la instalación del sistema para el caso de estudio (capítulo 6).

5.1. Servidor de catalogación

El servidor de catalogación se compone de un conjunto de aplicaciones desarrolladas utilizando tecnología Java sobre un servidor web y una base de datos relacional para la persistencia de los esquemas y metadatos. Contiene tres aplicaciones web: Administración, Catalogación y Búsqueda, las que pueden ser accedidas por un usuario tipo Administrador, Catalogador y Visitante, respectivamente.

Los requerimientos técnicos mínimos para instalar el servidor de catalogación son los siguientes:

- J2SDK 1.4.
- Jakarta Tomcat 4.1.
- PostgreSQL 7.2.

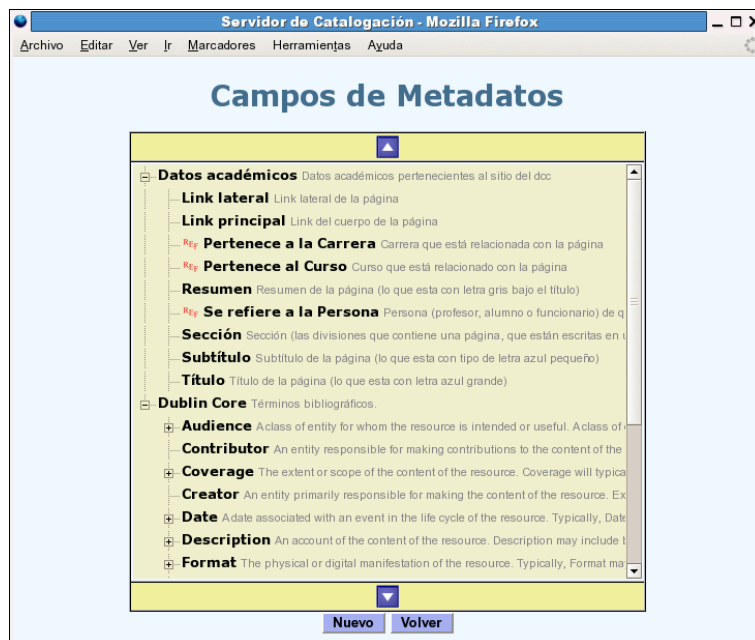


Figura 5.2: Listado de campos de metadatos utilizado para el esquema del caso de estudio.

5.1.1. Módulo de Administración

El módulo de administración permite definir los esquemas a utilizar para el registro de metadatos, es decir, permite mantener las ontologías de los metadatos del sistema. A este módulo sólo puede acceder un usuario administrador.

La figura 5.1 muestra el menú principal que contiene tres opciones para administrar los esquemas de metadatos: definir los campos de metadatos, definir los campos de definiciones de clase y definir los campos de grupos de referencia.

Definición de ontologías

La figura 5.2 muestra la visualización de los campos de metadatos para el esquema utilizado para el caso de estudio. La presentación de los campos es en forma de árbol, donde cada raíz corresponde a un esquema de metadatos diferente. El icono **REF** señala que el campo es de tipo Elección de Instancia en el caso de ser rojo o es de tipo Elección de Referencia en el caso de ser azul. Presionando sobre un campo se puede modificar su definición. Presionando sobre el botón *Nuevo* permite crear un nuevo campo raíz que corresponderá a una ontología.

La figura 5.3 muestra la pantalla de creación y modificación de campos. Los campos

Campo	
Nombre	Pertenece a la Carrera
Descripción	Carrera que está relacionada con la página
Tipo	Elección de Instancia Campo Ref.: Carrera
Puntos ranking	10
Campo padre	Datos académicos
Exportar RDF/XML	Nombre: carrera URL Namespace: http://www.dcc.uchile.cl/catalogo/ Nombre Namespace: cat
Exportar Meta-Tag	Nombre: CAT.carrera

Actualizar Borrar Nuevo Volver

Figura 5.3: Definición de campo del esquema de metadatos.

que componen cada esquema de metadatos deben pertenecer a alguno de los siguientes tipos: Ontología, Definición de Clase, Conjunto de Referencia, Valor de Referencia, Texto, Fecha, URL, Elección de Referencia y Elección de Instancia (ver sección 4.4). En el caso que el tipo de campo corresponda a Elección de Instancia o Elección de Referencia se debe seleccionar la clase a la que deben pertenecer las instancias o el grupo de referencias posibles, respectivamente, presionando sobre el link *Campo Ref.* (ver figura 5.4).

Para cada campo se debe definir también cual es su nombre de exportación a meta-etiqueta, cual es su nombre y namespace de exportación a RDF/XML, y cual es su valor de ranqueo para las búsquedas (ver sección 4.8).

Indicadores

La figura 5.5 muestra la pantalla donde se muestran los indicadores que presenta el sistema sobre los datos que contiene el catálogo. Estos son:

- Páginas catalogadas. El número absoluto de páginas que contiene el catálogo.
- Páginas sin metadato. El porcentaje de páginas que estando registradas en el catálogo, no contienen ningún tipo de metadatos. Un valor mayor a 0% significa que existen páginas que estando dentro del catálogo no serán encontradas en ninguna búsqueda.

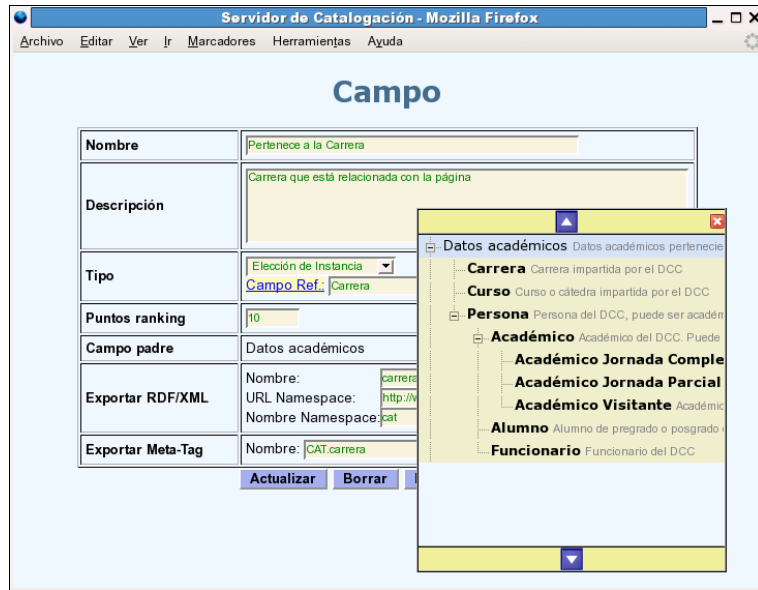


Figura 5.4: Selección de campos de referencia para campos de tipo Elección de Instancia o Elección de Referencia.

- Páginas sin metadatos de tipo textos. El porcentaje de páginas que estando registradas en el catálogo, no contienen metadatos de tipo Texto, Fecha o URL. Un valor mayor a 0 % puede significar un defecto en la catalogación, ya que normalmente todas las páginas tienen al menos un título o resumen.
- Páginas sin metadatos de tipo referencia ni instancia. El porcentaje de páginas que estando registradas en el catálogo, no contienen metadatos de tipo Elección de Instancia o Elección de Referencia. Cuando este valor se acerca a 100 % significa que el catálogo sólo contiene metadatos de texto lo que lo convierte en un buscador sintáctico.

Páginas	
Páginas catalogadas:	401
Páginas sin metadatos:	0%
Páginas sin metadatos de tipo textos:	0.249%
Páginas sin metadatos de tipo referencia ni instancia:	68.08%
Instancias definidas:	99
Instancias no usadas:	1.01%

[Volver](#)

Figura 5.5: Indicadores de cantidad y calidad de los datos en el catálogo.



Figura 5.6: Página principal módulo de catalogación.

- Instancias definidas. El número absoluto de instancias que se han definido dentro del catálogo.
- Instancias no usadas. El porcentaje de instancias que han sido definidas y que no han sido utilizadas como metadato de ninguna página. Un valor mayor a 0% significa que hay instancias que deben ser asignadas a páginas o que en caso contrario debiesen ser eliminadas.

5.1.2. Módulo de Catalogación

El módulo de catalogación permite crear, actualizar y eliminar los metadatos e instancias que existen en el sistema. A este módulo sólo puede acceder un usuario administrador o catalogador.

La figura 5.6 muestra el menú principal que permite ir al mantenedor de instancias y al mantenedor de metadatos de páginas.

Mantenimiento de instancias en una ontología

Al ingresar al mantenedor de instancias, se despliega inicialmente la lista completa de definiciones de clases del sistema ordenadas jerárquicamente. Cuando el usuario selecciona una clase se despliega una lista de todas las instancias que pertenecen a esa clase o a alguna de sus subclases ordenada en forma de árbol. Ver figura 5.7.

