2010-08-25

Desarrollo de herramientas de búsqueda en los archivos históricos de la CompañÍa de Jesús

Zaldívar-Carrillo, Víctor H.; Aranda-Sarabia, Juan A.; Alejandre-Marquez, Juan C.; Espinoza-Villarreal, Christian A.

Zaldívar-Carrillo, V.H.; Aranda-Sarabia, J.A.; Alejandre-Marquez, J.C.; Espinoza-Villarreal, C.A.(2010). "Desarrollo de herramientas de búsqueda en los archivos históricos de la Compañía de Jesús". 5° Congreso Internacional de Sistemas de Innovación para la Competitividad 2010 "Tecnologías Convergentes para la Competitividad". Consejo Estatal de Ciencia y Tecnología del Estado de Guanajuato; Celaya, Guanajuato, 25 al 27 de abril.

MT 6

Desarrollo de herramientas de búsqueda en los archivos históricos de la Compañía de Jesús

Víctor Hugo Zaldívar Carrillo[1]
Juan Alfonso Aranda Sarabia[2]
Juan Carlos Alejandre Márquez[3]
Christian Adán Espinoza Villarreal[4]
Andrés Torres García[5]

Resumen

El trabajo de los historiadores es pesado y difícil ya que consiste, fundamentalmente, en la localización, análisis e interpretación de documentos antiguos. Estos documentos, cuando están disponibles para su consulta, deben ser leídos uno por uno y todas las posibles referencias cruzadas, revisadas una por una. En el caso de la Compañía de Jesús, llegaron a la Nueva España en el siglo XV y trabajaron en colegios, misiones, parroquias y hospitales a lo largo de todo el continente hasta la expulsión de la Orden en 1767.

Durante este tiempo, la Compañía de Jesús puso en marcha un sistema de comunicación conocido como las Cartas Anuales o *Cartas Annua* que el Provincial enviaba al General de la Compañía en Roma. Cada una de las Cartas Anuales relataba las actividades

[1] Dr. en Ciencias en Informática por la Universidad de Ciencias y Técnicas de Laguedoc, Montpellier II. Actualmente profesor-investigador en el Departamento de Electrónica, Sistemas e Informática de la Universidad ITESO. victorhugo@iteso.mx
[2] Estudiante de Ingeniería en Sistemas Computacionales por la Universidad ITESO, presidente del Club Estudiantil de Robótica del Iteso. Is120676@iteso.mx
[3] Estudiante de Ingeniería en Sistemas Computacionales en la Universidad ITESO, Desarrollador de software para Nokia GmbH (Alemania) calo_11@hotmail.com.
[4] Estudiante de Ingeniería en Sistemas Computacionales en la Universidad ITESO. Caevchris_77@hotmail.com
[5]Estudiante de Ingeniería en Sistemas Computacionales por la Universidad ITESO. Miembro activo de la comunidad de estudiantes de Microsoft, Microsoft Student Partnet (MSP)

llevadas a cabo por miembros de la Compañía. Para ayudarse a redactar cada Carta y con el fin de circular la información en toda la Provincia de la Nueva España, cada uno de los encargados de las obras enviaba al Provincial una o varias cartas conocidas como *Puntos de Annua.*

La Compañía de Jesús, junto con el Archivo General de la Nación y el Instituto Nacional de Antropología e Historia se dio a la tarea de recopilar todas las Cartas Anuales escritas desde que los jesuitas llegaron a la Nueva España hasta su expulsión en 1767. Esta búsqueda resultó en más de 1300 folios que debieron paleografiarse para ayudar a la investigación histórica sobre la Compañía de Jesús.

Como resultado de este trabajo de recuperación y paleografía y para facilitar el acceso y manejo de los documentos, se decidió crear una herramienta informática que ofreciera a los investigadores un motor de búsqueda para hacer referencias cruzadas rápida y fácilmente, así como el acceso a los documentos a través de índices temáticos.

El desarrollo de esta herramienta presentó varios problemas técnicos que se han ido solucionando. Por ejemplo, la necesidad de que los documentos en formato electrónico siguieran fielmente las reglas de la paleografía. Este requerimiento hizo necesaria la realización de un filtro para manipular los archivos en formato XML. Se desarrolló también la interfaz gráfica para acceder a los índices temáticos y el motor de búsqueda para realizar referencias cruzadas en una o varias palabras, continuas o discontinuas en el texto del documento.

La herramienta ha sido bien recibida por los investigadores y se ha decido aumentarla con algoritmos de Text-Mining para aumentar las capacidades de la búsqueda así como la generación de referencias cruzadas entre los documentos.

