Software Público para la Digitalización y Divulgación de Acervos Antiguos

Reporte de Avance Técnico Enero-Junio 2005

Proyecto financiado por la Corporación Universitaria para el Desarrollo de Internet (CUDI)
Investigador responsable:
Dr. Alfredo Sánchez H.
Universidad de las Américas–Puebla (UDLA)
Investigador co-responsable:
M.C. María Magdalena González Agramón
Universidad de Sonora (USON)

Resumen

Este proyecto tiene como objetivo principal producir una versión de dominio público de un software para digitalización y acceso global a acervos antiguos. El proyecto se desarrolla como una colaboración entre la Universidad de las Américas, Puebla (UDLA) y la Universidad de Sonora (USON).

El contenido de este reporte semestral se ha organizado en los siguientes apartados que describen los avances del proyecto hasta el momento: (1) instalación, pruebas y estudio del software base, (2) análisis de requerimientos para el nuevo software, (3) manejo de datos, (4) arquitectura y diseño de componentes, (5) avances en colecciones, (6) avances en implementación, (7) documentos elaborados, y (8) actividades de colaboración. Se anexan referencias y directorio de los participantes.

1. Instalación, pruebas y estudio de CIText

El software a producir es una evolución del que se ha desarrollado previamente en la UDLA y que se ha denominado CIText. Para promover el intercambio de ideas para la realización de la nueva versión se ha instalado y promovido el uso de CIText en otras instituciones. La

"Biblioteca Central Universitaria" de la Universidad de Sonora cuenta ya con una versión de evaluación de CIText para poner a disposición del público sus colecciónes. Asimismo, desde 2004 la Biblioteca Lafragua de la Benemérita Universidad Autónoma de Puebla (BUAP) cuenta con una versión de CIText. Los detalles de instalación y problemas de configuración y desempeño están siendo considerados y atendidos para la siguiente versión.

Al mismo tiempo se realizó en la Universidad de las Américas Puebla un estudio de usabilidad sobre el sistema CIText instalado en la Biblioteca de la universidad. De este análisis se han recopilado varias recomendaciones y propuestas de mejoras a la interfaz [Muñoz 2005].

2. Análisis de requerimientos

Con base en las experiencias de uso de CIText y los estudios de usabilidad realizados en la UDLA se están contemplando los siguientes cambios en funcionalidad:

- Permitir imprimir y almacenar las imágenes del libro
- Mostrar opciones avanzadas de visualización con acercamientos, cambios en la imagen, rotación
- Reestructurar la vista del "árbol" de cada libro
- Agregar búsquedas por elementos específicos del libro y por términos
- Señalar las palabras encontradas resultantes de una búsqueda
- Mostrar la información bibliográfica completa de cada obra.
- Mostrar la temática del libro.
- Mostrar anotaciones o comentarios del material realizadas por otros usuarios
- Señalar inconsistencias del almacenamiento del material (páginas faltantes o sin contenido)
- Mostrar ayuda del sistema
- Personalizar la interfaz (colores, lenguaje)
- Reconocimiento de caracteres especializado para libros de fondo antiguo junto con un sistema de recuperación de información

3. Manejo de datos

Con base en una propuesta inicial de la UDLA y con retroalimentación de los colaboradores en USON y el Centro para el Estudio de Bibliotecas

Digitales de la Universidad de Texas A&M, se definieron los siguientes elementos para el manejo de datos en la nueva versión de CIText:

- Descripción de los libros mediante el estándar METS de metadatos con características propias de libros de fondo antiguo
- Administración de colecciones mediante un sistema de bases de datos XML nativo
- Un servidor de metadatos de acuerdo al estándar usando el protocolo definido por *Open Archives* (OAI-PMH)
- Una interfaz de consulta de imágenes de la colección basada en el concepto de "servicios web"

4. Arquitectura y diseño componentes

A partir de los requerimientos se diseñó una arquitectura de componentes como se ilustra en la Figura 1. Cada componente se describe brevemente a continuación.

