Gestión de recursos en la Red Académica

Autores: Javier Masa, Ingeniero de Aplicaciones, RedIRIS - javier.masa@rediris.es

Diego R. López, Coordinador de Aplicaciones, RedIRIS - diego.lopez@rediris.es

Resumen: Este artículo presenta una revisión de los mecanismos de gestión del conocimiento en la Red Académica española, discutiendo las tecnologías en uso, junto con su desarrollo y las razones para su empleo.

Palabras Clave: Directorio; LDAP; X.500; indexación; búsqueda; recurso; TIO; LIMS; centroide; web; Webber; Harvest; metainformación; Dublin Core; RDF; Red Temática; CVU; SARAC

Abstract: This paper presents a review of the mechanisms for knowledge management inside the Spanish NREN (National Research and Educational Network), discussing the technologies currently employed, their development and the rationale for their use.

Introducción

Dado que las redes académicas son las que iniciaron Internet y que uno de sus objetivos fundamentales es la difusión de la información que producen sus elementos, éstas contienen una ingente cantidad de recursos de información. Por desgracia, esta información se encuentra enormemente desestructurada y no puede hablarse propiamente de que constituyan una base de conocimiento científico y tecnológico.

Un objetivo durante largo tiempo acariciado por las organizaciones que gestionan estas redes (como RedIRIS en España) es ofrecer mecanismos de acceso coherentes y bien estructurados, que permitan tanto a los usuarios de estas redes como a la sociedad en general disponer de todo el conocimiento almacenado en ellas.

La naturaleza esencialmente distribuida y descentralizada de los servicios de información disponibles en Internet es, a la vez, causa de su enorme difusión y capacidad de comunicar información, y (por otro lado) causa de que la información tienda a ser difícilmente alcanzable incluso para los expertos. Se hace pues, necesario el establecimiento de mecanismos para la descripción de la información (metainformación) que permitan definir y realizar medios de acceso a los recursos disponibles en la Red.

Este artículo presenta una revisión de los mecanismos de gestión del conocimiento en la Red Académica española, discutiendo las tecnologías en uso, junto con su desarrollo y las razones para su empleo. Por último, incluimos una serie de aspectos sobre la evolución futura de estos servicios de información.

Indexado en los servidores de RedIRIS. Formatos de metadatos

En el año 1994 RedIRIS instaló su primer servidor web, y en 1996 tenía las suficientes páginas como para que surgiese la necesidad de poder realizar búsquedas sobre la información mantenida. Para ello era preciso indexar todo el contenido de dicho servidor web.

Casi todas las herramientas analizadas estaban destinadas a generar índices basados en el texto completo de los documentos. Se analizaron diferentes paquetes de dominio público y el elegido fue *Harvest* [3] porque cumplía varios de los requisitos que se habían impuesto:

- Permitir la construcción de índices a texto completo o generados a partir de una serie de palabras clave definidas por el administrador del sistema de indexación.
- Permitir indexar los recursos disponibles en otros servidores web distintos al servidor donde se ejecutaba el indexador.
- Programa de dominio público, con disponibilidad del código fuente.
- Permitir la exportación de la información indexada a otro servidor de indexado, de manera que éste último no tuviese que recorrer el servidor web para acceder a todos los documentos. Esta exportación se realizaría mediante una conexión especial en la que se transfiere todo el índice de una sola vez.

Se realizaron dos aproximaciones. La primera sirvió para comprobar que una indexación a texto completo no era todo lo operativa que se esperaba, ya que se obtenían cientos de documentos como resultado de cualquier consulta realizada. Esto no tenía mucha utilidad para un usuario normal que lo único que deseaba era rapidez y efectividad. En este tipo de índices (del estilo de los índices comerciales como *altavista*, *google* o *northern light*) el usuario ha de ser un buscador experimentado, capaz de especificar una consulta con todos los operadores lógicos necesarios para filtrar la gran cantidad de información que obtiene.