En esta pantalla se pueden modificar los atributos de una instancia presionando sobre ella, o se puede crear una nueva instancia presionando sobre la clase a la que pertenece.

La figura 5.8 muestra la pantalla de modificación de atributos de instancia. En la parte superior muestra la clase a la que pertenece la instancia (en el ejemplo “Académico

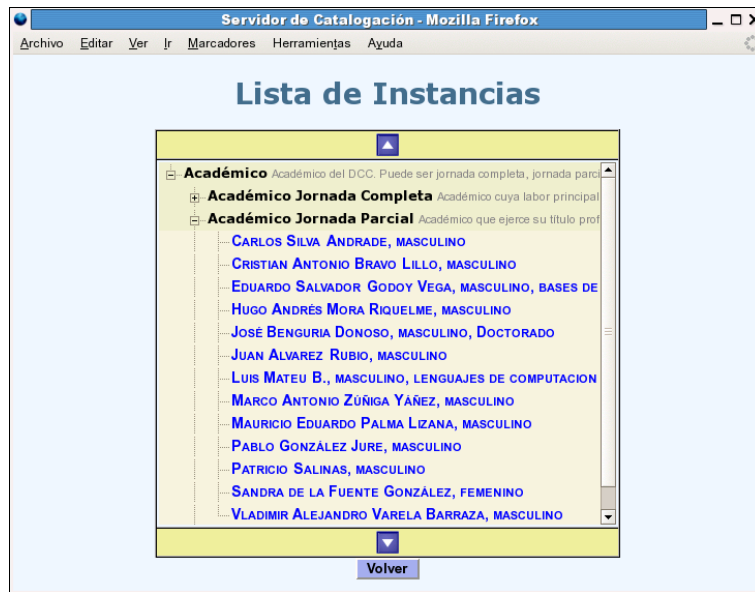


Figura 5.7: Lista jerárquicamente de clases e instancias definidas para una ontología.

Jornada Parcial”) y más abajo el nombre de la instancia, el cual es un resumen de todos los atributos de tipo texto. Cuando se realizan búsquedas sintácticas se utiliza el nombre de la instancia como el valor sobre el cual buscar.

En el caso de agregar o modificar atributos de tipo Elección de Instancia o Elección de Referencia se despliega un cuadro con la lista de posibles valores. En el ejemplo se muestra la creación de un atributo para el campo “Área de Investigación” donde se despliega un listado de posibles valores de referencia.

Un atributo puede ser eliminado primero seleccionándolo y luego presionando sobre el botón *Borrar*, actualizando automáticamente el nombre de la instancia. En el caso de eliminar el último atributo se elimina además la instancia misma, lo que es evitado en el caso que la instancia esté referenciada como valor de metadato de una página o como atributo en otra instancia.

Mantenimiento de metadatos de una página

Si el usuario catalogador desea modificar los metadatos asignados a una página o desea agregar una nueva página al catálogo debe seleccionar el mantenedor de metadatos de páginas desde el menú principal e ingresar la URL correspondiente.

En la figura 5.9 se muestra la pantalla de metadatos de una página. En la parte

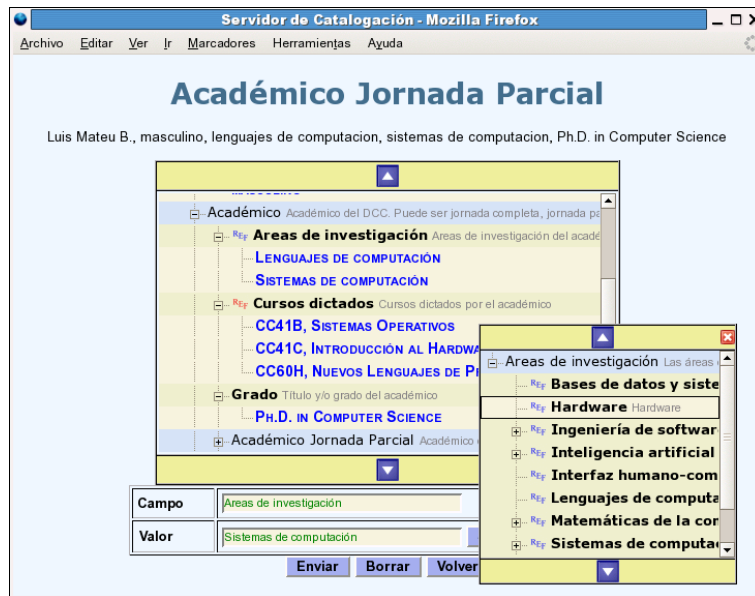


Figura 5.8: Modificación de los atributos de una instancia.

superior se despliega su URL, el título (que es el que se muestra en la pantalla de resultados de búsqueda), su valor de ranqueo base calculado (ver sección 4.8) y la fecha de última actualización de los metadatos. En el recuadro inferior se muestran los metadatos asignados a la página en una vista de árbol, los que pueden ser exportados a meta-etiquetas o RDF/XML o pueden ver vistos en la ficha de resumen de la página. Al presionar el botón *Borrar* se elimina el registro de la página junto con todos sus metadatos, actualizando el valor de ranqueo base de las páginas que hayan recibido referencias de ésta.

La figura 5.10 muestra la pantalla que permite modificar los metadatos de una página. En la parte superior se muestra el título de la página y en la parte central se muestra en forma de árbol el conjunto de campos de metadatos existentes en el sistema y como hijo de éstos los metadatos asignados para cada campo.

Para agregar un metadato a la página se debe seleccionar el campo al cual debe corresponder, y para modificar un metadato existente se debe presionar sobre él. Según el tipo del campo de metadato a agregar o modificar se muestra en la parte inferior de la página un cuadro de texto, un calendario, un listado de instancias o un listado de valores de referencia. Al modificar un campo tipo URL se actualizan los valores de ranqueo base a las páginas correspondientes. En el caso de eliminar todos los metadatos, el registro de la página no es eliminado sino que queda sin datos asociados, lo que queda de manifiesto en los indicadores del sistema.

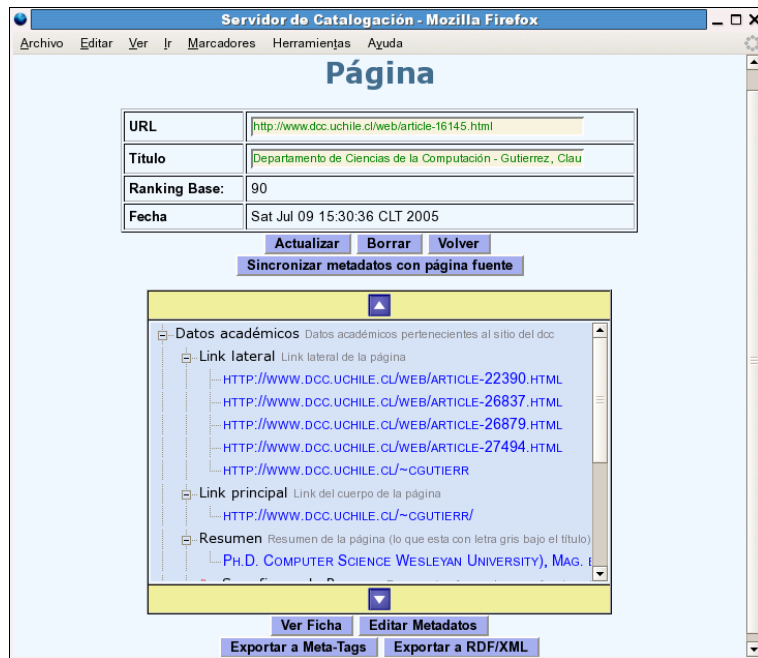


Figura 5.9: Visualización de metadatos asignados a una página.

La mantención de metadatos de página se puede realizar también con el cliente de catalogación, como se explica en la sección 5.2.

Sincronización de metadatos

La sincronización de metadatos permite definir un algoritmo para la obtención automática de metadatos a partir de una página web publicada. Este algoritmo es utilizado para realizar una carga inicial de metadatos en el catálogo y luego para realizar la mantención de metadatos en el tiempo.

Catalogo define una interfaz para la implementación de catalogadores automáticos, para lo cual se debe implementar un método que dado un recurso publicado retorne los metadatos correspondientes a éste. Cada catalogador automático debe definir además el conjunto de recursos sobre el cual aplica y los campos de la ontología del catálogo que automatiza. Catalogo provee la implementación de un catalogador automático genérico basado en expresiones regulares configurable a través de un archivo XML.

