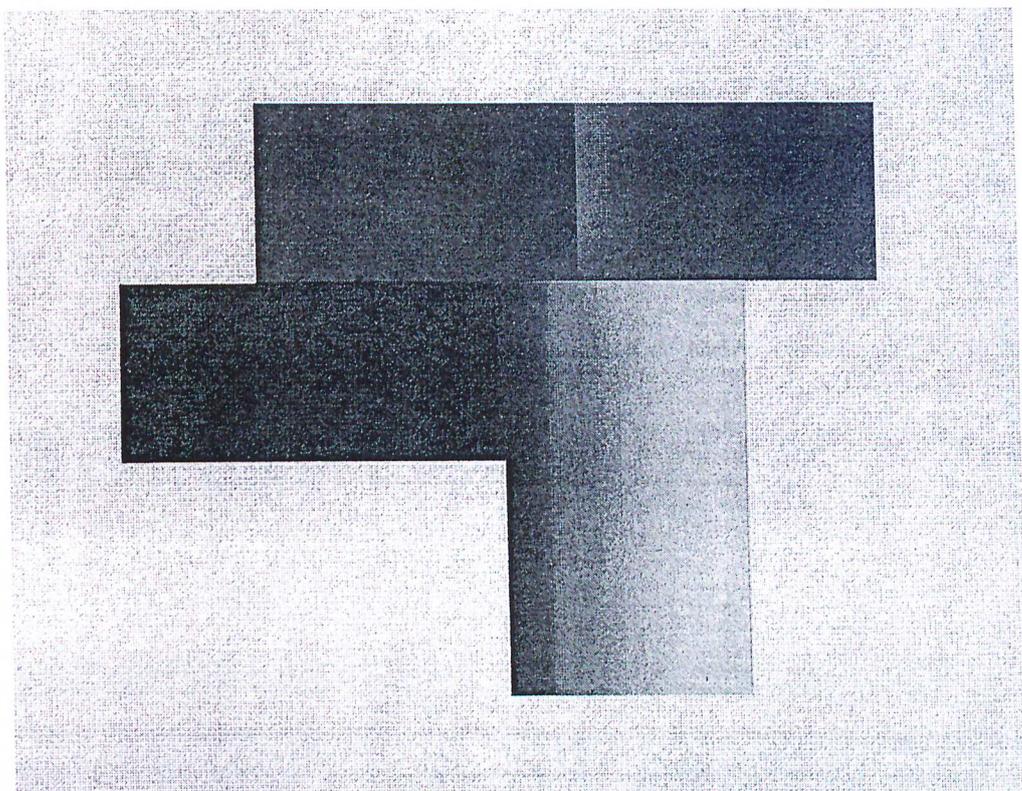


SISA

Sistema de Indización Semi-Automático



Guía de Usuario

Índice

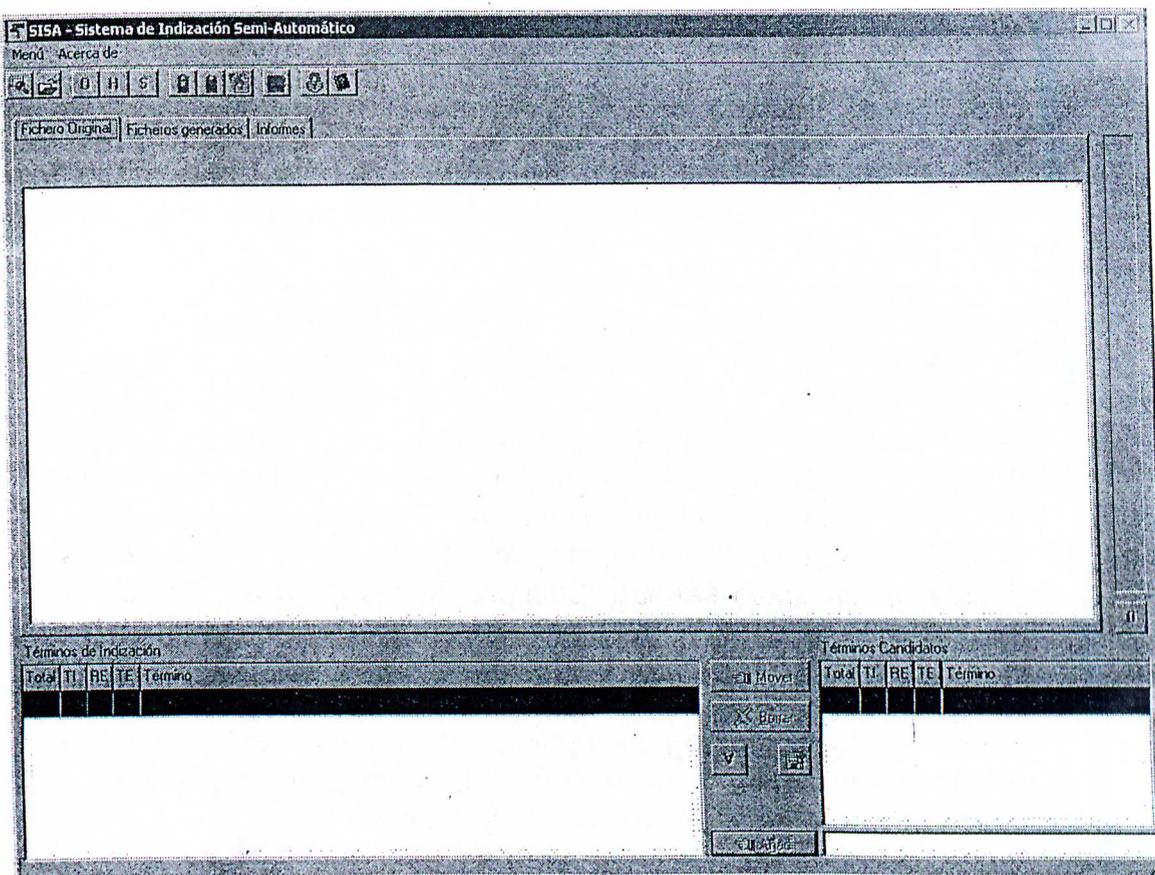
1	PREÁMBULO	3
2	INTRODUCCIÓN.....	4
3	NORMAS DE USO.....	6
4	OPCIONES DEL MENÚ	7
4.1	Menú de aplicación.....	7
4.1.1	Menú Artículo	7
4.1.2	Menú Informe	8
4.1.3	Menú Resultados	8
4.1.4	Menú etiquetas.....	8
4.2	Menú Acerca de.....	9
4.3	Barra de botones	9
5	CONFIGURACIÓN	10
6	SELECCIÓN DE FICHEROS	11
7	VOCABULARIO	12
8	EJEMPLO DE USO	13
9	EXPORTAR TÉRMINOS.....	16

