UNIVERSIDAD DE LA SALLE
Educar para Pensar, Decidir y Servir
Programa de Sistemas de Información y Documentación, Bibliotecología y Archivística

# MEMORIAS DE PRÁCTICAS EN EL USO DEL INDIZADOR SWISH-E

## SISTEMAS DE INDIZACIÓN Y RECUPERACIÓN DE LA INFORMACIÓN DIGITAL

Desarrollado Por: *Laureano Felipe Gómez D.* – *felipe.gomez3@gmail.com*

## UNIVERSIDAD DE LA SALLE

2009

UNIVERSIDAD DE LA SALLE
Educar para Pensar, Decidir y Servir
Programa de Sistemas de Información y Documentación, Bibliotecología y Archivística

**Proyecto de Asignatura**
**MEMORIAS DE PRÁCTICAS DE USO DEL INDIZADOR SWISH-E**

**ASIGNATURA: RECUPERACIÓN DE LA INFORMACIÓN DIGITAL**
**AUTOR(ES):** *Laureano Felipe Gómez Dueñas – felipe.gomez3@gmail.com*
**FECHA DE ELABORACIÓN:**
**NOTAS ADICIONALES:**

**PALABRAS CLAVES***: Sistemas de Recuperación de la Información, Indizadores, Software Libre*

**INDICE GENERAL**

# Introducción

El siguiente trabajo corresponde a un compendio de las prácticas realizadas utilizando el sistema de indización Swish-e realizadas en la asignatura "recuperación de la información digital".

El material asociado a las prácticas se pueden descargar del servidor llamado **BOMARZO** (http://bomarzo.rec.usal.es/ ), desde allí se puede acceder a todos los archivos generados y las colecciones de prueba utilizadas (http://bomarzo.rec.usal.es/swish-e/)

# Qué es Swish-e

Swish-e es un pequeño programa de indización y búsqueda de documentos no estructurados, este programa está basado en el sistema de indización SWISH que fue escrito originalmente por Kevin Hughes.

Swish-e no es una solución del tipo "Llave en Mano" que pueda descargar y utilizar cualquier usuario son conocimientos de informática, este programa fue desarrollado como una solución que sirva de **BASE** para el desarrollo de programas más avanzados, por lo cual necesita un nivel intermedio de conocimientos en informática y en algún lenguaje de programación (recomendado PERL).

Para esta práctica se ha decidido trabajar con un sistema operativo LINUX FEDORA CORE 7, todo el trabajo realizado s ha documentado mediante ejemplos y figuras, para que puedan ser realizados y analizados por cualquier usuario que disponga de un sistema operativo del tipo LINUX.

# Instalación de Swish-e en Fedora Core 7

Para el desarrollo de estas prácticas, se ha decidido instalar el programa Swish-e en un sistema operativo Linux, más concretamente utilizando la distribución **Fedora Core 7**. El motivo principal de esta decisión está dado en la facilidad de instalación, la preexistencia del lenguaje **PERL** instalado y su gran integración con el sistema operativo Linux (donde fue diseñado originalmente Swish-e).

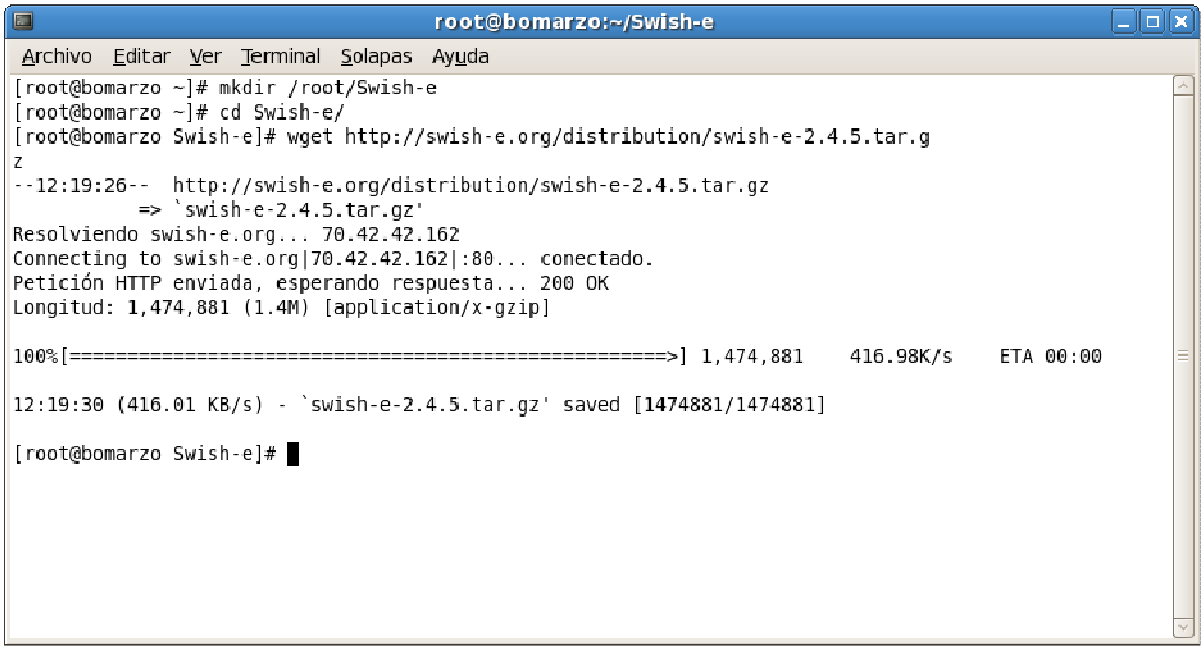
A continuación se detallan los pasos que se realizaron para la instalación del Software Swish-e

- **Creación de un directorio de trabajo:** Este paso consiste en crear un sitio donde se realizarán las prácticas asociadas, para esto decidimos crear una directorio llamado "Swish-e", en el directorio por defecto del usuario root. (**/root/Swish-e),** para lo cual ejecutamos el siguiente comando en una sesión de terminal (consola de comandos) y luego nos ubicamos en este directorio:

**[root@localhost ~]#mkdir /root/Swish-e**
**[root@localhost ~]#cd /root/Swish-e**

- **Descargamos el aplicativo Swish-e:** Para esto vamos a utilizar el comando ***wget***, el cual nos provee un mecanismo de interacción con los sitios web mediante una línea de comandos:

**wget http://swish-e.org/distribution/swish-e-2.4.5.tar.gz**




- **Instalamos el programa catdoc: Catdoc** es un programa que lee uno o más archivos de Microsoft Word y saca el contenido del texto dentro de ellos, este programa los utilizaremos para indizar archivos .DOC.

Actualmente la descarga de **Catdoc** (un software libre creado por un informático ruso) incorpora también la de **xls2csv** y **catppt**, dirigidos a extraer el texto contenido en los documentos Excel y Powerpoint, respectivamente

```
[root@bomarzo Swish-e]# yum install catdoc.i386
Setting up Install Process
Parsing package install arguments
Resolving Dependencies
--> Running transaction check
---> Package catdoc.i386 0:0.94.2-3.fc7 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved

================================================================
 Package              Arch      Version        Repository      Size
================================================================
Installing:
 catdoc               i386      0.94.2-3.fc7   updates         120 k

Transaction Summary
================================================================
Install      1 Package(s)
Update       0 Package(s)
Remove       0 Package(s)

Total download size: 120 k
Is this ok [y/N]: y
Downloading Packages:
(1/1): catdoc-0.94.2-3.fc 100% |=========================| 120 kB    00:00
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing: catdoc                ########################### [1/1]

Installed: catdoc.i386 0:0.94.2-3.fc7
Complete!
```

- **Instalamos el programa unrtf:** UNRTF es un programa que permite convertir archivos **RTF** a varios formatos entre ellos **HTML.**

```
[root@bomarzo trabajo2]# yum install unrtf.i386
Setting up Install Process
Parsing package install arguments
Resolving Dependencies
--> Running transaction check
---> Package unrtf.i386 0:0.20.2-2.fc7 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved

================================================================
 Package              Arch      Version         Repository      Size
================================================================
Installing:
 unrtf                i386      0.20.2-2.fc7    fedora          40 k

Transaction Summary
================================================================
Install      1 Package(s)
Update       0 Package(s)
Remove       0 Package(s)

Total download size: 40 k
Is this ok [y/N]: y
Downloading Packages:
(1/1): unrtf-0.20.2-2.fc7 100% |=========================|  40 kB    00:00
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing: unrtf                 ########################### [1/1]

Installed: unrtf.i386 0:0.20.2-2.fc7
Complete!
```

- **Instalamos los utilitarios xpdf:** Xpdf es un visor para el Formato Portable de Documento (PDF) libre de Adobe, que es rápido, pequeño y viene con algunas utilidades en línea de comandos. El paquete Xpdf contiene xpdf,

pdftops, pdftotext, pdftopbm, pdffonts, pdfimages y pdfinfo, algunas de estas utilidades los utilizaremos para extraer el texto de los archivos PDF.

```
[root@bomarzo Swish-e]# yum install xpdf
Setting up Install Process
Parsing package install arguments
Resolving Dependencies
--> Running transaction check
---> Package xpdf.i386 1:3.02-4.fc7 set to be updated
--> Processing Dependency: libXm.so.2 for package: xpdf
--> Processing Dependency: libt1.so.5 for package: xpdf
--> Processing Dependency: poppler-utils for package: xpdf
--> Running transaction check
---> Package poppler-utils.i386 0:0.5.4-8.fc7 set to be updated
---> Package t1lib.i386 0:5.1.1-7.fc7 set to be updated
---> Package lesstif.i386 0:0.95.0-20.fc7 set to be updated
--> Processing Dependency: libXp.so.6 for package: lesstif
--> Running transaction check
---> Package libXp.i386 0:1.0.0-8 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved

================================================
 Package              Arch      Version        Repository      Size
================================================
Installing:
 xpdf                 i386      1:3.02-4.fc7   updates         1.1 M
Installing for dependencies:
 lesstif              i386      0.95.0-20.fc7  updates         745 k
 libXp                i386      1.0.0-8        fedora          22 k
 poppler-utils        i386      0.5.4-8.fc7    updates         75 k
 t1lib                i386      5.1.1-7.fc7    updates         194 k

Transaction Summary
================================================
Install      5 Package(s)
Update       0 Package(s)
Remove       0 Package(s)

Total download size: 2.1 M
Is this ok [y/N]: y
Downloading Packages:
(1/5): lesstif-0.95.0-20. 100% |=========================| 745 kB    00:00
```

```
(2/5): libXp-1.0.0-8.i386 100% |=========================|  22 kB    00:00
(3/5): xpdf-3.02-4.fc7.i3 100% |=========================| 1.1 MB    00:00
(4/5): t1lib-5.1.1-7.fc7. 100% |=========================| 194 kB    00:00
(5/5): poppler-utils-0.5. 100% |=========================|  75 kB    00:00
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing: libXp                 ########################### [1/5]
  Installing: lesstif               ########################### [2/5]
  Installing: t1lib                 ########################### [3/5]
  Installing: poppler-utils           ########################### [4/5]
  Installing: xpdf                  ########################### [5/5]

Installed: xpdf.i386 1:3.02-4.fc7
Dependency Installed: lesstif.i386 0:0.95.0-20.fc7 libXp.i386 0:1.0.0-8 poppler-utils.i386 0:0.5.4-8.fc7 t1lib.i386 0:5.1.1-7.fc7
Complete!
```

- **Descomprimimos el archivo descargado de Swish-e:** Para este paso simplemente ejecutamos el siguiente comando en la ruta donde descargamos el archivo con el comando **wget.** Posteriormente nos ubicamos en el directorio creado.

