Uso de temáticas y palabras clave sugeridas por software para mejorar la recuperación de tesis electrónicas a través del catálogo

Victor M. Ferracutti, Fernando A. Martinez
Biblioteca Central, Universidad Nacional del Sur
vmferra@uns.edu.ar, fmartinez@uns.edu.ar

Resumen. El acceso libre a la información científica es esencial para llevar a cabo la labor científica y plasmar los resultados de la investigación en beneficios tangibles para la sociedad. En este sentido, el núcleo básico de la producción científica en las universidades lo constituyen las tesis y disertaciones de posgrado. La propuesta de la Universidad Nacional del Sur, utilizando tecnologías ampliamente distribuidas y proveyendo un punto de acceso único a través de su catálogo, facilita el procesamiento del material digital mejorando el acceso a la información científica promoviendo la cooperación. El trabajo colaborativo entre bibliotecarios e informáticos, apoyados por la experiencia, investigación y práctica docente, ha resultado en un prototipo automatizado (software) que sugiere temáticas y palabras clave de un texto dado utilizando una base de conocimiento compuesta por documentos científicos. Con el uso de este sistema se enriquece al objeto digital con metadatos (i.e. temáticas y palabras clave) a través de los cuales es posible relacionar diferentes documentos de distinto tipo (por ejemplo: libros, artículos de revistas, tesis y disertaciones) del catálogo, ampliando así las capacidades de recuperación -de contenidos digitales en particular- para los usuarios finales. Por otra parte, estas recomendaciones automatizadas reducen el tiempo de catalogación de las tesis y disertaciones guiando al catalogador en el uso de temáticas y palabras clave preexistentes.

Introducción

Diferentes iniciativas han surgido con el objetivo de mejorar el sistema tradicional de comunicación científica y facilitar el libre acceso a las publicaciones a través de Internet. Tal es el caso de la Iniciativa de Budapest para el Acceso Abierto[1], la Declaración de Bethesda sobre Publicación de Acceso Abierto[2], los *Principles and Strategies for the Reform of Scholarly Communication*[3] de la *American Library Association* y subsiguientes.

En este sentido y tal como lo menciona (Melero, 2005) el éxito del movimiento de Acceso Abierto (del inglés *Open Access*) no sólo radica en sus ventajas respecto a la disponibilidad y el acceso a las publicaciones electrónicas, sino también en el apoyo de una comunidad científica que avala esta concepción para la difusión e impacto de la producción científica. Esta comunidad, distribuida internacionalmente con relevantes ejemplos de grupos de trabajo en las universidades de Southampton, Michigan, Cornell y Old Dominium, así como en centros de investigación como el CERN, CNR (Italia), CNRS (Francia) o Max Plank (Alemania), ha generado una serie de herramientas y servicios para la

[1] http://www.soros.org/openaccess/translations/spanish-translation
[2] http://ictlogy.net/articles/bethesda_es.html
[3] http://www.ala.org/ala/mgrps/divs/acrl/publications/whitepapers/principlesstrategies.cfm
[4] http://www.arl.org/bm~doc/arlstat03.pdf
[5] http://www.stanforddaily.com/2004/02/06/fac-sen-discusses-journal-fees/

gestión editorial, el archivo, la recuperación y búsqueda de información en recursos de Acceso Abierto, incluso para medir su impacto, que sirven de apoyo para emprender nuevos proyectos en este contexto.

En la Argentina se ha creado recientemente un Sistema Nacional de Repositorios Digitales[4] (SNRD) en respuesta a una iniciativa del Ministerio de Ciencia, Tecnología e Innovación Productiva conjuntamente con el Consejo Interinstitucional de Ciencia y Tecnología (CICyT), con el propósito conformar una red interoperable de repositorios digitales en ciencia y tecnología y fomentando el acceso público y gratuito respetando el modelo de Acceso Abierto (Ministerio de Ciencia, Tecnología e Innovación Productiva de la República Argentina, 2011).