Palabras clave: Cartas Annuas, Compañía de Jesús, Provincia de la Nueva España, Motor de Búsqueda, Interfaz Gráfica, Herramientas Informáticas

Antecedentes y Contexto

Las Cartas Annuas

Los jesuitas llegaron a América en el siglo XV. A partir de entonces y hasta su expulsión en 1767 se encargaron de obras como colegios, misiones, parroquias y hospitales a lo

largo de todo el continente. Así, durante este periodo, se tiene información sobre las provincias de Perú, Nuevo Reino de Granada, Nueva España, Chile, Nueva Vizcaya, etc.
Las Cartas Annua son extensas cartas-informes, escritas por el provincial y enviadas a la Curia generalicia. En ellas se resumían todos los sucesos ocurridos durante el año en las casas de la Compañía en cada provincia. Las Cartas Annua tienen un contenido etnográfico importante pues una parte de ellas se dedicaba a las misiones (Fundación Histórica Tavera, 2010).
Las *Cartas Annuas* son fuentes históricas de primer orden ya que eran elaboradas cuidadosamente y con contenidos bien definidos y estructurados de acuerdo a unidades temáticas. (Fundación Histórica Tavera). Han sido utilizadas como fuente para realizar un sinnúmero de trabajos con temas desde los esclavos traídos de África (Navarrete, 2006) hasta el análisis literario del discurso religioso (Campbell, 1992).
Sin embargo, cuando los jesuitas fueron expulsados de la Nueva España en 1767, gran parte de sus bienes fueron confiscados, entre ellos, sus archivos. Así, los documentos de las Cartas Annuas y los Puntos de Annua se encuentran distribuidos en diferentes fondos y colecciones como el Archivo General de la Nación en México, el Archivo Nacional de Madrid, el Archivo Nacional de Chile, el Archivo de la Real Academia de la Historia, el Archivo Histórico de la Biblioteca Nacional de Antropología e Historia, el Archivo Histórico de la Compañía de Jesús, entre otros.

La recuperación de los documentos

Desde hace varios años, la Provincia Mexicana de la Compañía de Jesús y el Instituto Nacional de Antropología e Historia se han dado a la tarea de localizar y paleografiar los documentos de las Cartas Annua y los Puntos de Annua de la Provincia de la Nueva España.
La paleografía es una disciplina auxiliar de la historia que estudia la escritura y signos de los documentos antiguos (RAE, 2010). En general el trabajo de paleografía es necesario cuando los documentos que se están utilizando no son fácilmente accesibles. Consiste en transcribir el documento antiguo a un formato más moderno en papel o electrónico pero cuidando que el nuevo documento sea una representación fiel del documento antiguo. Esto es, guardando el mismo formato y ortografía y cuidando que se representen

características y peculiaridades de los documentos manuscritos como tachaduras y correcciones, y, a veces, notas agregadas por autores que no son el autor original del documento. Por ejemplo, en el caso de las Cartas Annuas, es común que estas cartas contengan al menos dos columnas, una con el texto de la carta y otra con anotaciones hechas, a veces, por los responsables de alguna de las obras de la Compañía.

Definición del proyecto

Una vez que se recuperaron todos los archivos posibles, se comenzó el proceso de paleografía con la intención de generar una publicación que sirva de consulta para los investigadores interesados en el periodo que abarcan las Cartas Annuas.

En este momento surge la inquietud de completar la obra impresa con un CD que contenga los archivos electrónicos paleografiados.

En este momento, se le solicita al ITESO que intervenga para contribuir a la publicación de esta obra con el desarrollo de una herramienta informática que funcione como interfaz del usuario para el acceso y manejo de los documentos. Durante las reuniones de trabajo con el socio del provincial de la Compañía de Jesús se definieron los objetivos que se muestran a continuación.