**tar zxvf swish-e-2.4.5.tar.gz**
**cd swish-e-2.4.5**


- **Compilación e instalación de Swish-e:** en este paso compilados el código fuente descargado para generar los archivos ejecutables (aplicativos) de Swish-e con los que vamos a trabajar y luego hacer la instalación automática de los archivos creados, para esto debemos ejecutar los siguientes comandos:

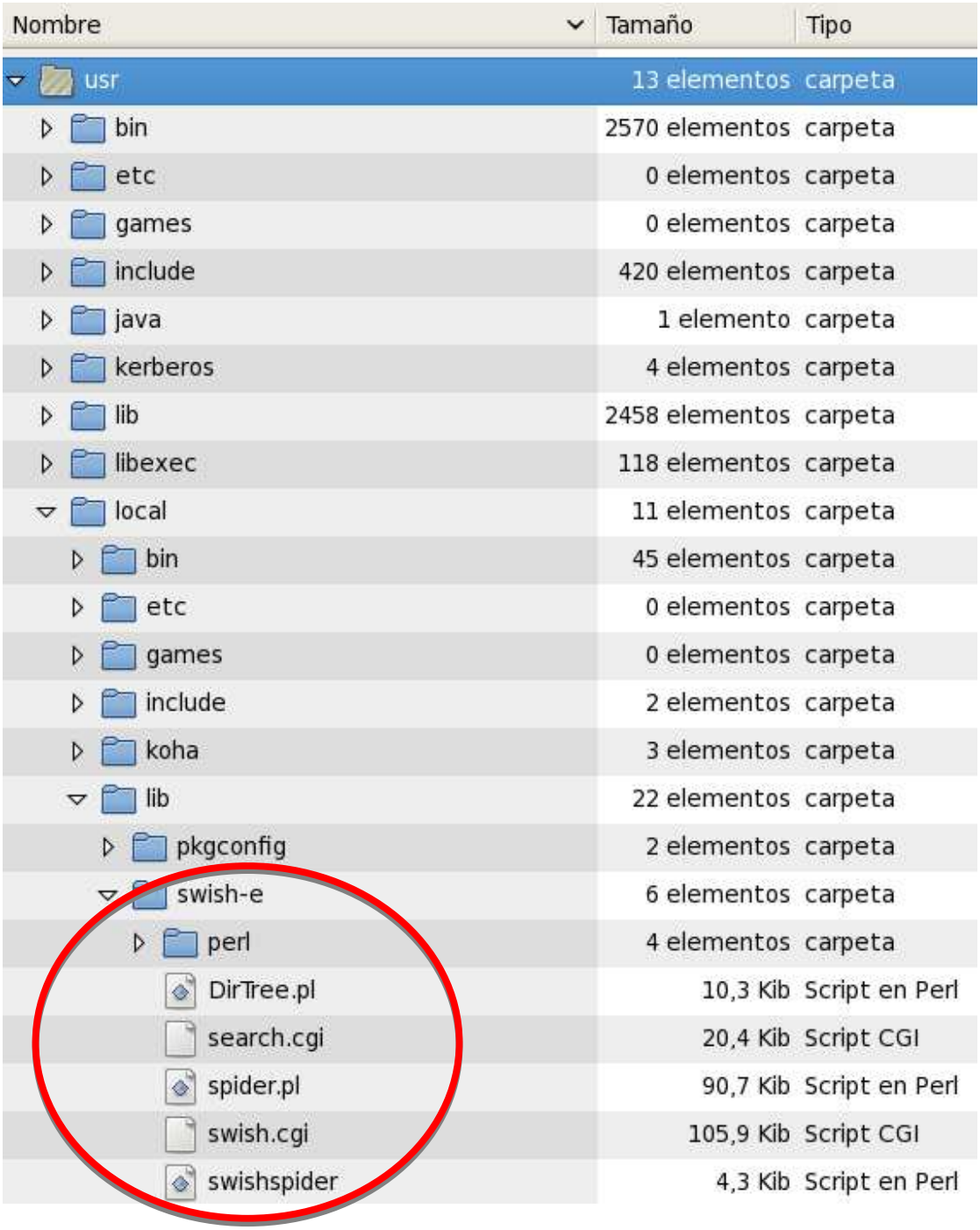
**./configure**
**make install**

- **Verificamos Instalación:** En este paso confirmamos donde quedo instalado el programa Swish-e, para esto ejecutamos el siguiente comando

**[root@bomarzo swish-e-2.4.5]# whereis swish-e**
**swish-e: /usr/local/bin/swish-e /usr/local/lib/swish-e**

**Este comando nos indica que:**

- El programa ejecutable de Swish-e se encuentra en: **/usr/local/bin/swish-e**
- Los archivos de configuración y otros se encuentran en: e **/usr/local/lib/swish-e**


Nombre
Tamaño
Tipo
usr 13 elementos carpeta
bin 2570 elementos carpeta
etc 0 elementos carpeta
games 0 elementos carpeta
include 420 elementos carpeta
java 1 elemento carpeta
kerberos 4 elementos carpeta
lib 2458 elementos carpeta
libexec 118 elementos carpeta
local 11 elementos carpeta
bin 45 elementos carpeta
etc 0 elementos carpeta
games 0 elementos carpeta
include 2 elementos carpeta
koha 3 elementos carpeta
lib 22 elementos carpeta
pkgconfig 2 elementos carpeta
swish-e 6 elementos carpeta
perl 4 elementos carpeta
DirTree.pl 10,3 Kib Script en Perl
search.cgi 20,4 Kib Script CGI
spider.pl 90,7 Kib Script en Perl
swish.cgi 105,9 Kib Script CGI
swishspider 4,3 Kib Script en Perl

- **Instalando los módulos de PERL:** Según el manual de instalación del programa Swish-e, para optimizar el funcionamiento del spider (Programa diseñado para recorrer la web siguiendo los enlaces entre páginas. Esta es la forma habitual empleada por los principales buscadores para encontrar las páginas que posteriormente forman parte de sus bases de datos), se requieren instalar las siguientes librerías del lenguaje de programación PERL:

    - **libwww-perl (LWP)** - librería estándar de PERL para web, esta librería es utilizada por el script spider.pl.
    - **URI –** Usada para identificar las URL y decodificarlas
    - **HTML-Tagset (HTML::Tagset)** - Usada por el spider.pl para identificar las etiquetas HTML
    - **HTML-Parser (HTML::Parser)** - Usada por el spider.pl para identificar el contenido de las páginas web HTML.
    - **MIME-Types (MIME::Types)** - Usada para filtrar documentos por su formato
    - **HTML-Template (HTML::Template)** – Usada para formatear la salida de los scripts CGI de Swish-e (swish.cgi)
    - **HTML-FillInForm (HTML::FillInForm)** - Utilizada para manejar plantillas web durante la presentación de resultados

    Para instalar estas librerías solo se requiere ejecutar los siguientes comandos desde una terminal (consola de comandos):

    **perl -MCPAN -e 'install "LWP"'**
    **perl -MCPAN -e 'install " URI "'**
    **perl -MCPAN -e 'install "Bundle::LWP"'**
    **perl -MCPAN -e 'install "MIME::Type"'**
    **perl -MCPAN -e 'install "Template"'**
    **perl -MCPAN -e 'install "HTML::FillInForm"'**
    **perl -MCPAN -e 'install "Test::More"'**
    **perl -MCPAN -e 'install "Spreadsheet::ParseExcel"'**
    **perl -MCPAN -e 'install "HTML::Entities"'**

Adicionalmente se recomienda instalar las siguientes librerías que incorporar procedimientos especializados para trabajar con documentos indizados con Swish-e:

**perl –MCPAN -e 'install "**`SWISH::API`**"'**
**perl –MCPAN -e 'install "**`SWISH::API::More`**"'**

- **Instalando las librerías para PHP:** El objetivo de utilizar el lenguaje de programación PHP para hacer búsquedas en los índices de Swish-e están dados en la facilidad para crear un catálogo muy completo y funcional, aprovechando la robustez de este lenguaje de programación, para instalar las librerías de PHP que permitan interactuar con Swish-e ejecute los comandos listados a continuación:

- **Instalar PEAR:** Este programa incluye los aplicativos necesarios para instalar nuevos módulos y librerías para PHP:

**yum install  php-pear**

```
[root@bomarzo ~]# yum install  php-pear
..
 (1/1): php-pear-1.5.0-3.n 100% |=========================| 401 kB    00:00
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing: php-pear                     ############################ [1/1]

Installed: php-pear.noarch 1:1.5.0-3
Complete!
```

- **Instalar librerías de desarrollo de PHP:** Estas librerías contienen los archivos necesarios para compilar y agregar nuevas funcionalidades al PHP:

**yum install  php-devel.i386**

```
[root@bomarzo ~]# yum install  php-devel.i386


Transaction Summary
================================================================
Install      1 Package(s)
Update       0 Package(s)
Remove       0 Package(s)
```

```
Total download size: 541 k
Is this ok [y/N]: y
Downloading Packages:
(1/1): php-devel-5.2.4-1. 100% |=========================| 541 kB    00:04
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing: php-devel                  ########################### [1/1]

Installed: php-devel.i386 0:5.2.4-1.fc7
Complete!
```

- **Compilar e Instalar PHP-Swish:** Mediante el programa **pecl** de PEAR, le indicamos a este las nuevas librerías que se deben descargar, compilar, agregar e instalar en nuestro PHP.

**pecl install channel://pecl.php.net/swish-0.3.0**

```
 [root@bomarzo ~]# pecl install channel://pecl.php.net/swish-0.3.0
downloading swish-0.3.0.tgz ...
Starting to download swish-0.3.0.tgz (71,837 bytes)
.....done: 71,837 bytes
4 source files, building
running: phpize
Configuring for:
PHP Api Version:         20041225
Zend Module Api No:      20060613
Zend Extension Api No:   220060519
 1. Please provide the path to swish-config : autodetect

1-1, 'all', 'abort', or Enter to continue:
building in /var/tmp/pear-build-root/swish-0.3.0
….
….
----------------------------------------------------------------------
Libraries have been installed in:
   /var/tmp/pear-build-root/swish-0.3.0/modules

If you ever happen to want to link against installed libraries
in a given directory, LIBDIR, you must either use libtool, and
```

```
specify the full pathname of the library, or use the `-LLIBDIR'
flag during linking and do at least one of the following:
   - add LIBDIR to the `LD_LIBRARY_PATH' environment variable
     during execution
   - add LIBDIR to the `LD_RUN_PATH' environment variable
     during linking
   - use the `-Wl,--rpath -Wl,LIBDIR' linker flag
   - have your system administrator add LIBDIR to `/etc/ld.so.conf'

See any operating system documentation about shared libraries for
more information, such as the ld(1) and ld.so(8) manual pages.
----------------------------------------------------------------------

Build complete.
Don't forget to run 'make test'.

running: make INSTALL_ROOT="/var/tmp/pear-build-root/install-swish-0.3.0" install
Installing shared extensions:     /var/tmp/pear-build-root/install-swish-
0.3.0/usr/lib/php/modules/
running: find "/var/tmp/pear-build-root/install-swish-0.3.0" -ls
28377401    8 drwxr-xr-x   3 root     root         4096 mar 11 12:30 /var/tmp/pear-build-
root/install-swish-0.3.0
28377422    8 drwxr-xr-x   3 root     root         4096 mar 11 12:30 /var/tmp/pear-build-
root/install-swish-0.3.0/usr
28377423    8 drwxr-xr-x   3 root     root         4096 mar 11 12:30 /var/tmp/pear-build-
root/install-swish-0.3.0/usr/lib
28377424    8 drwxr-xr-x   3 root     root         4096 mar 11 12:30 /var/tmp/pear-build-
root/install-swish-0.3.0/usr/lib/php
28377425    8 drwxr-xr-x   2 root     root         4096 mar 11 12:30 /var/tmp/pear-build-
root/install-swish-0.3.0/usr/lib/php/modules
28377421   68 -rwxr-xr-x   1 root     root        58030 mar 11 12:30 /var/tmp/pear-build-
root/install-swish-0.3.0/usr/lib/php/modules/swish.so

Build process completed successfully
Installing '/usr/lib/php/modules/swish.so'
install ok: channel://pecl.php.net/swish-0.3.0
configuration option "php_ini" is not set to php.ini location
You should add "extension=swish.so" to php.ini
Complete!
```

Ahora solo se debe editar el archivo "**php.ini"** que se encuentra en la ruta: **/etc/php.ini** y agregar la siguiente línea en el apartado de extensiones dinámicas:

**extension=swish.so**

```
;;;;;;;;;;;;;;;;;;;;
; Dynamic Extensions ;
;;;;;;;;;;;;;;;;;;;;
;
extension=swish.so
```

Una vez agregada esta línea, se puede observar en la información proporcionada por el PHP, que ha agregado información sobre el programa Swish-e:

**PHP Version 5.2.4**

| | |
|---|---|
| **System** | Linux bomarzo.rec.usal.es 2.6.23.15-80.fc7 #1 SMP Sun Feb 10 17:29:10 EST 2008 i686 |
| **Build Date** | Sep 18 2007 08:52:27 |
| **Configure Command** | './configure' '--build=i386-redhat-linux-gnu' '--host=i386-redhat-linux-gnu' '--target=i386-redhat-linux-gnu' '--program-prefix=' '--prefix=/usr' '--exec-prefix=/usr' '--bindir=/usr/bin' '--sbindir=/usr/sbin' '--sysconfdir=/etc' '--datadir=/usr/share' '--includedir=/usr/include' '--libdir=/usr/lib' '--libexecdir=/usr/libexec' '--localstatedir=/var' '--sharedstatedir=/usr/com' '--mandir=/usr/share/man' '--infodir=/usr/share/info' '--cache-file=../config.cache' '--with-libdir=lib' '--with-config-file-path=/etc' '--with-config-file-scan-dir=/etc/php.d' '--disable-debug' '--with-pic' '--disable-rpath' '--without-pear' '--with-bz2' '--with-curl' '--with-exec-dir=/usr/bin' '--with-freetype-dir=/usr' '--with-png-dir=/usr' '--enable-gd-native-ttf' '--without-gdbm' '--with-gettext' '--with-gmp' '--with-iconv' '--with-jpeg-dir=/usr' '--with-openssl' '--with-png' '--with-pspell' '--with-expat-dir=/usr' '--with-pcre-regex=/usr' '--with-zlib' '--with-layout=GNU' '--enable-exif' '--enable-ftp' '--enable-magic-quotes' '--enable-sockets' '--enable-sysvsem' '--enable-sysvshm' '--enable-sysvmsg' '--enable-track-vars' '--enable-trans-sid' '--enable-yp' '--enable-wddx' '--with-kerberos' '--enable-ucd-snmp-hack' '--with-unixODBC=shared,/usr' '--enable-memory-limit' '--enable-shmop' '--enable-calendar' '--enable-dbx' '--enable-dio' '--without-mime-magic' '--without-sqlite' '--with-libxml-dir=/usr' '--with-xml' '--with-apxs2=/usr/sbin/apxs' '--without-mysql' '--without-gd' '--without-odbc' '--disable-dom' '--disable-dba' '--without-unixODBC' '--disable-pdo' '--disable-xmlreader' '--disable-xmlwriter' '--disable-json' |
| **Server API** | Apache 2.0 Handler |
| **Virtual Directory Support** | disabled |
| **Configuration File (php.ini) Path** | /etc |

