Acervos antiguos digitales: hacia una red nacional mexicana

Patricia García Jiménez
José Alfredo Sánchez
Alberto García García
Biblioteca, Universidad de las Américas, Puebla (México)

0.1. Resumen

Este artículo presenta el desarrollo del proyecto "Red Nacional de Acervos Antiguos Digitalizados", parte del programa de Bibliotecas Digitales Universitarias para Todos (UDLA). El proyecto contempla la creación de una biblioteca digital federada de libros antiguos pertenecientes a colecciones mexicanas. UDLA se enfoca al aprovechamiento de la tecnología digital para la preservación, difusión y consulta de este tipo de documentos a través de Internet. Dentro de este contexto, se han diseñado e implementado componentes de software para facilitar la consulta, visualización y navegación de colecciones especiales digitalizadas, así como la recuperación de información a través de búsquedas en texto completo.

Palabras clave: Bibliotecas digitales. Acervos antiguos. Búsquedas en contenido completo. Recuperación de información. Méjico.

0.2. Abstract

This paper presents the development of the project "National Network of Digitized Old Book Repositories". This project is part of the University Digital Libraries for All (UDLA) program. The project focuses on the implementation of a federated digital library of antique books from Mexican collections. In addition, UDLA uses digital technology for preserving, disseminating an accessing this kind of documents over the Internet. In this context, software components have been designed and implemented in order to enable the access, visualization, browsing and content-based retrieval of digital special collections.

Keywords: Digital libraries. Old book repositories. Content-based queries. Information retrieval. México.

1. Introducción

La preservación de acervos bibliográficos históricos ha cobrado importancia mundial a partir de su valoración como elementos clave en el desarrollo de la

humanidad. La belleza de los libros antiguos, aunada a la riqueza de su contenido, los convierte en obras de arte dignas de admirar, conservar y estudiar. En México, diversas bibliotecas que cuentan con libros y archivos antiguos se han sumado a la gran tarea de rescatar, catalogar y difundir tan preciados materiales. Desafortunadamente, el gran valor y en muchas ocasiones el deterioro de estos libros hacen necesario restringir el acceso a los acervos rescatados y minimizar su manipulación física. Con ello, su uso y divulgación se logran solamente de manera muy limitada. Actualmente y gracias al desarrollo de la tecnología, los responsables de los repositorios que custodian colecciones históricas pueden evitar el daño material sobre documentos antiguos mediante proyectos de digitalización, que a la vez proporcionan acceso y difusión universal. A través de la implementación de las llamadas bibliotecas digitales, se crean herramientas que facilitan la consulta de documentos y su aprovechamiento para la investigación y posterior creación de conocimiento.

1.1. Bibliotecas Digitales UDLA-P

El programa de Bibliotecas Digitales Universitarias para Todos (UDLA por sus siglas en inglés) de la Universidad de las Américas, Puebla, (UDLA-P) tuvo sus orígenes en el año de 1999. Desde entonces, esta iniciativa se ha enfocado a la investigación y desarrollo en dos áreas principalmente: la creación de colecciones digitales y el desarrollo de servicios y ambientes para usuarios finales, con el propósito de transformar el concepto de biblioteca al incorporar los avances de las bibliotecas digitales y ponerlos al alcance de grandes comunidades de usuarios (Sánchez y Arias, 2003). En estas áreas, UDLA ha tenido avances significativos, y se dispone de productos finales para el acceso a colecciones de uso institucional y público en general, como, por ejemplo, el acceso a la reserva digital universitaria con más de 900 cursos registrados y unas 100.000 páginas digitalizadas, revisiones y anotaciones sobre documentos digitales, el acervo público de tesis digitales, así como herramientas que facilitan la recuperación de información, disponibles para otras instituciones. Merecen atención aparte aquellos desarrollos que facilitan el acceso a colecciones especiales antiguas pertenecientes a la Universidad de las Américas, Puebla, tal como CIText, un sistema que permite la consulta, visualización y navegación de documentos digitalizados y que forma parte crucial del proyecto aquí descrito.