Para configurar la sincronización de metadatos, se debe señalar las clases de los catalogadores a utilizar y sus parámetros de configuración (recursos sobre los que aplica, campos de ontología automatizados y otros particulares de cada implementación). El apéndice A con-

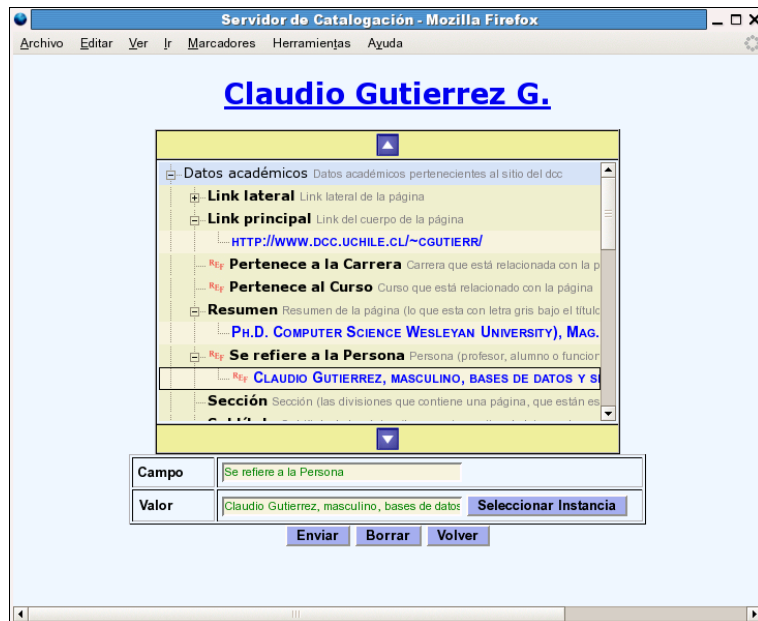


Figura 5.10: Modificación de metadatos de una página.

tiene un ejemplo del archivo XML utilizado por Catalogo para configurar la sincronización.

Una vez configurada la sincronización de metadatos, en la visualización de metadatos de una página se puede utilizar el botón *Sincronizar metadatos con página fuente* (ver figura 5.9) con el cual se procede a buscar un catalogador que aplique para la URL de la página seleccionada. En el caso de encontrar uno, el servidor de catalogación baja desde la red el recurso publicado e invoca el método del catalogador para obtener los metadatos del recurso. Si existen diferencias entre los metadatos calculados y los existentes dentro del catálogo se pide confirmación al usuario para actualizar los metadatos que difieren (ver figura 5.11).

Además de esta sincronización unitaria, se puede realizar una sincronización de metadatos en forma masiva. Para esto se debe ejecutar desde la línea de comandos la clase *catalogo.server.sincro.Sincronizador* entregando como parámetros el directorio local donde se encuentran las páginas a sincronizar y la URL base con la cual se encuentran publicadas. El directorio puede corresponder a una copia local del sitio publicado (por ejemplo, bajado con *wget*), o en el caso que el servidor de catalogación y el servidor web se encuentren físicamente juntos, al directorio mismo donde se publican las páginas. Esta sincronización masiva, al ser a través de la línea de comandos, puede fácilmente ser programable para ser realizada cada cierto periodo de tiempo.



Figura 5.11: Sincronización de metadatos entre un recurso publicado y los metadatos existentes dentro del catálogo.



Figura 5.12: Página principal módulo de búsquedas.

5.1.3. Módulo de Búsquedas

Permite que cualquier usuario pueda realizar búsquedas de los recursos catalogados y visualizar sus metadatos. La figura 5.12 muestra el menú principal que permite ir a los diferentes buscadores y visualizadores de metadatos.

Para los casos en que realizan búsquedas de texto se aplican las siguientes condiciones:

- Se realiza una búsqueda de todas las palabras en un mismo metadato.
- Para buscar una frase se debe ingresar entre los caracteres “ ”.
- La búsqueda no distingue entre mayúsculas y minúsculas.
- La búsqueda es sólo sobre palabras completas.

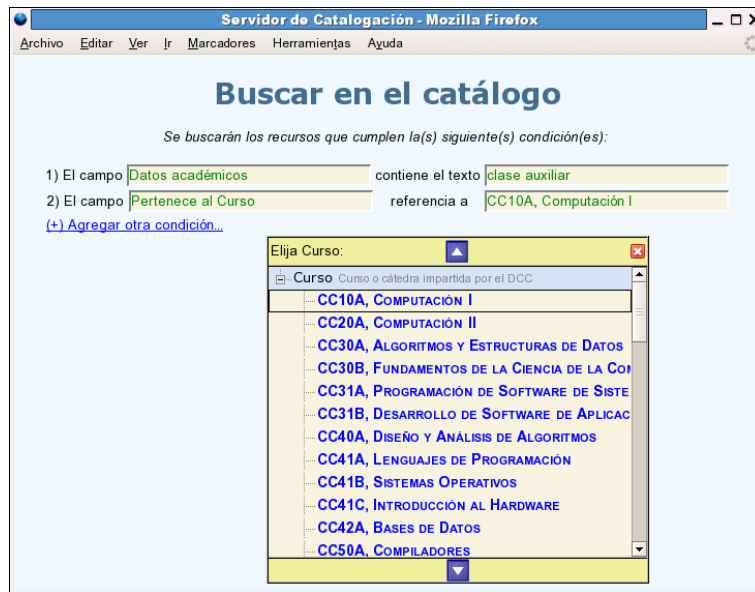


Figura 5.13: Buscador de recursos catalogados. En el ejemplo se buscan las páginas que posean algún metadato que contenga las palabras “clase” y “auxiliar” y que pertenezcan al curso “CC10A - Computación I”.

- Para las búsquedas se utilizan expresiones regulares por lo cual se pueden aplicar los caracteres especiales propios de éstas como¹: [] () . ? *.
- La búsqueda diferencia entre caracteres especiales o con tildes.

Buscar en el catálogo

La figura 5.13 muestra un buscador de los 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 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

¹No confundir con *wildcards*. Por ejemplo, la expresión regular `meta*` calza con los textos: met, meta, metaa, etc., mientras que la expresión regular `meta.*` calza con meta, metadato, metabúsqueda, etc.

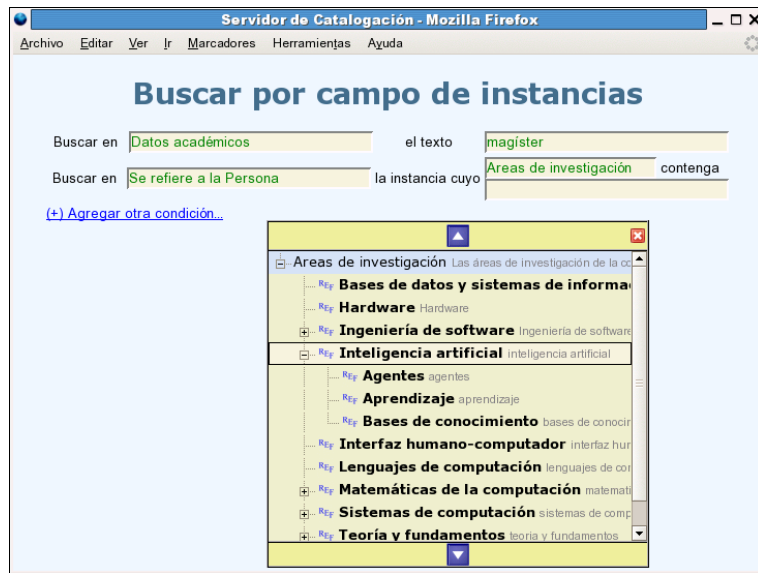


Figura 5.14: Buscador por campo de instancias. Permite anidar una búsqueda según un valor en los atributos de una instancia.

referencia a la instancia de nombre “CC10A - Computación I” de la clase Curso.

Como el listado de instancias puede llegar a crecer bastante, en algunos casos puede ser mejor presentar un buscador de instancias en vez del listado completo. Con este objetivo se realizó el buscador por campo de instancias.

Buscador por campo de instancias

Este buscador es similar al buscador por campo, salvo que en el caso que el tipo de campo seleccionado sea una Elección de Referencia no se despliega directamente la lista de instancias, sino que se puede realizar una búsqueda anidada según algún valor en los atributos de la instancia. En el ejemplo de la figura 5.14 se realiza una búsqueda de las páginas que contengan un metadato bajo el esquema *Datos Académicos* que contiene la palabra “magíster” y contengan en el campo *Se refiere a la Persona* una referencia a una instancia de la clase Académico que contiene en el atributo “Áreas de investigación” el valor de referencia “Inteligencia Artificial”.