La segunda aproximación consistía en realizar el indexado basado en palabras clave, ya fuesen extraídas automáticamente de ciertas etiquetas HTML del documento como **<TITLE>**, **<H1>**, **<H2>**, **<B>**, etc., o las que se podían extraer de la etiqueta **<META>** destinada a incorporar metainformación a los documentos HTML. El tag **<META>** tiene la estructura ***<META NAME="N" CONTENT="V">*** donde **N** hace referencia al nombre del metadato y **V** a su valor. Por ejemplo para especificar la descripción de un documento podría utilizarse esta línea:

<META NAME="description" CONTENT="Documento que describe los diferentes sistemas de indexación utilizados en RedIRIS">

RedIRIS optó inicialmente por un sistema basado en palabras clave por ser el único modo de evitar índices inmensos no demasiado útiles para el usuario. Estas palabras clave serían extraídas de las etiquetas **<TITLE>**, **<H1>** y **<B>** del documento, de acuerdo con ciertos estudios que establecían que la parte susceptible de ser catalogada en un documento HTML era la que se escribía dentro del título **<TITLE>**, las cabeceras de párrafo **<H1>** y las palabras en negrita **<B>**.

Una vez realizadas las pruebas de indexado con estas especificaciones fue necesario un reajuste de las mismas debido a que los resultados no fueron los esperados. Las páginas se escribían sin una norma concreta en cuanto a la utilización de las etiquetas seleccionadas para la extracción de las claves y los términos que se indexaban no eran los correctos.

Para evitar este problema, el sistema volvió a ser modificado para extraer también información de la etiqueta **<META>**. Se decidió usar el conjunto estándar de metadatos *Dublin Core* [2] que, formado por 15 elementos, permitía introducir metainformación para catalogar perfectamente cada documento .

Usando metainformación se amplían las posibilidades de las consultas siendo posible realizar búsquedas por campos concretos como

- Materias
- Autor del documento
- Palabras clave
- Descripción

Cada una de las páginas web que RedIRIS mantiene en sus servidores contiene un subconjunto de los 15 metadatos en formato Dublin Core que son utilizados al realizar el indexado.

Uno de los problemas planteados en 1997 cuando se tomó la decisión de indexar el servidor web usando exclusivamente metainformación consistía en la forma de introducir estos metadatos en los documentos. Era impensable introducir 15 elementos extra en cada página HTML que se diseñara. Después de un análisis de toda la información disponible en el servidor, surgió la necesidad de crear una herramienta que ayudase a la introducción de esta metainformación. Así nació *Webber* [8].

Webber era un pequeño script en lenguaje Perl que tomaba como fuente varios elementos:

- El documento con código HTML y algunas líneas con código específico de Webber con formato **#VARIABLE=VALOR**.
- Ficheros de tipo plantilla almacenados en unos directorios determinados

Con toda esta información Webber generaba la página HTML final que estaría visible en el servidor web y a su vez podía ser indexada adecuadamente por el robot de indexado.

Las páginas se encontraban agrupadas por temas afines en diferentes directorios. Estos directorios estaban agrupados a su vez en otros directorios de forma que existía una relación entre la materia tratada en la página web y el lugar físico en la estructura de directorios. Esta estructura jerárquica se adaptaba muy bien al funcionamiento de Webber.

De esta forma se podían incluir palabras clave genéricas de cada materia en el fichero de plantilla en cada uno de los directorios. Webber recorría hacia arriba todos los directorios encontrados desde el que contenía al fichero tratado, hasta la raíz del servidor web, buscando estos ficheros de plantilla. Estos ficheros eran leídos y se extraían todas las palabras clave. A este grupo de palabras clave debía añadir las encontradas en el propio fichero tratado.

Esta herramienta ha sido ampliamente usada y ha evolucionado bastante desde entonces hasta que en el año 2000 RedIRIS la puso a disposición del público. Actualmente RedIRIS está trabajando en módulos inteligentes que sean capaces de extraer la metainformación del contenido del propio texto que escribe el usuario.