En consonancia con estas propuestas que fomentan el acceso abierto a la producción científica, desde la Universidad Nacional del Sur (UNS) -a través de su Biblioteca Central- se han elaborado proyectos para la digitalización de la producción universitaria. El primero de estos proyectos, cuya fecha de comienzo data de 2001, se refiere a las tesis y disertaciones de posgrado, entendiendo a las mismas como el activo principal de la producción científica de las universidades. En este proyecto se decidió la adopción del modelo propuesto por ETD-Net de UNESCO[5], reconociendo que representa un eslabón en la cadena que constituye la Biblioteca Digital en una universidad y que es natural insertar el concepto de tesis y disertaciones electrónicas en una universidad (Ferracutti, Herrera, & Piccotto, 2005).

Cabe aclarar que estatutariamente la Biblioteca Central de la UNS (BC UNS), en su misión de preservar las tesis y la memoria académica de la institución, es depositaria de todas las tesis y disertaciones de postgrado de esta institución. En este sentido, los objetivos son:

- Facilitar la diseminación y el acceso al conocimiento producido por la institución;
- Integrar iniciativas nacionales de catalogación y publicación electrónica de tesis y disertaciones;
- Ofrecer productos y servicios a nivel nacional;
- Proveer acceso integrado a referencias o textos completos de tesis y disertaciones;
- Preservar la producción.

Luego de una instancia de apropiación de la tecnología para registrar los objetos digitales correspondientes a las tesis y disertaciones, y en función de la dificultad de clasificación de los mismos por la especificidad de los temas tratados en esos documentos, se observó la necesidad de mejorar el mecanismo de clasificación de los mismos en el catálogo de la biblioteca.

[4] http://repositorios.mincyt.gob.ar/
[5] http://portal.unesco.org/ci/en/ev.php-URL_ID=1580&URL_DO=DO_TOPIC&URL_SECTION=201.html

Como consecuencia de esa necesidad, el trabajo colaborativo entre bibliotecarios e informáticos, apoyados por la experiencia, investigación y práctica docente, ha resultado en herramientas en estado prototipo que usan ejemplos anteriores para generar sugerencias para la catalogación de recursos digitales de entrada y con la capacidad de reflejar un contexto temático durante el proceso de búsqueda. La herramienta propuesta aplica el razonamiento basado en casos en combinación con técnicas de recuperación de información para tomar ventaja de un amplio conjunto de recursos que previamente fueron clasificados con el fin de apoyar la tarea del catalogador.

De esta manera, la colaboración buscada se ha materializado en el trabajo con docentes y alumnos del Departamento de Ciencia e Ingeniería de la Computación de la UNS reflejado en (Delgado, Maguitman, Ferracutti, & Herrera, 2011; Dini, Varela, Antúnez, Maguitman, & Herrera, 2010; N. L. Mitzig & Mitzig, 2012) en donde se presentan herramientas que facilitan el procesamiento (clasificación e indización) automático de las tesis de posgrado.

Por otra parte, esta mejora en la clasificación facilita la recuperación de los materiales de una biblioteca a través del catálogo en línea (OPAC por sus siglas en inglés), por medio de descriptores (temáticas y palabras clave) compartidos entre diversos materiales. Este es un paso preliminar para la posibilidad de búsqueda contextualizada como elemento de los catálogos de próxima generación (Breeding, 2010; Yee, 2005).

En ese trabajo se presenta la aplicación de un prototipo automatizado de sugerencias de temáticas y palabras clave relevantes para la clasificación de tesis y disertaciones de posgrado, resultado de la colaboración entre bibliotecarios, infromáticos e investigadores que facilitará la búsqueda contextualizada de dichos recursos por parte de los usuarios a través de los catálogos.

Este documento se organiza de la siguiente manera. En la sección de Desarrollo se presenta la estructura que relaciona los diversos materiales gestionados por la Biblioteca Central, el prototipo automatizado de sugerencias y la verificación y validación inicial de su funcionamiento sobre la colección de tesis y disertaciones. En la sección de Conclusiones se identifican las lecciones aprendidas, se enuncia la validez inicial de la aplicación y se enumera el trabajo futuro.