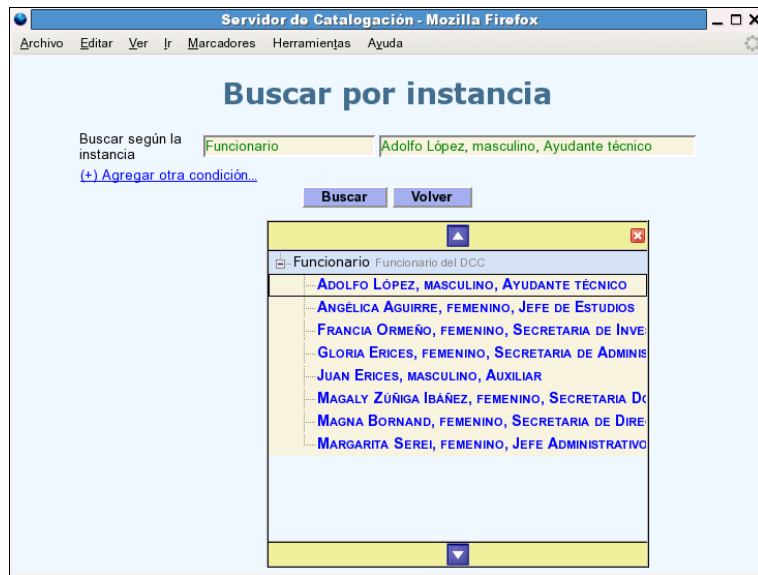


Figura 5.15: Buscador por instancia. En el ejemplo se buscan las páginas que contienen una referencia a la instancia de nombre “Adolfo López”.

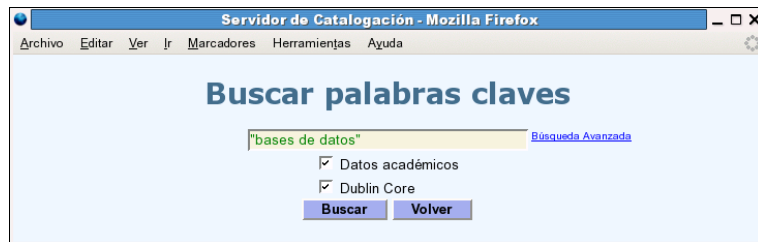


Figura 5.16: Buscador de recursos por palabras claves. En el ejemplo se buscan las páginas que posean algún metadato que contenga la frase “bases de datos”.

Buscador por instancia

La figura 5.15 muestra el buscador por instancia. Permite hacer búsquedas de todos los recursos que contengan cierta instancia en alguno de sus metadatos. En el ejemplo se realizará una búsqueda de todas las páginas que contengan alguna referencia a la instancia de nombre “Adolfo López” que pertenece a la clase Funcionario.

Buscar usando palabras claves

La figura 5.16 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

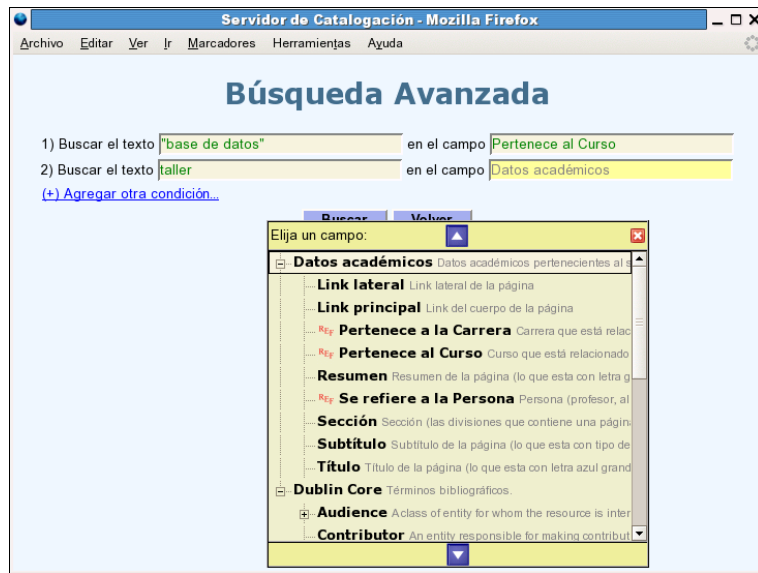


Figura 5.17: Búsqueda avanzada de palabras claves. En el ejemplo se buscan las páginas que contengan la frase “base de datos” para algún metadato asignado *Pertenece al Curso* y la palabra “taller” para algún metadato bajo la ontología *Datos Académicos*.

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.

Búsqueda avanzada de palabras claves

La figura 5.17 muestra el buscador avanzado de palabras claves. Permite hacer una o más búsquedas de un conjunto de palabras restringido a los metadatos ingresados bajo cierto campo y sus subcampos. Se aplican las mismas condiciones para las búsquedas de texto descritas anteriormente. En el ejemplo se buscan las páginas que contienen la frase “base de datos” en algún metadato asignado bajo el campo *Pertenece al Curso* y la palabra “taller” para algún metadato bajo la ontología *Datos Académicos*.

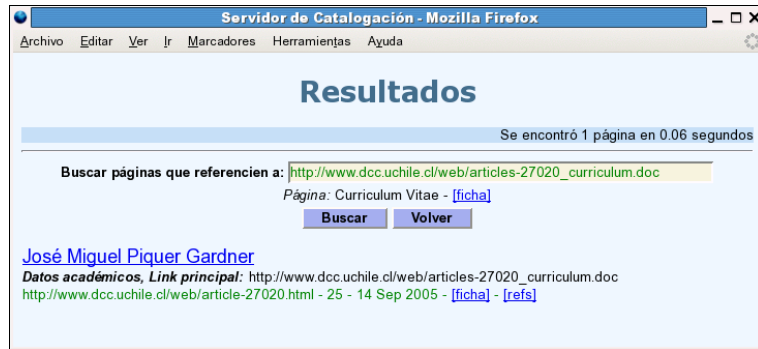


Figura 5.18: Buscador de referencias a una página. En el ejemplo se listan todas las páginas que contienen al menos una referencia al archivo *articles-27020_curriculum.doc*.



Figura 5.19: Navegador de instancias. Lista jerárquicamente el modelo de clases definido para las ontologías.

Búsqueda de referencias a una página

La figura 5.18 muestra el buscador de referencias a una página. Permite buscar todas las páginas que tienen al menos una referencia a una URL dada. Sólo se muestran aquellas páginas que contienen en un metadato de tipo URL el mismo texto que el ingresado, distinguiendo mayúsculas y minúsculas. Con este buscador se puede encontrar el contexto semántico al cual pertenece un recurso, lo que es especialmente útil para el caso de archivos binarios.

Se encontraron 8 registros		
Nombre (A)	Sexo	Cargo
Adolfo López	masculino	Ayudante técnico
Angélica Aguirre	femenino	Jefe de Estudios
Francia Ormeño	femenino	Secretaria de Investigación.
Gloria Erices	femenino	Secretaria de Administración
Juan Erices	masculino	Auxiliar

1 2 [Siguiente >>]

Figura 5.20: Navegador de instancias. Lista las instancias ingresadas para la clase Funcionario.

Navegación de instancias

El navegador de instancias de la figura 5.19 despliega todas las clases existentes para cada ontología en un visor de árbol. Al ser seleccionada una clase se dirige a la pantalla siguiente donde se muestran todas las instancias que pertenecen a la clase seleccionada o a sus subclases. El despliegue de las instancias es en forma de tabla, donde cada una se muestra en una fila y las columnas corresponden a los atributos definidos para la clase seleccionada. La tabla puede ser ordenada ascendentemente y descendentemente según el valor de algún atributo. La tabla muestra hasta cinco filas simultáneamente, en el caso que la lista sea mayor se divide en grupos de cinco donde se pueden avanzar o retroceder entre los grupos. Al seleccionar una instancia se despliega en la parte inferior el conjunto de páginas que tienen al menos un metadato asociado con ésta.

En la figura 5.20 se muestra un navegador para instancias de la clase Funcionario, el cual se encuentra ordenado ascendentemente según el atributo Nombre. Al seleccionar una instancia se realiza una búsqueda de todos los recursos que contienen alguna referencia a la instancia dada.

Ver datos de una página

La figura 5.21 muestra la ficha de resumen de una página catalogada. Esta se compone de un *frame* el cual lista en la parte superior los metadatos asociados a la página y en la parte inferior enlaza directamente con la página catalogada. Permite exportar los metadatos existentes para la página como meta-etiquetas o RDF/XML y, en caso de ser un usuario catalogador, enlaza con la página de mantención de metadatos.



Figura 5.21: Ficha de una página. La parte superior contiene los metadatos asociados a la página y en la parte inferior muestra la página catalogada.