**swish**

| swish support | Enabled |
|---|---|
| **source available from** | http://swish-e.org |

## Comenzando con Swish-e

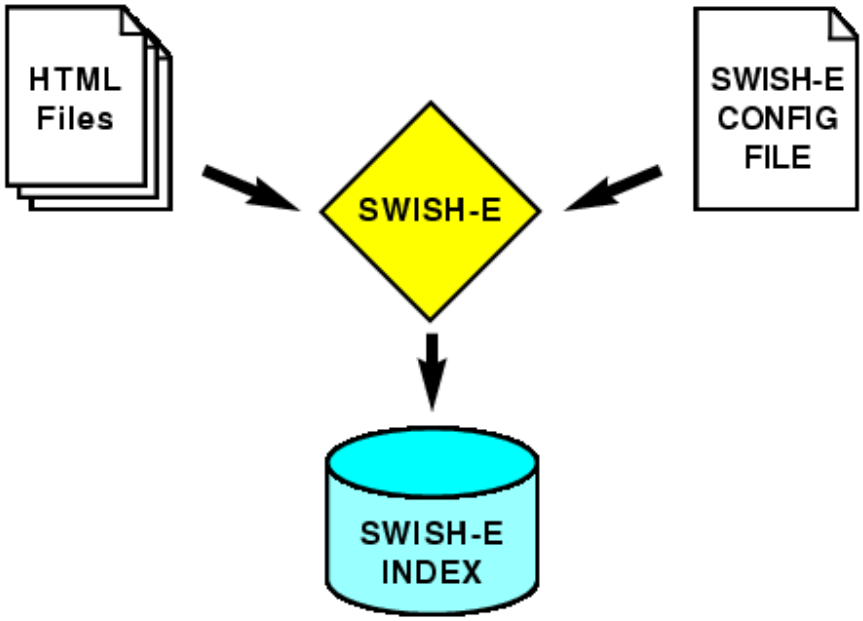
Para utilizar Swish-e, se debe tener inicialmente una colección de documentos, estos deberán ser indizados en Swish-e previamente a ser consultados. La indización se puede realizar de dos maneras distintas, si los archivos de la colección son muy sencillos y sin ninguna estructura (documentos de texto como se presenta en la colección que se encuentra ubicada en: **/var/www/html/□stán-e/Solo_Texto/Colección/**), lo más rápido será indizarlos mediante la línea de comandos, utilizando el siguiente comando:

```
swish-e –I /var/www/html/swish-e/Solo_Texto/Coleccion/
Indexing Data Source: “File-System”
Indexing “/var/www/html/□stán-e/Solo_Texto/Colección/”
Removing very common words...
no words removed.
Writing main index...
Sorting words ...
Sorting 5,462 words alphabetically
Writing header ...
Writing index entries ...
  Writing □stá text: Complete
  Writing □stá hash: Complete
  Writing □stá data: Complete
5,462 unique words indexed.
4 properties sorted.
100 files indexed.  158,058 total bytes.  24,954 total words.
Elapsed time: 00:00:00 CPU time: 00:00:00
Indexing done!
```

Un listado completo de los parámetros de ejecución de Swish-e mediante la línea de comandos se puede encontrar en http://swish-e.org/docs/swish-run.html

Observe que aparecen dos archivos, los cuales contienen todos los términos de indización encontrados, por defecto el programa Swish-e nombra los archivos de índices como **index.swish-e**, **index.swish-e.prop** y los ubica en el directorio

donde fue ejecutado el aplicativo, la siguiente figura[1] ilustra el proceso de indización de Swish-e con unos archivos HTML y la generación de los archivos de índices asociados:




En el caso que la colección sea un poco más compleja, y esta maneje múltiples tipos (formatos) de archivo con diferentes características, lo más recomendable es elaborar un archivo de configuración que se adecue a la colección y permita optimizar el entorno de trabajo y la posterior recuperación de los documentos asociados. Un archivo de configuración no es más que un archivo de texto donde se especifican un grupo de parámetros al programa Swish-e, estos parámetros generalmente se escriben mediante la expresión:

**Parámetro valor asociado**

Un listado completo de los parámetros de configuración de Swish-e se pueden encontrar en http://swish-e.org/docs/swish-config.html

Un ejemplo de un archivo básico de configuración se muestra a continuación:

[1] Rabinowitz, Josh. How to Index Anything. Linux Journal. Julio 2003. <http://www.linuxjournal.com/article/6652>

```
# Ejemplo Básico del archive de configuración

   # El indice se llamara prueba1.index
   IndexFile /var/www/html/swish -e/Varios_Formatos/prueba1.index

   # Definir la ruta de los archivos a indizar
   # IndexDir puede contener un directorio o un listado de archivos
   IndexDir /var/www/html/swish -e/Varios_Formatos/

   # Indizar únicamente archivos .html
   IndexOnly .html

   # Muestre información básica del proceso de indización
     IndexReport 1
```

En este caso se han utilizado los siguientes parámetros:

- **IndexFile:** Indica el nombre de los archivos que componen el índice
- **IndexDir:** Indica la ruta donde se encuentra la colección que se desea indizar (en este caso es el directorio "**/var/www/html/swish-e/Varios_Formatos/**").
- **IndexOnly:** Indica los tipos de formato de archivo que debe tener en cuenta para realizar la indización en este caso únicamente debe tomar archivos con extensión .html
- **IndexReport:** Indica el tipo de información que le debe mostrar al usuario que esta ejecutando el proceso de indización

Para ver todos los términos indizados, se puede utilizar la opción **–k** que permite ver el contenido del archivo de índices y devuelve todas las palabras que comienzan con la letra proporcionada (utilice un * si desea ver todo el diccionario de términos).

```
swish-e –k t –f index.swish-e
# SWISH format: 2.4.5
index.swish-e: t □stán□□ tal talando tales talla taller talones talón □stán□□ también tampoco tan tanto tarde tareas tasas tass □stán□□ tecnología □stán□□n□n□ telefoneara telefónica telefónicas □stán□□n□n televisivo televisión televión temblores teme temer temperaturas temporal tendido tenemos tener tenga tenido tenía tenían tercer tercera tercero terence teresa terminado terminal terminales terminarán terminó terraplén terremoto terremotos terreno terrestres territorio terrorista terroristas tesorería testificó testigo testigos testimonio tez tg the thomas □stán□□n tibia tiempo tienden tiene tienen
```

tierra tierras tijerino timo tipo tirando tiro tiros tiroteado titulada titular titulares tlc tobillo tocoron tocorón toda todas todavía todelar todo todos stán toma toman tomar tomen tomó toneladas tonelaje topaba tormentas torno toro torre torrecilla torrenciales torrentoso torres torácico total totalidad totalmente town trabaja trabajaban trabajador trabajadores trabajar trabajo trabajos trabajó traca tracas traficantes trafico tragedia tramo tramos tranquilidad tranquilo transbordador transcurso transitaba transportaba transporte transportes tras trasladada trasladado trasladados trasladar trasladarse trasladará trasladó trasobares trata trataba tratado tratamiento tratan tratar tratara trataron tratarse trate tratos trató traumatismo través trayecto trece tren tres trevín tribunal tribunales tripulación tripulantes troncos tropezó tropical tráfico trágica trágico trámites tránsito tt stán stán turismo tuvieron tuviesen tuvo stán tv tver stán técnica técnicos término título títulos túnel

Observe que se utilizó adicionalmente la opción **-f** para indicar el nombre de los archivos que contienen el índice de términos (**-f index.swish-e**), aunque solo es obligatorio si el nombre del índice no es el asignado por defecto.

Para hacer una búsqueda en la colección que previamente se ha indizado se utiliza la opción **–w,** posteriormente se coloca la expresión de búsqueda:

```
swish-e  –f index.swish-e  –w talando
# SWISH format: 2.4.5
# Search words: talando
# Removed stopwords:
# Number of hits: 1
# Search time: 0,000 seconds
# Run time: 0,011 seconds
1000 /var/www/html/swish-e/Solo_Texto/Coleccion/221.txt “221.txt” 3423
```

## Indizando distintos formatos de archivos

Swish-e es netamente un sistema indizador de **archivos de texto** (Los archivos de texto plano (en inglés “plain text”)), son aquellos que están compuestos únicamente por texto sin formato, sólo caracteres. Estos caracteres se pueden codificar de distintos modos dependiendo de la lengua usada. Algunos de los sistemas de codificación más usados son: ASCII, ISO-8859-1 o Latín-1, Unicode, etc.)[2], Aunque Swish-e indiza archivos de texto, el programa provee una gran alternativas de opciones para indizar texto que incluya marcas (documentos

[2] Colaboradores de Wikipedia. Archivo de texto [en línea]. Wikipedia, La enciclopedia libre, 2008 [fecha de consulta: 29 de febrero del 2008]. Disponible en <http://es.wikipedia.org/w/index.php?title=Archivo_de_texto&oldid=15484971>.

HTML, y XML) y otros tipo de documentos almacenados nativamente en formato binario (Un Archivo binario es un archivo informático que contiene información de cualquier tipo, codificada en forma binaria para el propósito de almacenamiento y procesamiento en ordenadores. Por ejemplo los archivos informáticos que almacenan texto formateado o fotografías)[3]. Las prácticas asociadas a los diferentes formatos de archivo se encuentran en:
**"/var/www/html/swish-e/Varios_Formatos"** que se pueden observar en la URL: **http://bomarzo.rec.usal.es/swish-e/Varios_Formatos/**

**Archivos HTML/XML**

La ventaja al procesar documentos estructurados (Texto con marcas de estructura), se encuentra en que puede extraer información adicional al contenido textual, la cuales se pueden utilizar en la etapa de recuperación para facilitar la tarea, aumentar las prestaciones y mejorar la relevancia de los documentos recuperados. Tal como lo comentan algunos estudios[4] la recuperación de información sobre documentos estructurados es mucho más eficaz que sobre documentos no estructurados[5]. Algunos aspectos que se deben tener en cuenta para respaldar estas afirmaciones están dadas por:

- La etiqueta <title>, que se encuentra en la cabecera de las páginas web, generalmente contiene información muy relevante respecto al contenido del documento.

```
<html>
<head>
<title>Anales de Documentación, Revista de Biblioteconomía y Documentación</title>
</head>
```

[3] Colaboradores de Wikipedia. Archivo binario [en línea]. Wikipedia, La enciclopedia libre, 2008 [fecha de consulta: 28 de febrero del 2008]. Disponible en <http://es.wikipedia.org/w/index.php?title=Archivo_binario&oldid=15467087>.
[4] CALLAN, J. Passage-level evidence in document retrieval. Conference on Research and Development in Information Retrieval Dublin, 1994
– Macleod, I. Storage an retrieval of structured documents. Information Processing and Managements, 26(2), 1990
– SALTON,G., ALLAN, J. y BUCKLEY, C. Approach to passage retrieval in full text information systems. Conference on Research and Development in Information Retrieval., Pittsburgh 1993
[5] García Martínez, Ana Maria. Definición y estilo de los objetos de información digitales y metadatos para la descripción. Universidad de Extremadura. Boletín de la Asociación Andaluza de Bibliotecarios, nº 63, Junio-2001, pp. 23-47

- El uso de las meta etiquetas en la cabecera de las páginas web, contribuye a mejorar la descripción del documento por medio del uso de sus metadatos, aunque el uso de estas metaetiquetas no es muy extendido en los sitios web, cuando se usan, generalmente están normalizados usando el estándar Dublin Core.

```
<html>
<head>
<title>Anales de Documentación, Revista de Biblioteconomía y Documentación</title>
<!--Metadatos en Dublin Core especialmente diseñados para una página Web en XHTML-->
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<meta name="DC.title" lang="es" content="Anales de Documentación: Revista de Biblioteconomía y Documentación" />
<meta name="DC.subject" lang="es" content="Documentación" />
<meta name="DC.subject" lang="es" content="Archivística" />
<meta name="DC.subject" lang="es" content="Biblioteconomía" />
<meta name="DC.subject" lang="es" content="Restauración Documental" />
<meta name="DC.subject" lang="es" content="Gestión de Unidades de Información" />
<meta name="DC.subject" lang="es" content="Alfabetización Informacional" />
<meta name="DC.subject" lang="es" content="Estudios de Usuarios" />
<meta name="DC.subject" lang="es" content="Mercado de la Información" />
<meta name="DC.subject" lang="es" content="Planificación aplicación y evaluación de proyectos y servicios de las unidades de información" />
<meta name="DC.subject" lang="es" content="Información científica" />
<meta name="DC.description" lang="es" content="Anales de Documentación es una revista anual editada por el Departamento de Información y Documentación y la Facultad de Ciencias de la Documentación de la Universidad de Murcia. La revista tiene dos ediciones complementarias: una edición impresa, que se publica en el mes de abril de cada año, y otra edición electrónica, en la dirección http://www.um.es/fccd/anales. En esta dirección están consultables las normas de publicación, suscripción, intercambios, así como la composición del Consejo Editor y del Comité de Redacción." />
<meta name="DC.publisher" content="Escuela Universitaria de Biblioteconomía y Documentación, Servicio de publicaciones : Universidad de Murcia." />
<meta name="DC.date" content="2007-12-12" />
<meta name="DC.date" content="Vol. 1 (1998)-" />
<meta name="DC.type" scheme="DCTERMS.DCMIType" content="Collection" />
<meta name="DC.format" content="text/html" />
<meta name="DC.format" content="9860 bytes" />
<meta name="DC.identifier" scheme="DCTERMS.URI" content="http://www.um.es/fccd/anales/" />
<meta name="DC.identifier" scheme="DCTERMS.URI" content="E-ISSN: 1697-7904 Anales de Documentación (Internet) [edición web - web edition]" />
```

```
<link rel="DC.source" href="http://www.um.es/fccd/anales/" />
<meta name="DC.language" scheme="DCTERMS.ISO639-2" content="spa" />
<meta name="DC.relation" content="ISSN: 1575-2437 Anales de Documentación [edición impresa - print edition]" />
<link rel="DC.relation" href="http://www.um.es/ http://eprints.rclis.org/" />
<link rel="DC.relation" href="http://www.erevistas.csic.es/portal/" />
<link rel="DC.relation" href="http://temaria.net/" />
<meta name="DC.rights" content="© Facultad de Comunicación y Documentación, http://www.um.es/f-comunicacion" />
</head>
```

