Universidad Nacional de La Plata (UNLP)
La Plata - Argentina

Facultad de Informática

Representación de los Recursos dentro de una Biblioteca Digital

Propuesta Técnica al Doctorado en Ciencias Informática

Resolución 7ma. - 24/06/2010. Expediente: 3300-2042/10-000

Autor:
Jose Daniel Texier
dantexier@gmail.com

Directora:
Profa. Marisa De Giusti

6 de junio de 2011

1. Tema Principal (Objetivo General)

Representación de los recursos dentro de una biblioteca digital.

2. Introducción

El desarrollo de Internet y de las nuevas tecnologías de la información nos hace pensar inminentemente en el cambio de época que estamos viviendo hoy en día en la denominada, "Era de la Información". Estas transformaciones son una novedad para nuestra sociedad, la última ocurrió hace aproximadamente más de 200 años cuando la Revolución Industrial condujo a la humanidad del agrarianismo al industrialismo. La génesis del cambio actual se puede observar en lo cultural, social, económico y tecnológico. En estos años se observa que los niños y jóvenes crecen con un computador en sus manos y que los adultos no, y estos adultos han vivido el antes y el después de la aparición de la Internet. Por ello hoy se puede decir que quien posee la información tiene el poder. En comScore[1] se muestran algunas estadísticas importantes:

- Un 63.0 % de los usuarios de Internet son menores de 40 años.
- Un 15.5 % de la población mundial tiene acceso a Internet, en Argentina está en 48.3 % y en Venezuela 31.1 %.
- Los sitios más visitados a nivel mundial en estricto orden son: Google, Facebook y Youtube.
- Un 12.6 % del comercio electrónico que se realiza es el de libros y revistas, convirtiéndose éstos en los de mayor consumo.
- Sólo el 6.2 % de la consultas en Internet son académicas y/o científicas.

Estas estadísticas nos comprueban el gran auge de la Internet. Sin embargo, las bibliotecas tradicionales ofrecen los documentos en soportes físicos que se facilitan a los usuarios a través de servicios de préstamos y consultas. Esta tradicional política esta alejando a la población de las bibliotecas porque las consultan que realizan de los ejemplares que a veces no se pueden encontrar y muchas veces si se encuentran no están actualizados porque existen ediciones viejas. Por ello, han comenzado a surgir desde hace más de una década el diseño y la creación de las bibliotecas digitales[2], es decir, se inició la automatización de las bibliotecas tradicionales convirtiéndolas en bases de datos para ser procesadas por sistemas o herramientas informáticas, cuyo propósito está dirigido a la población en general.

Estamos evolucionando con respecto a la búsqueda y transmisión de la información, puesto que se están sufriendo contracciones económicas, así como también concepciones

de vida distintos en el mundo permitiendo generar políticas de conservación del ambiente, soluciones más económicas, un mejor acceso a la información, etc., lo que conlleva dejar atrás el paradigma de buscar un libro y solicitarlo en la biblioteca. Una posibilidad que concita interés en un grupo importante de personas es el acceso a contenidos (libros, artículos, etc) a través de internet, con las comodidades que esto implica, a la vez que este tipo de acceso posibilita una mejor preselección del material.

La tecnología ha permitido que el conocimiento se traslade a los soportes digitales y eso ha permitido la consolidación de las bibliotecas digitales. Para cumplir con su rol en la sociedad una BD debe:

1. Tener una colección de datos.
2. Poseer un sistema de ordenamiento de documentos.
3. Brindar servicios a una comunidad de usuarios.

El primer elemento consiste en tener una colección de documentos que si están o no en formato electrónico es otra cosa; segundo se debe contar con un sistema de ordenamiento para los documentos, además de un sistema para almacenar y recuperar importantes cantidades de información. La colección de documentos y el sistema de ordenamiento pueden estar en forma electrónica pero eso no garantiza la existencia de una biblioteca digital. Tercero se necesita una comunidad de usuarios a quienes servir, que tengan unas necesidades y características determinadas. Si se tienen estos tres elementos se reconoce una biblioteca digital.