1 PREÁMBULO

El presente manual pretende ser una guía para que cualquier usuario, familiarizado o no con SISA, descubra toda la funcionalidad que le ofrece la aplicación y pueda obtener los mejores resultados en su uso.

Se trata de la primera versión que aparece específicamente para Windows, donde se ha procurado primar la facilidad de uso y los interfaces de usuario sencillos. No es necesario decir que se mantiene toda la funcionalidad creada en la versión original, junto con mejoras procedentes de todos aquellos usuarios que han tenido la oportunidad de utilizar la aplicación, y cuya crítica ha servido para hacerla más ágil, eficiente y, sobre todo, práctica.

A continuación se muestra la pantalla inicial de la aplicación que servirá de punto de partida para descubrir las utilidades de indización documental que nos ofrece esta nueva versión.



2 INTRODUCCIÓN

El Sistema de Indización Semi-Automático (SISA) que se ha desarrollado, es una herramienta cuyo objetivo es colaborar en la tarea de la indización de los documentos que recibe como entrada.

Para su desarrollo se ha tomado como base la tesis doctoral de D. Isidoro Gil Leiva, aunque es preciso constatar que no se han seguido todos los procesos descritos en la misma. Veamos pues, a continuación, las diferencias entre la citada tesis doctoral y la aplicación resultante:

1. Una de las tareas necesarias en el proceso de indización es la horizontalización del texto. Para este proceso hemos supuesto que una línea no superará nunca la cantidad de 100 palabras, por razones de eficiencia en la implementación del algoritmo. Sin embargo, es una suposición más que razonable, puesto que difícilmente tendremos un texto con 100 o más palabras consecutivas sin un punto, coma, punto y coma o cualquier otro tipo de separador.
2. En segundo lugar, el algoritmo propuesto por D. Ignacio Gil Leiva permitía una profundidad de hasta cuatro palabras para la definición de los términos preferentes. La práctica ha demostrado que es posible utilizar más palabras, con lo que se ha ampliado dicha profundidad a 10.
3. A su vez, el algoritmo propone dos tipos de comparaciones en función del número de palabras del texto y del número de términos propuestos. Si tras la horizontalización y la eliminación de palabras vacías tenemos más palabras en el texto que términos propuestos, compararemos cada uno de los términos con las palabras. En caso contrario, la comparación se hará a partir de las palabras con los términos propuestos. En la aplicación final, esta comparación se realiza en un único sentido. Siempre se compara cada uno de los términos propuestos con las palabras del texto. El motivo es la implementación elegida: los términos propuestos siempre se almacenan en listas, para mejorar la eficiencia del proceso. Y además los términos propuestos pueden contener 'USES', con lo que es necesario definir una estructura de uso concreta que impide el funcionamiento en ambos sentidos de la aplicación.

El resto del proceso se corresponde exactamente con la implementación desarrollada en la tesis doctoral. Se llevan a cabo los procesos de horizontalización, eliminación de palabras vacías, búsqueda de términos propuestos y proposición de términos de indización. En cuanto a éste último punto, también es necesario realizar algunas matizaciones:

1. Los términos propuestos serán presentados como posibles términos de indización tal y como se propone en la tesis.
2. Las palabras vacías se presentarán como posibles términos de indización sólo en los siguientes supuestos:
 - Si la palabra semivacía aparece en el título, resumen y texto del artículo.
 - Si la palabra semivacía aparece diez (10) veces o más en el artículo.

Es necesario destacar una extensión implementada en la aplicación: se introducen términos generales que permiten categorizar los términos incluidos en el vocabulario con nuevos descriptores. Esto permite una jerarquización de los términos que responde a la posible necesidad de agrupación de términos.

Finalmente, sólo queda destacar los aspectos más relevantes de la interfaz gráfica. En la misma, es posible introducir nuevos términos de indización, eliminar los existentes o incluir términos de indización a partir de los términos propuestos por la herramienta construida.