Resultados de una búsqueda

Para cada uno de los buscadores detallados anteriormente, en el caso de existir páginas que cumplan con sus condiciones, retornan un conjunto de páginas que conforman el resultado de la búsqueda.

La figura 5.22 muestra un conjunto de resultados para la búsqueda de la frase “bases de datos”. En la parte superior se despliega el número de páginas que componen el conjunto de resultado y el tiempo que tomó la búsqueda. Luego se presenta el mismo buscador para permitir realizar una nueva búsqueda a partir de los valores anteriores, y continuación se presenta la lista de resultados ordenado del más relevante al menos relevante. La lista muestra hasta diez resultados simultáneamente, en el caso que la lista sea mayor se divide en grupos de este tamaño.

Para cada página del resultado se despliega el título (que contiene el link con la página encontrada) y luego se listan todos los metadatos que cumplen con alguna condición de la búsqueda. En la parte inferior se escribe la URL de la página, el puntaje otorgado por el algoritmo de ranqueo para la búsqueda en particular (en el ejemplo 115 puntos para el más relevante) y luego la última fecha de actualización de metadatos de la página.

Para cada página se presentan además dos links. El primero enlaza con la ficha de



Figura 5.22: Resultados de una búsqueda. Lista el conjunto de recursos que contienen algún metadato con la frase “bases de datos”.

la página (figura 5.21) y el segundo enlaza con el buscador de referencias a una URL (figura 5.18) el que permite buscar información relacionada con el recurso.

5.2. 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 en el navegador como plug-in. 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. Una vez instalado el cliente se debe agregar una referencia en la barra lateral del navegador a la dirección *chrome://cliente/content/panel.xul*.

El plug-in permite dos tareas: Ver metadatos y Modificar metadatos, las que son descritas a continuación.

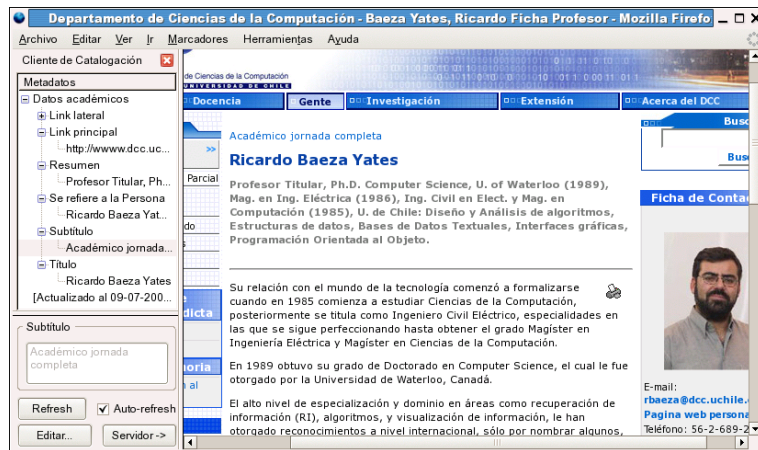


Figura 5.23: Visualización de los metadatos de un recurso catalogado utilizando el cliente de catalogación.

5.2.1. Ver metadatos

Permite obtener desde el servidor de catalogación los metadatos definidos para la página que se esté visitando y desplegarlos en la barra lateral. Esto puede ser realizado por cualquier tipo de usuario.

La figura 5.23 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. Al marcar el recuadro *auto-refresh* se habilita el modo automático, donde cada 10 segundos se realiza la acción del botón *Refresh*.

El botón *Ir al sitio* permite que en la pantalla aparezca la página de búsquedas del servidor de catalogación. El botón *Editar Metadatos* permite acceder a la pantalla de modificación de metadatos para la página abierta en la ventana central.

5.2.2. Modificar metadatos

Permite modificar los metadatos para la página que se esté visitando, agregando nuevos y modificando o eliminando los existentes. Cada cambio es enviado al servidor de catalogación donde los datos son actualizados. Esto puede ser realizado sólo por un usuario catalogador. La utilización de este mantenedor es similar al mantenedor existente en el servidor de catalogación descrito en la sección 5.1.2.

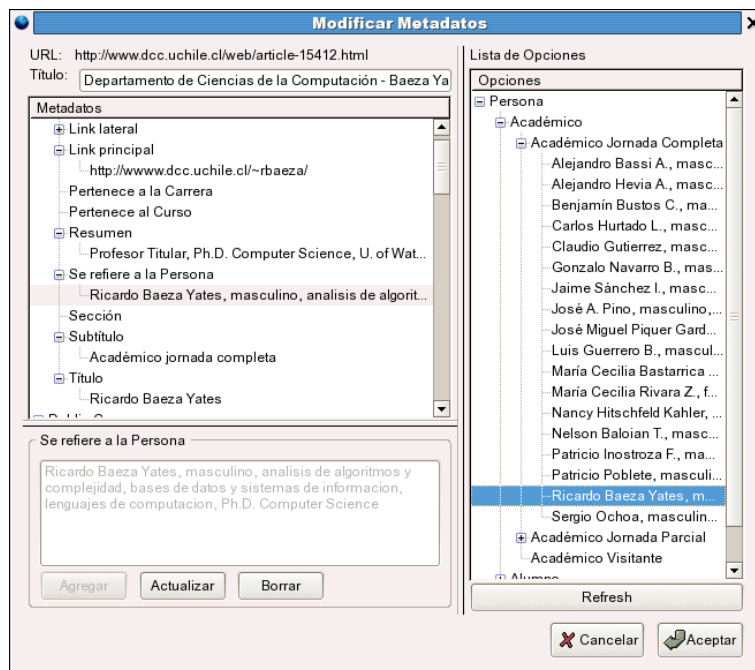


Figura 5.24: Modificación de metadatos de un recurso utilizando el cliente de catalogación.

La figura 5.24 muestra la modificación de metadatos para una página. En la parte superior se muestra la URL de la página en mantención y permite ingresar el título que tendrá (la cual es por defecto el título existente en el HTML). Bajo esto se muestran los metadatos correspondientes a la URL, donde al seleccionar uno se despliega su valor en el recuadro inferior para su actualización. En el caso que el tipo del metadato sea una Elección de Instancia o Elección de Referencia se despliega en el costado derecho la lista de instancias o referencias correspondientes para su selección.

Para ingresar un metadato se debe seleccionar un campo de metadato de entre la lista de metadatos, ingresar su valor, y presionar sobre el botón *Agregar*. Para eliminar un metadato, se debe seleccionar desde la lista de metadatos y luego presionar sobre el botón *Borrar*.

En el siguiente capítulo se muestra un caso de estudio donde se utilizó el sistema *Catalogo* en un sitio web real para realizar búsqueda de los recursos publicados.

Capítulo 6

Caso de estudio

En este capítulo se presentará como caso de estudio la instalación y uso del sistema *Catalogo* sobre un sitio web. El objetivo es producir una guía para la instalación del sistema, explicar los pasos que se deben seguir, dejar un sistema de uso público para probar las características de la herramienta desarrollada, evaluar su comportamiento y guiar posibles trabajos futuros.

Para el caso de estudio 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>.

6.1. Procedimiento

Siguiendo los pasos descritos en la sección 4.9 para la instalación de un sistema de catalogación, se realizaron las siguientes tareas:

1. Se decidió catalogar todas las páginas del sitio web corporativo del DCC, esto es, sin incluir las páginas personales. Se decidió además catalogar sólo las páginas de publicación de información y no páginas de enlace entre páginas de información.
2. Se estudió el sitio y su estructuración. Éste es altamente estructurado por ser generado a través de un sistema de publicación llamado *Newtonberg*, lo que facilitó la carga inicial de datos. No se catalogaron páginas de portada de secciones ni versiones imprimibles.
3. Se definió el esquema de metadatos basado en la ontología del proyecto *Depmark* y en

la estructura visual del sitio (sección 6.2).

4. Se ingresaron manualmente las instancias correspondientes a las clases definidas en la ontología del sitio.
5. Se hizo una carga inicial de datos obteniendo una copia de las páginas web publicadas y luego utilizando el catalogador proporcionado por el sistema se insertaron los metadatos al catálogo (sección 6.3).
6. Se realizó el marcado manual del sitio, donde se verificó la carga automática, se asociaron las páginas con las instancias creadas y se evaluó el funcionamiento del cliente de catalogación.
7. Se utilizaron los buscadores genéricos existentes en el sistema desarrollado.
8. Se ajustaron los puntos de ranqueo de los campos hasta obtener resultados satisfactorios en las búsquedas.
9. Se instaló en producción y se realizó el proceso de mantención de metadatos en forma manual y automática (sección 6.4 y 6.5).

El cuadro 6.1 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.

Tarea	<i>t</i>
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 16 días</i>	

Cuadro 6.1: Resumen del tiempo para la puesta en marcha del sistema.