1.2. Las colecciones especiales UDLA

La sala Porfirio Díaz, ubicada en la biblioteca central de la UDLA, alberga una serie de documentos, archivos, periódicos y libros antiguos que constituyen parte importante de las colecciones especiales de dicha institución, entre los que cabe mencionar la correspondencia presidencial del general Porfirio Díaz, una de las fuentes de información más valiosas para la historia de México. Una parte

importante de esta colección, conformada principalmente por cartas y telegramas, está ya disponible en versión digital.

Recientemente se ha incorporado a la biblioteca digital la Biblioteca Franciscana, como resultado de un convenio entre la Universidad de las Américas, Puebla, y la Provincia Franciscana del Santo Evangelio de México, con el fin de preservar y hacer ampliamente disponible un acervo proveniente de las colecciones de seis conventos franciscanos de México: San Andrés Calpan y San Francisco (Puebla), San Gabriel (Cholula), Coyoacán y Santa Úrsula (ciudad de México) y Orizaba (Veracruz). El acervo se aloja en el Portal de Peregrinos del Convento de San Gabriel, en Cholula, Puebla, el cual ha sido restaurado y acondicionado para este fin.

1.3. Acervos antiguos de la Biblioteca Franciscana

La importancia del acervo de la Biblioteca Franciscana radica en su naturaleza histórica, como fuente documental de la Provincia Franciscana y de la historia novohispana mexicana en general. El acervo comprende más de 24.000 volúmenes de obras impresas desde el siglo XVI hasta el XIX. Dentro de los temas principales que abarca la colección se encuentran, junto con obras generales de consulta, los siguientes: Historia de la Iglesia, Patrística, Teología Moral y Dogmática, Filosofía Cristiana, Hagiografía, meditación y espiritualidad, Historia Natural y de la Ciencia, Literatura Española y Francesa. Actualmente, la Biblioteca Franciscana trabaja en conjunto con la UDLA-P en el proyecto denominado "Red Nacional de Acervos Antiguos Digitalizados", cuyas características se presentan a continuación.

2. Red Nacional de Acervos Antiguos Digitalizados

Este proyecto tiene como objetivo primordial la construcción de un acervo digital que concentrará las obras más importantes de las bibliotecas mexicanas con libros antiguos y los hará disponibles globalmente a través de Internet para su apreciación como objetos de arte y para su uso por estudiosos e investigadores. Los problemas de estos acervos que motivaron fuertemente la realización de este proyecto son su preservación, el acceso restringido, su difusión limitada, los grandes espacios de almacenamiento y la falta de disponibilidad del material en cuanto a tiempo o número de copias.

2.1. Fases del proyecto

La ejecución del proyecto se ha planteado en tres fases. Dentro de su primera etapa, se contemplan como acervos pilotos el perteneciente a la Biblioteca Franciscana y la Biblioteca "José María Lafragua" de la Benemérita Universidad Autónoma de Puebla. Se tiene como meta subsiguiente la inclusión de otras

colecciones del país durante la segunda fase, en la que se aplicarán la experiencia y procesos desarrollados durante la primera etapa. En la fase tercera se establecerán los mecanismos para asegurar la presentación uniforme e integrada de los acervos distribuidos, para generar finalmente una red de acervos antiguos.

2.2. Procesos

Para hacer posible la utilización y recuperación del acervo, fue necesario contemplar una serie de procesos, desde la selección del material a digitalizar, hasta el desarrollo de sendas herramientas que permitieran tanto la creación de la colección digital, como el acceso al material para su consulta y navegación.

La selección se ha basado en los estatutos generados por diversas asociaciones culturales y bibliotecas, que han serivido de guías para seleccionar el material más apropiado para digitalizar. Estos parámetros toman en cuenta el grado de deterioro físico del volumen, su importancia histórica y artística, el uso recurrente del material, el valor físico e intelectual, así como la misión de cada institución. La Biblioteca Franciscana por su parte, ha seleccionado inicialmente cien obras para su digitalización.