- En HTML se encuentran las etiquetas de encabezados <h>, esta etiqueta viene acompañada de un número, desde el 1 hasta el 6, predefiniendo éstos el tamaño del encabezado. Así, <h1> sería el encabezado más grande mientras que <h6> sería el más pequeño. Generalmente en el contenido los encabezados (h1, h2, h3, h4, h5, h6) se encierran entre las etiquetas <H1>, <H2>, <H3>, corresponderían con información que tiene un mayor grado de relevancia respecto al resto de contenidos. Los encabezados son:

```
• <h1> Texto muy grande</h1>
• <h2>Texto grande</h2>
• <h3>Texto algo más grande de lo normal</h3>
• <h4>Texto normal</h4>
• <h5>Texto pequeño</h5>
• <h6>Texto muy pequeño</h6>
```

- Cuando en una página web se encuentra componentes de texto en **negrillas** ó resaltado en **otro color,** se puede presumir que este texto correspondería con términos especiales en el documento que tienen un mayor significado respecto al texto convencional.

El documento de configuración generado para indicar la colección de documentos HTML que se encuentra en **"/var/www/html/swish-e/trabajo1" (http://bomarzo.rec.usal.es/swish-e/trabajo1),** se presenta a continuación**:**

```
# ----- Configuración Trabajo 1 Indizando archivos de marcas ----------
# ----- Configuración Trabajo 1 Indizando archivos de marcas ----------
#
```

**# Versión: 1.0**
**# Licencia: MPL 1.1/GPL 2.0/LGPL 2.1**
#
# Desarrollo del archivo de configuración del punto 1:
# 1) Ventajas o no de trabajar directamente con ficheros html, o de convertir esos ficheros html primero a
# texto puro y luego trabajar con este texto. La respuesta debe ir acompañada de la solución práctica, con
# una pequeña colección de documentos en la que tengamos ambas situaciones y la solución de indización
# planteada. Las consultas realizadas deben dejar constancia de las ventajas o no del sistema.
#
**# Realizado por: Laureano Felipe Gomez Dueñas**
**# Universidad de Salamanca**
**# 2008**
#
# ***** END LICENSE BLOCK *****
#
##
## **Información del Script**
##
#############################################################################
IndexName "Trabajo1"
IndexDescription "Este índice corresponde a una colección de documentos HTML."
IndexPointer "http://bomarzo.rec.usal.es/swish-e/trabajo1/"
IndexAdmin "SIB-Manager (felipe.gomez3@gmail.com)"

#
##
## **Parámetros del Sistema**
##
#############################################################################

**# Tomando los índices en la carpeta donde están ubicadas las colecciones**
**# El índice se llamara indice.index *("index.swish-e")**
IndexFile /var/www/html/swish-e/trabajo1/indice.index

**# Selecciono los directorios donde se harán las indizaciones de documentos.**
IndexDir "/var/www/html/swish-e/trabajo1/Coleccion/"

**# Corresponde con los tipos de archivos únicos que debe indizar**
IndexOnly .htm .html

```
# Seguir enlaces simbólicos
FollowSymLinks yes

# Almaceno una pequeña parte del documento en el índice, las primeras 500 letras
StoreDescription HTML <body> 500

#convierto las entidades &; en caracteres indizables
ConvertHTMLEntities yes

# Aplico cuales caracteres *(entidades) deben ser transformados,
#esto generará todos los términos en Minúsculas
TranslateCharacters áéíóúÁÉÍÓÚüÜ aeiouaeiouuu

# Agrego el listado de palabras vacías
IgnoreWords "/var/www/html/swish-e/trabajo1/vacias.txt"

# Selecciono una técnica de análisis semántico de términos
FuzzyIndexingMode Stemming_es

# Now, specify which meta name to include in the index.
MetaNames author title description

# No indizar otras meta etiquetas encontradas
UndefinedMetaTags ignore
```

Cuando se realiza una búsqueda en la colección de documentos HTML, inicialmente equivaldría a realizar la búsqueda en archivos de texto plano, tal como lo demuestra la siguiente búsqueda (Expresión= **listado**), donde nos recupera **11** documentos que contienen la expresión “listado” en cualquier parte del documento.

```
[root@bomarzo trabajo1]# swish-e  -f  indice.index -w listado
# SWISH format: 2.4.5
# Search words: listado
# Removed stopwords:
# Number of hits: 11
# Search time: 0,000 seconds
# Run time: 0,011 seconds

1000 /var/www/html/swish-e/trabajo1/Coleccion/Ministerio de EducaciÃ³n PÃºblica de la RepÃºblica de Costa Rica.htm "Ministerio de EducaciÃ³n PÃºblica de la RepÃºblica de Costa Rica" 37923
```

```
1000 /var/www/html/swish-e/trabajo1/Coleccion/Recursos ElectrÃ³nicos CarloIII.htm "Listado a-z de recursos electrónicos" 21655

772 /var/www/html/swish-e/trabajo1/Coleccion/recweblogs.html "@bsysnet.com - Weblogs" 26605

526 /var/www/html/swish-e/trabajo1/Coleccion/Portal de revistas electrÃ³nicas de la UCM.htm "Portal de revistas electrónicas de la UCM" 54136

333 /var/www/html/swish-e/trabajo1/Coleccion/La educaciÃ³n en medios_ Webs para universitarios, recursos educativos.html "La educaciÃ³n en medios: Webs para universitarios, recursos educativos" 33448

333 /var/www/html/swish-e/trabajo1/Coleccion/Ministerio de EducaciÃ³n y Ciencia.htm "Ministerio de Educación y Ciencia" 15220

333 /var/www/html/swish-e/trabajo1/Coleccion/ASOCIACIONES-ESTUDIANTES-HUMANIDADES-EDUCACION-ESTUDIOS.htm "ASOCIACIONES-ESTUDIANTES-HUMANIDADES-EDUCACION-ESTUDIOS" 43190

333 /var/www/html/swish-e/trabajo1/Coleccion/Recursos educativos_2.htm "Recursos educativos" 19112
333 /var/www/html/swish-e/trabajo1/Coleccion/ENLACES EDUCATIVOS.htm "ENLACES EDUCATIVOS" 110592
333 /var/www/html/swish-e/trabajo1/Coleccion/__ SecretarÃ-a de EducaciÃ³n de Honduras _.htm "..:: Secretaría de Educación de Honduras ..::" 59943
333 /var/www/html/swish-e/trabajo1/Coleccion/Educasites.htm "Educasites.net - Guía de Recursos Educativos" 53577
.
```

Sin embargo, si aprovechamos las marcas presentes en los documentos HTML/XML (en concreto sobre las meta etiquetas), se puede buscar un documento especializado en cuyo título aparezca el término "**listado**" (solo nos recupera **1** documento):

```
[root@bomarzo trabajo1]# swish-e  -f  indice.index -w title=listado
# SWISH format: 2.4.5
# Search words: title=listado
# Removed stopwords:
# Number of hits: 1
# Search time: 0,000 seconds
# Run time: 0,011 seconds
```

```
1000 /var/www/html/swish-e/trabajo1/Coleccion/Recursos ElectrÃ³nicos CarloIII.htm "Listado a-z de recursos electrónicos" 21655
```

## Buscando en Varias Colecciones Simultáneamente

Swish-e permite buscar en varias colecciones simultáneamente, para ello solamente se tiene que incluir en los parámetros de búsqueda los nombres (ubicación) de los archivos de los índices asociados a cada colección.
Para esta actividad se han generado dos archivos de configuración que se encuentran en la ruta (/var/www/html/swish-e/trabajo2) http://bomarzo.rec.usal.es/swish-e/trabajo2/.

En esta ubicación se encuentran dos colecciones: CIENCIA y DEPORTES, estas colecciones contienen cada una 100 noticias en formato texto plano **(TXT)**.

El archivo de configuración de la colección CIENCIA es ciencia.conf

```
# ----- Configuración Trabajo 2 Indizando varias colecciones ----------
#
# Versión:  1.0
# Licencia: MPL 1.1/GPL 2.0/LGPL 2.1
#
#  Desarrollo del archivo de configuración del punto 2:
#  2) ¿Podemos trabajar con más de un índice a la vez al realizar una consulta?
#  Se debería acompañar de los ficheros que den respuesta a esta pregunta y
#  consultas realizadas al sistema.
#
#  Realizado por: Laureano Felipe Gomez Dueñas
#  Universidad de Salamanca
#  2008
#
# ***** END LICENSE BLOCK *****
#
##
##  Información del Script
##
################################################################################
IndexName       "Trabajo2"
IndexDescription "Este índice corresponde a una colección de documentos TXT especialziados en CIENCIA."
IndexPointer    "http://bomarzo.rec.usal.es/swish-e/trabajo2/"
```

```
IndexAdmin      "SIB-Manager (felipe.gomez3@gmail.com)"

#
##
## Parámetros del Sistema
##
################################################################################


# Tomando los índices en la carpeta donde están ubicadas las colecciones
# El índice se llamara indice.index *("index.swish-e")
IndexFile /var/www/html/swish-e/trabajo2/ciencia.index

# Selecciono los directorios donde se harán las indizaciones de documentos.
IndexDir "/var/www/html/swish-e/trabajo2/CIENCIA/"

# Corresponde con los tipos de archivos únicos que debe indizar
IndexOnly .txt

# Determino cuales caracteres *(entidades) deben ser transformados, esto generará todos los términos en Minúsculas
TranslateCharacters áéíóúÁÉÍÓÚüÜ aeiouaeiouuu

# Agrego el listado de palabras vacías
IgnoreWords "/var/www/html/swish-e/trabajo2/vacias.txt"

# Selecciono una técnica de análisis semántico de términos
FuzzyIndexingMode Stemming_es
```

El archivo de configuración de la colección DEPORTES es deportes.conf

```
# ----- Configuración Trabajo 2 Indizando varias colecciones ----------
#
# Versión:  1.0
# Licencia: MPL 1.1/GPL 2.0/LGPL 2.1
#
#  Desarrollo del archivo de configuración del punto 2:
#  2) ¿Podemos trabajar con más de un índice a la vez al realizar una consulta?
#  Se debería acompañar de los ficheros que den respuesta a esta pregunta y
#  consultas realizadas al sistema.
#
#  Realizado por: Laureano Felipe Gomez Dueñas
#  Universidad de Salamanca
```

```
#  2008
#
# ***** END LICENSE BLOCK *****
#
##
##  Información del Script
##
############################################################################
IndexName       "Trabajo2"
IndexDescription "Este índice corresponde a una colección de documentos TXT especialziados en DEPORTES."
IndexPointer    "http://bomarzo.rec.usal.es/swish-e/trabajo2/"
IndexAdmin      "SIB-Manager (felipe.gomez3@gmail.com)"

#
##
## Parámetros del Sistema
##
############################################################################


# Tomando los índices en la carpeta donde están ubicadas las colecciones
# El índice se llamara indice.index *("index.swish-e")
IndexFile /var/www/html/swish-e/trabajo2/deportes.index

# Selecciono los directorios donde se harán las indizaciones de documentos.
IndexDir "/var/www/html/swish-e/trabajo2/DEPORTES/"

# Corresponde con los tipos de archivos únicos que debe indizar
IndexOnly .txt

# Determino cuales caracteres *(entidades) deben ser transformados, esto generará todos los términos en Minúsculas
TranslateCharacters áéíóúÁÉÍÓÚüÜ aeiouaeiouuu

# Agrego el listado de palabras vacías
IgnoreWords "/var/www/html/swish-e/trabajo2/vacias.txt"

# Selecciono una técnica de análisis semántico de términos
FuzzyIndexingMode Stemming_es
```