Desarrollo

La BC UNS respeta un diseño (ver Ilustración 1. Relación entre distintos tipos de documentos) en donde cada tipo de material (referencial, texto completo, libros, revistas, videos, etc.) es tratado convenientemente con un software particular para cada caso. En todos los casos se aporta la descripción o metadatos al catálogo tal de facilitar la búsqueda y recuperación al usuario final. El catálogo es considerado el punto de acceso preferido para los usuarios.

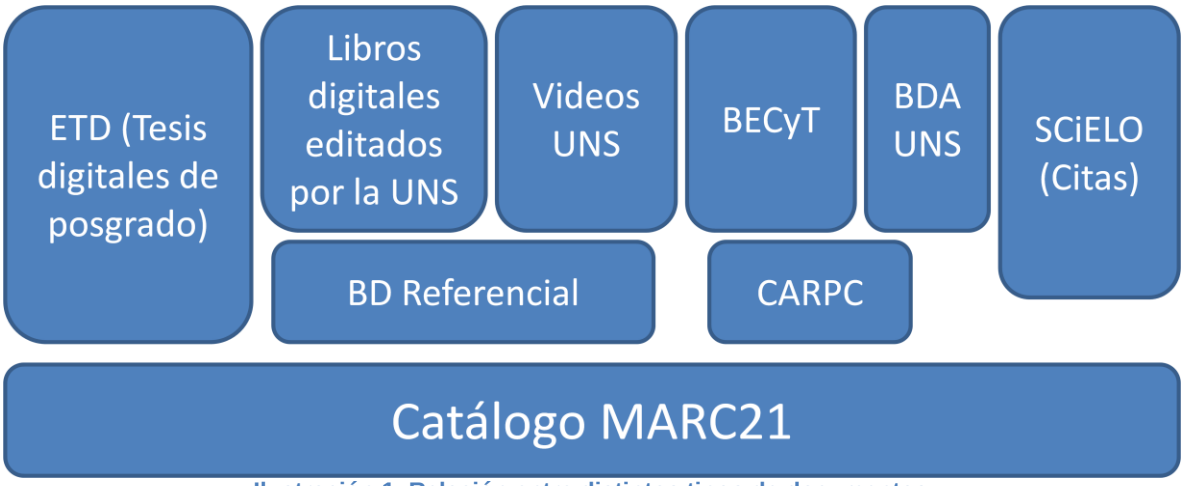

Ilustración 1. Relación entre distintos tipos de documentos

Dado que para el material impreso la BC UNS registra la información bibliográfica de acuerdo al formato de intercambio MARC21[6] y la normativa internacional de catalogación AACR2, ha optado por seguir con la misma forma de trabajo para los diferentes objetos digitales a los que les da tratamiento. Entre los objetos digitales se destacan las tesis y disertaciones de posgrado y las revistas editadas por la propia universidad.

Dadas las características propias de cada tipo de material y los metadatos que de cada uno de ellos se registran, una forma de facilitar la recuperación de materiales relacionados y en particular los objetos digitales es a través del uso de temáticas y palabras clave comunes que permitar una búsqueda temática más completa y pertinente.

Cabe aclarar que una buena parte de los metadatos (por ejemplo: título, autor, asesor, resumen) están explícitamente indicados en el propio objeto digital, mientras que otros (como las áreas temáticas y los descriptores) deben ser inferida por el catalogador.

Esto supone que utilizar una herramienta automatizada para ayudar al catalogador en el proceso de asignación de los datos que faltan reduce el tiempo involucrado en la catalogación promoviendo el uso de un vocabulario común.

Como organizar información que abarca diversos tópicos es una tarea difícil y costosa para el catalogador, quien probablemente no se encuentra familiarizado con la temática heterogénea y especializada de los recursos a clasificar, en (N. L. Mitzig & Mitzig, 2012) se presenta un prototipo automatizado para facilitar la catalogación de recursos electrónicos.