Es importante destacar que en el caso de la modificación de los términos de indización propuestos, para recuperar los originales presentados por la herramienta será necesario volver a indizar el artículo, puesto que el mismo no guarda una 'memoria' de los resultados obtenidos inicialmente.

Como conclusión es necesario hacer hincapié en dos aspectos:

1. La aplicación ha tratado de respetar al máximo posible los algoritmos propuestos por D. Isidoro Gil Leiva en su tesis doctoral. Las modificaciones introducidas y comentadas anteriormente tienen como objetivo la mayor eficiencia algorítmica y/o flexibilidad en el uso de la herramienta.
2. El lenguaje utilizado para su desarrollo ha sido Delphi por las facilidades que ofrece en cuanto al interfaz gráfico, mantenimiento y futura extensión de la funcionalidad de la aplicación que ofrece.

3 NORMAS DE USO

Para el buen uso de la aplicación, es necesario cumplir una serie de normas que se enumeran a continuación:

Normas generales:

- 1) Los ficheros que contienen el vocabulario controlado, las palabras vacías y el artículo, deben ser todos ellos archivos de texto. (Por ejemplo: "articulo.txt").
- 2) Además, los archivos anteriores deben estar presentes dentro del mismo directorio de la aplicación, llamado 'fuentes'.

Fichero de palabras vacías:

- 1) Sólo puede tener una palabra por cada línea del fichero.
- 2) No puede contener la palabra "USE".

Fichero de términos preferentes:

- 1) Cada línea debe empezar con un número de término preferente seguido de un punto. (Por ejemplo: "5.")
- 2) Sólo puede aparecer un término "USE" por línea.
- 3) Un término "USE" debe aparecer entre dos términos diferentes. "5. USE CSIC" sería un ejemplo de línea incorrecta.
- 4) "USE" no puede ser un término del vocabulario.

Fichero del artículo:

- 1) Debe contener las siguientes marcas:
 - #CTI# y #FTI# para identificar el título.
 - #CR# y #FR# para identificar el resumen.
 - #CTE# y #FTE# para identificar el texto.
- 2) Las marcas deben aparecer en mayúsculas y en líneas diferentes.
- 3) Una línea con una marca no puede contener ningún otro tipo de texto.

Consejos de uso:

Si el artículo contiene la expresión "sustantivo+adjetivo", la aplicación entenderá que es una única palabra. Para que se identifiquen como palabras distintas es necesario escribir "sustantivo + adjetivo".

Del mismo modo, cuando aparece una palabra entre comillas, corchetes, etc. en el artículo, como por ejemplo 'casa' o [casa] o <casa>, si queremos que la aplicación sea capaz de encontrar el término **casa** deberemos escribirlo como ' casa ', [casa] o < casa >, es decir, dejando espacios entre el término y los símbolos que lo acompañan.

Finalmente, no hay que hacer distinciones entre palabras con o sin acento, tildes o diéresis porque la aplicación lo hace por nosotros.

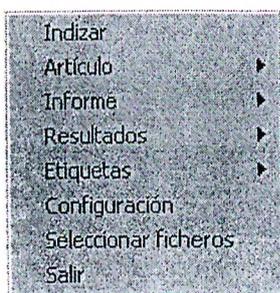
4 OPCIONES DEL MENÚ

En el diseño de SISA se han incluido dos menús. El primero abarca toda la funcionalidad de la aplicación, mientras que el segundo sólo contiene el clásico *Acerca de*, presente en la mayoría de aplicaciones, para conocer la versión que se está utilizando.

4.1 Menú de aplicación

El menú de aplicación tiene como entradas directas las siguientes:

1. Indizar: permite la indización de un artículo.
2. Configuración: abre la ventana de configuración, descrita más adelante.
3. Seleccionar ficheros: como en el caso anterior, abre una ventana descrita más adelante que permite seleccionar los ficheros a indizar.
4. Salir: cierra la aplicación.

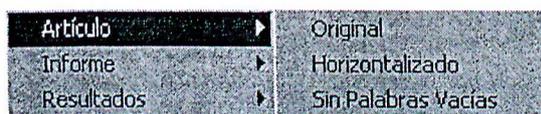


Aparte de los menús comentados, existen los de artículo, informe, resultados y etiquetas, que contienen otros menús asociados y cuya funcionalidad se describe en detalle para cada uno de ellos.

4.1.1 Menú Artículo

Una vez se ha indizado el artículo, el menú del mismo nombre ofrece las siguientes opciones:

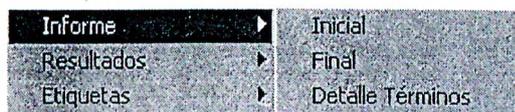
1. Original: permite ver el artículo original.
2. Horizontalizado: muestra el fichero tras el proceso de horizontalización.
3. Sin palabras vacías: es el fichero resultante de los procesos de horizontalización y eliminación de palabras vacías sobre el original.



4.1.2 Menú Informe

Como en el caso anterior, tras la indización del artículo, el menú Informe ofrece las siguientes opciones:

4. Inicial: es el resultado de la indización tras el proceso automático, sin ninguna intervención por parte del usuario.
5. Final: puesto que el usuario puede añadir o eliminar términos para utilizarlos como descriptores del artículo, el informe final muestra el resultado tras los cambios introducidos por el usuario.
6. Detalle de términos: al seleccionar un término, este aparece marcado en rojo sobre el artículo y, además, en esta opción se pueden ver única y exclusivamente las líneas en las que aparece el mismo.