Luego al ejecutar el script de indización llamado **Crear_Indices.sh.txt** genera la siguiente salida:

```
[root@bomarzo trabajo2]# chmod 777 Crear_Indices.sh.txt
```

```
[root@bomarzo trabajo2]# ./Crear_Indices.sh.txt
Indexing Data Source: "File-System"
Indexing "/var/www/html/swish-e/trabajo2/CIENCIA/"
Removing very common words...
no words removed.
Writing main index...
Sorting words ...
Sorting 4,580 words alphabetically
Writing header ...
Writing index entries ...
  Writing word text: Complete
  Writing word hash: Complete
  Writing word data: Complete
4,580 unique words indexed.
4 properties sorted.
100 files indexed.  266,929 total bytes.  41,407 total words.
Elapsed time: 00:00:00 CPU time: 00:00:00
Indexing done!
Indexing Data Source: "File-System"
Indexing "/var/www/html/swish-e/trabajo2/DEPORTES/"
Removing very common words...
no words removed.
Writing main index...
Sorting words ...
Sorting 4,245 words alphabetically
Writing header ...
Writing index entries ...
  Writing word text: Complete
  Writing word hash: Complete
  Writing word data: Complete
4,245 unique words indexed.
4 properties sorted.
100 files indexed.  181,694 total bytes.  28,352 total words.
Elapsed time: 00:00:00 CPU time: 00:00:00
Indexing done!
```

Ahora para hacer una búsqueda en varias colecciones solo tengo que indicarle Swish-e los nombres de los archivos de índices a consultar utilizando la opción **–f (swish-e -f ciencia.index deportes.index -w listado),** los resultados se presentan a continuación:

```
[root@bomarzo trabajo2]# swish-e -f ciencia.index deportes.index -w listado
# SWISH format: 2.4.5
```

```
# Search words: listado
# Removed stopwords:
# Removed stopwords:
# Number of hits: 6
# Search time: 0,000 seconds
# Run time: 0,021 seconds
1000 /var/www/html/swish-e/trabajo2/CIENCIA/2787.txt "2787.txt" 2110
1000 /var/www/html/swish-e/trabajo2/CIENCIA/3227.txt "3227.txt" 1470
1000 /var/www/html/s
wish-e/trabajo2/DEPORTES/518.txt "518.txt" 1308
1000 /var/www/html/swish-e/trabajo2/DEPORTES/191.txt "191.txt" 3376
1000 /var/www/html/swish-e/trabajo2/DEPORTES/373.txt "373.txt" 5341
1000 /var/www/html/swish-e/trabajo2/DEPORTES/278.txt "278.txt" 2612
```

## Utilizando una colección más compleja

Para este trabajo se decidió utilizar una colección de prueba consistente en **657** archivos organizados en **58** directorios, que fueron recolectados en el desarrollo de un estudio de usuarios contratado por la Superintendencia de Servicios Públicos Domiciliarios (http://http.superservicios.gov.co/) en el desarrollo de un proyecto para la creación de una biblioteca especializada en recursos digitales[6]

[6] Gomez Dueñas, Laureano Felipe. Manual de ingreso de contenidos: Biblioteca Especializada en Recursos Digitales/ CWIS. 2007. <http://cwis.superservicios.gov.co/SPT--FullRecord.php?ResourceId=50 >


Los archivos se encuentran ubicados en la ruta

**/var/www/html/swish-e/trabajo3/Colección**

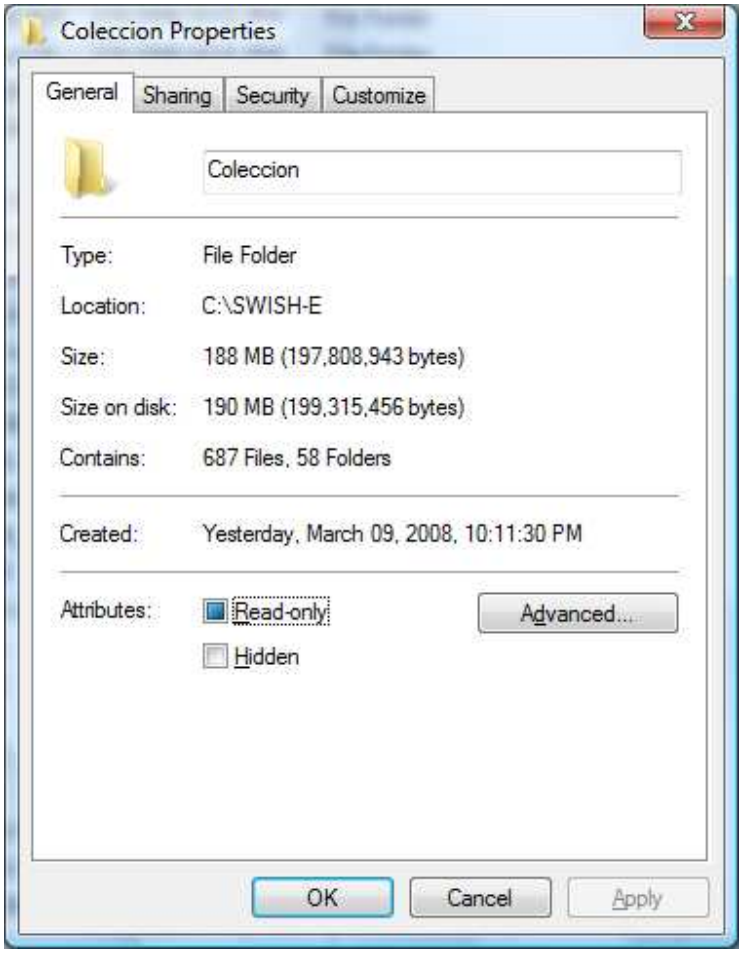
Estos documentos pueden ser visualizados en Internet mediante la dirección

http://bomarzo.rec.usal.es/swish-e/trabajo3/Colección/

| Nombre | Tamaño | Tipo |
|---|---|---|
| www | 6 elementos | carpeta |
| cgi-bin | 0 elementos | carpeta |
| error | 21 elementos | carpeta |
| html | 9 elementos | carpeta |
| catalis | 8 elementos | carpeta |
| cwis | 200 elementos | carpeta |
| docuwiki | 12 elementos | carpeta |
| joomla | 27 elementos | carpeta |
| ocs | 21 elementos | carpeta |
| swish-e | 1 elemento | carpeta |
| Coleccion | 46 elementos | carpeta |
| CONSULTAS | 8 elementos | carpeta |
| CONTROL INTERNO | 4 elementos | carpeta |
| documentos gas natural | 3 elementos | carpeta |
| Leyes y Decretos | 0 elementos | carpeta |
| mme | 7 elementos | carpeta |
| NORMATIVA DE GLP | 81 elementos | carpeta |
| paginas complemetarias | 9 elementos | carpeta |
| SENTENCIA DEBIDO PROCESO | 2 elementos | carpeta |
| SENTENCIA RESERVA LEGAL | 1 elemento | carpeta |
| SENTENCIAS | 13 elementos | carpeta |
| SENTENCIAS CORTE CONSTITU... | 2 elementos | carpeta |
| SISTEMA GESTION CALIDAD | 28 elementos | carpeta |
| TELECOMUNICACIONES | 6 elementos | carpeta |
| TELECOMUNICACIONES 1 | 6 elementos | carpeta |
| universidad de california | 6 elementos | carpeta |
| CONVENIO ARCHIVO DE BOG... | 105,5 Kib | Documento de Word |
| Encuesta ESTUDIO_DE_USUAR... | 39,5 Kib | Documento de Word |
| ENLACES JAIME GUERRA.doc | 21,5 Kib | Documento de Word |
| ESQUEMA PLANTAS GLP.pdf | 175,1 Kib | Documento PDF |
| INF_SECT_03_06_V1_gestion.xls | 161,0 Kib | Hoja de cálculo de Excel |

Entre las características de esta colección de documentos encontramos que:

- Contienen **687** archivos organizados en **58** directorios




- Los archivos ocupan **197.808.943** bytes (188 MB)
- Contiene diferentes tipos de formatos de archivos con las siguientes características:

| Formato | Cantidad | Peso |
|---|---|---|
| Documentos de texto enriquecido **RFT** | 8 | 2.90 MB (3,045,784 bytes) |
| Presentaciones en PowerPoint **PPT** | 5 | 3.50 MB (3,675,648 bytes) |
| Imágenes Digitalizadas **TIFF** | 7 | 9.33 MB (9,792,848 bytes) |

| | | |
|---|---|---|
| Archivos de páginas Web complementarias MHTML (**MHTML** significa MIME HTML (Multipurpose Internet Mail Extension HTML o en español protocolo de transferencia de hipertexto Multiuso de la Extensión del Correo del Internet). Es un estándar para incluir recursos que en páginas HTTP usualmente están enlazados externamente, tal como los archivos de imágenes y sonido, en el mismo archivo como el Código de Protocolo de Transferencia de Hipertexto)[7] | 6 | 2.38 MB (2,506,540 bytes) |
| Archivos de código en java script **JS** | 4 | 10.6 KB (10,909 bytes) |
| Imágenes JPEG **JPG** | 175 | 1.68 MB (1,772,360 bytes) |
| Páginas Web **HTML** | 16 | 1.42 MB (1,490,262 bytes) |
| Hojas de Cálculo Excel **XLS** | 20 | 4.36 MB (4,581,376 bytes) |
| Imágenes **GIF** | 53 | 80.6 KB (82,554 bytes) |
| Documentos de Word **DOC** | 143 | 33.2 MB (34,815,335 bytes) |
| Hojas de estilo **CSS** | 3 | 28.9 KB (29,617 bytes) |
| Imágenes Mapas de BITS **BMP** | 2 | 1.54 MB (1,616,556 bytes) |
| Documentos Acrobat Reader PDF | 245 | 128 MB (134,389,154 bytes) |

## Parametrizando la colección en Swish-e

De acuerdo a la información que se ha analizado de la colección seleccionada, adicionalmente al haber estudiado los comandos más utilizados en Swish-e, se presenta un modelo del archivo de configuración ideal, el cual debería contemplar como mínimo los siguientes aspectos:

- Los archivos JS, CSS, pese a contener texto y por tal motivo ser indizables, no son documentos que posean información relevante al usuario (a menos que se construya una colección de códigos fuente), sin embargo son muy

[7] Colaboradores de Wikipedia. MHTML [en línea]. Wikipedia, La enciclopedia libre, 2008 [fecha de consulta: 5 de enero del 2008]. Disponible en <http://es.wikipedia.org/w/index.php?title=MHTML&oldid=14099905>.

relevantes en el momento de la presentación de los resultados, para darle mayor fidelidad al documento original.

- Las imágenes GIF, BMP, TIFF y JPEG, que acompañan las páginas web (HTM, HTML, MHTML,ASP,PHP, JSP, etc..), no son útiles en el momento de la indización (aparte que no se podrían indizar propiamente como texto), sin embargo, son muy relevantes en el momento de la presentación de los resultados, para darle mayor fidelidad al documento original.

- El aplicativo **unrtf** transforma los documentos RTF al formato HTML, así que se tomará esta consideración al momento de crear el archivo de configuración.

- La colección se encuentra en Inglés, castellano y alemán, ya que corresponde primordialmente a información recolectada de internet por los empleados de la superintendencia de Servicios Públicos.

- Se tomarán el archivo de Palabras Vacías (**STOPWORDS**) provisto por la herramienta integrando los diversos idiomas existentes en la colección.

- Aunque el uso de técnicas de **stemming** permiten representar de un mismo modo las distintas variantes de un término, a la vez que reducen el tamaño del vocabulario y mejoran, en consecuencia, la capacidad de almacenamiento de los sistemas y el tiempo de procesamiento de los documentos, en el caso de esta colección **no** se aplicaran técnicas de **stemming**, ya que la colección se encuentra en múltiples idiomas y el sistema Swish-e actualmente no puede identificar cual modelo de idioma aplicar en cada documento. Adicionalmente debido a la complicada estructura semántica del castellano, considero que esta técnica no mejora substancialmente la recuperación y si aumenta el ruido en los documentos recuperados, tal como lo presentan Mari Vallez y Rafael Pedraza-Jimenez, “estos algoritmos presentan el inconveniente de no agrupar en ocasiones palabras que deberían estarlo, y viceversa, mostrar como iguales palabras que realmente son distintas”[8].

