RISI 7(2), 2010 (69-75)
Revista de Investigación de Sistemas e Informática
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos

ISSN 1815-0268 (versión impresa)
ISSN 1816-3823 (versión electrónica)

Propuesta de desarrollo de un repositorio digital de documentos de investigación para la FISI utilizando software libre

Development proposal of a repository of digital research papers for the FISI using free software

Nora La Serna Palomino, Augusto Cortez Vásquez, Fernando Gómez Jaime

Universidad Nacional Mayor de San Marcos
Facultad de Ingeniería de Sistemas e Informática

nlasernap@unmsm.edu.pe, cortez_augusto@yahoo.fr

RESUMEN

El trabajo que se presenta en este artículo se desarrolla en el marco de los repositorios virtuales para contenidos digitales. En particular, se presenta una propuesta de requerimiento y diseño de implementación de un repositorio digital para el Instituto de Investigación de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos. El repositorio permitirá organizar, mantener y consultar documentos relacionados a la producción de la investigación en el Instituto. Se ha hecho un estudio y selección de la herramienta libre que permitirá la construcción del repositorio, esta técnica es el Lucene. En este trabajo, además se presentan los requerimientos y un diseño de la aplicación.

Palabras clave: repositorios virtuales, D-Space, Lucene, protocolo OAI-PMH, software libre

ABSTRACT

The work that is presented in this article develops within the framework of the virtual repositories for digital content. In particular, presented a proposal for injunction and design of implementation of a repository digital for the Research Institute of the Faculty of Engineering Systems and Informatics at the National University of San Marcos. The repository will organize, maintain and documents related to the production of research at the Institute. It has become a study and selection of the free tool that will allow construction of the dump, this technique is the Lucene. In this work, it was also presented the requirements and an application design.

Keywords: Virtual repositories, D-Space, Lucene protocol ADMINISTRATORS-STATE, free software

Virtual repo
ADMINI

1. INTRODUCCIÓN

Un repositorio digital es un depósito o archivo en un sitio web centralizado, en donde se almacena y mantiene información digital, en bases de datos o archivos informáticos. Los archivos pueden estar en su servidor o referenciar desde su web al alojamiento originario. Los repositorios generalmente son de carácter académico e institucional y tienen por objetivo organizar, archivar, preservar y difundir la producción intelectual de la organización [1]. Algunas de las herramientas libres más utilizadas que permiten la implementación de éstos repositorios son: el DSPACE, E-print, LUCENE, Protocolo OAI-PMH, etc.

Los recursos mencionados permiten a organizaciones, tanto públicas como privadas, construir sus repositorios digitales, de tal manera que estas pueden organizar su información y administrarla. La información, que básicamente son documentos como informes, proyectos, artículos, etc. pueden ser vistos y actualizados por los miembros de la organización si tienen permisos para hacerlo. Otros usuarios de la web, que no son miembros de la organización, pueden acceder a documentos que se han registrado para ser compartidos. Por ejemplo, artículos de investigación elaborados por miembros de la organización.

Son cientos de organizaciones públicas y privadas a nivel mundial que utilizan estos recursos informáticos con la finalidad de organizar los documentos que manejan, en particular las instituciones académicas. En el Perú, hay varias instituciones que cuentan con repositorios digitales, la Universidad Nacional Mayor de San Marcos, cuenta con uno en la Biblioteca Central; sin embargo, no se conoce de alguna Facultad de la universidad que maneje estos recursos. Se presenta esta propuesta para la Facultad de Ingeniería de Sistemas e Informática, en concreto para el Instituto de Investigación, que es el de construir un repositorio de documentos utilizando alguna de las herramientas libres mecionadas en el primer párrafo.

Los objetivos principales que se describen en este trabajo son: a) realizar un estudio de las herramientas libres para la construcción de un repositorio digital; y b) presentar una propuesta de desarrollo de un repositorio de documentos de investigación que se producen en el Instituto de Investigación de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos.

La estructura del presente artículo es la siguiente: En la sección 2 se bosqueja el marco organizacional y tecnológico que corresponde a los repositorios digitales; en la secciones 3 se presentan los requerimientos y diseño de la aplicación para la implementación de un repositorio digital; mientras que la sección 4 corresponde a las Conclusiones del trabajo realizado y se proponen tareas futuras para su implementación; y finalmente en 6 se presentan las referencias bibliográficas y bibliografía utilizada.

2. ORGANIZACIÓN Y TECNOLOGÍA UTILIZADA EN REPOSITORIOS DIGITALES

A partir del año 1990 se dieron varios movimientos de intelectuales, en torno al acceso libre al conocimiento, a partir de sus encuentros se forma el Open Access community, comunidad que defiende y promueve el acceso gratuito y sin barreras al conocimiento científico. Estos movimientos permitieron entre otras actividades, definir los estándares para la creación de los repositorios digitales.