Tal como se observa en la Ilustración 2. Estructura del prototipo, esta herramienta está compuesta por:

[6] http://www.loc.gov/marc/

- un *Web Crawler*: recorre y descarga las temáticas asociadas a las revistas junto con todos los artículos en formato XML de revistas de los diferentes sitios SciELO, que conforman la base de conocimiento o entrenamiento;
- un clasificador o asignador de temáticas: se encarga de asignar la temática a los artículos de las revistas de acuerdo a la/s temática/s de la revista a la que pertenecen. Como resultado se genera un índice con los resultados obtenidos;
- un motor de sugerencias: utiliza el índice generado por el clasificador para proveer servicio (*Web Service*). Este utiliza el índice para buscar contenidos de artículos similares a un texto dado, retornando las tres mejores sugerencias de temáticas y palabras clave obtenidas;
- un complemento para Mozilla Firefox[7]: sirve de interfaz para el usuario haciendo uso del *Web Service*.

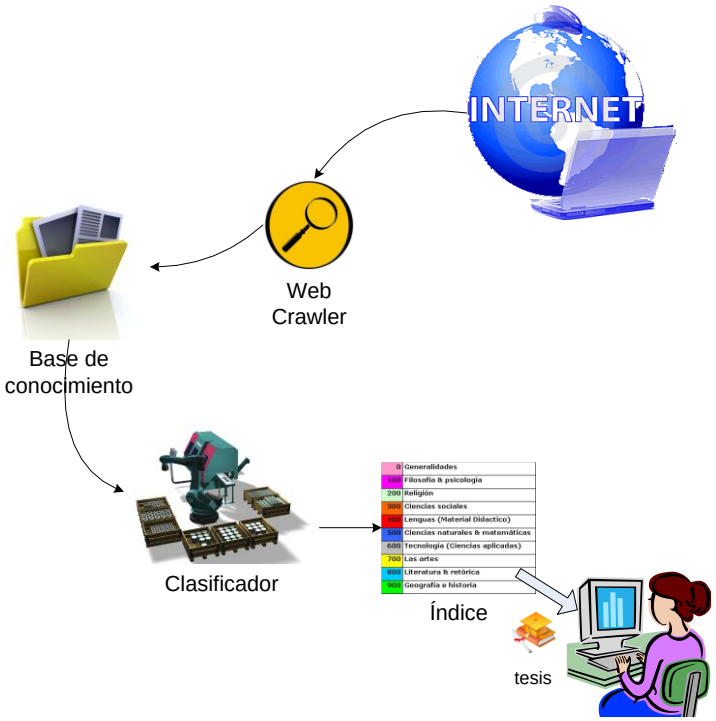

Ilustración 2. Estructura del prototipo

De acuerdo a las pruebas realizadas en (N. L. Mitzig & Mitzig, 2012) la calidad de las sugerencias depende de la cantidad y calidad de palabras del texto de entrada y la calidad de la base de conocimiento utilizada.

Respecto de la primera cuestión, en el caso particular de tesis y disertaciones como objetos digitales, es posible utilizar el resumen de las mismas como texto de entrada; pudiendo incluso utilizar también la introducción de las mismas como extensión del resumen.

[7] http://www.mozilla.org/en-US/firefox/new/

En cuanto a la base de conocimiento utilizada, es posible ampliarla con el conjunto de tesis y disertaciones existentes en la propia institución, asi como también con diversas bases de documentos científicos disponibles a texto completo.

Este *Web Service* ha sido verificado a través de la medición de su desempeño en cuanto a la precisión de sugerencias obtenidas, y validado en una fase exploratoria entre los catalogadores de la BC UNS, arrojado en ambos casos resultados positivos que sugieren continuar con el desarrollo.

Resultados

El *Web Service* presentado en este trabajo es consecuencia de la efectiva colaboración entre bibliotecarios, informáticos e investigadores en el área de la recuperación de la información. Teniendo una base de conocimientos adecuada, la herramienta ofrece sugerencias adecuadas que pueden ser utilizadas directamente como metadatos del material tratado. Por otra parte, existen indicios de que la herramienta disminuye el tiempo de catalogación. Esto es notorio para las tesis y disertaciones de posgrado, que tienen temas demasiado específicos para las capacidades de analisis y clasificacion de los catalogadores.