[8] Mari Vallez y Rafael Pedraza-Jimenez. El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines [on line]. "Hipertext.net", núm. 5, 2007. <http://www.hipertext.net> [Consulta: 21/02/2008]. ISSN 1695-5498

El archivo de configuración generado para la colección de prueba es el siguiente:

```
# ----- Configuración Trabajo 3 Utilizando una Colección Grande ----------
#
# Versión:  1.0
# Licencia: MPL 1.1/GPL 2.0/LGPL 2.1
#
#  Desarrollo del archivo de configuración del punto 3:
#   3) Crear una colección documental (entorno a unos 100 ficheros) con diferentes situaciones (documentos
#   pdf, doc, html, txt), a lo mejor en diferentes idiomas, etc. y plantear el fichero de configuración ideal para
#   dar adecuada solución a la indización. Crear diferentes preguntas, algunas muy complejas que demuestren
#  que efectivamente se ha dado solución a los problemas que pueda tener la colección.
#
#  Realizado por: Laureano Felipe Gomez Dueñas
#  Universidad de Salamanca
#  2008
#
# ***** END LICENSE BLOCK *****
#
##
##  Información del Script
##
############################################################################
IndexName       "Trabajo3"
IndexDescription "Este índice corresponde a una colección mixta de documentos (>100 Docs)"
IndexPointer    "http://bomarzo.rec.usal.es/swish-e/trabajo3/"
IndexAdmin      "SIB-Manager (felipe.gomez3@gmail.com)"

#
##
## Parámetros del Sistema
##
############################################################################


# Tomando los indices en la carpeta donde estan ubicadas ls colecciones
# El indice se llamara indice.index *("index.swish-e")
IndexFile /var/www/html/swish-e/trabajo3/indice.index
```

```
# Selecciono los directorios donde se haran las indizaciones de documentos.
IndexDir "/var/www/html/swish-e/trabajo3/Coleccion/"

# Aplico cuales catacteres deben ser transformados, esto generará todos los términos en Minusculas
TranslateCharacters áéíóúÁÉÍÓÚüÜ aeiouaeiouuu

# Agrego el listado de palabras vacias Supercompleto
IgnoreWords "/root/Swish-e/colecciones/vacias.txt"

# Selecciono cuales son los caracteres indizables
WordCharacters ABCDEFGHIJKLMNÑOPQRSTUVWXYZabcdefghijklmnñopqrstuvwxyz

# No Selecciono ninguna técnica de analisis semantico de términos
# FuzzyIndexingMode Stemming_es

# Seleccionos los diferentes formatos de archivo que se deben incluir en el indice
IndexContents TXT* .pdf .doc .xls .ppt .log
IndexContents HTML .htm .html .shtml .rtf
IndexContents XML* .xml

#No tener en cuenta los siguientes formatos de archivo
NoContents .jpg .gif .jpeg .png .tif .bmp .eps .tiff .js .css

# Indizar solo los siguientes formatos de archivo
IndexOnly .htm .html .txt .xls .doc .ppt .pdf .shtml .rtf

#Contenidos por Defecto
DefaultContents HTML*

# We allow a period and a dash within words, but strip them
# from the beginning or end of a word.  This is done after
# WordCharacters above is used to split words.
IgnoreFirstChar .-
IgnoreLastChar  .-

# Parametros de aplicación de porgramas que transforman documentos binarios en documentos de texto
FileFilter .pdf pdftotext   "'%p" -'
FileFilter .doc catdoc '-s8859-1 -d8859-1 "%p"'
FileFilter .xls xls2csv '-s8859-1 -d8859-1 "%p"'
FileFilter .ppt catppt '-s8859-1 -d8859-1 "%p"'
FileFilter .rtf unrtf '--html "%p"'
```

```
#############################################################################
#COnfiguracion para archivos HTML
# Seguir enlaces simbólicos
FollowSymLinks yes

# Almaceno una pequeña parte del documento en el índice, las primeras 500 letras
StoreDescription HTML <body> 500
StoreDescription TXT* 500

#convierto las entidades &; en caracteres indizables
ConvertHTMLEntities yes

# Ahora especifico las meta etiquetas a incluir en el índice
MetaNames author description title

# Por defecto no indizo otras meta etiquetas
UndefinedMetaTags ignore
```

Luego al ejecutar el script de indización llamado **./trabajo3.sh.txt, el sistema** genera la siguiente salida:

```
Sorting words ...
Sorting 50,739 words alphabetically
Writing header ...
Writing index entries ...
  Writing word text: Complete
  Writing word hash: Complete
  Writing word data: Complete
50,739 unique words indexed.
5 properties sorted.
437 files indexed.  182,323,007 total bytes.  3,670,178 total words.
Elapsed time: 00:02:26 CPU time: 00:00:09
Indexing done!
```

Para probar que la indización ha salido correctamente, se han optado por realizar las siguientes pruebas:

- **Impresión de los términos indizados:** Para este caso se aplica la opción (**-k)** en el comando de búsqueda así:

**#Desplegar el diccionario de términos**
swish-e -f indice.index **-k***

Allí se puede observar una gran cantidad de términos sin normalizar, los resultados generados se presentan a continuación:

```
# SWISH format: 2.4.5
indice.index: a aa aaa aaaa aaac aac aacero aachca aactualizar aaplicada aarhus aas aañ ab
aba abad abajo abandona abandonada abandonadas abandonado abandonados abandonan abandonando
abandonar abandoned abandonen abandono abanico abaratar abarbanel abarc abarca abarcadas
…….
…….
```

- **Búsqueda por Frase:** en este caso se utilizará el operador de frase (**“ ”**) para buscar un texto especifico en la colección:

  **# Búsqueda por frase - utilizando comillas**
  swish-e -f indice.index -w **'El Presidente de la República'**

  En esta búsqueda se puede observar que se recuperan todo tipo de documentos (HTML,RTF, DOC, PDF, etc..), los resultados generados se presentan a continuación:

```
# SWISH format: 2.4.5
# Search words: El Presidente de la República
# Removed stopwords:
# Number of hits: 76
# Search time: 0,020 seconds
# Run time: 0,030 seconds
1000 /var/www/html/swish-e/trabajo3/Coleccion/SENTENCIA LEY DE GARANTIAS.rtf "I" 1593325
889 /var/www/html/swish-e/trabajo3/Coleccion/SENTENCIAS CORTE CONSTITUCIONAL CARGOS DE ACCESO.doc "SENTENCIAS CORTE CONSTITUCIONAL CARGOS DE ACCESO.doc" 220672
861 /var/www/html/swish-e/trabajo3/Coleccion/CONTROL INTERNO/NORMAS CONTROL INTERNO/Ley 42 de 1993.rtf "D.O. CXXVIII, No. 40732, enero 27, 1993, P.1" 65448
852 /var/www/html/swish-e/trabajo3/Coleccion/NORMATIVA DE GLP/LEY 142-94 CONCORDADA.pdf "LEY 142-94 CONCORDADA.doc" 797184
```

```
…..
…..
```

- **Búsqueda utilizando truncamiento, operadores Booleanos y de proximidad:** En este caos se intentará aprovechar varios tipos de operadores de búsqueda para delimitar el conjunto de documentos a recuperar:

  **# Búsqueda avanzada utilizando operadores**
  swish-e -f indice.index -w (servicio* **near1** publi*) **and** ((Acueducto or Alcantarillado) **not** gas*)

  En esta búsqueda se puede observar que recupera dos archivos, un documento DOC y un documento HTML, los resultados generados se presentan a continuación:

```
# SWISH format: 2.4.5
# Search words: (servicio* near1 publi*) and ((Acueducto or Alcantarillado) not gas*)
# Removed stopwords:
# Number of hits: 2
# Search time: 0,046 seconds
# Run time: 0,057 seconds
1000 /var/www/html/swish-e/trabajo3/Coleccion/SENTENCIAS/PROCEDENCIA DE ACCION POPULAR.doc "PROCEDENCIA DE ACCION POPULAR.doc" 385536

994 /var/www/html/swish-e/trabajo3/Coleccion/paginas complemetarias/_SuperIntendencia de Servicios PÃºblicos Domiciliarios - SSPD_archivos/index.htm ":. Superintendencia de Servicios Públicos Domiciliarios - SSPD .:" 47519
.
```

- **Búsqueda en contexto (Documentos HTML):** Es este ejemplo se aprovechará el etiquetado de los documentos HTML indizados para buscar información en su contenido, brindando mayor relevancia a algunas etiquetas especificas:

  **# Buscar únicamente los términos que aparezcan en el título, cabeceras web y etiquetas de énfasis visual (negrillas, itálicas, subrayados, etc..)**

swish-e f indice.index -**t the** -w "agua"

En este caso se observa que únicamente recupera archivos HTML, los resultados generados se presentan a continuación:

```
# SWISH format: 2.4.5
# Search words: agua
# Removed stopwords:
# Number of hits: 1
# Search time: 0,000 seconds
# Run time: 0,012 seconds
1000 /var/www/html/swish-e/trabajo3/Coleccion/SENTENCIA DEBIDO PROCESO/t-270-04.htm "t-270-04.htm" 934567
.
```

## Indizando contenido externo (Spidering)

Un **Spidering**, se puede definir como un proceso por el que un aplicativo denominado **"araña",** recopila información de múltiples sitios en Internet para su almacenamiento, hasta que un indizador procesos utiliza esta información recolectada para su posterior búsqueda y recuperación. Para este trabajo los archivos de configuración y la colección de documentos se encuentra ubicados en la ruta:

**/var/www/html/swish-e/trabajo4/**

Estos documentos pueden ser visualizados en Internet mediante la dirección

http://bomarzo.rec.usal.es/swish-e/trabajo4/

Aunque Swish-e contiene un método para leer directamente documentos de Internet, leerlos e indizarlos (spidering), tal como lo haría un sistema crawler, ya sea desde la línea de comandos ó por medio de un archivo de configuración, se ha optado por utilizar un programa externo llamado **HTTrack**.

**HTTrack** (http://www.httrack.com) es una aplicación informática de Software libre con licencia GPL, multilenguaje y multiplataforma que actúa como un navegador Offline, cuyo fin es la captura Web, este programa permite descargar páginas web en el disco duro para luego poder navegar por ellas sin tener que estar conectado a Internet. Con **HTTrack** se puede descargar un sitio entero de Internet, esta descarga puede incluir las páginas HTML, imágenes, directorios y otros archivos

que se encuentren referenciados en las páginas descargadas. **HTTrack** mantiene la estructura de enlaces del sitio original y permite navegar por el sitio descargado como si del sitio original se tratase. Normalmente este tipo de programas distinguen entre enlaces internos (enlaces a archivos contenidos en el mismo sitio web de la página inicial) y externos (enlaces a archivos contenidos en otros sitios web).

Para el trabajo se decidio tomar el sitio de exlibris como pundo de partida de indexador y se ha configurado para que realice busqueda en prufundidad de seis niveles y tome todos los contenidos externos vinculados desde este sitio hasta completar los niveles señalados, esto se realizò mediante la siguiente instrucciòn:

***httrack http://exlibris.usal.es*** -r6

Los ducumentos recuperados por el indizador se encuentran en la direcciòn:

**http://bomarzo.rec.usal.es/swish-e/trabajo4/httrack/**


Despues de realizar en analisis y descarga de sitio senalado y los sitios asociados (proceso que duró tres horas), el programa generó un directorio por cada sitio recolectado. El total de archivos descargados ocupa **2.7 GB** tal como lo presenta el siguiente comando:

**[root@bomarzo trabajo4]# du -k -c httrack/**
**2.774.276 httrack/**

El archivo **LOG** del Spider generó la siguiente información:

```
[root@bomarzo httrack]# cat hts-log.txt | more
```

```
HTTrack3.42-noV6+libhtsjava.so.2 launched on Mon, 10 Mar 2008 19:34:50 at exlibr
is.usal.es +*
(httrack exlibris.usal.es +* -r6 )

Information, Warnings and Errors reported for this mirror:
note:   the hts-log.txt file, and hts-cache folder, may contain sensitive inform
ation,
        such as username/password authentication for websites mirrored in this p
roject
        do not share these files/folders if you want these information to remain
 private

19:34:50        Info:   Note: due to exlibris.usal.es remote robots.txt rules, l
inks begining with these path will be forbidden: /administrator/, /cache/, /comp
onents/, /editor/, /help/, /images/, /includes/, /language/, /mambots/, /media/,
 /modules/, /templates/, /installation/ (see in the options to disable this)
19:36:01        Info:   Note: due to www.w3.org remote robots.txt rules, links b
egining with these path will be forbidden: /2004/ontaria/basic, /Team, /Project,
 /Web, /Systems, /History, /Out-Of-Date, /2002/02/mid, /mid/, /2004/08/W3CTalks,
 /2007/11/Talks/search, /People/all/, /RDF/Validator/ARPServlet, /2003/03/Transl
ations/byLanguage, /2003/03/Translations/byTechnology, /2005/11/Translations/Que
ry, /2003/glossary/subglossary/, /2000/06/webdata/xslt, /2000/09/webdata/xslt, /
2005/08/online_xslt/xslt, /Bugs/, /Search/Mail/Public/, /2006/02/chartergen (see
 in the options to disable this)
19:36:23        Warning:        File has moved from www.usal.es/ to http://www.u
sal.es/web-usal/
19:41:49        Warning:        Redirected link is identical because of 'URL Hac
k' option: www3.usal.es/robots.txt and www.usal.es/robots.txt
19:41:49        Warning:        File has moved from www3.usal.es/robots.txt to h
ttp://www.usal.es/robots.txt
19:41:49        Warning:        Redirected link is identical because of 'URL Hac
k' option: www3.usal.es/~socrates/ and www.usal.es/~socrates/
19:41:49        Warning:        File has moved from www3.usal.es/~socrates/ to h
ttp://www.usal.es/~socrates/
```