Campo	Descripción	Tipo
Título	Corresponde al título principal de la página, el de mayor tipografía.	Texto
Subtítulo	Corresponde al texto de menor tipografía que se encuentra sobre el título.	Texto
Resumen	Corresponde al texto en negrita ubicado bajo el título y que es un resumen de la página.	Texto
Sección	Corresponde a cada uno de los títulos de las secciones en que está dividido el texto de una página.	Texto
Link principal	Corresponde a los links que se encuentran dentro del texto central de la página.	URL
Link lateral	Corresponde a los links que existen en los costados de la página y que dirigen a páginas relacionadas. No incluye a los links de navegación del sitio.	URL
Pertenece a la Carrera	Instancia de la clase Carrera que es parte del tema de la página.	Elección de Instancia
Pertenece al Curso	Instancia de la clase Curso que es parte del tema de la página.	Elección de Instancia
Se refiere a la Persona	Instancia de la clase Persona que es parte del tema de la página.	Elección de Instancia

Cuadro 6.2: Campos de metadatos que componen la ontología a utilizar en el sitio.

6.2. Esquema de metadatos

Para el registro de metadatos se utilizó un esquema particular al sitio. Para definirlo se estudió el conjunto de páginas a catalogar y las búsquedas que se esperaba que se realizaran. El cuadro 6.2 describe los campos de metadatos en la ontología.

La figura 6.1 contiene el diagrama de clases utilizado por el esquema de metadatos, el cual está basado en la ontología utilizada por el proyecto *Depmark* [MG02]. Está compuesto de tres clases bases: *Persona*, *Curso* y *Carrera*. La clase *Persona* contiene seis subclases que la especifican según el tipo de persona que representa: *Alumno*, *Funcionario*, *Académico*, *Académico Jornada Completa*, *Académico Jornada Parcial* y *Académico Visitante*. Los atributos “Carrera que cursa” y “Cursos dictados” de las clases *Alumno* y *Académico* corresponden a una elección de instancia de las clases *Carrera* y *Curso*, respectivamente. Se definieron además dos campos tipo Conjuntos de Referencia:

- Áreas de Investigación, el cual contiene las áreas donde se realiza investigación en el DCC organizadas jerárquicamente. Actualmente corresponde a una lista de 27 áreas,

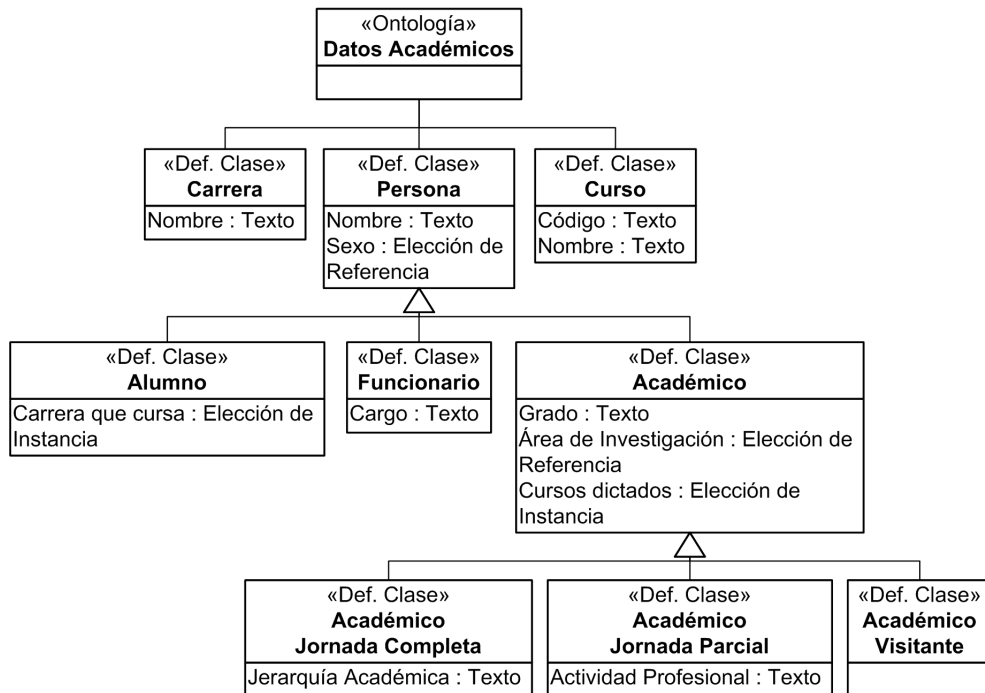


Figura 6.1: Diagrama de clases para la ontología *Datos Académicos*.

las que pueden ser asignadas como valor al atributo “Área de Investigación” de las instancias de la clase Académico.

- Valores para sexo, el cual contiene los valores “masculino” y “femenino”, que pueden ser asignados como valor al atributo “Sexo” de las instancias de la clase Persona.

Para la asignación de puntos de ranqueo de cada campo se inició con una base de 10. Luego se realizaron ajustes para mejorar el resultado de las búsquedas, donde se aumentó a 30 para el título, a 20 para el subtítulo, a 15 para el resumen y se disminuyó a 5 para el link lateral.

Además del esquema particular al sitio se utilizó el esquema de datos definidos por Dublin Core con sus elementos básicos y sus calificadores¹, agregando al sistema 51 campos para esta ontología. Más detalles de Dublin Core en la sección 2.5.

¹Según lo definen los documentos en las URLs <http://purl.org/dc/elements/1.1/>, <http://purl.org/dc/terms/> y <http://purl.org/dc/dcmitype/>.

6.3. Catalogación inicial

La creación del catálogo del sitio se realizó en tres fases: primero una catalogación automática del sitio a través de *scripts* especialmente desarrollados, luego la creación de las instancias para las clases definidas en la ontología, y finalmente una catalogación manual ingresando los metadatos que no fueron automatizables.

Para realizar la catalogación automática primero se realizó una copia local del sitio web publicado utilizando el comando *wget*. 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 la sincronización de metadatos para las páginas del sitio. Se utilizó el sincronizador basado en expresiones regulares proporcionado por el sistema (ver sección 5.1.2 y apéndice A) para capturar los metadatos de tipo Texto y URL de la ontología (*Título, Subtítulo, Resumen, Sección, Link principal y Link lateral*).

Después se realizó manualmente la creación de instancias para el modelo de clases definido en la figura 6.1 con sus correspondientes atributos y relaciones. En total se ingresaron 99 instancias: 4 para Carrera, 51 para Curso y 44 para Persona y sus subclases.

Una vez ingresadas las instancias se procedió a registrar manualmente las asociaciones entre instancias y páginas a través de los campos de tipo Elección de Instancia (*Pertenece a la Carrera, Pertenece al Curso y Se refiere a la Persona*).

El conjunto total de datos ingresados, sumando metadatos de páginas y atributos de instancia, 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.

6.4. Instalación del sistema

El sistema se encuentra instalado en un servidor GNU/Linux que es independiente al sitio web catalogado. Para la instalación del sistema se utilizó J2SDK 1.4.2, PostgreSQL 7.2.1 y Jakarta Tomcat 5.0.24.

El sistema se encuentra actualmente disponible para uso público en la dirección web <http://putu.dcc.uchile.cl/catalogo/>. Es necesaria una autenticación con nombre de usuario y password para poder acceder al menú de administración de metadatos y de ontologías.

6.5. Mantenimiento de metadatos

La mantención de metadatos del catálogo se compone de una parte automática y una parte manual. La mantención automática se realiza a través del mismo proceso de sincronización de metadatos definido para la carga inicial. Esto es, se utiliza un proceso ejecutado a través de la línea de comandos que recolecta las páginas publicadas y las sincroniza con los metadatos del catálogo. El proceso se invoca periódicamente a través del programador de tareas del sistema operativo del servidor, *cron*.

La mantención manual tiene por objetivo actualizar los campos de la ontología no automatizados e ingresar metadatos para los recursos que no se encuentren dentro del conjunto sincronizado. Esta tarea es realizada por el administrador del catálogo y por una persona externa, sin conocimientos iniciales del sistema ni de las páginas a catalogar, a la cual se le hizo una breve explicación del objetivo de la catalogación y del sistema, explicándole el cliente de catalogación y su uso.

Después de realizar la mantención manual de metadatos utilizando el cliente de catalogación, se obtuvieron un conjunto de observaciones o mejoras sobre la herramienta:

- La herramienta podría tener modos de uso. Un modo para el ingreso de metadatos, donde la interfaz (en particular el foco de los botones) esté adaptada para agregar datos rápidamente, y un modo para la mantención de metadatos, donde la interfaz esté adaptada para modificar metadatos.
- El registro de los links es la tarea más difícil y demorosa, tomando prácticamente el doble de tiempo que un metadato de texto, y en cantidad, corresponden a cerca de la mitad de todos los metadatos registrados. Es necesario agregar a la herramienta alguna funcionalidad para simplificar la catalogación de los links de una página.
- La interfaz de la herramienta hace difícil catalogar los archivos binarios. Es necesario investigar una forma para catalogar los binarios a través de la herramienta y no solamente a través del servidor de catalogación.