Indexado del servidor Web de RedIRIS (http://www.rediris.es)

Ya se ha comentado que el proceso de indexado se realiza con el software de dominio público *Harvest*. Fue elegido porque permite que el índice generado pueda ser exportado a otro servidor de índices para que éste no tenga que entrar a recorrer nuestro servidor buscando las páginas y extrayendo la información de ellas. Con esto conseguimos evitar una sobrecarga innecesaria del servidor web y también que el índice que genere ese robot de indexado no sea el correcto (o el que RedIRIS habría deseado).

Harvest está formado por dos componentes: el recolector (*gatherer*) y el indexador (*broker*). El primero se encarga de entrar en los servidores web y recolectar las páginas una detrás de otra extrayendo la información necesaria para que el indexador pueda generar el índice. El recolector deja esa información en un lugar determinado de forma que otros indexadores puedan entrar y llevársela de una vez sin necesidad de volver a recorrer el servidor web que se desea indexar.

Inicialmente se crearon dos recolectores, uno para las páginas existentes en castellano y otro para las existentes en inglés. Se crearon también dos indexadores, uno para cada idioma, y dos interfaces de consulta. Con el paso del tiempo se llegó a la conclusión de que no era necesario duplicar esfuerzos. Bastaba con tener un recolector y un indexador. Se modificó el interface de búsqueda para que permitiese consultar por un texto en los documentos en uno de los dos idiomas disponibles. A cada consulta realizada se le añadía

automáticamente una condición del tipo **"AND DC.Language=es**" o **"AND DC.Language=en**" (donde **DC.Language** es el metacampo que identifica en Dublin Core al idioma del documento) si el usuario así lo decidía o no se le añadía nada si el usuario decidía realizar la búsqueda con independencia del idioma en el que estuviese la página. Actualmente se permiten también búsquedas en legua gallega y catalana, de las que hay algunas páginas en el servidor de RedIRIS.

Indexado de la red I+D española

RedIRIS inició un grupo de trabajo llamado *iris-index* para crear una estructura jerárquica que permitiera el intercambio de información indexada. De esta manera, se pretendía disponer de un índice del contenido de la red I+D española.

Cada centro participante indexaba su servidor mediante el software Harvest. Recolectaban los documentos que creían necesarios y dejaban la información accesible a una máquina central en RedIRIS que acumulaba todos los índices para formar otro mayor. Se realizaron pruebas con varios centros y los resultados fueron satisfactorios. Para comprobar el rendimiento del sistema, RedIRIS creó un interface web de consultas al índice generado, y los usuarios realizaban sus peticiones en una máquina de RedIRIS.

El tamaño del índice central era la suma de los tamaños de todos los índices que se habían generado en esta red de pruebas. Esta razón llevó a pensar que el crecimiento de esta estructura sería limitado, ya que a medida que se incorporaran nuevos servidores aumentaría el tamaño del índice central y llegaría un momento en el que no dispondríamos de espacio en disco para almacenar tanta información. Se decidió buscar una nueva estructura que permitiera solventar este problema.

Uno de los requisitos de la nueva estructura elegida era que ocupase poco espacio y que basase su funcionamiento en la red de servidores, pero no para aglutinar los índices en uno solo sino para encaminar las consultas a cada uno de ellos. En lugar de mantener el índice global, se trataba de mantener exclusivamente un listado de los servidores de indexado que conformasen la red. Cuando un usuario realizase una consulta al servidor central de RedIRIS, la consulta debía ser distribuida a todos los servidores de la red.

Esta aproximación permitía aumentar el número de servidores a costa de tener que preguntar a todos ellos cada vez que se formulase una consulta. Se realizaron pruebas con otro interface encargado de realizar esta distribución de consultas y de recopilar los resultados para mostrarlos al usuario con bastante éxito. Sin embargo, este mecanismo de encaminamiento de consultas es una estrategia poco eficiente debido a que para resolver cada consulta hay que interrogar a todos los servidores de la red independientemente de que tengan o no resultados para dicha consulta.