Esta propuesta pretende dar un aporte científico y académico original para la estandarización y fortalecimiento de las bibliotecas digitales para la sociedad, y aunque existen numerosos proyectos relacionados con el tema, este estudio se centrará en el almacenamiento y recuperación de recursos digitales en repositorios y en la indexación de la información.

La propuesta se desarrollará bajo la filosofía del Open Access[3] porque refleja un adelanto en el área de las bibliotecas digitales e intenta abrir el acceso a recursos académicos. La importancia radica en la necesidad de: (a) establecer y mantener estándares que permitan un crecimiento de ellas y (b) evitar redundancia de trabajo e información lo que representaría un mejor servicio.

Gracias a iniciativas como Open Archives Initiative (OAI)[4], Online Computer Library Center (OCLC), International Federation of Library Associations (IFLA), Networked Digital Library of Theses and Dissertations (NDLTD) se han logrado establecer:

- Estándares de interoperabilidad.
- Estándares para la catalogación de recursos.
- Estándares para la preservación de los objetos digitales.

Este trabajo comenzará con la investigación en el Servicio de Difusión de la Creación Intelectual (SeDiCI)[5], creado en el 2003, de la Universidad Nacional de la Plata (UNLP) junto con los cursos del Doctorado que se irán tomando para mi formación académica. En la actualidad, SeDiCI tiene como objetivo principal socializar el conocimiento generado en las diferentes áreas académicas de la UNLP con el fin devolver a la comunidad sus esfuerzos mediante la creación de un repositorio institucional que consiste en una base de datos compuesta por un grupo de servicios destinados a capturar, almacenar, ordenar, preservar y redistribuir la documentación en formato digital que brinda un acceso abierto a la documentación académica ofrecida.

SeDiCI está realizando investigaciones y desarrollando herramientas adecuadas en las siguientes áreas: lenguajes naturales, compresión e indexación de archivos, catalogación de documentos, almacenamiento y recuperación de recursos, repositorios semánticos, gestión de herramientas multidiomas, portales Web, derechos de autor, etc.; lo cual demuestra que están trabajando constantemente para seguir creciendo con la tecnología y en acuerdo con la cantidad de información proveniente de los diferentes repositorios externos (aproximadamente 45 repositorios para abril del 2010) registrados internacionalmente y la producción académica de la UNLP.

El tema principal de este estudio estará relacionado directamente con los repositorios y las siguientes áreas: base de datos, modelado de datos, metadatos, almacenamiento e indexación de los metadatos, XML y datos semiestructurados. La propuesta mejorará la calidad del funcionamiento sintáctico y estructural del repositorio de SeDiCI lo que redundará en beneficio del usuario a la hora de realizar búsquedas de recursos y por tanto mejores servicios.

3. Áreas

Bases de Datos (BD) convencionales y no convencionales; BD deductivas, BD Activas, BD Objetuales y Objeto-Relacionales, BD Móviles, BD temporales y espaciales; BD documentales y multimedia; Recuperación de información, Indexación y BD en Web; Modelado de datos; XML y Datos Semiestructurados; Web Semántica y Ontologías.

4. Objetivos Específicos

1. Analizar el problema de la representación de recursos dentro de una biblioteca digital.

2. Analizar las ventajas y desventajas de los distintos paradigmas de bases de datos en cuanto a la representación de los recursos.

3. Proponer un modelo de datos flexible para representar los recursos dentro de una biblioteca digital, considerando al menos las siguientes características:
    - distintos formatos de metadatos para distintos tipos de recursos;
    - varias catalogaciones de un mismo recurso utilizando distintos formatos de metadatos;
    - representación de entidades abstractas de forma independiente, permitiendo identificarlas y reutilizarlas;
    - relaciones entre recursos;
    - relaciones entre entidades abstractas;
    - relaciones entre entidades abstractas y recursos.
4. Proponer una arquitectura en capas simple y clara que provea los niveles de abstracción adecuados según los niveles de acceso requeridos.
5. Analizar las técnicas de preservación digital y determinar cuál es la más adecuada para garantizar la preservación tanto de los recursos como de los objetos digitales.
6. Determinar la forma de indexación más adecuada para que la recuperación de los recursos sea eficiente y de buen redimiento. El espacio ocupado por el repositorio debe mantenerse dentro de parámetros aceptables.
7. Desarrollar un prototipo para la representación de los recursos dentro de una biblioteca digital en SeDiCI

5. Metodología de Investigación

La presente investigacion es exploratoria y empírica[6], dado que el tema escogido ha sido poco estudiado hasta el momento, por ello se realizará una descripción sistemática del problema para llegar al estado del arte del tema tratado.