El archivo de configuración utilizado para realizar la indización de los contenidos antes mencionados (**trabajo4.config**) contiene las siguientes instrucciones:

```
# ----- Configuración Trabajo 4  Trabajando con un Spider -------------
#
# Versión:  1.0
# Licencia: MPL 1.1/GPL 2.0/LGPL 2.1
#
#  Desarrollo del archivo de configuración del punto 3:
#  4) Esta cuestión es optativa. Preparar swish-e para poder realizar una recogida de información con el
#  spider. Podemos tener la posibilidad de usar el spider de swish-e y en este caso debemos
```

```
ajustar la
#  configuración para poder hacer una buena recogida; o bien podemos utilizar un crawler externo, descargar
#  las páginas deseadas y sobre ellas realizar el proceso de indicación y consulta.
#   Hay que entregar todos los datos utilizados, ficheros, índices, ficheros de configuración, consultas
#  realizadas, etc.
#
#  Realizado por: Laureano Felipe Gomez Dueñas
#  Universidad de Salamanca
#  2008
#
# ***** END LICENSE BLOCK *****
#
##
##  Información del Script
##
#############################################################################
IndexName        "Trabajo4"
IndexDescription "Este índice corresponde a una colección creada por el programa HTTRACK"
IndexPointer     "http://bomarzo.rec.usal.es/swish-e/trabajo4/"
IndexAdmin       "SIB-Manager (felipe.gomez3@gmail.com)"

#
##
## Parámetros del Sistema
##
#############################################################################


# Tomando los índices en la carpeta donde están ubicadas las colecciones
# El índice se llamara indice.index *("index.swish-e")
IndexFile /var/www/html/swish-e/trabajo4/httrack.index

# Selecciono los directorios donde se haran las indizaciones de documentos.
IndexDir "/var/www/html/swish-e/trabajo4/httrack/"

# Aplico cuales caracteres deben ser transformados, esto generará todos los términos en Minusculas
TranslateCharacters áéíóúÁÉÍÓÚüÜ aeiouaeiouuu

# Agrego el listado de palabras vacías Supercompleto
IgnoreWords "/var/www/html/swish-e/trabajo4/vacias.txt"

# Selecciono cuales son los caracteres indizables
WordCharacters ABCDEFGHIJKLMNÑOPQRSTUVWXYZabcdefghijklmnñopqrstuvwxyz

# No Selecciono ninguna técnica de análisis semántico de términos
# FuzzyIndexingMode Stemming_es
```

```
# Seleccionos los diferentes formatos de archivo que se deben incluir en el indice
IndexContents TXT* .pdf .doc .xls .ppt .log
IndexContents HTML .htm .html .shtml .rtf
IndexContents XML* .xml

#No tener en cuenta los siguientes formatos de archivo
NoContents .jpg .gif .jpeg .png .tif .bmp .eps .tiff .js .css

# Indizar solo los siguientes formatos de archivo
IndexOnly .htm .html .txt .xls .doc .ppt .pdf .shtml .rtf

#Contenidos por Defecto
DefaultContents HTML*



# We allow a period and a dash within words, but strip them
# from the beginning or end of a word.  This is done after
# WordCharacters above is used to split words.
IgnoreFirstChar .-
IgnoreLastChar  .-

# Parámetros de aplicación de programas que transforman documentos binarios en documentos de
texto
FileFilter .pdf pdftotext   '"%p" -'
FileFilter .doc catdoc '-s8859-1 -d8859-1 "%p"'
FileFilter .xls xls2csv '-s8859-1 -d8859-1 "%p"'
FileFilter .ppt catppt '-s8859-1 -d8859-1 "%p"'
FileFilter .rtf unrtf '--html "%p"'


#############################################################################
#Configuración para archivos HTML
# Seguir enlaces simbólicos
FollowSymLinks yes

# Almaceno una pequeña parte del documento en el índice, las primeras 500 letras
StoreDescription HTML <body> 500
StoreDescription TXT* 500

#convierto las entidades &; en caracteres indizables
ConvertHTMLEntities yes

# Ahora especifico las meta etiquetas a incluir en el índice
MetaNames author description title

# Por defecto no indizo otras meta etiquetas
UndefinedMetaTags ignore
```

Una vez creados los índices se realizaron varias búsquedas para comprobar la efectividad del proceso realizado y la velocidad de búsqueda del programa Swish-e, las consultas realizadas se muestran a continuación (http://bomarzo.rec.usal.es/swish-e/trabajo4/busquedas.sh.txt ):

```
#!/bin/sh
#
# ***** Script Básico de Busquedas Realizadas  *****
# ----- Configuración Trabajo 4 Utilizando un Spider HtTrack------
#
# Versión: 0.1 beta
# Licencia: MPL 1.1/GPL 2.0/LGPL 2.1
# Contributor(s): Laureano Felipe Gomez Dueñas
#
# Descripción
# Se pretende ejecutar Múltiples Búsquedas Avanzadas
#
# ***** VARIABLES *****


#
##
## Línea de Comandos
##
################################################################################


# 1 - Desplegar el diccionario de términos por la letra t
swish-e -f  httrack.index  -kt > Resultados_Busqueda1.txt

# 2 - Búsqueda por frase - utilizando comillas
swish-e -f  httrack.index   -w 'Biblioteca' > Resultados_Busqueda2.txt

# 3 - Búsqueda avanzada utilizando operadores
swish-e -f   httrack.index    -w "(biblioteca* near1 digital*) or repositorio* or ('Biblioteca virtual') " >
Resultados_Busqueda3.txt

# 4 - Buscar únicamente los términos que aparezcan en el título, cabeceras web y etiquetas de
énfasis visual (negrillas, itálicas, subrayados, etc..)
swish-e -f  httrack.index   -t the  -w "educacion" > Resultados_Busqueda4.txt
```

## Crear un catálogo Web con Swish-e

Existen múltiples alternativas para crear un catálogo Web, que permita recuperar los contenidos indizados por Swish-e, en este caso se han optado por desarrollar dos alternativas de trabajo:

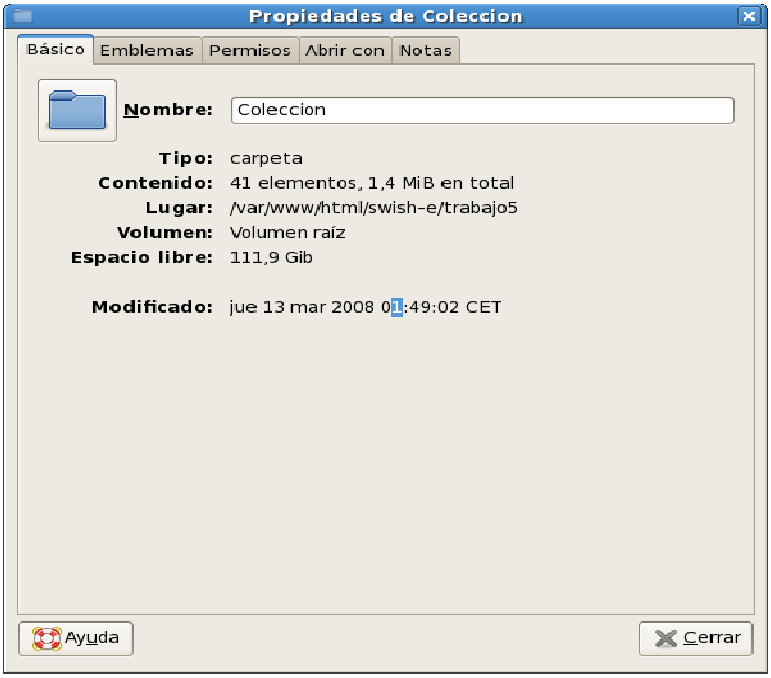
- Utilizar los Scripts de PERL suministrados por el mismo programa Swish-e, los cuales se encuentran en la ruta **/usr/local/lib/swish-e/,** los archivos creados para este ejemplo se pueden ver en: **http://bomarzo.rec.usal.es/swish-e/trabajo5/cgi-swish-e/**
- Utilizar las librerías instaladas de PHP, y crear un catálogo a partir de estas, los archivos creados para este ejemplo se pueden ver en: **http://bomarzo.rec.usal.es/swish-e/trabajo5/catalogo_PHP/**

Para este trabajo, se decidió utilizar una colección compuesta por 41 documentos que corresponden con páginas web, estas corresponden con las páginas iníciales (portadas) de recursos en educación, principalmente los sitios de los ministerios de educación de los países iberoamericanos. Los archivos se encuentran ubicados en la ruta

**/var/www/html/swish-e/trabajo5/Colección**

Estos documentos pueden ser visualizados en Internet mediante la dirección

http://bomarzo.rec.usal.es/swish-e/trabajo5/Colección/



El archivo de configuración utilizado para realizar la indización de los contenidos antes mencionados (**trabajo5.config**) contiene las siguientes instrucciones:

```
# ----- Configuración Trabajo 5 Indizando archivos HTML para un catalogo ----------
#
# Versión:  1.0
# Licencia: MPL 1.1/GPL 2.0/LGPL 2.1
#
#  Desarrollo del archivo de configuración del punto 5:
#  5) Esta cuestión es optativa y solamente se puede completar si también se hace la cuestión 4. Tenemos
#  que implementar un interfaz web para poder consultar la información indizada con swish-e. Se puede
#   utilizar el CGI que viene con la propia distribución swish-e o bien localizar otros sistemas alternativos. La
#   mejor forma de poderlo probar es montarlo sobre XAMPP. Aquí se debe entregar todo el material que
#  permita valorar la adecuada consecución del objetivo.
#
#  Realizado por: Laureano Felipe Gomez Dueñas
#  Universidad de Salamanca
#  2008
#
# ***** END LICENSE BLOCK *****
#
##
##  Información del Script
##
#############################################################################
IndexName       "Trabajo5"
IndexDescription "Este índice corresponde a una colección de documentos HTML que serán consultadas vía WEB."
IndexPointer    "http://bomarzo.rec.usal.es/swish-e/trabajo5/"
IndexAdmin      "SIB-Manager (felipe.gomez3@gmail.com)"

#
##
## Parámetros del Sistema
##
#############################################################################

# Tomando los índices en la carpeta donde están ubicadas las colecciones
# El índice se llamara indice.index *("index.swish-e")
IndexFile /var/www/html/swish-e/trabajo5/indice.index

# Selecciono los directorios donde se harán las indizaciones de documentos.
IndexDir "/var/www/html/swish-e/trabajo5/Coleccion/"
```

```
# OJO en la variable swishdocpath quito la ubicacion fisica del doc y solo dejo el nombre del archivo
ReplaceRules remove "/var/www/html/"

# Corresponde con los tipos de archivos únicos  ( Por defecto) que debe indizar
DefaultContents HTML*
IndexOnly .htm .html

# Seguir enlaces simbólicos
FollowSymLinks yes

# Almaceno una pequeña parte del documento en el índice, los primeros 500 caracteres
StoreDescription HTML* <body> 200

#convierto las entidades &; en caracteres indizables
ConvertHTMLEntities yes

# Aplico cuales caracteres *(entidades) deben ser transformados, esto generará todos los términos en Minúsculas
TranslateCharacters áéíóúÁÉÍÓÚüÜ aeiouaeiouuu

# Agrego el listado de palabras vacías
IgnoreWords "/var/www/html/swish-e/trabajo5/vacias.txt"

# Selecciono una técnica de análisis semántico de términos
FuzzyIndexingMode Stemming_es

# Ahora especifico las meta etiquetas a incluir en el índice
MetaNames author description title swishdocpath swishtitle

# Por defecto no indizo otras meta etiquetas
UndefinedMetaTags ignore
```

Nótese que se han agregado un nuevo parámetro llamado **ReplaceRules remove,** el cual nos permite modificar los valores de la variable **swishdocpath (**variable de indización asociada a los documentos**),** removiendo algunos texto fijos que contienen datos de la ubicación FISICA de los documentos que no son útiles en el momento de visualizarlos LOGICAMENTE vía Web.

Una vez creado los índices, se procederá a configurar el servidor web, nótese que Swish-e no incluye un servidor Web, este debe estar previamente instalado en el equipo donde se desea configurar el catálogo web, se recomienda utilizar el servidor **WEB APACHE**, en este caso, en FEDORA viene pre instalado el servidor web, la ruta donde se encuentran las  páginas WEB es:

**/var/www/html**

## *Catálogo con Scripts en PERL*

Para utilizar el sistema de búsqueda por defecto de Swish-e basado en scripts CGI de PERL, se requiere utilizar un directorio que permita la ejecución de scripts, El directorio por defecto de **APACHE** donde se encuentran los scripts web es **CGI-BIN**:

**/var/www/cgi-bin**

Una vez ubicado el directorio de scripts, se debe copiar a este el archivo **"swish.cgi"**, el cual está ubicado en los archivos por defecto de instalación del Swish-e, en este caso, se encuentra instalado en la ruta:

**/usr/local/lib/swish-e/**

Para copiar los archivos simplemente se ejecuta el comando CP, así:

**cp /usr/local/lib/swish-e/swish.cgi /var/www/cgi-bin/**

Una vez copiado el archivo "**swish.cgi**", en el directorio de los scripts **CGI-BIN**, se puede comprobar la interfaz del buscador proporcionada por el programa, para esto hay que dirigirse al siguiente URL:

**http://bomarzo.rec.usal.es/cgi-bin/swish.cgi**


Aunque aparezca un formulario, no se podrá realizar ninguna búsqueda hasta que no se le indique al script "**swish.cgi**", donde ubicar los índices y los archivos que componen la colección, para esto, se requiere crear un archivo de configuración **("swishcgi.conf")** el cual es presentado a continuación:

```
return {
      title       => 'Busqueda en Swish-e',
      swish_binary => '/usr/local/bin/swish-e',
      swish_index  => '/var/www/html/swish-e/trabajo5/indice.index',
     }
```

Este archivo de debe crear en la misma ruta donde se encuentre el archivo "**swish.cgi**", posteriormente al realizar nuevamente la búsqueda, se puede observar que el catálogo ya permite buscar sobre los índices anteriormente creados.