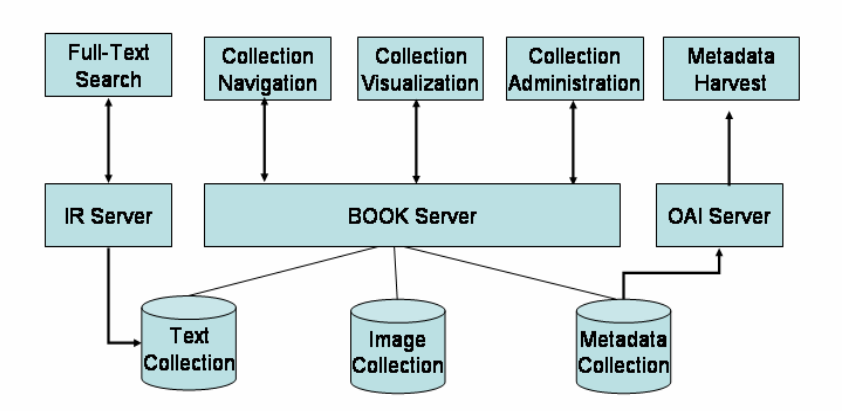

Figura 1. Componentes del software para manejo de acervos antiguos

- *Collection Administration*
  Este componente se encarga de la adquisición de imágenes, reconocimiento de texto y estructura de archivos para la generación de metadatos. Asimismo, del registro de la estructura de metadatos que describen los libros.

- *Book Server*

Parte central del sistema recupera los libros almacenados de acuerdo a su estructura para los servicios de administración, navegación y visualización

- *Information Retrieval Server*
  Motor de búsqueda sobre la colección que permita encontrar información sobre el texto reconocido en las imágenes digitalizadas.

- *Interoperability Server* (OAI)
  Servidor que permite compartir los metadatos de la colección mediante el protocolo OAI-PMH definido por la iniciativa de *Open Archives*.

- *Collection Navigation Service*
  Servicio de navegación sobre la colección y la estructura de los libros. Incluye un servicio de búsqueda que búsqueda sobre metadatos y contenido

- *Collection Visualization Service*
  Servicio de visualización de los libros con opciones de estructura y manipulación de las imágenes

5. Avances en colecciones

UDLA

Adicionalmente del equipo requerido para la captura a color; cámara digital y mesa de copiado con que ya se cuenta, se adquirió un nuevo escáner aéreo *Bookeye A2* con financiamiento de la Agencia para el Desarrollo Internacional de los Estados Unidos (USAID).

En el caso de digitalización se está trabajando con los libros que se encontraban en formato blanco y negro para obtener su versión a color. Estos 12 libros equivalen a 37,486 páginas. Adicionalmente se cuenta con nuevas adquisiciones que se incorporarán a la versión de CIText en línea. Esta colección servirá de prueba para los prototipos del nuevo software

USON

En el mes de febrero se acondicionó un área de digitalización en la Biblioteca Central Universitaria de esta institución, que incluye una mesa de copiado, cámara digital y escáner. Para apoyar esta tarea se cuenta con dos computadoras y software especializado para procesamiento de imágenes y catalogación. A cargo del área de digitalización se encuentra un supervisor, dos asistentes y un programador.

Para seleccionar el material a digitalizar se conformó un grupo de especialistas para determinar la prioridad de digitalización del acervo. En este grupo se encuentran bibliotecarios, investigadores y personal del departamento de Historia y de Letras y Lingüística de esta institución. Los principales criterios tomados en consideración para la selección son: relevancia del material, grado de uso, grado de deterioro y año de edición, entre otros. Actualmente se han digitalizado 5 libros del fondo antiguo para un total de 2,815 páginas. Cada uno de los libros cuenta con su catalogación y una versión PDF disponible en la página de la biblioteca.

Para fin de año se contempla la digitalización de un total de 15 títulos (8,815) páginas. De igual manera se pretende contar con dichos libros disponibles desde la interfaz de CIText-USON y posteriormente a través del nuevo software.

6. Avances en implementación

Al momento se han revisado las colecciones de ambas instituciones y se ha utilizado CIText para almacenarlas. Paralelamente se han realizado análisis y pruebas sobre base de datos XML y el estándar XMLDB. Para promover la interoperabilidad se han modelado colecciones de documentos en XML y distribuido a través del estándar de *Open Archives* OAI-PMH. En el caso del almacenamiento y modelado de libros se han realizado pruebas de modelado de acuerdo al estándar de metadatos METS. En el caso del procesamiento de imágenes se está analizando la automatización de estas tareas. Los avances en el reconocimiento de caracteres (OCR) se han diseñado estrategias que permitan mejorar el trabajo realizado en el proyecto de Antique-OCR . Se han desarrollado prototipos de interfaces basados en estudios de usabilidad de la interfaz de CIText. Las interfaces se han modelado para clientes con mínimos requerimientos hasta interfaces más complejas utilizando Java Applets o el estándar de vectores para Web SVG. El análisis y pruebas de otros proyectos relacionados como la iniciativa de la Biblioteca de la Universidad de Virginia y la Universidad de Cornell denominada

FEDORA que modela documentos usando el estándar METS y el proyecto Navimages del Ministerio de Cultura de Francia para la organización procesamiento y navegación de imágenes crean un contexto para el diseño e implementación de este proyecto.