4.1.3 Menú Resultados

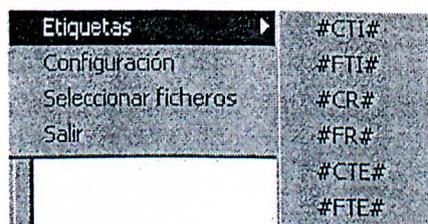
Siempre teniendo en cuenta que se ha indizado el artículo, el menú Resultados ofrece las siguientes opciones:

1. Incluir términos: permite abrir una nueva ventana donde se muestran los términos del vocabulario y, si está marcada la opción de incluir términos generales en la configuración, también los citados términos, para que se puedan seleccionar los que se desea incluir como descriptores.
2. Exportar: Los resultados obtenidos los lleva a un fichero de texto, cuya ruta debe estar previamente definida en la configuración.



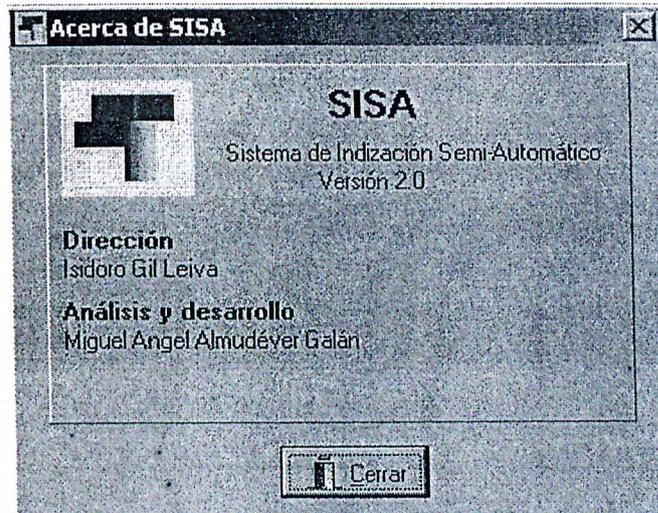
4.1.4 Menú etiquetas

Este menú muestra las etiquetas para que se puedan incluir en el artículo con un simple clic. Si las etiquetas ya están incluidas, muestra una marca delante de la etiqueta correspondiente para indicarlo.



4.2 Menú Acerca de

Al pinchar sobre el menú Acerca de, aparece un único elemento seleccionable, denominado SISA, y que muestra una ventana como la siguiente:



4.3 Barra de botones

Para agilizar el uso de la aplicación, se ha incluido en la misma una barra de botones, que no son más que accesos directos a las funciones ofrecidas por SISA. A continuación se muestra la citada barra y se describe la utilidad de cada uno de los botones, de izquierda a derecha.

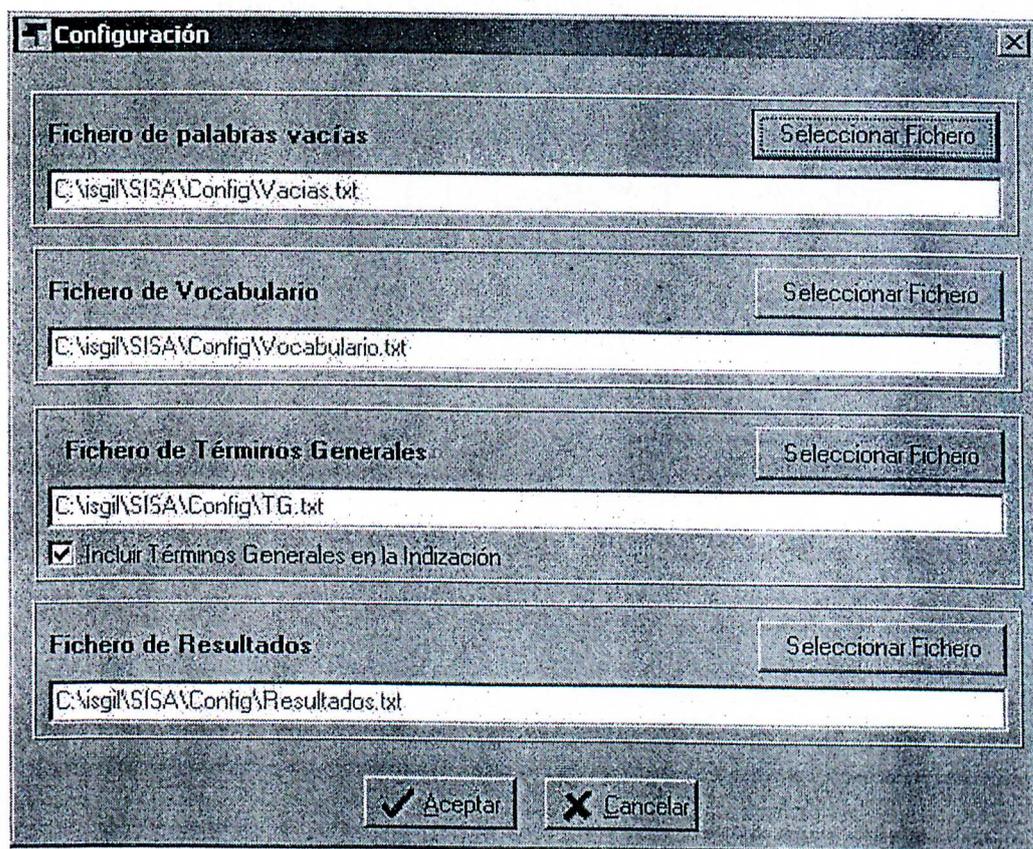


- Botón de indizar.
- Seleccionar ficheros.
- Fichero Original.
- Fichero Horizontalizado.
- Fichero Sin palabras vacías.
- Informe inicial.
- Informe final.
- Detalle de términos.
- Exportar resultados.
- Configuración.
- Acerca de SISA.