Para la digitalización, se establecieron dos criterios a partir de las características y estado del material seleccionado: utilizar un escáner de cama plana para aquellos libros cuyo estado de conservación fuera bueno o, por el contrario, emplear un escáner aéreo para los libros con un deterioro considerable, con la finalidad de minimizar su manipulación física.

Una fase importante de los procesos es la recuperación del texto, lo cual es indispensable para permitir a los usuarios del acervo la búsqueda de volúmenes y páginas específicas. Las páginas digitalizadas, por lo tanto, pasarán por un proceso de reconocimiento óptico de caracteres (OCR, por las siglas de su término inglés *Optical Character Recognition*) para obtener toda la información textual posible, la cual será utilizada por los mecanismos de búsqueda en contenido completo. Las páginas digitalizadas son revisadas y editadas con herramientas especiales para limpiarlas y mejorar su aspecto y calidad visual, dado que en el momento de su captura se transfieren a la imagen manchas o rupturas que afectan la legibilidad del mismo.

Una vez que se ha cubierto cada uno de los pasos anteriormente mencionados, el material es registrado en una base de datos para finalmente hacerlo accesible al público vía Internet.

2.3. Alcance

Al llevarse a cabo el presente proyecto se lograrán avances en múltiples frentes. Las obras de gran valor artístico, literario y científico que se encuentran actualmente en acervos de acceso restringido se pondrán al alcance del público

general de una manera integrada. Se dará difusión a los acervos existentes y se promoverá su importancia como patrimonio cultural universal. Se dará un paso importante en el proceso de garantizar que las fuentes primarias de información sean preservadas y accesibles para generaciones presentes y futuras de estudiantes, profesores e investigadores. Se proporcionarán los medios para recuperar y utilizar los documentos digitales sin arriesgar la integridad de los documentos originales. Se aprovecharán las ventajas del medio digital para facilitar el acceso a la información según el contenido—y no solamente la descripción— de las obras. Los acervos antiguos estarán disponibles permanentemente sin restricciones de número de copias, lugar físico u horario de consulta. Se construirá una biblioteca digital federada en México de los acervos documentales históricos de sus colecciones.

3. CIText

CIText (Consulta a Imágenes Textuales) es un componente de UDLA desarrollado con la finalidad de proporcionar acceso a colecciones especiales antiguas, facilitando a los usuarios su consulta, navegación y visualización a través de Internet. Algunas de las funcionalidades que este sistema permite son la visualización de páginas digitalizadas, el recorrido secuencial, la selección de páginas específicas, las búsquedas en contenido completo a nivel general (acervo) o específico (por libro) y la opción de agregar anotaciones o comentarios.

Para facilitar la integración de CIText con otros componentes de UDLA así como para permitir su uso en diversas plataformas, se utilizó el lenguaje de programación Java. Se empleó el manejador de bases de datos MySQL así como el estándar de conexión a bases de datos JUDBC (Java Universal DataBase Connectivity), librería definida previamente para UDLA. En su construcción se consideraron indispensables dos componentes, uno dirigido al administrador y otro al usuario final, los cuales se definen en las siguientes secciones.

El empleo de tecnología OCR, utilizada para la recuperación del texto completo de cada libro, implicó la introducción de errores de reconocimiento causados por algunas características propias del desgaste que presentaba el material, como manchas y rupturas, así como por el uso de diferentes tipografías. Para la búsqueda de información, tomando en cuenta este tipo de errores, se implementaron tres algoritmos: *soundex*, *similarex* y *clases de caracteres*, los cuales consideran dentro de su funcionamiento este tipo de fallas, permitiendo así recuperar mayor información de la que sería posible sin su utilización. Los tres algoritmos se basan en la codificación de cada palabra en el texto, la cual finalmente es almacenada en una base de datos para su posterior uso en la búsqueda de información. Al igual que el texto, la consulta hecha por el usuario se codifica utilizando estos algoritmos.