ID	***Objetivos de negocio de la empresa que aplican al proyecto***	***Fuente***	***Prioridad***
B03	*Desarrollar un motor de búsqueda para el contenido de las CARTAS ANUALES o LITTERAE ANNUAE y PUNTOS DE ANUAE*	*Carta de Eugenio Gómez, S.J.*	*Alta*
B04	*Integrar las CARTAS ANUALES o LITTERAE ANNUAE y PUNTOS DE ANUAE y el motor de búsqueda en un DVD*	*Carta de Eugenio Gómez, S.J.*	*Media*

Tabla 1: Objetivos generales del proyecto

Así como una serie de requisitos funcionales del sistema, que se muestran en la siguiente tabla:

ID	Requisitos de productos/servicios	Prioridad
F01	*El motor de búsqueda deberá ser capaz de realizar búsquedas por palabra*	*Alta*
F02	*El motor de búsqueda deberá ser capaz de utilizar operadores lógicos*	*Alta*
F03	*Se deberá desarrollar un índice por cada palabra relevante del documento*	*Alta*
F04	*Se deberá integrar el Índice onomástico al motor de búsqueda*	*Alta*
F05	*Se deberá integrar el Índice toponímico al motor de búsqueda*	*Alta*
F06	*Se deberá integrar el Índice de gentilicios al motor de búsqueda*	*Alta*
F07	*Se deberá integrar el Índice de lenguas e idiomas al motor de búsqueda*	*Alta*
F08	*Se deberá integrar Índice de enfermedades al motor de búsqueda*	*Alta*
F09	*Se deberá integrar Índice de seminarios al motor de búsqueda*	*Alta*
F10	*Se deberá integrar Índice de colegios al motor de búsqueda*	*Alta*
F11	*Se deberá integrar Índice de residencias al motor de búsqueda*	*Alta*
F12	*Se deberá integrar Índice de misiones al motor de búsqueda*	*Alta*
F13	*Se deberá integrar Índice de haciendas al motor de búsqueda*	*Alta*
F14	*Se deberá integrar Índice de templos e iglesias al motor de búsqueda*	*Alta*
F15	*Se deberá integrar Índice de congregaciones y cofradías al motor de búsqueda*	*Alta*
F16	*Los documentos solo serán de lectura, lo cual impedirá que se hagan modificaciones a las fuentes originales*	*Alta*
F17	*Se desarrollarán los algoritmos necesarios para posteriormente utilizar google desktop*	*Alta*
F18	*Se utilizará google desktop como interfaz de búsqueda*	*Alta*
F19	*Se deberá integrar el glosario de términos jesuíticos que será entregado por los Jesuitas*	*Media*
E01	*El total CARTAS ANUALES o LITTERAE ANNUAE y PUNTOS DE ANUAE están transcritos en archivos mayores o igual a la versión 2003 de Ms Excel*	*Alta*
E02	*El total CARTAS ANUALES o LITTERAE ANNUAE y PUNTOS DE ANUAE deberán ser entregados al Asesor Técnico a más tardar el 29 de noviembre de 2009*	*Alta*
E03	*Como resultado del trabajo paliográfico, se respetará el idioma en que fueron escritos los documentos originales*	*Alta*
E04	*El DVD con los documentos y el motor de búsqueda podrá ser utilizado bajo la plataforma Windows XP o superior*	*Alta*
E05	*El DVD con los documentos y el motor de búsqueda podrá ser utilizado bajo la plataforma MAC OS 10.5 o superior*	*Baja*

Tabla 2: Requerimientos funcionales de la herramienta

El equipo de trabajo decidió que el trabajo del ITESO en este proyecto consistiría en el desarrollo de una herramienta informática que consta de 3 partes:

- Interfaz Gráfica para el acceso a los documentos.
- Motor de Búsqueda para acceder a documentos a través de la búsqueda de palabras o conjuntos de palabras.

- Filtrado y conversión de archivos. Como mencionamos anteriormente, es necesario que los documentos paleografiados representen fielmente los originales, este filtro es el que se encarga de dar formato a los documentos.

La siguiente figura muestra un esquema de esta herramienta:

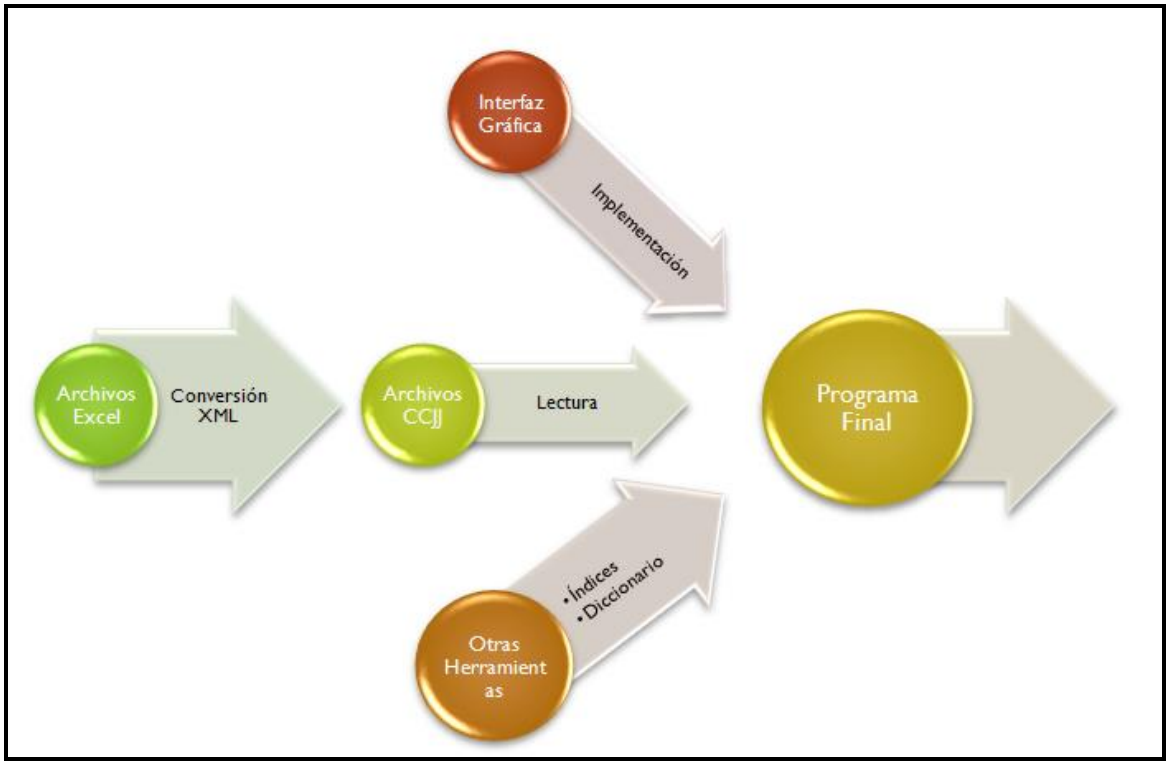

Figura 1: Esquema de trabajo en el proyecto.

Descripción del trabajo realizado

En esta sección vamos a presentar el trabajo que se realizó en cada una de estas etapas.

Conversión XML

El primer paso consistió en la creación de una herramienta de filtrado y conversión de archivos ya que los documentos paleografiados fueron capturados en hojas electrónicas (archivos de Excel). La intención de crear este filtro fue pasar del formato de archivos de Excel a uno más sencillo en XML.

Decidimos utilizar XML porque el archivo resultante es más ligero al tratarse de un archivo en texto plano y también porque el tratamiento de dicho archivo es relativamente simple utilizando Hojas de Estilos para su presentación y manipulación.

El desarrollo de este filtro nos obligó a crear un estándar para el formato en XML convenido como los archivos CCJJ. El proceso de conversión requería de un "Parser" que leyera los archivos de Excel y los escribiera en el formato CCJJ.

Este estándar consiste en un conjunto de etiquetas definidas por el equipo de trabajo y que tienen una semántica ad hoc para este caso.

Este proceso consiste básicamente en la recolección de información de los archivos paleografeados tales como los formatos, la alineación, el tipo de letra, la columna donde se encontraba, para después al momento de ser escrito en el nuevo archivo, cada renglón contuviese a su vez toda la información indispensable correspondiente en forma de etiquetas de XML.

Este proceso de conversión se aplica a todos los documentos que son transcripciones de las Cartas Annuas así como a los documentos de los diversos índices que se usarán para las búsquedas(los formatos en XML para los índices son distintos a los CCJJ sin embargo pasan por un proceso similar a estos).

Lectura de archivos CCJJ

Una vez que tenemos los archivos CCJJ, la interfaz gráfica podrá leerlos de manera eficiente minimizando el tiempo requerido para desplegar esta información. Lo anterior se debe a la simpleza de los formatos así como de la facilidad de lectura que proveen los mismos.

Lectura de índices

Los índices tienen un formato aún más sencillo puesto que el contenido de estos para fines prácticos es el tipo/categoría de índice (onomástico, seminarios, etc.) así como una breve concordancia de sus conceptos y los lugares donde fueron encontrados (hablando en términos de Las cartas Jesuitas y sus respectivas hojas). Esto nos permitirá hacer búsquedas en un futuro mucho más precisas y rápidas.