Se puede evitar este problema usando información adicional a la lista de servidores que existen en la red. Esta información extra indica al servidor central qué servidores mantienen los datos que se están solicitando. De esta forma se puede interrogar únicamente a los servidores que devolverán datos. Para ello cada servidor participante en la red tiene que generar una base de conocimientos de los índices que mantiene. Este índice reducido (centroide) es el que debe enviar al servidor central.

Cuando el servidor central recibe una consulta, lo primero que debe hacer es interrogar su base de conocimiento, que ha ido recopilando de todos los servidores, para decidir a qué servidores ha de distribuir la petición del usuario. Esta base de conocimiento tiene la siguiente estructura:

- Datos sobre el servidor.
- Tipo de atributo y todos los valores que ese servidor tiene para ese atributo.

Por ejemplo:

> ***Servidor** AAA **con dirección** A11111*
>
> ***claves:** pez, agua, mar, ambiental, ...*
>
> ***autor:** Juan López, Antonio Romero, Alfonso Macías, ...*
>
> ***título:** Contaminación marina, Ingeniería del aprovechamiento del agua del mar, Dinámica de las poblaciones marinas,...*
>
> ***Servidor** BBB **con dirección** B22222*
>
> ***claves:** directorio, ldap, indexación, ...*
>
> ***autor:** Diego R. López, Javi Masa, Antonio Fuentes, ...*
>
> ***título:** Servicio de directorio LDAP, Indexado en RedIRIS, ...*

Estos dos índices se transmiten a la máquina central de RedIRIS. Cuando alguien pregunte por la palabra “pez”, es posible redirigir esta consulta al servidor AAA con la total garantía de que la consulta va a devolver resultados satisfactorios. Si alguien pregunta por “autor: Antonio”, se obtienen dos servidores a los que reenviar la consulta.

Indexado de las Redes Temáticas (Comunidades Virtuales de Usuarios)

RedIRIS proporciona un servicio de indexación para las Redes Temáticas [6] que mantiene. Estas Redes Temáticas se encuentran clasificadas de forma jerárquica por materias. Por ejemplo, dentro del área de “ciencias médicas” se encuentra “neurociencias”, y dentro de ésta “Neurología”

Este servicio permite buscar en cada una de las redes temáticas, independientemente del nivel en el que se encuentren en la estructura jerárquica. Es posible buscar sólo en los recursos de una subárea, un área raíz o en todas las áreas a la vez.

La indexación se realiza basándose en la metainformación que contienen los recursos. Por un lado, se usan los campos definidos en el estándar de Dublin Core. Por otro lado, se emplean unos metacampos adicionales que ayudan a distinguir el área en el que se encuentra el recurso. Estos metacampos adicionales son incluidos de forma automática por el sistema de carga de páginas, diseñado para este fin, en cada una de las páginas de una Red Temática específica. La estructura de uno de estos campos es de esta forma: CVU.Code = a01b01c03. Las letras indican la profundidad en la estructura jerárquica de áreas temáticas y los números identifican a la subárea dentro del área.

La indexación se realiza con el software Harvest y el índice generado se exporta al piloto iris-index que hemos visto en el apartado anterior.

Para consultar este índice temático se ha creado una estructura jerárquica de interfaces de consulta de manera que sea posible navegar por el índice de materias hasta que se elige el punto desde donde consultar. A los términos de búsqueda que se introducen se les añade automáticamente la expresión “AND CVU.Code=a01b02...” correspondiente al lugar desde donde se está consultando. De esta forma con un solo índice se pueden realizar consultar sobre múltiples áreas.