7. Documentos elaborados

En el transcurso de este proyecto se han realizado los trabajos que a continuación se presentan.

Tesis

Córdova J.M. 2005 Generalización en la implementación del protocolo de OAI en colecciones digitales. Tesis de Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales. Escuela de Ingeniería, Universidad de las Américas Puebla. Mayo.

López O. 2005 Filtro binario para segmentación de matrices. Tesis de Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales. Escuela de Ingeniería, Universidad de las Américas Puebla. Mayo.

Tesis en desarrollo

Muñoz J.I. 2005. Interfaz para el proyecto CIText de Recuperación de Información en Acervos Antiguos. Tesis de Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales. Escuela de Ingeniería, Universidad de las Américas Puebla. Diciembre.

Reportes

Mendoza A. 2005. Primer informe del proyecto "Software público para la digitalización y divulgación de acervos antiguos". Reporte Técnico. Biblioteca Central Universitaria. Universidad de Sonora. Mayo.

Muñoz J.I. 2005. Estudio de usabilidad de CIText. Reporte Técnico CUDI05-03 Laboratorio de Tecnologías Interactivas y Cooperativas. Universidad de las Américas - Puebla. San Andrés Cholula, Puebla. Mayo.

Razo A. 2005. Estudio de modelado en METS. Reporte Técnico Laboratorio CUDI05-02 Laboratorio de Tecnologías Interactivas y Cooperativas ICT. Universidad de las Américas - Puebla. San Andrés Cholula, Puebla. Mayo.

8. Actividades de colaboración

Para facilitar la colaboración entre la Universidad de las Américas, Puebla y la Universidad de Sonora se definió un sitio de web que permite compartir la información del estado del proyecto, realizar pruebas de prototipos y estructurar la información generada. La página se encuentra disponible en la siguiente dirección:
http://ict.udlap.mx/projects/cudi/udlasonora

El foro de la reunión de primavera de CUDI sirvió de contexto para revisar avances y determinar actividades a realizar el segundo semestre en la cual están planeadas visitas a las respectivas instituciones para pruebas e instalación de CIText y el nuevo software.

Referencias

Corporación Universitaria para el desarrollo de Internet A.C. (CUDI)
http://www.cudi.edu.mx

Universidad de las Américas, Puebla
http://www.udlap.mx

Laboratorio de Tecnologías Interactivas y Cooperativas ICT
http://ict.udlap.mx

Universidad de Sonora
http://www.uson.mx

Página del proyecto
http://ict.udlap.mx/projects/cudi/udlasonora/

Metadata Enconding and Transmission Standard
http://www.loc.gov/standards/mets

Open Archives Initiative
http://www.openarchives.org

Proyecto FEDORA
http://www.fedora.info

Proyecto Navimages
http://sdx.archivesdefrance.culture.gouv.fr/gpl/navimages

Participantes

Directores del Proyecto

Dr. J. Alfredo Sánchez Huitrón
Coordinador del Programa de Bibliotecas Digitales
Profesor Titular de Ingeniería en Sistemas Computacionales

Correo electrónico: *alfredo@mail.udlap.mx*
Página personal: *http://ict.udlap.mx/people/alfredo*
Teléfono: +52(222)229-2666
Fax: +52(222)229-2431

Universidad de las Américas Puebla.
Sta. Catarina Mártir s/n, San Andrés Cholula.
Puebla, México CP 72820 *http://www.udlap.mx*

M.C. María Magdalena González Agramón
Directora de Desarrollo Académico

Correo electrónico: *magali@guaymas.uson.mx*
Teléfono: +52 (662) 2 59 21 51

Universidad de Sonora
Blvd. Luis Encinas y Rosales s/n
Col. Centro, Hermosillo
Sonora, México CP 83000 *http://www.uson.mx*

UDLAP

Investigadores

M.C. Antonio Felipe Razo Rodríguez
Ing. Lourdes Fernández Ramírez

Investigador asistente

Ing. Omar López Rincón

Estudiantes de Licenciatura

Juan Ignacio Muñoz Campos

USON

Investigadores

Lic. Alfonso Mendoza
Lic. René Molina F.