Generación del Diccionario

A la par que se crean los archivos CCJJ, el diccionario de las palabras (de 4 letras o más) se irá creando junto con su respectiva concordancia. Aunque este proceso implica aumentar el tiempo de conversión de documentos implica un ahorro de tiempo significativo al final ya que de esta manera no ahorramos lecturas futuras a los documentos CCJJ; esto es ya que hacemos la mayor parte del proceso en la primer lectura de los archivos Excel y en la escritura de los archivos CCJJ.

Reduciendo así la cantidad de procesos requeridos para la generación de los recursos para nuestras herramientas.

Interfaz Gráfica

La interfaz gráfica es la encargada de interactuar con el usuario, esta le permitirá al usuario visualizar los documentos que solicite a través de los índices de documentos así como de los que solicite por medio de una búsqueda de palabras.

Las búsquedas serán sobre todos los documentos CCJJ y podrán hacerse como en cualquier motor de búsquedas como Google, AltaVista, etc. Con sus funcionalidades básicas.

La interfaz gráfica propia desarrollada en la plataforma .NET usando C# (Deitel & Deitel, 2007 ; Ceballos, 2007) como lenguaje de programación. Esta interfaz le permite al usuario interactuar con las funcionalidades implementadas en el sistema. A partir del menú principal se puede acceder a las 4 secciones principales:

- Diccionario
- Búsquedas
- Índices
- Documentos

A continuación se muestran algunas ventanas de la herramienta desarrollada y se explican estas secciones.

Diccionario

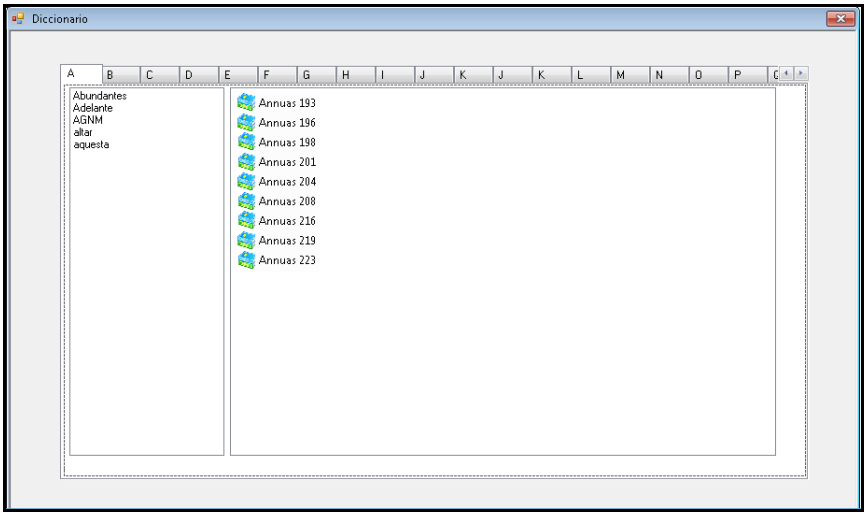

Figura 2 – Diccionario

La sección de "Diccionario" incluye una pestaña para cada letra del alfabeto y dos ventanas. En la ventana de la izquierda, se muestra un listado ordenado de las palabras que comienzan con la letra seleccionada. Cuando el usuario selecciona una de estas palabras, en la ventana de la derecha se despliegan los documentos en donde aparece al menos una vez dicha palabra. El usuario tiene la posibilidad de abrir cada uno de estos documentos.

Búsquedas

El módulo de búsquedas permite realizar búsquedas sobre cada una de los documentos contenidos en la herramienta Estos documentos han sido transformados a formato XML. Según la palabra, frase o expresión lógica que haya sido ingresada por el usuario, la herramienta genera una consulta en LINQ (Language Integrated Query, Microsoft 2010) para realizar búsquedas sobre cada uno de los documentos y así mostrar aquellos documentos o cartas que contengan la frase o palabra buscada.