Indexado del Catálogo Recursos Científicos. Páginas Amarillas de Ciencia y Tecnología

Este catálogo está formado por los sistemas de información de centros, departamentos, grupos y proyectos de investigación situados en la red de investigación española, clasificados por códigos de la UNESCO sobre ciencia y tecnología [1]. RedIRIS mantiene una ficha por cada recurso que, almacenado en una base de datos, es susceptible de búsquedas por palabras clave. Estas palabras clave son introducidas por las personas que gestionan el recurso en cuestión al darlo de alta en el catálogo. Estas palabras clave dan lugar a un metadiccionario de términos científicos que se va agrandando a medida que se registran recursos.

Para introducir este catálogo en el sistema de indexación usado en RedIRIS (Harvest) es preciso extraer la información necesaria de estas fichas para que puedan ser recolectadas e indexadas por Harvest. De esta forma es posible buscar estos recursos interrogando la base de datos o al índice del grupo de trabajo de indexación comentado en apartados anteriores.

Servicio de Directorio

Uno de los primeros servicios que RedIRIS inició en el año 1990 fue el Servicio de Directorio X.500 [7]. El Directorio era considerado como una base de datos distribuida geográficamente por el mundo en la que cada nodo mantenía información de una determinada zona. Estaba basado en servidores de directorio (X.500) interconectados entre sí. Estos servidores mantenían datos de objetos, que generalmente correspondían a organizaciones, grupos o departamentos y personas. Existía una estructura jerárquica bajo España (al igual que en otros países) en la que aparecían las organizaciones y debajo de ellas unidades de organización y personas.

Cada servidor gestionaba una pequeña parte del Directorio y tenía conocimiento de la parte del Directorio que gestionaban el resto de servidores. Cuando un usuario realizaba una consulta a alguno de los servidores, ésta se podía resolver sin problemas ya que, o bien el servidor tenía los datos solicitados, o bien conocía la dirección del servidor que los tenía y podía preguntarle para devolver los datos al usuario. RedIRIS gestionaba la raíz de España, con un servidor que servía de interconexión con los servidores raíces del resto de países, de forma que a los ojos del usuario parecía como si existiese una única base de datos mundial.

Con el transcurso de los años, Internet ha llegado a casi todos los usuarios y, cada vez más, se demanda un directorio donde se pueda encontrar algún dato de una persona, como por ejemplo su dirección de correo electrónico.

Con la estructura inicial, cuando un usuario preguntaba al servidor de directorio de RedIRIS por datos de una persona de la Universidad de Cádiz, tenía que producirse la comunicación entre los dos servidores pero el de RedIRIS era el que tenía que recoger los datos y enviarlos al usuario final. Si la consulta era algo más ambiciosa como "*la dirección de correo de Antonio Romero que trabaja en algún centro de España*" el servidor de RedIRIS tenía que preguntar a todos los servidores de directorio españoles y esperar a que le respondieran para enviar los datos al usuario. A medida que cada centro instalaba su propio servidor de directorio, este número de peticiones se iba incrementando y la carga de esa máquina se hacía cada vez mayor. Imaginemos una consulta similar pero a nivel mundial. Se podría producir un caos en los servidores de directorio. Debido a esto, inicialmente se prohibieron este tipo de búsquedas, filtrándolas en los servidores, y solo se permitieron las que englobaban a un único servidor.

Los servidores de directorio han evolucionado y ahora la mayoría soportan el protocolo de comunicaciones LDAP, que permite que un programa cliente se comunique con un servidor de directorio de una forma muy simple. Debido a esto la mayoría de los navegadores y clientes de correo soportan consultas a servidores LDAP.

Basándonos en la experiencia recogida en el área de la indexación de servidores web, se ha planteado hacer algo similar con el Directorio. Se ha creado

un servicio piloto de indexación del directorio con una estructura de servidores LDAP. Cada servidor de la red crea una base de conocimiento sobre las entradas que mantiene y la envía al servidor central. La estructura de esta base de conocimiento tiene formato TIO (*Tagged Index Object*) [4], un formato diseñado para el intercambio de información indexada, especialmente preparado para poder realizar actualizaciones de los datos del índice de forma individual.