- Se producen muchos errores tipográficos en la escritura de los metadatos de texto. Además, existe una tendencia a repetir los metadatos entre campos (el texto del título era repetido como resumen). Es necesario que la herramienta realice una verificación sintáctica de los datos ingresados por el usuario para evitar estos problemas.
- El cliente de catalogación podría contener un botón para gatillar la sincronización de los metadatos de la página visualizada, evitando tener que invocar la sincronización desde el servidor.

Capítulo 7

Discusión y Conclusiones

El trabajo realizado ha consistido en desarrollar una herramienta que permita utilizar conceptos de la web semántica, como son los metadatos y la catalogación, para mejorar las búsquedas en un sitio web. Las conclusiones de este trabajo están dadas principalmente por la experiencia del desarrollo del sistema y de su utilización.

7.1. 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, si no 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 basado en palabras claves 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 semántico 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 sólo un

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 7.1: Comparación entre el sistema de catalogación y los buscadores sintácticos de un sitio. 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 7.1 presenta una comparación según diferentes aspectos entre el modelo de catalogación propuesto y los buscadores sintácticos de un sitio.

7.2. Recomendaciones

Con la experiencia lograda en el caso de estudio se creó un conjunto de recomendaciones para la creación de ontologías, la catalogación de páginas y la búsqueda, las que son resumidas a continuación.

7.2.1. Creación de ontologías

Al crear una ontología para realizar búsqueda en un sitio, se deben tener en cuenta los siguientes aspectos:

- La ontología debe contener sólo los campos que interesen para hacer búsquedas. Por ejemplo, si no se tiene pensado (o no otorga ninguna utilidad) hacer una búsqueda por edad de una persona, no tiene sentido agregar ese campo a la ontología. Ese dato puede ser importante y debe estar en la página que contenga su información, pero no como parámetro de búsqueda. Una ontología que contenga demasiados campos sólo aumenta su complejidad lo que se traduce en:
 - Mayor dificultad de llenar y mantener los metadatos, y mayor cantidad de campos de metadatos sin uso y menor actualización.
 - Mayor probabilidad de cometer errores en la creación de metadatos, y menor calidad en los metadatos y en los resultados de las búsquedas.
 - Mayor dificultad para ser entendido para el usuario final, y menor utilización de las búsquedas.
- Los esquemas de metadatos deben ser agrupaciones lógicas de metadatos con un sentido coherente. En el caso que se desee catalogar diferentes aspectos del sitio se deben crear ontologías distintas, cada una especializada en un ámbito particular.
- Los metadatos no deben contener la información buscada ni información temporal, sino que deben ser una referencia a la página que contiene estos datos. Por ejemplo, si se

espera que los usuarios busquen el número telefónico de una persona no debe existir un campo cuyo nombre sea teléfono y su valor de metadato sea el número buscado, sino que basta que el número se encuentre dentro de una página que puede ser encontrada con el nombre de la persona. Así la información se mantiene publicada sólo en un lugar y los metadatos se utilizan para buscar información, evitando posibles inconsistencias o desactualizaciones del metadato.

- Al definir un *script* de sincronización se debe tratar de incluir el máximo posible de elementos de la ontología, ya que simplificará después la mantención de los metadatos. Sin embargo, no se debe caer en el extremo de utilizar sólo metadatos de texto ya que las búsquedas serían sólo sintácticas disminuyendo el valor semántico que puede agregar el catálogo a las búsquedas (en ese caso sería mejor utilizar un motor de búsqueda tradicional).

7.2.2. Catalogación de páginas

Como consecuencia de la catalogación y revisión manual de las páginas, los usuarios catalogadores del sistema cobran gran importancia en los resultados de las búsquedas. Su tarea no es automatizable, ya que se basa en su capacidad de extraer elementos importantes de una página, en su criterio de agregar cada metadato y en su capacidad de entender y resumir un texto. Para llevar a cabo su trabajo deben tener en cuenta los siguientes puntos:

- Deben conocer el contexto sobre el cual se realiza la catalogación, es decir, deben saber cuál será el uso que se les dará a los metadatos creados.
- Deben rescatar sólo los aspectos que definen una página, prestando atención en las palabras claves. No deben catalogar el texto completo de una página (a menos que se desee obtener un buscador sintáctico).
- Deben asociar una página con una instancia sólo cuando ésta tiene una relación temática con el texto, no sólo cuando aparece nombrada. Recordar que no se quiere conseguir la mayor cantidad de resultados sino que los mejores resultados.
- No rescatar los links que correspondan a navegación del sitio, a funciones JavaScript o a direcciones de correo electrónico.
- Al catalogar archivos binarios (doc, ppt, pdf, zip y otros) se debe catalogar además el link de la página que lo referencie.

- Conocer la terminología del área a la que pertenece el sitio que se cataloga y utilizarla correctamente.
- No repetir los textos de la página en más de un metadato.
- Evitar cometer errores de tipografía. Se estudiará la posibilidad de agregar un chequeo de ortografía al cliente de catalogación.

7.2.3. Búsqueda de recursos

El usuario buscador debe reconocer las características del buscador semántico y las diferentes capacidades que un buscador sintáctico. Para esto deben:

- Intentar restringir las búsquedas con instancias o valores de referencia ya que permite aprovechar el valor semántico de los metadatos.
- Utilizar el buscador de links a una URL para conocer el contexto en el cual se encuentra una página y donde se puede obtener mayor información.
- En el caso de encontrar metadatos equivocados, notificar el hecho a un usuario catalogador. Se estudiará la posibilidad de automatizar esto en el cliente de catalogación.

7.3. Trabajo futuro

Los aspectos futuros que pueden continuar esta investigación han sido divididos en dos áreas: investigación del modelo de catalogación y desarrollo del software *Catalogo*.

7.3.1. Investigación del modelo de catalogación

- Estudiar los problemas asociados a la unicidad de una página según su URL. En particular estudiar los casos de páginas dinámicas¹, páginas por defecto² y *mirrors* de sitios.

¹Por ejemplo, para el caso de tipo GET, el contenido de una página puede depender (o no depender) de parámetros que le sean enviados como parte de la URL con los caracteres especiales ?, & y =.

²Por ejemplo, para las URLs del tipo *http://xxx/* el servidor web responde con una página que tiene otra URL, como por ejemplo *http://xxx/index.htm*, *http://xxx/web/index.html*

- 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 previo pueda hacer uso de ella.
- Estudiar formas para agregar mayor conocimiento del servidor de catalogación sobre el nivel de dinamismo de las páginas catalogadas, para evitar la catalogación de páginas de enlace, como páginas de portada.
- Estudiar formas de integración entre catálogos, y por consiguiente, estudiar la integración entre diferentes esquemas de metadatos.
- Estudiar el algoritmo de ranqueo y sus posibles mejoras. En particular el de ranqueo estático puede ser mejorado para asemejarse al algoritmo utilizado por Google.
- 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.

7.3.2. Desarrollo del sistema Catalogo

Dentro del trabajo futuro para este software se encuentran:

- Agregar una funcionalidad para reportar errores en metadatos por parte de usuarios visitantes a usuarios catalogadores.
- Mejorar la mantención de metadatos, en particular reutilizando la etapa de carga automática para crear un monitor de modificaciones de páginas catalogadas.
- Implementar búsquedas de palabras similares, para realizar búsquedas que ignoren los acentos, los plurales y los sinónimos (a través de un tesoro). Por ejemplo tomar como iguales las palabras “semantica” y “semántica”, “dato” y “datos”, etc.