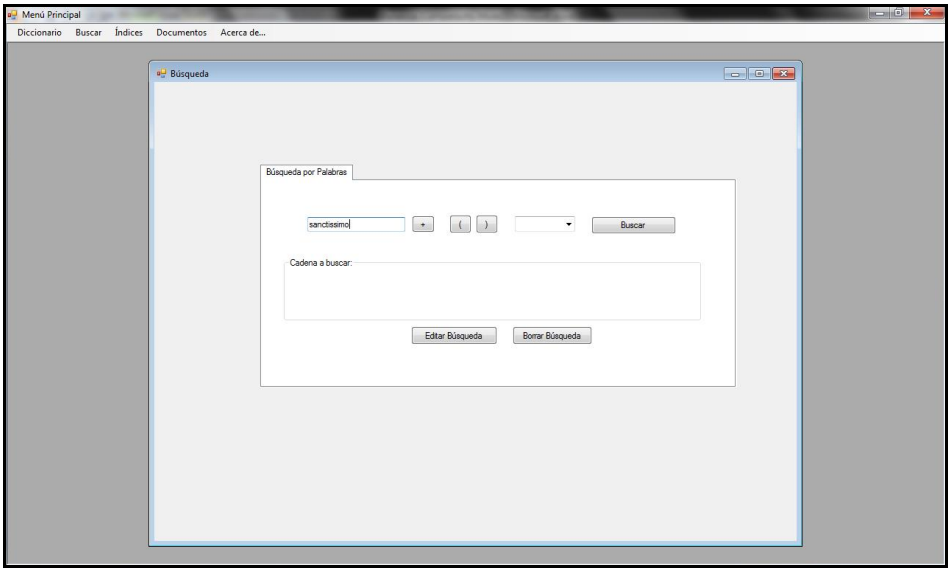

Figura 3 - Búsquedas

Por ejemplo si se busca la palabra "sanctissimo" y el resultado es el que se muestra en la figura siguiente:

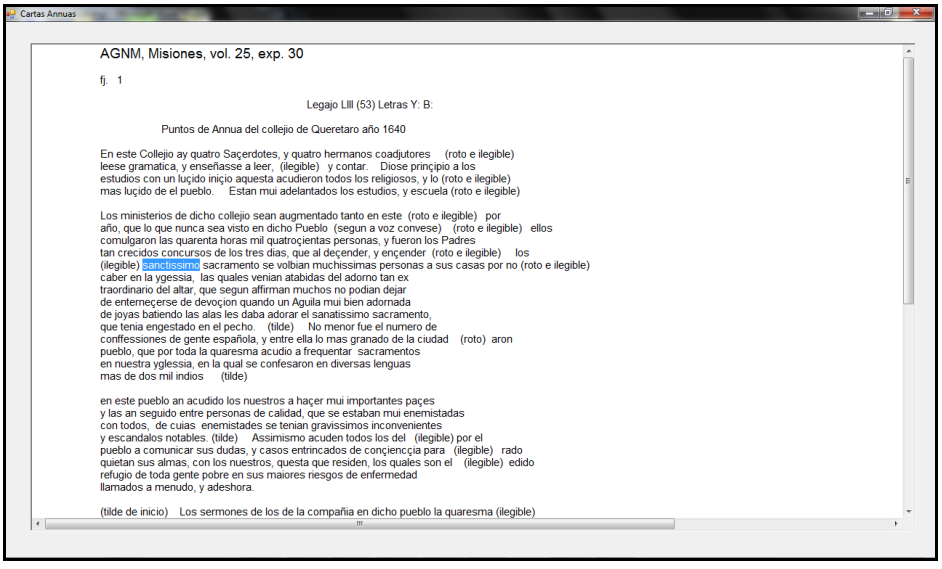

Figura 4 – Documento resultado de la búsqueda

Por otro lado, la herramienta de búsquedas cuenta con un editor de expresiones lógicas que son utilizadas para hacer búsquedas más sofisticadas utilizando operadores lógicos como AND, OR y NOT y paréntesis para lograr una expresión más compleja.

Por otro lado, el usuario tendrá en todo momento acceso a un editor manual de búsquedas. Este editor es útil cuando se realizan búsquedas con expresiones complejas a las que solamente se le quiera hacer una pequeña modificación.

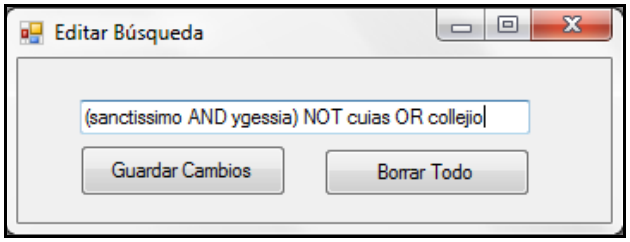

Figura 5 – Editor de búsquedas

Índices

El módulo de Índices funciona de manera parecida a los diccionarios. Las pestañas de la parte de arriba permiten seleccionar un índice temático a utilizar. En la ventana de la izquierda aparecen los términos del índice seleccionado y en la ventana de la derecha los documentos en los que aparece dicho término.