Si bien aún no se ha realizado una evaluación de usabilidad formal de la herramienta, existe aceptación inicial de la misma por parte de los catalogadores; habiéndose probado satisfactoriamente sobre diferentes versiones del navegador Mozilla Firefox.

Por otra parte, la implementación de la herramienta bajo el concepto de *Web Service*, permite utilizarla sobre diferentes aplicaciones siempre y cuando funcionen sobre un navegador Web en una computadora con conexión a Internet. Si bien los requerimientos para alojar y procesar la base de conocimientos son importantes en cuanto espacio de disco y memoria, es suficiente una sola computadora que puede ser el servidor donde luego se alojará el *Web Service*.

Conclusiones

En las bibliotecas actuales se procesan materiales de distintos tipos, incluyendo los tradicionales registros bibliográficos referenciales y también documentos digitales. Este procesamiento puede llevarse adelante utilizando sistemas automatizados heterogéneos por lo que existe una necesidad para que utilicen al menos descriptores homogéneos que faciliten la recuperación de materiales relacionados a través del OPAC.

La diversidad de necesidades y la variedad de tecnologías disponibles hacen necesaria la cooperación entre bibliotecarios, informáticos e investigadores, generando lenguajes que faciliten actividades de desarrollo e innovación.

En cuanto al trabajo futuro se observa la necesidad de explorar implementaciones de los modelos de referencia para bibliotecas digitales (por ejemplo: DELOS y 5S) para el desarrollo e integración de los distintos sistemas de acceso abierto en instituciones universitarias.

En cuanto al prototipo actualmente desarrollado, es menester ampliar base de conocimiento para proveer vocabulario compartido tal que se contemplen mayor diversidad de temáticas. Una posibilidad es utilizar los documentos disponibles a través de las tesis y disertaciones accesibles a través de la *Networked Digital Library of Theses and Dissertations* (NDLTD).

Por otra parte, deben realizarse diferentes *tests* de usabilidad para completar la validación de la herramienta.

También es necesario desarrollar la contraparte del recomendador, que se refiere a la búsqueda contextualizada a través del OPAC. Esto consiste en la utilización de uno o varios párrafos (extraídos de algún documento de texto representativo del tema objeto de la necesidad de información) para dar lugar a una búsqueda temática más completa y pertinente.

Bibliografía

Breeding, M. (2010). *Next-gen library catalogs*. New York: Neal-Schuman Publishers.

Delgado, P. H., Maguitman, A. G., Ferracutti, V. M., & Herrera, L. A. (2011). Using thematic contexts and previous solutions for maintaining and accessing institutional repositories. *Journal of Computer Science & Technology, 11*(2), 61-67. Retrieved from http://journal.info.unlp.edu.ar/journal/journal31/papers/JCST-Oct11-1.pdf

Dini, M. A., Varela, M. A., Antúnez, V., Maguitman, A. G., & Herrera, L. A. (2010). Soporte inteligente para el mantenimiento y acceso contextualizado a repositorios institucionales. Paper presented at the Buenos Aires.

Ferracutti, V. M., Herrera, L. A., & Piccotto, N. I. (2005). Maturing towards the digital library : Implementation of the electronic thesis and dissertations. *ETD2005 : Evolution through Discovery : 8th International Symposium on Electronic Theses and Dissertations,*

Melero, R. (2005). Significado del acceso abierto (open access) a las publicaciones científicas: definición, recursos
copyright e impacto. *El profesional de la información, 15*(4), 255-266.

Resolución MINCyT nº 469/11, (2011).

Mitzig, N. L., & Mitzig, M. S. (2012). *Sugerencias SciELO. sistema de apoyo para la catalogación en repositorios institucionales.* Universidad Nacional del Sur).

Yee, M. M. (2005). Guidelines for OPAC displays. In B. Sleeman, & P. Bluh (Eds.), *From catalog to gateway: Charting a course for future access: Briefings from the ALCTS catalog form and function committee* (pp. 83-90). Chicago: