Manual técnico

Sistema de Sugerencias de Palabras Clave

Lic. Mitzig, Mónica S., Lic. Mitzig, Natalia L.
Lic. Martinez, Fernando A. y Lic. Piriz, Ricardo A.

diciembre del 2013

Índice

Índice

Indice de tablas

1. Introducción

Este trabajo surge como una extensión y mejora del prototipo *Sugerencias SciELO: Sistema de apoyo para la catalogación en Repositorios Institucionales*, resultado de la Tesis de grado de *Mitzig, M. - Mitzig, N.*[1].

Se implementó un complemento para el navegador *Mozilla Firefox*, el sistema es *multiplataforma, fácil de usar y amigable,* pues con sólo tener texto seleccionado desde un sitio web o un *.pdf* que se abra con el navedador y hacer un click con el botón derecho del mouse, se tendrá la posibilidad de elegir la opción de "**Sugerir palabras clave**" más similares de acuerdo al texto seleccionado.

El *Sistema de Sugerencias de Palabras clave* se desarrolló a través de un complemento para *Mozilla Firefox*[2], que se comunica con un *Servicio Web* que reside en un servidor de la Biblioteca Central de la Universidad Nacional del Sur.

Un requisito fundamental para hacer uso de la herramienta propuesta, es que el usuario del sistema de sugerencias debe utilizar ***Mozilla Firefox*** como navegador.

2. Herramientas utilizadas

Para el funcionamiento del sistema desarrollado se requieren de las siguientes herramientas y bibliotecas/librerías:

- Servidor Web: *Apache Tomcat*[3],
- *Java(TM) SE Development Kit 7*[1], ya que las versiones anteriores no cuentan con el paquete *java.nio.file.File.*,
- *Apache Lucene* [2] para java y
- *javax* [3]

3. Trabajo realizado

3.1. Cosecha de metadatos

Para la cosecha/descarga de metadatos se utilizó un software libre desarrolado por ***Public Knowledge Project (PKP)***[4] ***denominado Open Harvester Systems (OHS)***[5], el cual es un sistema de indexación de metadatos gratuito. *OHS* permite crear un índice de búsqueda de los metadatos a partir del protocolo *OAI (Open Archives Initiative)*[4].

Con dicho software, se realizó la cosecha en formato ***Dublin Core (DC)***[5] de todos los repositorios listados en la *BDU2*[6] y de todos los sitios *SciELO oficiales*[7].

En el ***Anexo I*** se puede observar el listado de todos los sitios que se cosecharon utilizando la aplicación *OHS*.

[1] *http://www.oracle.com/technetwork/java/javase/downloads/jdk-7u2-download-1377129.html*
[2] lucene-core-3.0.1.jar y lucene-queries-3.0.1.jar
[3] javax.ejb.jar, `http://www.java2s.com/`
[4] `http://www.openarchives.org/`
[5] `http://dublincore.org/`
[6] `http://bdu.siu.edu.ar/cgi-bin/reporpt.pl`
[7] `http://www.scielo.org/`

En la siguiente tabla, se puede observar la cantidad de recursos digitales cosechados/descargados por repositorio:

Tabla 1: Cantidad de recursos digitales descargados por Repositorio.

Repositorios	***Cantidad de recursos descargados***
RI de la Universidad de Belgrano	1349
Bib. Dig. Fac. de Cs. Exactas y Naturales - UBA.	276
SciELO Portugal	630
SciELO Sudáfrica	6656
SciELO España	24665
Bib. Dig. Académica - UNS	462
SciELO Chile	10800
RI de la Univ.de Cs. Empresariales y Sociales	1600
Bib. Virtual UNL	6246
SciELO Perú	1260
INIDEP@OceanDocs	4563
SciELO Brasil	9159
Rep. OAI Bib. Dig. UNCUYO	1783
Red de Bibs. Virtuales de Cs. Sociales (CLACSO)	9333
Bib. Dig. Univ. Católica Arg.	2478
Cor-ciencia	1200
Cartapacio	1325
SciELO Venezuela	8688
Rep. Dig. UDESA	803
SciELO Cuba	5996
Nülan	1886
NDLTD	76000
SciELO México	19400
SciELO Argentina	12310
SciELO - Salud Pública	22664
SEDICI-UNLP	24799
FLACSOAndes	4373
Rep. Dig. UFASTA	142
RI del Ministerio de Educ. de la Nación	33204
Memoria Académica - FaHCE	20503
Rep.Hip.UNR	1208
SciELO Costa Rica	5054
SciELO Social Sciences English Edition	665
SciELO Colombia	3218
Total de recursos descargados	**324698**

3.2. Generación del índice de registros

Una vez finalizada la cosecha de todos los repositorios que se encuentran en la *BDU2* y de todos los sitios *SciELO oficiales*, se desarrolló una aplicación para generar un archivo *XML* para cada uno de los registros cosechados que

cumplan con las siguientes características:

- Recurso cuyo idioma sea *Español*
- *Título* del recurso digital,
- *Palabras clave* y
- *Descripción* o *resumen* del recurso.

En el siguiente cuadro, se puede observar la cantidad de recursos digitales que cumplieron con las todas las condiciones especificadas anteriormente, los cuales formarán parte de nuestra base de conocimiento. Por ende, todos estos recursos se utilizaron para la generación de un índice denominado *índice de registros*.

Tabla 2: Cantidad de recursos digitales indexados por Repositorio.

Repositorios	***Cantidad de recursos indexados***
RI de la Universidad de Belgrano	1289
Bib. Dig. Fac. de Cs. Exactas y Naturales - UBA.	97
SciELO España	13489
Bib. Dig. Académica - UNS	249
SciELO Chile	5557
RI de la Univ.de Cs. Empresariales y Sociales	1106
Bib. Virtual UNL	1768
SciELO Perú	930
INIDEP@OceanDocs	1839
SciELO Brasil	11
Rep. OAI Bib. Dig. UNCUYO	939
Bib. Dig. Univ. Católica Arg.	1858
Cartapacio	1142
SciELO Venezuela	6097
Rep. Dig. UDESA	551
SciELO Cuba	5085
Nülan	1315
NDLTD	9
SciELO México	11298
SciELO Argentina	7155
SciELO - Salud Pública	2592
SEDICI-UNLP	16666
FLACSOAndes	25
Rep. Dig. UFASTA	137
RI del Ministerio de Educ. de la Nación	1295
Memoria Académica - FaHCE	9259
Rep.Hip.UNR	605
SciELO Costa Rica	2651
SciELO Colombia	2173
Total de recursos indexados	**97187**

En la siguiente tabla, se indica la cantidad de recursos que no se tuvieron en cuenta para la generación del índice de recursos y se descartaron por no cumplir con algunos de los requisitos especificados anteriormente, ya sea porque el idioma del recurso no es español, por no contar con palabras clave, título o descripción/resumen.

Tabla 3: Cantidad de recursos digitales descartados por Repositorio.

Repositorios	***Cantidad de recursos descartados***
RI de la Universidad de Belgrano	60
Bib. Dig. Fac. de Cs. Exactas y Naturales - UBA.	179
SciELO Portugal	630
SciELO Sudáfrica	6656
SciELO España	11176
Bib. Dig. Académica - UNS	213
SciELO Chile	5243
RI de la Univ.de Cs. Empresariales y Sociales	494
Bib. Virtual UNL	4478
SciELO Perú	330
INIDEP@OceanDocs	2724
SciELO Brasil	9148
Rep. OAI Bib. Dig. UNCUYO	844
Red de Bibs. Virtuales de Cs. Sociales (CLACSO)	9333
Bib. Dig. Univ. Católica Arg.	620
Cor-ciencia	1200
Cartapacio	183
SciELO Venezuela	2591
Rep. Dig. UDESA	252
SciELO Cuba	911
Nülan	571
NDLTD	75991
SciELO México	8102
SciELO Argentina	5155
SciELO - Salud Pública	20072
SEDICI-UNLP	8133
FLACSOAndes	4348
Rep. Dig. UFASTA	5
RI del Ministerio de Educ. de la Nación	31909
Memoria Académica - FaHCE	11244
Rep.Hip.UNR	603
SciELO Costa Rica	2403
SciELO Social Sciences English Edition	665
SciELO Colombia	1045
Total de recursos descartados	**227501**

Para mayor información, consultar la siguiente documentación: `../ManualTecnico/javadocGeneradorXML/index.html`

3.3. Servicio web

Se desarrolló un motor de búsqueda a través de un servicio web denominado *Web Service Sugerencias* basado en *SOAP*, el entorno utilizado fue NetBeans IDE 7.4[6], utilizando el lenguaje de programación *Java* [8] y como servidor de aplicaciones utilizamos *Tomcat Apache*[9].

Java provee utilidades que permiten crear archivos *WSDL* basado en el código Java del servicio web. Este archivo se expone en la red para ser utilizado por los clientes interesados en usar el servicio. Los mensajes se intercambian en formato *SOAP*.

Además, para el desarrollo de dicha aplicación, se utilizaron las siguientes bibliotecas/librerías *JDOM*[10], *Apache Lucene* [11], *Commons Digester* [12] y por último *MoreLikeThis*[7] [13] para realizar búsqueda por *similitud*.

El principal método de nuestro servicio web es *obtenerSugerencias*, el cual, dado un texto retorna las tres primeras sugerencias obtenidas con *MoreLikeThis* como resultado de consultar el *índice de registros* generado anteriormente, de acuerdo al texto dado.

Por lo tanto, para poder ejecutar correctamente el servicio web: *Web Service Sugerencias*, necesitamos que en el servidor se encuentren los siguientes elementos:

- El *.war* correspondiente al servicio web: *WebServiceSugerencias.war*,
- Una carpeta llamada *IndiceRecodrs*, la cual contiene el índice de registros y
- El archivo correspondiente a los *stop words*.

Para más información, consultar: `../ManualTecnico/javadocWebServiceSugerencias/index.html`

3.4. Complemento Sugerencias de Palabras clave

Luego, se desarrolló un complemento para *Mozilla Firefox*[2] denominado *Sugerencias de Palabras clave*, que utiliza el servicio web explicado anteriormente, para obtener las tres mejores sugerencias de palabras clave según el texto seleccionado por el usuario del complemento. Dicho complemento, funciona tanto para texto seleccionado desde una página web o desde un archivo .pdf que se abra con el lector de *.PDF* de *Firefox*.

La implementación de dicho complemento para *Mozilla Firefox*, se realizó utilizando *Add−ons Builder*[14], el cual proporciona una potente *API*[2] y herramientas para desarrollar complementos utilizando *HTML, CSS* y *JavaScript*. Utilizamos la librería *Addons SDK 1.14*.

Lo primero que hicimos fue crear un *Context-Menu* utilizando la *API Context-Menu*, para agregar un ítem *"Sugerir Palabras Clave"* al menú contextual de la página. De esta manera, cuando el usuario selecciona texto de alguna página/formulario web o un .pdf que se abra con el lector de *.PDF* de *Firefox*, al hacer click con el botón derecho del mouse, aparecerá en el menú la opción a seleccionar: *Sugerir Palabras Clave*.

De esta forma, cuando el usuario elige dicha funcionalidad, se mostrará en un panel las ***tres mejores sugerencias*** obtenidas de haber consultado el índice de registros, según el texto seleccionado por el usuario. La información que se muestra en el panel es obtenida a través del servicio web: *Web Service Sugerencias*, explicado anteriormente.