Cuando se desee cambiar la interfaz del catálogo web ó traducir alguno de sus componentes, solamente hay que modificar el archivo "**swish.cgi**", mediante el uso de un editor de texto, para obtener mayor información de los parámetros que componen este script, se puede consultar la documentación que se encuentra en http://swish-e.org/docs/swish.cgi.html

```
# prepend this path to the filename (swishdocpath) returned by swish.  This is used to
# make the href link back to the original document.  Comment out to disable.

prepend_path    => 'http://bomarzo.rec.usal.es/',

# This is the property that is used for the href link back to the original
# document.  It's "swishdocpath" by default
```

```
link_property   => 'swishdocpath' ,
```

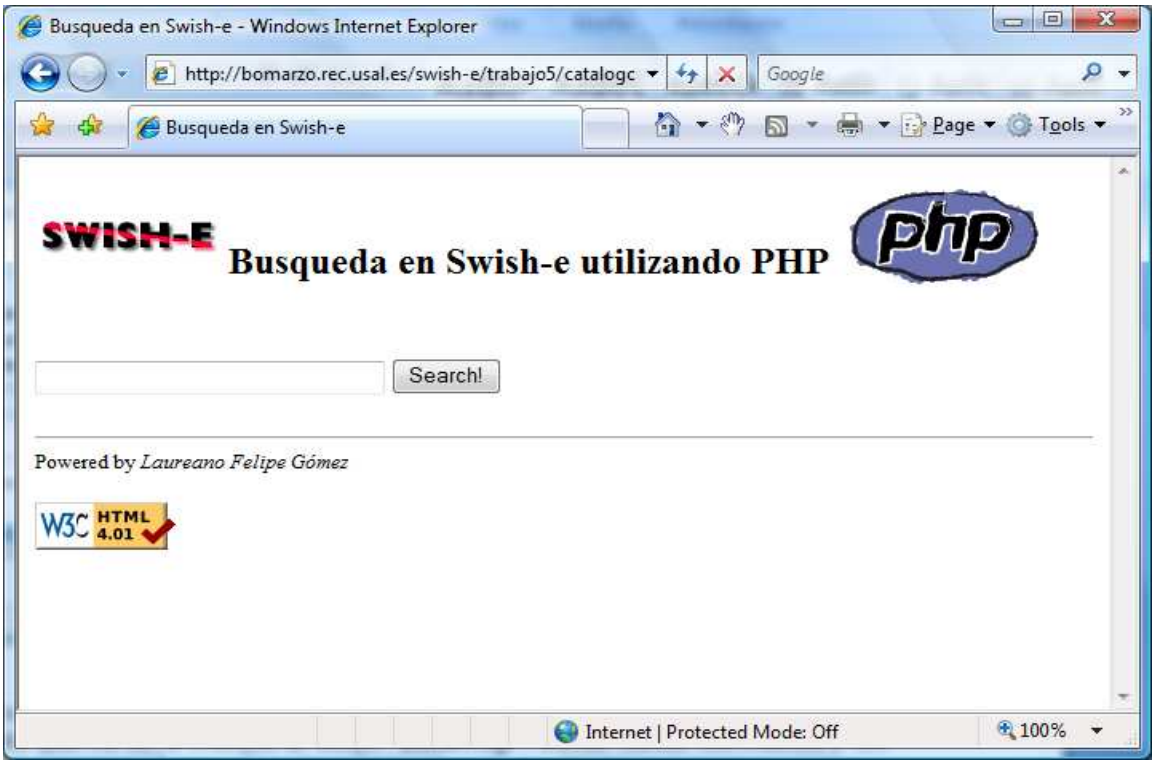
Para ajustar los enlaces a los documentos cuando se presentan los resultados de búsqueda en el catálogo, se requiere ajustar las variables **prepend_path y link_property** tal como se muestra en la tabla anterior, la primera variable contiene la ruta "Lógica" Web donde se encuentran los documentos, mientras que la segunda contiene la ubicación de los documentos indizados según los parámetros dados por el archivo de configuración (**IndexDir y ReplaceRules remove**).

## *Catálogo en PHP*

En la página http://es2.php.net/manual/es/ref.swish.php puede consultar todas las funciones que se pueden incorporar en los scripts web al utilizar la librería **PHP-Swish-e**

En este ejemplo se ha trabajado con dos archivos:

- **catalogo.html:** Corresponde a un página Web sencilla que contiene un formulario con una caja de búsqueda, la cual al ser ejecutada invoca a la página dinámica "buscar.php"



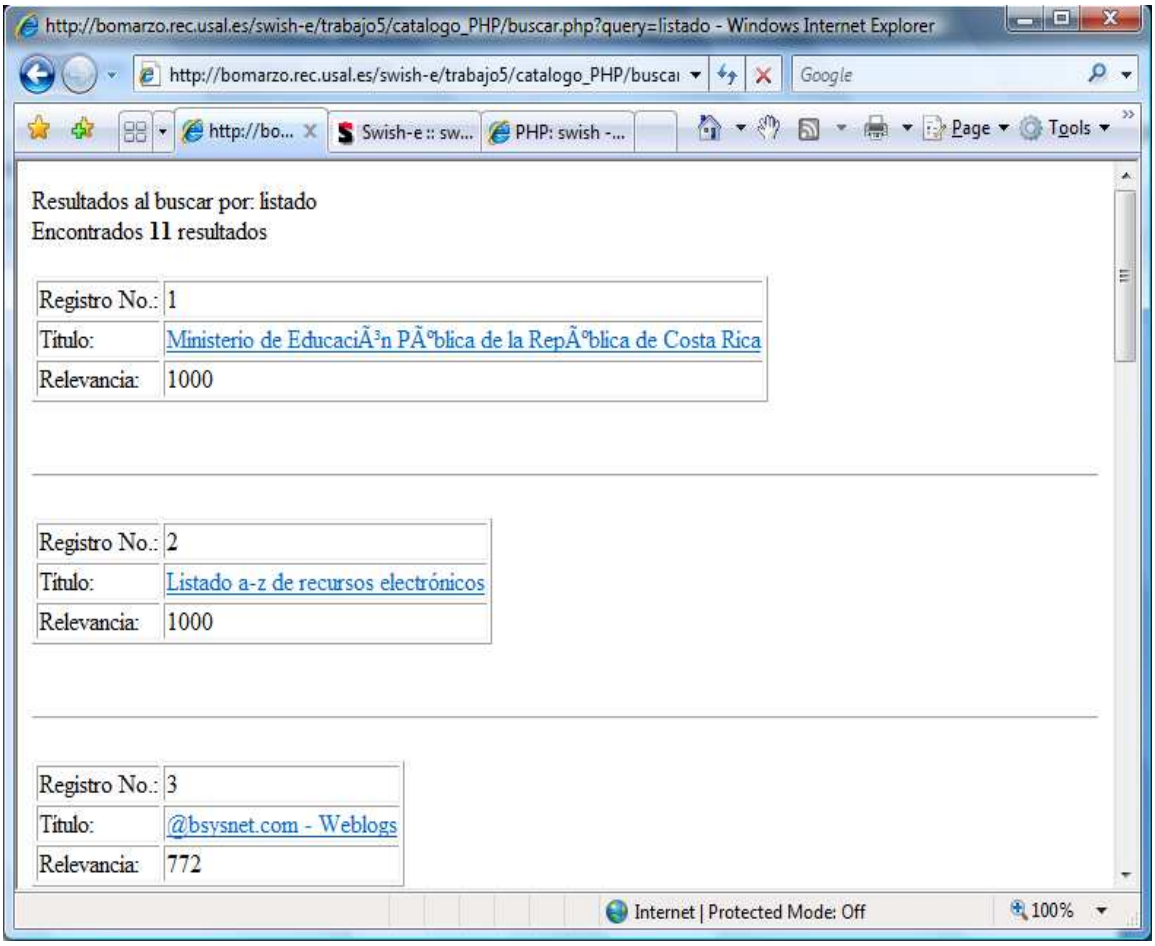

- **buscar.php:** Esta página recibe un parámetro con la expresión de búsqueda, y se encarga de abrir los índices de la colección desarrollada y realizar la búsqueda sobre los mismos, si encuentra resultados, los muestra en pantalla en forma de fichas (tablas), con los siguientes campos: Título, ubicación, ranking, y posición de búsqueda. El archivo "**buscar.php"** contiene las siguientes instrucciones:

```
<?php

//traigo variable por GET
$query = $_GET['query'];

//pregunto si viene algo, de lo contrario lo envió a catalogo.html
if (!@$query)
    header("location:http://bomarzo.rec.usal.es/swish-e/trabajo5/catalogo_PHP/catalogo.html");

//defino variables

//variable del for
```

```
$i=0;
//variables de resultados
$result= NULL;
//variable del Path_Archivo;
$Path_Archivo = NULL;

//inicio pagina WEB
echo "<html><body>";

//la muestro
echo "Resultados al buscar por: " . $query . "<br>";

//intento ejecutar busqueda
try
{
  //creo una variable del tipo Swish-e
  $swish = new Swish("/var/www/html/swish-e/trabajo5/indice.index");

  //realizo la busqueda
  $results = $swish->query($query);
  //muestro resultados encontrados
  echo "Encontrados <strong>", $results->hits, " </strong> resultados\n <BR>";

    //navego por los resultados si los hay??
    for ($i;$i< $results->hits;$i++)
    {
      //adelanto en el siguiente resultado
      $result = $results->nextResult();

      //comienzo una tabla
      echo "<br><table border=1>";
      //muestro Identificador
      echo "<tr><td>Registro No.:  </td><td> "  . $result->swishreccount . "</td></tr>";
      //creo el enlace al documento
      echo  '<tr><td>T&iacute;tulo:  </td><td>  <a  href="http://bomarzo.rec.usal.es/'  .  $result-
>swishdocpath .  '">';
      //muestro el titulo
      echo  $result->swishtitle  . '</a></td></tr>';
      //muestro RANK
      echo "<tr><td>Relevancia:    </td><td> "  . $result->swishrank     . "</td></tr>";
      //muestro RESUMEN
      echo "<tr><td>Resumen:    </td><td> "  . $result->swishdescription     . "</td></tr>";
      //finalizo tabla
      echo "</table><br><br> <hr>";
    } //fin FOR



}
```

```
catch (SwishException $e) //en caso de algún error
{
   echo "Error: ", $e->getMessage(), "\n";
} //fin try/catch

//Termino página WEB
echo "</body></html>";

?>
```

Cuando es ejecutada una búsqueda desde la página **"catalogo.html"**, la página dinámica **"buscar.php"** muestra los siguientes resultados:



## Ventajas y Desventajas de Swish-e

**Ventajas:** Entre las principales ventajas al utilizar Swish-e encontramos:

- Como tal, el sistema Swish-e es una herramienta muy poderosa con la cual se puede indizar cualquier tipo de archivo que contenga texto.
- Al ser una herramienta con licencia de software libre, permite su portabilidad a cualquier sistema operativo, para los cuales se encuentran múltiples versiones en el sitio web de Swish-e.
- Es una herramienta muy rápida en Indización y Búsqueda
- Permite indizar colecciones enormes de documentos, sin afectar su funcionalidad.

**Desventajas:** las principales desventajas al utilizar el programa Swish-e se encontraron las siguientes:

- No es un sistema de Indización llave en mano, como si lo es Google Desktop Search.
- No permite la indización optima de archivos semi-estructurados que no manejen las marcas de HTML/XML, por ejemplo no se podría indizar archivos del tipo MARC21.
- No maneja ningún protocolo de Interoperabilidad para recuperar la informaciòn en forma normalizada (por ejemplo z39.50, OAI-PMH, etc..).
- Actualización del software: Tal como se puede observar en la página de Swish-e, la última versión (2.5.4) es del 29 de Enero de 2007,

| Source Packages | | |
|---|---|---|
| swish-e-2.4.5.tar.gz | Mon, 29 Jan 2007 19:59:04 UTC | 1.4M |

# Apéndices

## *Anexo 1. Estructura del trabajo desarrollado*

Como el trabajo se realizó directamente sobre un servidor web, los archivos de prácticas pueden ser descargados desde: http://bomarzo.rec.usal.es/swish-e/, a continuación encontrará un resumen del contenido de estos directorios:

| | |
|---|---|
| Solo_Texto/10-Mar-2008 19:01 - | Ejemplos de indizaciòn bàsica archivos Texto Plano TXT |
| Varios_Formatos/11-Mar-2008 12:51 | Ejemplos de indizaciòn bàsica varis formatos de archivos |
| trabajo1/12-Mar-2008 19:35 - | Corresponde con el primer trabajo, en el que se solicita indizar pàginas HTML y hacer bùsquedas sobre su estructura |
| trabajo2/12-Mar-2008 19:46 - | Corresponde con el segundo trabajo en el cual se realiza la bùsqueda sobre varias colecciones documentales simultaneamente |
| trabajo3/12-Mar-2008 23:55 - | Corresponde con el tercer trabajo en el cual se crea una colecciòn compleja que contiene mùltiples formatos de archivos con una organizaciòn jerarquica. |
| trabajo4/13-Mar-2008 01:43 - | Corresponde con el cuarto trabajo en el cual se ha utilizado un SPIDER para indizar contenido externo que luego es indizado por Swish-e |
| trabajo5/13-Mar-2008 01:49 - | Corresponde con el quinto trabajo,en el cual se solicita realizar un catàlogo WEB para consultar los archivos |

## Licencia de este documento