5 CONFIGURACIÓN

Con el objetivo de hacer más flexible el uso de la aplicación, se ha incorporado una ventana de configuración, con las rutas a los siguientes elementos:

- fichero de palabras vacías
- fichero de vocabulario
- fichero de términos generales
- fichero de resultados

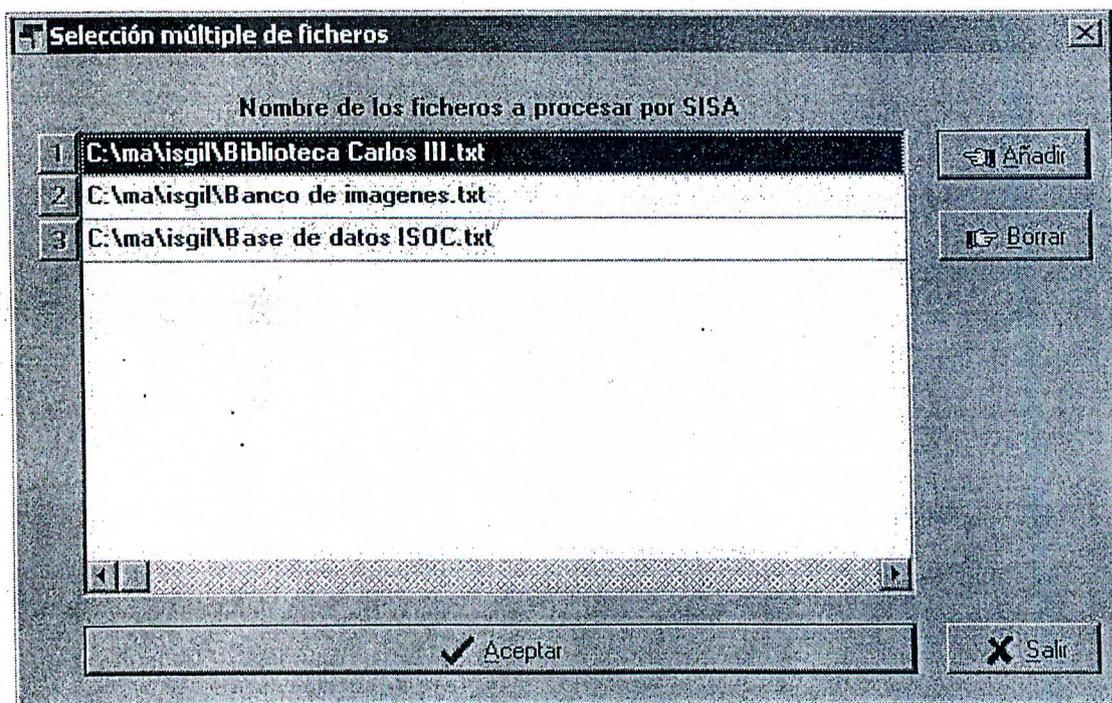


Por defecto, con la instalación de la aplicación, estos ficheros se crean en una carpeta llamada *Config*, que se encuentra dentro del directorio de instalación de SISA. Sin embargo, cada usuario puede modificar esta configuración para utilizar los ficheros que desee. Por ejemplo, se podría tener varios ficheros de vocabulario en función del contenido de los documentos a indizar (ecología, medicina, astrología, informática...) y cambiar la configuración para ajustarla a cada uno de estos temas.

Además, se ofrece la utilidad de incorporar un nivel de jerarquía superior con la inclusión de los términos generales de la indización, que son una especie de agrupación de los términos que pueden aparecer en el vocabulario.

6 SELECCIÓN DE FICHEROS

SISA permite la indización simultánea de hasta diez ficheros. Para ello, tiene una ventana de selección de ficheros, donde se lleva a cabo la selección de ficheros que se van a indizar. La citada ventana se muestra a continuación:



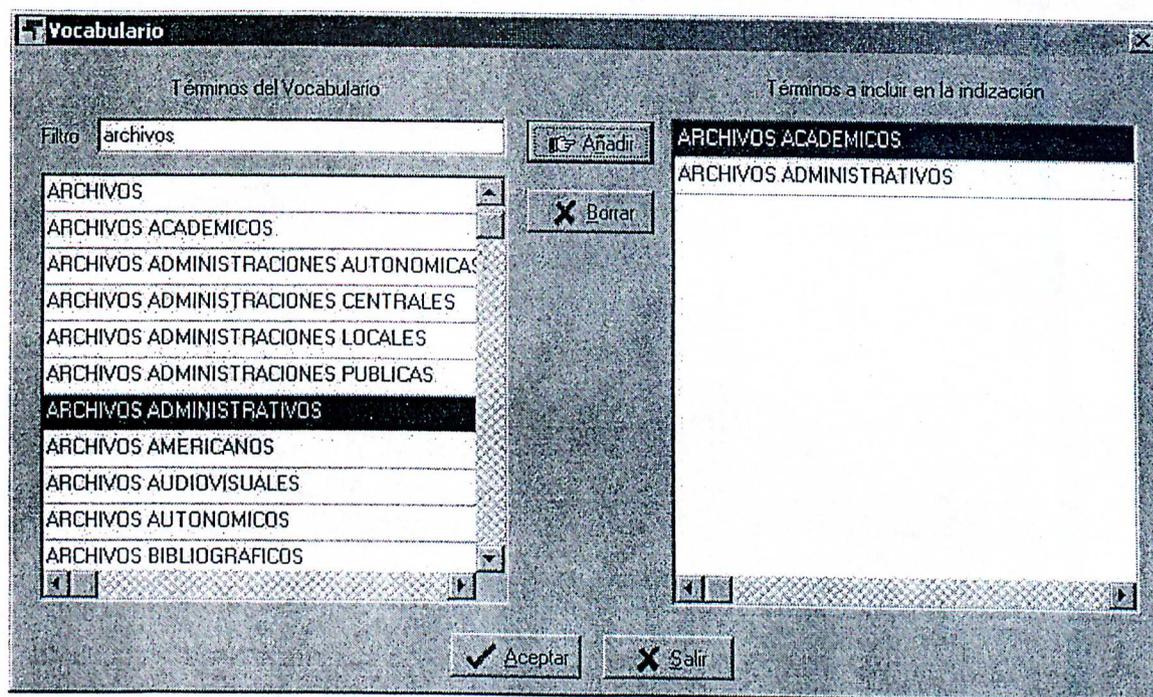
Las características más importantes de esta ventana son:

- Al pulsar sobre el botón *Añadir*, aparece una nueva ventana para la selección de ficheros, desde donde se pueden incorporar varios simultáneamente.
- En caso de error, se pueden *Borrar* los ficheros no deseados y, como en el caso anterior, se pueden elegir varios a la vez.
- El botón *Aceptar* hace que se vuelva a la ventana principal de la aplicación, para poder así indizar los ficheros.
- Mediante el botón *Salir* cierra la ventana sin seleccionar ficheros.

7 VOCABULARIO

Tras la indización de ficheros, se puede incluir nuevos términos como descriptores. Estos términos se pueden seleccionar entre los candidatos, ser añadidos directamente por el usuario o, como vamos a ver a continuación, seleccionados de entre los términos del vocabulario y los generales.

Como matiz, cabe destacar que los términos generales sólo aparecerán si en la configuración se ha seleccionado la opción de incluir términos generales en la indización de los documentos.

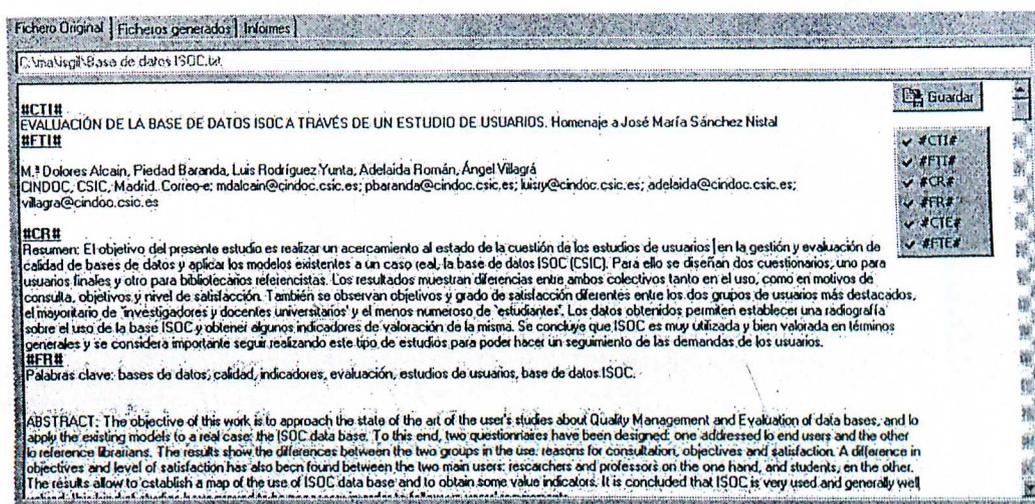


Con el fin de agilizar la búsqueda de términos se ha incluido un filtro. Tal y como se van añadiendo letras al filtro, la ventana va cambiando el contenido de los términos del vocabulario para mostrar solamente aquellos términos que empiezan por las letras o palabras introducidas en el filtro. Los términos no están ordenados: en primer lugar, si procede, se muestran los términos generales en el mismo orden en que se han escrito en su fichero asociado, y lo mismo ocurre con los términos del vocabulario.