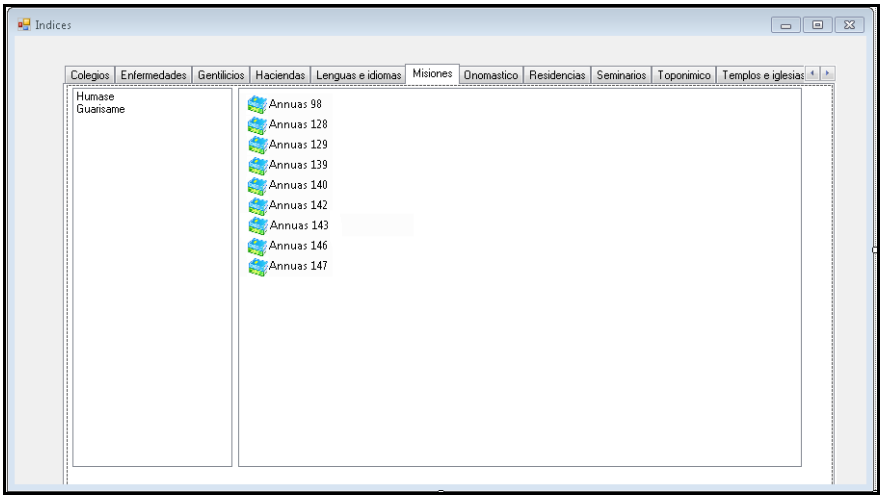

Figura 7 – Índices

Los índices temáticos que se incluyeron en esta versión de la herramienta son:

- Colegios
- Enfermedades
- Gentilicios

- Haciendas
- Lenguajes e Idiomas
- Misiones
- Onomástico
- Residencias
- Seminarios
- Toponímicos
- Templos e Iglesias

Documentos

Al seleccionar la pestaña de "Documentos" se abre una ventana como la de figura que le permite seleccionar cualquiera de los documentos incluidos en la herramienta. Este módulo se desarrolló porque algunos investigadores (usuarios de la herramienta) prefieren abrir directamente los documentos para leerlos y tomar notas.

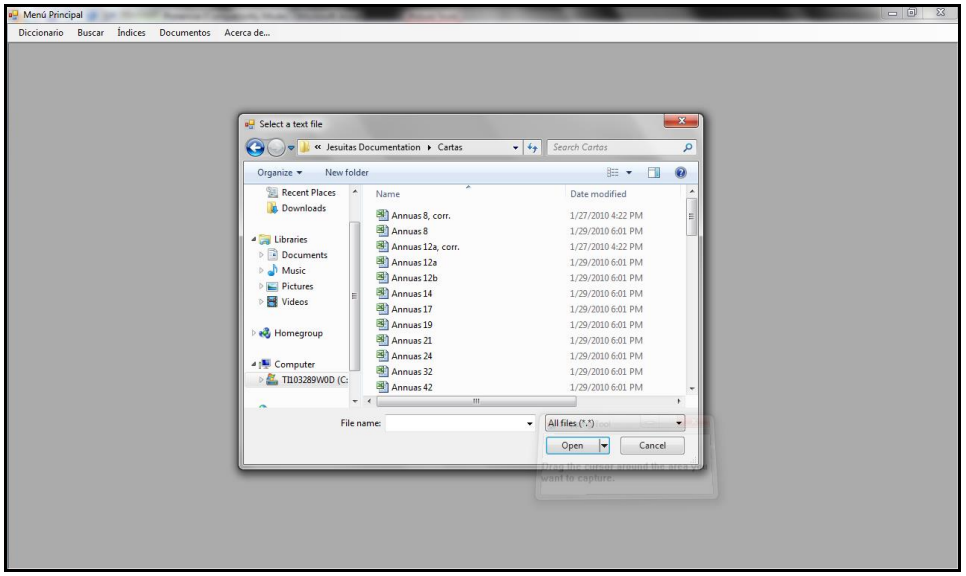

Figura 8 - Documentos

Al elegir la carta, se abrirá un documento similar a este, como se puede observar, mantiene el mismo formato de "3 columnas" que los documentos originales.