- Mejorar la interfaz genérica en los buscadores y navegadores de metadatos.
- Desarrollar la internacionalización³ del software, para permitir el uso de diferentes idiomas según lo requiera el usuario.
- Configurar un enlace con un buscador sintáctico para redirigir sobre éste en el caso que no se existan resultados en el catálogo.
- Asignar propiedades a conjuntos de URLs para facilitar la mantención de metadatos. Por ejemplo, se podría impedir la asignación metadatos a las portadas de secciones para que no exista la posibilidad de que algún usuario catalogador las registre por error.
- Desarrollar nuevos indicadores que permitan conocer el estado de los metadatos. Se proponen los siguientes:
 - % de páginas no asociadas a instancias. En el caso de tener un valor alto significa que el buscador es muy sintáctico y no aprovecha las características semánticas de los metadatos.
 - % de instancias sin usar. Al aumentar su valor significa que faltan páginas por catalogar o que existen instancias por eliminar.
 - % de valores de referencia no usados. Al aumentar su valor significa que el conjunto de valores de referencia debe ser simplificado o se necesita mejorar la calidad de la catalogación.
- Realizar mejoras en el cliente de catalogación como:
 - Poder personalizar el cliente, para configurarlo en modo de ingreso donde la interfaz se optimiza para el ingreso rápido de metadatos, y configurarlo en modo de mantención donde la interfaz permite hacer actualizaciones de metadatos.
 - Implementar el manejo de páginas que contengan *frames* u otro tipo de incrustaciones donde cada componente puede contener sus propios metadatos.
 - Estudiar la factibilidad de hacer revisión ortográfica antes de enviar metadatos, con el fin de evitar errores de escritura en los metadatos.
 - Estudiar la posibilidad de realizar asignaciones de metadatos a un conjunto de páginas simultáneamente, en especial la asignación de instancias y valores de referencia.

³Característica comúnmente conocida como *i18n*.

Agradecimientos

El autor agradece financiamiento al Proyecto FONDECYT 1030810, “Metadatos para describir y consultar la Web Oculta”.

Bibliografía

- [Ame04] American Customer Satisfaction Index. *Second Quarter Scores: Manufacturing/Durable Goods & E-Business: Search Engines*, Agosto 2004. <http://www.theacsi.org>.
- [Atk02] Susan Atkey. *Issues in Cataloguing the Web*. School of Library, Archival and Information Studies/UBC, Diciembre 2002.
- [BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. *The Semantic Web*. Scientific American, Inc, Mayo 2001.
- [CHHL99] 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.
- [CV01] Fidel CACHEDA and Angel VINA. *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.
- [DFKO01] Ying Ding, Dieter Fensel, Michel Klein, and Borys Omelayenko. *The Semantic Web: Yet Another Hip?* Data and Knowledge Engineering, Octubre 2001.
- [DH97] Lorcan Dempsey and Rachel Heery. *A review of metadata: a survey of current resource description formats*. UKOLN Metadata Group, Marzo 1997.
- [DK03] Stefan Decker and Vipul Kashyap. *The Semantic Web: Semantics for Data on the Web*. VLDB 2003, Septiembre 2003.
- [Doc01] Cory Doctorow. *Metacrap: Putting the torch to seven straw-men of the meta-utopia*, Agosto 2001. <http://www.well.com/~doctorow/metacrap.htm>.
- [HHL03] Jeff Heflin, 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.
- [Key05] Keynote Systems. *Yahoo! Search and MSN Search Close the Gap with Google*. Press Release 05-01-13, Enero 2005. <http://www.keynote.com>.
- [KK01] JosÃ© Kahan and Marja-Riita Koivunen. *Annotea: An Open RDF Infrastructure for Shared Web Annotations*. World Wide Web Consortium, Mayo 2001.
- [MG02] Ernesto Krsulovic Morales and Claudio GutiÃ©rrez. *Building Yearbooks with RDF*. Centro de Investigacion de la Web. Departamento de Ciencias de la Computacion. Universidad de Chile, Diciembre 2002.

- [MÃ©02] Eva MÃ©ndez. *Metadatos y recuperacion de informacion: Estandares, problemas y aplicabilidad en bibliotecas digitales*. Departamento de BiblioteconomÃ­a y Documentacion de la Universidad Carlos III de Madrid. Ediciones Trea. ISBN: 84-9704-055-4, Junio 2002.
- [Nie02] Jakob Nielsen. *Intranet Usability: The Trillion-Dollar Question*. Useit.com Alertbox, Noviembre 2002.
- [Nie04] Jakob Nielsen. *When Search Engines Become Answer Engines*. Useit.com Alertbox, Agosto 2004.
- [NM01] 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.
- [Nun03] Juan Manuel Barrios Nunez. *Presentacion Tema de Tesis para Magister en Ciencias, mencion Computacion: Catalogacion semantica de sitios web*. Departamento de Ciencias de la Computacion. Universidad de Chile, Diciembre 2003.
- [Sea01a] SearchTools.com. *Review of Remote Search Services*. <http://www.searchtools.com/analysis/remote-search/>, 2001.
- [Sea01b] SearchTools.com. *Web Admin's Guide to Site Search Tools*. <http://www.searchtools.com/guide/>, 2001.
- [Ste99] Dick Stenmark. *A Methodology for Intranet Search Engine Evaluation*. In Kakola, T. (ed.), Proceedings of IRIS22, August 7-10, Department of CS/IS, University of Jyvaskyla, Finland, Agosto 1999.
- [Sul02a] Danny Sullivan. *Death Of A Meta Tag*. Search Engine Watch, Octubre 2002.
- [Sul02b] Danny Sullivan. *How To Use HTML Meta Tags*. Search Engine Watch, Diciembre 2002.
- [Sul02c] Danny Sullivan. *Meta Tag Lawsuits*. Search Engine Watch, Febrero 2002.
- [Wor99] World Wide Web Consortium. *Resource Description Framework Model and Syntax Specification. W3C Recommendation*, Febrero 1999.
- [Wor02] World Wide Web Consortium. *Annotea Protocols. W3C Draft*, Diciembre 2002.
- [Wor03] World Wide Web Consortium. *OWL Web Ontology Language Guide. W3C Recommendation*, Agosto 2003.

Apéndice A

Código Fuente de Ejemplo

A.1. Sincronizador.xml

El Sistema Catalogo permite configurar la sincronización de metadatos a través del archivo *sincronizador.xml*. En este archivo se declaran las clases java de los catalogadores automáticos que debe utilizar el sistema (cada clase debe implementar la interfaz *catalogo.server.sincro.Catalogador*), junto con sus parámetros de configuración (por ejemplo, los recursos sobre los que aplica y los campos de ontología automatizados).

El siguiente listado presenta el archivo *sincronizador.xml* utilizado para la instalación del sistema en el caso de estudio del capítulo 6.

```
<sincronizacion>

  <catalogador class="catalogo.server.sincro.CatalogadorImplRegEx">

    <!-- url que acepta este catalogador -->
    <url encoding="ISO-8859-1">
      <regex type="match" ci="false"><![CDATA[
        http://www.dcc.uchile.cl/web/article.*\.html
      ]]></regex>
    </url>

    <!-- el titulo de la pagina, puede ser <title>([<]*)</title> -->
    <title>
      <regex type="find" ci="false"><![CDATA[
        <td class="NormalT1">([<]*)</td>
      ]]></regex>
    </title>

    <!-- titulo: texto azul grande -->
    <text-metadata campo-id="213">
      <regex type="find" ci="false"><![CDATA[
        <td class="NormalT1">([<]*)</td>
      ]]></regex>
    </text-metadata>
  </catalogador>
</sincronizacion>
```

```

    ]]></regex>
</text-metadata>

<!-- subtítulo: texto azul chico sobre el título -->
<text-metadata campo-id="214">
    <regex type="find" ci="false"><![CDATA[
        <td class="NormalT2">([<]*)</td>
    ]]></regex>
</text-metadata>

<!-- resumen: texto gris bajo el título -->
<text-metadata campo-id="215">
    <regex type="find" ci="false"><![CDATA[
        <td class="NormalPP">([<]*)</td>
    ]]></regex>
</text-metadata>

<!-- secciones: los puntos azules -->
<text-metadata campo-id="222">
    <regex type="find" ci="false"><![CDATA[
        <li class="NormalP"><a class="NormalPLINK" href="[<"]*">([<]*)</a></li>
    ]]></regex>
</text-metadata>

<!-- los links con class="BajoImpacto -->
<url-metadata campo-id="243">
    <regex type="find" ci="false"><![CDATA[
        class="BajoImpacto
    ]]></regex>
</url-metadata>

<!-- los links con class="Normal -->
<url-metadata campo-id="242">
    <regex type="find" ci="false"><![CDATA[
        class="Normal
    ]]></regex>
</url-metadata>

<!-- los links sin class -->
<url-metadata campo-id="242">
    <regex type="!find" ci="false"><![CDATA[
        class=
    ]]></regex>
</url-metadata>

<!-- no registrar links a versiones imprimibles -->
<url-reject>
    <regex type="find" ci="false"><![CDATA[
        /printer-
    ]]></regex>
    <regex type="find" ci="false"><![CDATA[
        /articles-28198_imagen_barra.gif
    ]]></regex>
</url-reject>

```

```
]]></regex>
<regex type="match" ci="false"><![CDATA[
    http://www.dcc.uchile.cl/?
]]></regex>
<regex type="match" ci="false"><![CDATA[
    http://www.dcc.uchile.cl/web/?
]]></regex>
</url-reject>
</catalogador>

</sincronizacion>
```