Esta investigación permitirá determinar los requerimientos necesarios para hacer el análisis y diseño respectivo, el cual conducira a la implementación del prototipo. De esta manera se estará dando aporte original al tema.

Para la construcción del prototipo, se estudiaran las metodologías del Proceso Unificado de Desarrollo de Software[7], del Desarrollo de Software Dirigido por Modelos[8] y del Modelo para el Diseño de Hipermedia Orientado a Objetos (OOHDM)[9].

Una vez seleccionada la metodología de desarrollo de software, se procederá a la construcción de cada una de las etapas para luego tener los resultados y conclusiones con base en los objetivos específicos propuestos y las pruebas realizadas.

6. Desarrollo

El trabajo que se realizará consiste en el desarrollo de un prototipo que permita la representación de los recursos dentro de una biblioteca digital en SeDiCI basado en el cumplimiento de las siguientes premisas:

- Lograr un modelo flexible y expandible de representación de los recursos de una Biblioteca Digital.
- Realizar pruebas para determinar el modo óptimo de almacenamiento e indexación de recursos.

El sistema tendrá un gran impacto en la implementación de búsqueda y recuperación de información de la UNLP, otras universidades y los miembros de consorcios a los que pertenece la UNLP.

7. Actividades por Objetivos Específicos

7.1. Analizar el problema de la representación de recursos dentro de una biblioteca digital

- Definir el concepto de recursos dentro de una biblioteca digital.
- Estudiar cada una de las características en la representación de recursos.
- Clasificar en categorías los problemas en la representación de recursos existentes en SeDiCI.
- Describir los problemas categorizados en la representación de recursos existentes en SeDiCI.

7.2. Analizar las ventajas y desventajas de los distintos paradigmas de bases de datos en cuanto a la representación de los recursos

- Estudiar cada uno de los paradigmas de bases de datos existentes.
- Realizar una matriz comparativa sobre la base de las ventajas y desventajas de los paradigmas relacionados con la representación de los recursos.
- Seleccionar el paradigma más idóneo para el dominio tratado.

7.3. Proponer un modelo de datos flexible para representar los recursos dentro de una biblioteca digital

- Diseñar el modelo de datos para representar los recursos mediante la descripción de las partes que lo compondrán.

7.4. Proponer una arquitectura en capas simple y clara que provea los niveles de abstracción adecuados según los niveles de acceso requeridos

- Estudiar las diferentes opciones de arquitectura de capas existentes.
- Diseñar un modelo de arquitectura de capas.

7.5. Analizar las técnicas de preservación digital y determinar cuál es la más adecuada para garantizar la preservación tanto de los recursos como de los objetos digitales

- Estudiar las alternativas y técnicas de almacenamiento para los objetos digitales (PDF, videos, fotos, etc.)
- Exponer las razones de garantia de preservación entre los objetos digitales y los recursos que los representan.

7.6. Determinar la forma de indexación más adecuada para que la recuperación de los recursos sea eficiente y de buen redimiento

- Estudiar las diferentes formas de indexación.
- Definir los parametros para una correcta recuperación de los recursos.
- Seleccionar la forma de indexación adecuada para la propuesta.

7.7. Desarrollar un prototipo para la representación de los recursos dentro de una biblioteca digital en SeDiCI

- Seleccionar la metodología para la construccion del prototipo.
- Diseñar el prototipo.

- Implementar el prototipo.
- Realizar pruebas con el prototipo.