El algoritmo *Soundex* consiste en la codificación basada en la similitud fonética de las palabras para reducirlas a una forma canónica común. Se ha utilizado principalmente en aplicaciones de búsquedas de nombres de personas, en sistemas de reservación aérea, censos y en otras que presentan problemas de errores de escritura debidos a la similitud fonética (Knuth, 1975; Myka y Güntzer, 1995). El algoritmo original fue desarrollado para el lenguaje inglés, por lo que ha sido necesario realizar modificaciones de acuerdo al idioma a utilizar, adecuando la agrupación por similitud fonética de las letras en diferentes grupos o clases (Pfeifer et al., 1995). Para este proyecto nuestro grupo produjo una variante del algoritmo en español, debido a que una buena parte del material que se encuentra en la Biblioteca Franciscana está en dicho idioma. Por otra parte, es necesario considerar que muchos libros del conjunto están escritos en español antiguo.

El algoritmo *Similarex* codifica la similitud en apariencia de los caracteres reconocidos, es decir, el hecho de que algunas letras suelen confundirse con otras debido a los trazos comunes que presentan. Se toman en cuenta a su vez series o patrones de caracteres que en grupo pueden aparentar otras letras. En su implementación, se utilizaron las tablas de confusión originales propuestas por los autores (Myka y Güntzer, 1995), dado que éstas no dependen del idioma para su aplicación, haciendo algunas adaptaciones necesarias para este proyecto. Al igual que Similarex, el algoritmo *Clases de caracteres* genera una codificación basándose en la similitud en apariencia de los caracteres; sin embargo, no toma en cuenta series o patrones durante el proceso. La tabla de confusión generada para este proyecto se basó en la propuesta por los autores (Myka y Güntzer, 1995), realizando las adaptaciones necesarias para su aplicación al material utilizado.

El módulo administrador está dirigido al usuario administrador de la colección. Una de las funciones que permite desempeñar es agregar material nuevo —libros completos— al acervo digital, lo que implica tanto los datos de catalogación como las imágenes y textos completos correspondientes a cada una de las páginas del material digitalizado. Los formatos empleados son "JPEG" y "TIFF" para imágenes a color o blanco y negro respectivamente, y archivos de texto sin formato para incluir los resultados de los algoritmos implementados para permitir la búsqueda en texto completo. El almacenamiento se efectúa en directorios diferentes para cada libro, cada uno de los cuales contiene toda la información necesaria (imágenes y textos). Es durante este proceso cuando se generan automáticamente los archivos codificados. El sistema muestra una forma de captura con campos de texto para introducir la información de catalogación necesaria así como para seleccionar los directorios respectivos. Este módulo, a su vez, permite revisar la información existente en la base de datos con la finalidad de comprobar si efectivamente el material se almacenó de forma correcta, así como de validar y liberar aquellos comentarios o anotaciones hechas por los usuarios a los libros.

El módulo de consulta está orientado a los usuarios finales para el acceso, visualización y navegación del acervo digital. El sistema muestra dos formas de realizar el proceso de consulta: mediante la selección de los títulos disponibles a través de una lista que muestra los libros existentes en el acervo (figura 1), o a través de la búsqueda en texto completo en toda la colección de términos específicos de interés para el usuario (figura 2). Una vez ubicado y seleccionado el material, el sistema muestra el libro completo (figura 3) y el usuario podrá recorrerlo de forma secuencial a través de los controles de navegación —ubicados en la parte inferior de la interfaz mostrada en la figura 3— o mediante la selección de páginas específicas a través de las listas de páginas disponibles y páginas preeliminares localizadas a la derecha de la misma interfaz.