El punto central está basado en un servidor LDAP especial llamado LIMS que solo permite realizar consultas. LIMS, desarrollado por Roland Hedberg, funciona de la siguiente forma; para cada petición que recibe de un usuario mira en su base de conocimiento y obtiene la lista de los servidores de directorio a los que hay que preguntar. Llegados a este punto puede hacer dos cosas; realizar las consultas a los servidores y devolver los datos a los usuarios o enviar la lista con las referencias de servidores al programa del usuario y dejar que sea éste el que realice las consultas mediante LDAP a cada uno de los servidores. De esta forma se libera de carga este punto central de consultas.

Evolución futura

Dentro de RedIRIS percibimos tres líneas fundamentales en el desarrollo de los sistemas de gestión del conocimiento de la Red Académica

1. Uno de los elementos claves es el proyecto SARAC (Servicio de Acceso a Recursos de Alta Calidad), un proyecto que se ha iniciado hace unos meses y que pretende construir un catálogo de recursos calificados y revisados por expertos en las diferentes áreas temáticas y construido por un equipo de documentalistas utilizando el formato de descripción de recursos RDF. Actualmente se trabaja en la construcción de los interfaces de gestión y de acceso al catálogo.
2. La generalización del empleo de metadatos dentro de la Red Académica constituye otro objetivo básico. Para ello contamos con la difusión de herramientas como Webber (que ya mencionamos con anterioridad) para facilitar la asignación de metainformación a los recursos a medida que se vayan creando. Por otra parte, intentamos también exportar el concepto de metainformación a otras áreas diferentes de la información documental (como el Directorio o los recursos multimedia). En esta área uno de los proyectos a corto plazo es la inclusión de metainformación en las entradas del Directorio, de manera que sea posible construir un índice que constituya el directorio de expertos accesibles en la Red Académica.
3. Por último, consideramos fundamental ofrecer un interface de acceso único a todos estos mecanismos de localización de recursos. Para ello se están definiendo mecanismos de *federación* que permitan acceder a representaciones heterogéneas del conocimiento, como pueden ser sistemas basados en los objetos TIO mencionados anteriormente o estructuras que

utilicen RDF. Estos mecanismos deben englobar los servicios descritos aquí y ser capaces de incorporar nuevas fuentes de información como catálogos de bibliotecas, listas de correo, servidores de news o servidores de vídeo bajo demanda.

Referencias

1. UNESCO. *Clasificación por códigos códigos de ciencia y tecnología de la Unesco.* [En línea]. Disponible en <http://www.rediris.es/recursos/unesco.txt>
2. DUBLIN CORE METADATA INITIATIVE (DMCI). *Dublin Core Meta Data Element Set, Version 1.1: Reference Description*. [En línea]. Disponible en <http://purl.org/dc/documents/rec-dces-19990702.htm> 1999
3. HARVEST PROJECT. *Harvest Web Indexing*. [En línea]. Disponible en <http://www.tardis.ed.ac.uk/harvest/>
4. HEDBERG, R. y otros. *A Tagged Index Object for use in the Common Indexing Protocol*, RFC 2654. [En línea]. Disponible en <ftp://ftp.rediris.es/docs/rfc/26xx/2654> 1999.
5. REDIRIS. *Servicio de búsquedas*. [En línea]. Disponible en <http://www.rediris.es/busquedas/>
6. REDIRIS. *Servicio de indexación de las Redes Temáticas*. [En línea]. Disponible en <http://www.rediris.es/cvu/buscar/>
7. REDIRIS. *Servicio de Directorio*. [En línea]. Disponible en <http://www.rediris.es/sdir/>
8. REDIRIS. Webber. [En línea]. Disponible en <http://www.rediris.es/app/webber/>