8. Temas de Investigación

En la investigación, que se estará llevando a cabo para el desarrollo de la tesis doctoral, se tienen los siguientes temas:

1. Base de Datos (BD):
    - BD convencionales y no convencionales.
    - BD deductivas.
    - BD Activas.
    - BD Objetuales y Objeto-Relacionales.
    - BD Móviles.
    - BD temporales y espaciales.
    - BD documentales y multimedia.
    - Modelado de datos.
    - NoSQL.
2. Bibliotecas Digitales:
    - Metadatos.
    - Repositorios.
    - Indexación.
    - Catalogación.
    - Recuperación de información.
    - Open Access.
    - Metadatos.
    - XML y datos semiestructurados.
    - Sistemas de información federados.
    - Indexación y BD en Web.
    - Arquitectura Orientada a Servicios y/o Web Services.
    - Web semántica y ontologías.

9. Posibilidades de Realización en el Ámbito del Tesista

El inicio del Doctorado está planificado, tentativamente, para el 06 de septiembre del 2010. La propuesta contempla una primer etapa de cursada y obtención de los 45 créditos en dos años. Luego (en los dos años siguientes) desarrollar y culminar la tesis. Para ello, contaré con mi formación como Ingeniero en Informática en la Universidad Nacional Experimental del Táchira (UNET) y con una Maestría en Computación de la Universidad de los Andes (ULA), en la cual trabajé en el área de Lenguajes de Programación y Compiladores. Asimismo, he trabajado como docente en las áreas de Lenguajes de Programación de Java y C/C++ en cursos de pregrado.

En esta etapa estaré trabajando en conjunto con las personas que sostienen el SeDiCI lo que facilitará la realización de investigación y desarrollo en el área de interés y que esto a posteriori se vuelque para el mejoramiento de otros sistemas similares.

10. Antecedentes de la Dirección

La Ing. De Giusti propuso la creación del Servicio de Difusión de la Creación Intelectual, SeDiCI a la UNLP en el año 2002, trabajó durante un año con la Comisión de Interpretación y reglamento para la configuración del proyecto y actualmente dirige este servicio. Participa en la organización de un repositorio a nivel nacional con el Ministerio de Ciencia, Tecnología e Innovación Productiva de la Argentina. Las bibliotecas digitales son el centro de sus temas de investigación, particularmente en lo relativo a la puesta en marcha, selección de contenidos y áreas prioritarias. Cuenta con numerosas publicaciones vinculadas a la temática, los cuales pueden verse en: http://www.sedici.unlp.edu.ar/difusion/difusion.php

La Dra. Silvia Gordillo tiene una gran experiencia en bases de datos que constituyen un área central a esta propuesta, particularmente en bases de datos relacionales y orientadas a objeto. La Profesora también cuenta con antecedentes profusos en las áreas de recuperación de información y minería de datos. Cuenta con numerosas publicaciones en estas áreas y en tópicos marcados como centrales en esta propuesta.

11. Actividades de Formación Propuestas

Los cursos que se realizarán según la oferta presentada por el postgrado y la propuesta de formación planteada por la Directora y Codirectora son:

1. Año 2010:
    - Curso: Aspectos Legales de Gobierno Electrónico, Facultad de Informática UNLP.

- Curso: Monitorización y Optimización de Rendimiento en Sistemas de Cómputo de Altas Prestaciones, Facultad de Informática UNLP.
- Curso: Metodología de la Investigación científica, Facultad de Informática UNLP.
- Curso: Computación Móvil: Arquitectura y Aplicaciones, Facultad de Informática UNLP.
- Curso: Introducción al Reconocimiento automático de Patrones, Facultad de Informática UNLP.

2. Año 2011:

   - Curso: Seminario / Taller de Elaboración de Proyectos y Tesis Técnicas y Herramientas, Facultad de Informática UNLP.
   - Curso: Gestión de Calidad según normas ISO. Aplicaciones a Software, Facultad de Informática UNLP.
   - Curso: Base de Datos, Facultad de Informática UNLP.
   - Curso: Técnicas de Computación para la Web y Escalabilidad, Facultad de Informática UNLP.
   - Curso: Ingeniería del Conocimiento, Facultad de Informática UNLP.
   - Curso: Administración de Proyectos de Software, Facultad de Informática UNLP.
   - Curso: Tópicos de Ingeniería, Facultad de Informática UNLP.