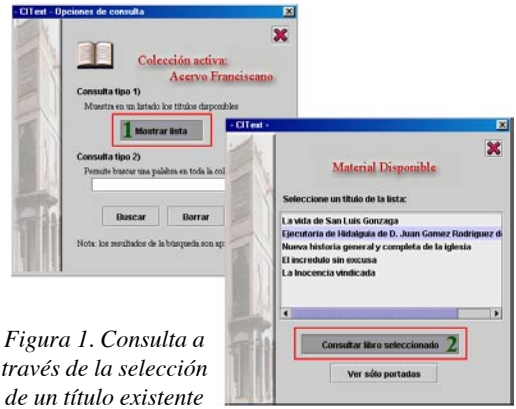

Figura 1. Consulta a través de la selección de un título existente

Figura 2. Interfaz para la selección de tipo de consulta

Además, el usuario puede efectuar búsquedas en el texto completo de ese libro, y desplazarse a aquellas páginas dónde se encuentra el dato buscado simplemente seleccionándolas en la lista de resultados. Los pasos para este proceso se muestran numerados en la figura 4: primero, el usuario debe introducir el término a buscar; en segundo lugar, dar inicio al proceso de búsqueda presionando el botón "buscar"; finalmente, el sistema mostrará una lista con los números de página que contienen la información solicitada.

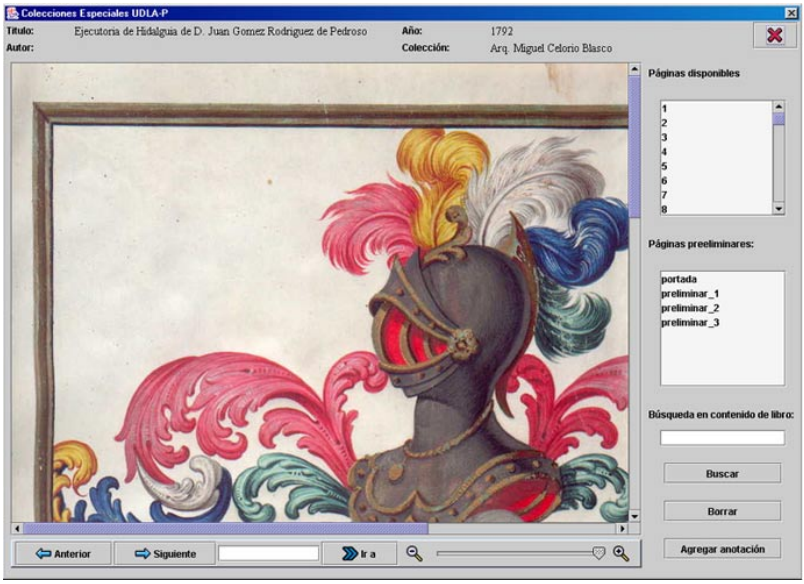

Figura 3. Interfaz para consulta de libros digitalizados

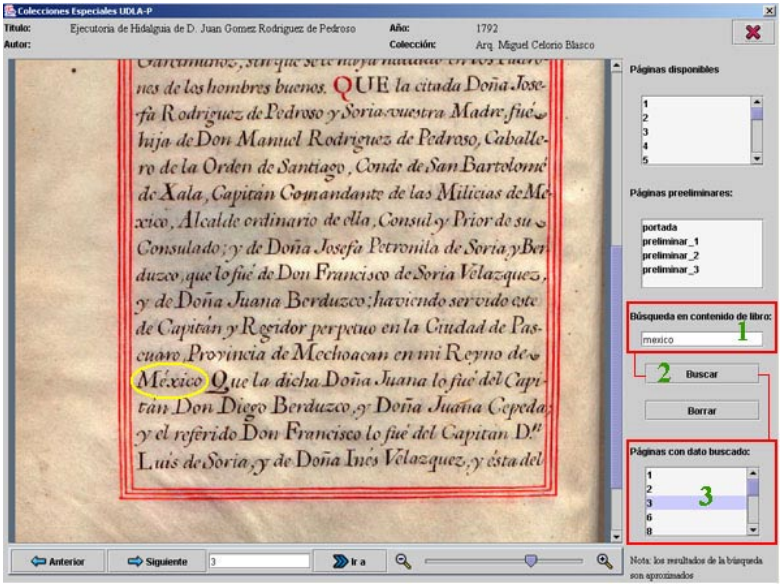

Figura 4. Ejemplo de búsqueda en contenido completo en libro específico

4. Trabajo en curso y propuestas futuras

Actualmente, se ofrecen al público dos colecciones, una de periódicos institucionales y la colección de libros antiguos de la Biblioteca Franciscana, que cuenta por el momento con dos libros completos y algunas páginas de otros libros que han sido seleccionados para realizar pruebas técnicas y mostrar el potencial del proyecto (http://biblio.udlap.mx). En cuanto a los avances de la primera fase de este desarrollo, se han digitalizado hasta el momento 13 libros de los 100 propuestos por la Biblioteca Franciscana, los cuales están en proceso de revisión y edición para mejorar su calidad de visualización y finalmente hacerlos disponibles en Internet a través de CIText.