El usuario del complemento ***Sugerencias de Palabras Clave*** podrá seleccionar una *extensa cantidad de texto*, ya que se utiliza el método *POST* para pasar el texto seleccionado al servicio web. Además, el sistema funciona correctamente cuando el texto seleccionado por el usuario contiene *caracteres especiales*, como por ejemplo: @, ±, −, etc .

[8] La elección se debe al hecho de que el trabajo de Tesis de *Mitzig, M. - Mitzig, N.*[1] fue desarrollado utilizando el mismo lenguaje de programación.

[9] `http://tomcat.apache.org/`

[10] *http://www.jdom.org*

[11] *http://lucene.apache.org/*

[12] *http://commons.apache.org/digester*

[13] *http://lucene.apache.org/core/old_versioned_docs/versions/3_0_1/api/contrib-queries/org/apache/lucene/search/similar/MoreLikeThis.html*, versión 3.0.1

[14] `https://builder.addons.mozilla.org/`

Anexo I

A. Listado de Repositorios cosechados

En el siguiente listado, se pueden observar todos los repositorios y de sitios *SciELO* oficiales, de los cuales se cosecharon/descargaron recursos digitales en formato OAI.

- Repositorio OAI Biblioteca Digital Universidad Nacional de Cuyo - `http://bdigital.uncu.edu.ar/OAI/index.php`
- Scientific Electronic Library Online (SciELO Argentina) - `http://www.scielo.org.ar/oai/scielo-oai.php`
- Scientific Electronic Library Online (SciELO Chile) - `http://www.scielo.cl/oai/scielo-oai.php`
- Biblioteca Digital Academica - Universidad Nacional del Sur `http://bibliotecadigital.uns.edu.ar`
- SciELO Social Sciences English Edition - `http://socialsciences.scielo.org/oai/scielo-oai.php`
- SciELO - Salud Pública - `http://www.scielosp.org/oai/scielo-oai.php`
- Memoria Académica - Facultad de Humanidades y Ciencias de la Educación - `http://www.memoria.fahce.unlp.edu.ar/oaiserver.cgi`
- Repositorio Institucional del Ministerio de Educacion de la Nacion - `http://repositorio.educacion.gov.ar:8080/oai/request`
- Nülan - Universidad Nacional de Mar del Plata - Facultad de Ciencias Económicas y Sociales - `http://nulan.mdp.edu.ar/cgi/oai2`
- Biblioteca Digital Universidad Catolica Argentina - `http://bibliotecadigital.uca.edu.ar/greenstone/cgi-bin/oaiserver.cgi`
- NDLTD - `http://union.ndltd.org:8080/union.OAI-PMH/`
- Scientific Electronic Library Online (SciELO Perú) - `http://www.scielo.org.pe/oai/scielo-oai.php`
- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista: Cartapacio de Derecho y secciones: `http://www.cartapacio.edu.ar/ojs/index.php/ctp/oai/`
- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista: Investigacion y Docencia: `http://www.cartapacio.edu.ar/ojs/index.php/iyd/oai/`
- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista:Revista del Centro de Investigaciones en Filosofia Juridica y Filosofia Social: `http://www.cartapacio.edu.ar/ojs/index.php/centro/oai`
- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista: Revista Jurídica del Centro: `http://www.cartapacio.edu.ar/ojs/index.php/RJC/oai`
- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista: Revista del Centro de Investigaciones en Ciencias Sociales: `http://www.cartapacio.edu.ar/ojs/index.php/rcicso/oai`
- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista: Bioética y Bioderecho: `http://www.cartapacio.edu.ar/ojs/index.php/byb/oai`

- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista: Libros de Integrativismo Trialista: http://www.cartapacio.edu.ar/ojs/index.php/mundojuridico/oai
- Cartapacio - Universidad Nacional del Centro - Facultad de Derecho - Revista: Trabajos del Centro: http://www.cartapacio.edu.ar/ojs/index.php/tdc/oai
- Cor-ciencia - Ministerio de Ciencia y Tecnología - Córdoba (Provincia) - http://www.corciencia.org.ar/cgi/oai2
- FLACSOAndes - Sede Argentina FLACSO - http://www.flacsoandes.org/oai/request
- INIDEP@OceanDocs - Ministerio de Economía de la Nación - http://www.oceandocs.org/odin-oai/request
- Red de Bibliotecas Virtuales de Ciencias Sociales (CLACSO) - Consejo Latinoamericano de Ciencias Sociales - http://biblioteca.clacso.edu.ar/gsdl/cgi-bin/oaiserver.cgi
- Rep.Hip.UNR - Repositorio Hipermedial de la Universidad Nacional de Rosario - http://rephip.unr.edu.ar/oai/request
- Repositorio Digital UDESA - Universidad de San Andrés - http://repositorio.udesa.edu.ar/oai/request
- Repositorio Digital UFASTA - http://redi.ufasta.edu.ar:8080/oai/request
- Repositorio Institucional de la UCES - Universidad de Ciencias Empresariales y Sociales - http://dspace.uces.edu.ar:8180/oai/request
- Repositorio Institucional de la Universidad de Belgrano - Universidad de Belgrano - http://repositorio.ub.edu.ar:8080/oai/request
- SEDICI-UNLP - Repositorio Institucional de la UNLP - http://sedici.unlp.edu.ar/oai/request
- Scientific Electronic Library Online (SciELO Colombia) - http://www.scielo.org.co/oai/scielo-oai.php
- Scientific Electronic Library Online (SciELO Costa Rica) - http://www.scielo.sa.cr/oai/scielo-oai.php
- Scientific Electronic Library Online (SciELO España) - http://scielo.isciii.es/oai/scielo-oai.php
- Biblioteca Virtual - Tesis - Universidad Nacional del Litoral - http://bibliotecavirtual.unl.edu.ar:8180/tesis-oai/request
- Scientific Electronic Library Online (SciELO Sudáfrica) - http://www.scielo.org.za/oai/scielo-oai.php
- Scientific Electronic Library Online (SciELO Venezuela) - http://www.scielo.org.ve/oai/scielo-oai.php
- Scientific Electronic Library Online (SciELO México) - http://www.scielo.org.mx/oai/scielo-oai.php
- Scientific Electronic Library Online (SciELO Cuba) - http://scielo.sld.cu/oai/scielo-oai.php
- Scientific Electronic Library Online (SciELO Portugal) - http://www.scielo.oces.mctes.pt/oai/scielo-oai.php
- Scientific Electronic Library Online (SciELO Brasil) - http://www.scielo.br/oai/scielo-oai.php
- Biblioteca Digital Facultad de Ciencias Exactas y Naturales - Universidad de Buenos Aires - http://digital.bl.fcen.uba.ar/gsdl-282/cgi-bin/oaiserver.cgi

De todos los recursos que se descargaron de los repositorios mencionados, sólo se utilizaron los recursos que cumplieron con los requisitos especificados anteriormente.

Referencias

[1] Mitzig, Mónica S. y Mitzig, Natalia L. (2012) *Tesis de Licenciatura - Sugerencias SciELO: Sistema de apoyo para la Catalogación en Repositorios Institucionales* Universidad Nacional del Sur.

[2] *Facilidad para crear complementos para Firefox.* `https://addons.mozilla.org/es-ES/developers/`

[3] Servidor web *Apache Tomcat* `http://tomcat.apache.org/`

[4] Public Knowledge Project (PKP) `http://pkp.sfu.ca/`

[5] Documentación del Software Open Harvester Systems (OHS) `http://pkp.sfu.ca/ohs/ohs_documentation/`

[6] *Introducción a Web Services.* `http://netbeans.org/kb/docs/websvc/intro-ws.html`

[7] Clase *MoreLikeThis* de la librería *Lucene* `http://lucene.apache.org/core/old_versioned_docs/versions/3_0_1/api/all/org/apache/lucene/search/package-summary.html`