12. Planificación

La investigación se desarrollará, principalmente, en tres momentos, despues de finalizados los cursos de formación:

12.1. MOMENTO 1: Estado del Arte

La revisión teórica y de las investigaciones que fundamenten el tema planteado en la propuesta se hará a partir de la elaboración de un apartado de cada uno sobre la base de:

- El problema de representación de recursos en la biblioteca digital.
- Diferentes paradigmas de bases de datos en la biblioteca digital.
- Técnicas de preservación digital.

Tiempo: 8 meses (desde diciembre 2011 hasta julio 2012) Producto: una publicación en una revista especializada.

12.2. MOMENTO 2: Desarrollo de Modelos

Los modelos y propuestas de los subproblemas del tema se desarrollarán a través de:

- Un modelo de datos flexible para representar los recursos dentro de una biblioteca digital.
- Una arquitectura en capas simple y clara para el manejo de la información digital de los recursos.
- Implementación de la técnica de preservación digital más adecuada para garantizar la preservación.
- Forma de indexación más adecuada para que la recuperación de los recursos sea eficiente.

Para cada uno de las propuestas anteriores se desarrollará un software usando las técnicas más apropiadas.

Tiempo: 4 meses (desde agosto 2012 hasta noviembre 2012) Producto: una publicación en una revista especializada.

12.3. MOMENTO 3: Prototipo Final

Finalmente, se procede a demostrar a tráves de un desarrollo tecnológico único, la solucion del problema principal presentado, esto permitirá dar un aporte original al estado del arte en el tema específico seleccionado. Por ello se desarrollará un prototipo para la representación de los recursos dentro de una biblioteca digital en SeDiCI.

Tiempo: 6 meses (desde diciembre 2012 hasta mayo 2013). Producto: una publicación, como mínimo, en una revista especializada.

13. Revistas - Journals

- D-Lib Magazine. http://www.dlib.org/.
- ALA Tech Source.
- El Profesional de la Información (EPI) y ThinkEPI.
- Internet Computing Online - IEEE.

- International Journal on Digital Libraries.
- Information Retrieval. ISSN: 1386-4564 (print version) ISSN: 1573-7659 (electronic version). Journal no. 10791. IMPACT FACTOR: 1.841 (2009) Journal Citation Reports, Thomson Reuters
- Journal of Computer Science and Technology (JCS&T).
- Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología (TE&ET).

14. Centros de Investigación

- Public Knowledge Project (PKP). http://pkp.sfu.ca/research.
- DSpace@MIT. http://libraries.mit.edu/dspace-mit/info/dspace-help.html.
- OAICat. http://www.oclc.org/research/activities/oaicat/default.htm.
- OAIHarvester2. http://www.oclc.org/research/activities/past/orprojects/-harvester2/harvester2.htm.
- SRW/U. http://www.oclc.org/research/activities/srw/default.htm.

Referencias

[1] "comScore.com." `http://www.comscore.com`, Abril 2010.

[2] X. Agenjo and F. Hernández, "Tendencias internacionales en el desarrollo funcional de la recuperación de la información: Linked Open Data (LOD)," 2010.

[3] "Open access initiative." `http://www.soros.org/openaccess`, Abril 2011.

[4] "Open Archives Iniciative." `http://www.openarchives.org`, Abril 2011.

[5] "SeDiCI - Servicio de Difusión de Creación Intelectual." `http://sedici.unlp.edu.ar`.

[6] C. Sabino, "Como hacer una tesis," 1998.

[7] I. Jacobson, G. Booch, and J. Rumbaugh, *El proceso unificado de desarrollo de software*. Addison Wesley, 2000.

[8] C. Pons, R. Giandini, and G. Pérez, "Desarrollo de software dirigido por modelos. conceptos teóricos y su aplicación práctica," *Editorial: EDUNLP and McGraw-Hill Education*, 2010.

[9] D. Schwabe, G. Rossi, and S. Barbosa, "Systematic hypermedia application design with oohdm," in *Proceedings of the the seventh ACM conference on Hypertext*, pp. 116–128, ACM, 1996.