Como trabajo a futuro, se pretenden incluir algunos otros componentes de UDLA, como Hermes (Maldonado et al., 2003), que permite recuperar información relevante mediante el uso de diferentes modelos de recuperación de información; así como herramientas que permitirán determinar los métodos de segmentación más útiles para el desarrollo de un reconocedor de diversas tipografías, incluyendo letra manuscrita.

Finalmente, para asegurar la consistencia y la calidad en el intercambio de información, se utilizarán protocolos definidos para tal efecto, como los propuestos por la Iniciativa de Archivos Abiertos (OAI, por sus siglas en inglés, http://www.openarchives.org/index.html), organización dedicada a promover la interoperabilidad en Bibliotecas Digitales.

5. Conclusiones

Los acervos antiguos constituyen una gran riqueza cultural y artística que puede y debe estar al alcance de investigadores y público en general para su apreciación y aprovechamiento. Actualmente la problemática que enfrentan este tipo de colecciones en cuanto a su preservación, difusión y consulta pone en riesgo esta posibilidad. A través de la realización de este proyecto, se aprovecharán los avances en bibliotecas digitales para divulgar los acervos antiguos de nuestro país y será posible que múltiples usuarios los consulten simultáneamente desde cualquier parte del mundo.

El surgimiento de las bibliotecas digitales ha permitido vislumbrar soluciones a esta problemática, al facilitar el acceso y la diseminación de información de manera organizada. El medio digital, aunado al desarrollo de nuevas y mejores tecnologías así como de iniciativas que trabajen en el rescate y difusión del libro antiguo, permitirá asegurar el patrimonio cultural de la humanidad en beneficio de presentes y futuras generaciones.

6. Referencias

García, J. P. (2002). Consulta a textos digitalizados: implementación y análisis en el contexto de las colecciones especiales de la UDLA. Universidad de las Américas-Puebla, Cholula, Puebla, México, 2002. Tesis Licenciatura - Ingeniería en Sistemas Computacionales.

Knuth, D. (1975). Sorting and Searching. // The art of computer programming. Reading, MA: Addison Wesley, 1975.

Maldonado-Naude, F.; Sánchez, J. A.; Baeza-Yates, R. (2003). Using Hermes-F: Experiences with a framework for developing information retrieval applications. // Fourth Mexican International Conference on Computer Science (ENC 2003, Sept. 8-12, Tlaxcala, Mexico). Los Alamitos, Calif.: IEEE Computer Society Press, 2003. 101-108. ISBN 0-7695-1915-6.

Myka, A.; Güntzer U. (1995). Fuzzy-full text searches in OCR databases. // Nabil R. Adam, Bharat K. Bhargava, Milton Halem, Yelena Yesha (eds.). Digital Libraries, Research and Technology Advances, ADL '95 Forum, McLean, Virginia, USA, May 15-17, 1995, Selected Papers. Springer, 1996. 130-145. (Lecture Notes in Computer Science; 1082).

Pfeifer, U.; Poersch, T., Fuhr, N. (1995). Searching proper names in databases. // R. Kuhlen; Rittberger, M. (eds.). Proceedings HIM´95, Konstanz: Universitatsverlag, 1995. 259-275.

Sánchez, J. A.; Arias, J. A. (2003). Fourth-phase digital libraries: Pacing, linking, annotating and citing in multimedia collections. Proceedings of the Joint Conference on Digital Libraries (JCDL 2003, Houston, Tex., May).