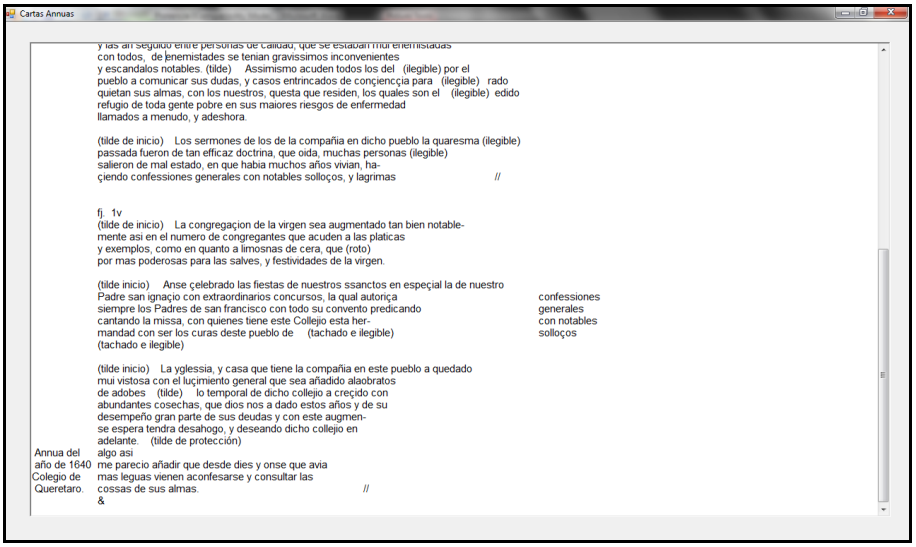

Figura 9 – Ejemplo de carta con 3 columnas

La siguiente figura muestra el mismo documento anterior, pero en el formato XML antes de ser procesado para su despliegue por la interfaz.

Figura 10 – Documento en XML

Conclusiones y trabajo a futuro

En este documento hemos presentado una herramienta informática que se desarrolló para acompañar la obra impresa resultante del proceso de paleografiado de documentos de las Cartas Annuas y Punto de Annua de la Compañía de Jesús que comprenden el periodo entre siglo XV y el siglo XVII.

Esta herramienta será un complemento de la obra impresa y le permitirá a los usuarios e investigadores tener acceso más fácil y rápido a los documentos, así como realizar búsquedas a través de los índices temáticos incluidos o bien utilizando la herramienta para generar consultas con palabras o frases.

La obra deberá ser publicada para finales de 2010 e incluirá la primera versión de esta herramienta en un CD.

En cuanto al trabajo a futuro, la posibilidades son enormes. Nuestra intención es utilizar el corpus de texto de las Cartas Annua (13000 folios) para probar algoritmos de minería de datos que permitan encontrar correlaciones entre términos. También vamos a utilizar este corpus como base de prueba para otro proyecto en el que pretendemos generar ontologías de manera automática utilizando otros algoritmos de minería de datos.

Estamos también en pláticas con el equipo del INAH para portar esta herramienta a otras colecciones de documentos.

Bibliografía

Binstock, C., Peterson, D., Smith, M., et al. (2002). The XML Schema Complete Reference. Boston:Addison-Wesley.

Campbell, Y (1992, agosto). *Aspectos literarios del discurso religioso de Nueva Vizcaya: Cartas Annuas del siglo XVII*, ponencia presentada en el XI Congreso de la Asociación Internacional de Hispanistas, Irvine, CA, EE.UU

Ceballos, F. J. (2007), Enciclopedia *de Microsoft Visual C#*. México: Alfaomega-Ra-Ma.

Deitel, H. M. & Deitel, P. J. (2007) *C#, Cómo programar*, 2ª ed., México: Pearson-Prentice Hall.

Eguíluz Pérez, J., (2008). *Introducción a CSS*. España:Librosweb.

Eguíluz Pérez, J. (2009). *CSS Avanzado*. España:Librosweb.

Larman, C. (2003). UML y Patrones, 2ª ed., Madrid:Pearson/Prentice Hall.

Navarrete M. C. (2006). *La representación jesuítica de los etíopes del siglo XVII desde las Cartas Annuas*. Memoria & Sociedad, 10 (21), 85-106

http://lanic.utexas.edu/project/tavera/italia/jesus.html, Consultado el 10 de junio de 2010

http://www.rae.es/rae.html, Consultado el 10 de junio de 2010.

http://msdn.microsoft.com/en-us/netframework/aa904594.aspx, Consultado el 10 de junio de 2010

http://www.es.masterbase.com/recursos/glosario.asp#w, Consultado el 9 de junio de 2010

http://www.observatoriouned.org/index.php/actualidad/plataformas-lms-y-similares/114-sakai.html?showall=1, consultado el 14 de mayo del 2010

http://moodle-vs-sakai.blogspot.com/ consultado el 14 de mayo del 2010

http://joomdle.com/, consultado el 14 mayo del 2010

http://www.gradekeeper.com/, consultado el 6 de mayo 2010.