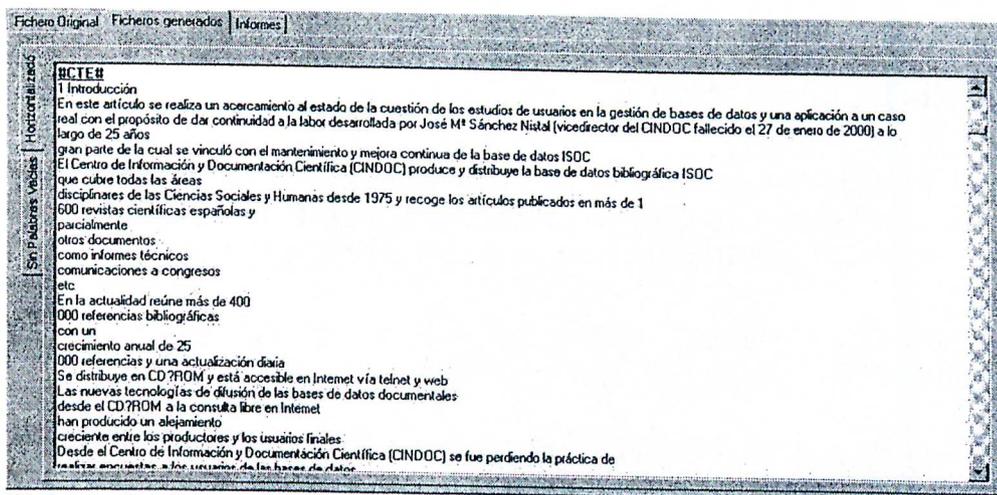
8 EJEMPLO DE USO

En este apartado vamos a desarrollar lo que sería el recorrido que puede llevar a cabo un usuario de la aplicación para descubrir todas las opciones que le ofrece la misma. Paso a paso, la navegación podría ser así:

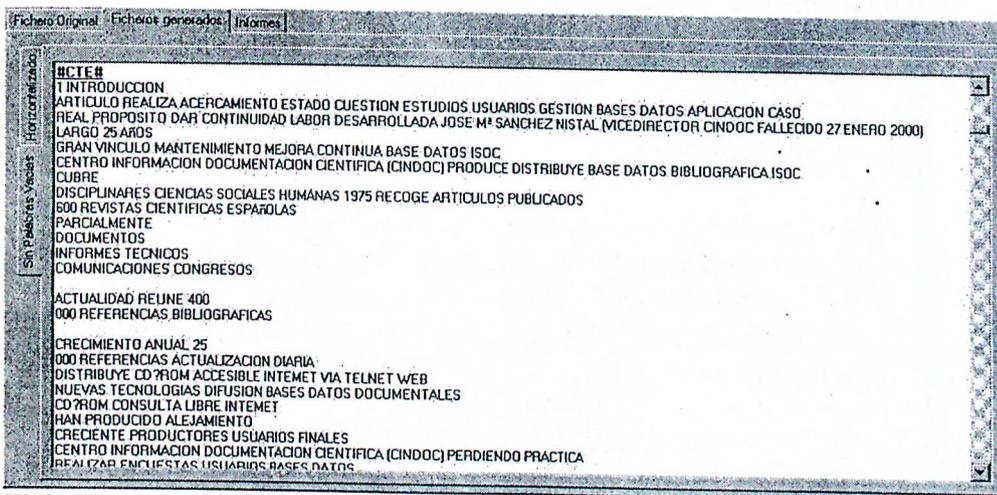
1. Iniciar la aplicación, donde aparece la ventana que tenemos en el preámbulo del presente manual.
2. Configurar las rutas de los archivos de vocabulario, palabras vacías, términos generales y resultados.
3. Seleccionar los ficheros a indizar. Una vez tenemos los ficheros seleccionados, si es necesario podemos introducir las marcas necesarias para el proceso de indización; en el ejemplo posterior, como se puede observar, todas las marcas están contenidas en el fichero. Tenemos dos opciones:
 - a. **Guardar:** activa cuando se ha modificado el documento.
 - b. **Menú de marcas de fichero:** aparece haciendo clic sobre el botón derecho del ratón cuando el cursor está sobre el artículo. Las marcas del fichero aparecen seleccionadas cuando ya están incluidas en el artículo.



4. Cuando los artículos – o el artículo, si sólo es uno – ya están preparados, procedemos a indizarlos. Una vez indizados, la aplicación nos permite seleccionar el fichero cuyos resultados queremos ver. A partir de aquí, podemos acceder a toda la funcionalidad de la aplicación:
 - a. **Pestaña de ficheros generados:** donde podremos ver el artículo original horizontalizado y sin palabras vacías. Se muestran en las siguientes imágenes:

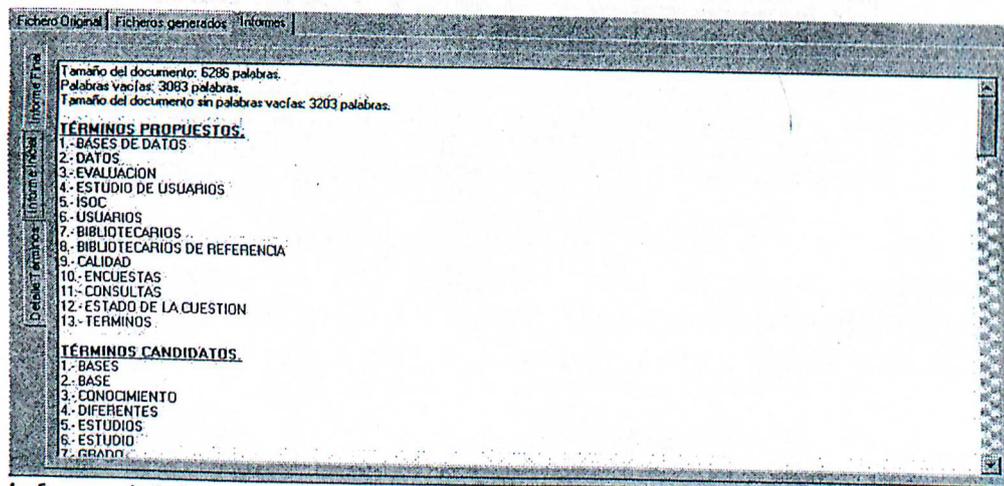


Fichero horizontalizado

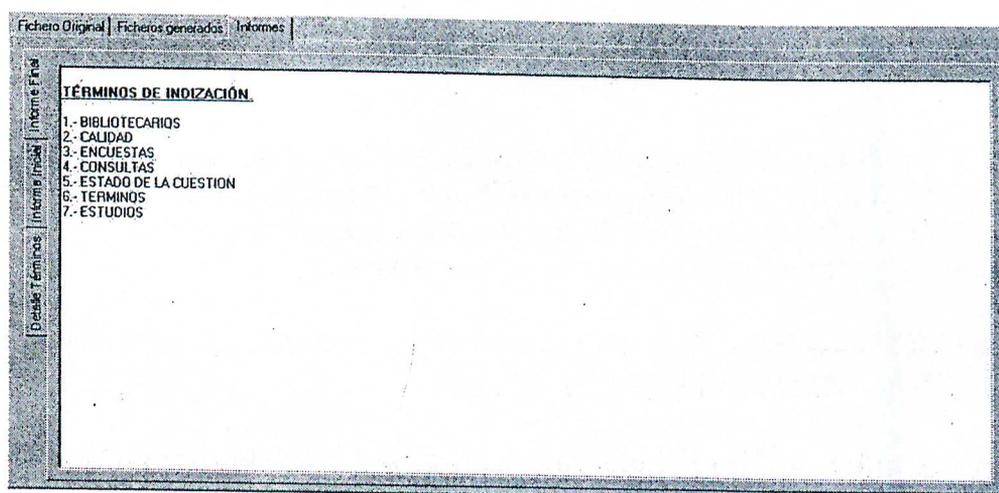


Fichero sin palabras vacías

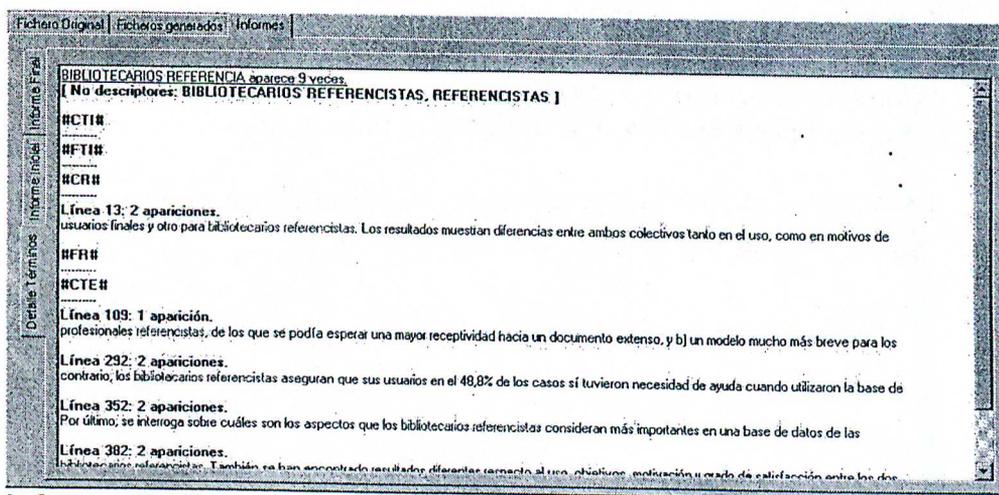
b. **Pestaña de informes:** desde esta opción tenemos tres posibilidades:



Informe inicial



Informe final



Informe detalle

5. Tras analizar todas las opciones disponibles, podemos decidir insertar nuevos términos desde el vocabulario, incluir como descriptores a alguno de los términos candidatos, o eliminar algunos de los descriptores por no considerarlo adecuado.
6. Finalmente y como se explica en el último apartado, se pueden exportar los términos a un fichero de texto, para su posterior uso.

9 EXPORTAR TÉRMINOS

El fichero de resultados es un archivo de texto cuya ruta se indica en la configuración. Tras la indización de un artículo, podemos exportar los descriptores al citado artículo, para guardarlos como resultado del proceso o para llevarlos más fácilmente a cualquier otra aplicación.

Para que se muestren los resultados del proceso de indización en el fichero de resultados se siguen las siguientes normas:

1. Escribir en una línea la etiqueta #FICHERO#, donde FICHERO se corresponde con la ruta del fichero indizado.

Ejemplo: si el artículo contiene una línea como la siguiente, #C:\ejemplo\SISA\Articulo.txt#, y se ha indizado un fichero cuya ruta es la que aparece entre almohadillas, entonces en la línea inmediatamente posterior se mostrará el resultado de la indización.

2. Escribir en una línea la etiqueta #TODOS#, hace que se muestren los resultados de todos los ficheros indizados.