En JISC, 2005, se define a los repositorios digitales institucionales como “Sistema en red de hardware y software, que proporciona servicios referidos a una colección de objetos digitales. Estos pueden ser recuperados, compartidos, exportados con diferentes propósitos y contextos”.

OpenDoar es un directorio de repositorios académicos de acceso abierto. Además, proporciona información estadística de estos repositorios y permite la búsqueda de repositorios o sus contenidos [13]. En la Figura 1 se puede observar uno de los reportes que emite OpenDoar, en el que se muestra el porcentaje de los tipos de producción intelectual que almacenan las organizaciones en sus repositorios.

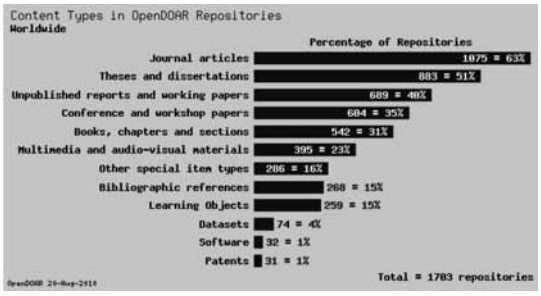

Figura N.° 1. Repositorios por herramientas utilizadas. Fuente: (OpenDOAR, 2010)

Algunas de las tecnologías y herramientas libres más utilizadas que permiten la implementación de éstos repositorios son: el DSPACE, E-prints, LUCENE, Protocolo OAI-PMH, etc. Algunos de estos se presentan brevemente en las siguientes subsecciones.

2.1. DSpace

Dspace es un repositorio digital que captura, guarda, indexa y permite la consulta de la producción intelectual de los grupos y centros de investigación de Universidades. Creado por el Instituto Tecnológico de Masachuset y Hewlett-Packard, en la actualidad es una herramienta libre disponible para instituciones de investigación a nivel mundial. Su uso se ha extendido también a instituciones privadas [1,2]. En la actualidad tienen registrado repositorios de alrededor de 963 empresas públicas y privadas a nivel mundial, la mayor parte de ellas instituciones académicas.

2.2. E-prints

Eprints es un software libre que facilita la creación de repositorios virtuales, creado por la universidad de Southampton. Creado con la finalidad de crear un repositorio institucional de edición electrónica para la investigación académica, pero puede ser usado para otros propósitos [3]. Está diseñado con el objetivo de ser fácil, rápido de instalación y gratuito. Eprints se distribuye bajo la licencia GNU, lo cual significa que el código fuente es accesible y modificable por cualquier programador, con la condición que las modificaciones se hagan también accesibles públicamente. Eprints puede funcionar en cualquier ordenador con sistema operativo Linux.

2.3. OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting)

OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting) es un protocolo para la transmisión de contenidos en internet, creado por investigadores a nivel mundial a partir de su primera reunión en Octubre de 1999 en Nuevo México, USA. Cuya finalidad es desarrollar y promover estándares de interoperabilidad para facilitar la difusión eficiente de contenidos en Internet [5].

Su arquitectura es basada en el modelo cliente-servidor. Los clientes son los archivos que proporcionan la información (proveedores de datos), y los servidores son los recolectores o servicios que toman los datos, con el objetivo de incorporar algún valor añadido y presentarlos a los usuarios finales (proveedores de servicios). En las Figura 3 se presentan la arquitectura de OAI-PMH, en el que se destacan sus funciones como proveedores de servicios y datos.

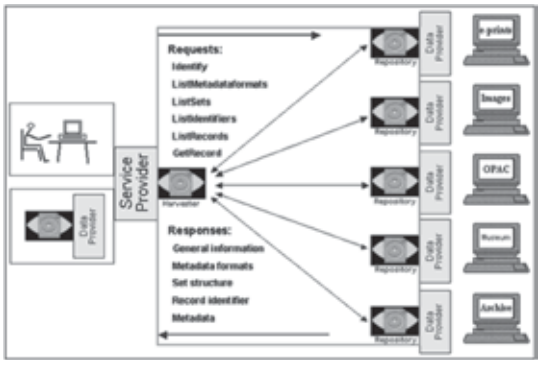

Figura N.° 2. Arquitectura de OAI-PMH [5].

2.4. Lucene

Lucene es un software que permite crear buscadores de contenidos, básicamente permite indexación y busqueda de documentos. Utilizado por numerosos proyectos, y es software libre respaldado por la fundación Apache [4].

Define un modelo de clases compacto y de fácil comprensión, permitiendo que una implementación inicial completa de búsqueda e indización se puede realizar con muy pocas líneas de código y pocas instancias de objetos de Lucene. Es una librería que permite incorporar capacidades de indexación y búsqueda a las aplicaciones. La Figura 3 muestra la Arquitectura de Lucene, en donde se muestran los módulos de indexación y búsqueda de Lucene, y su interfaz con el ambiente de la aplicación.

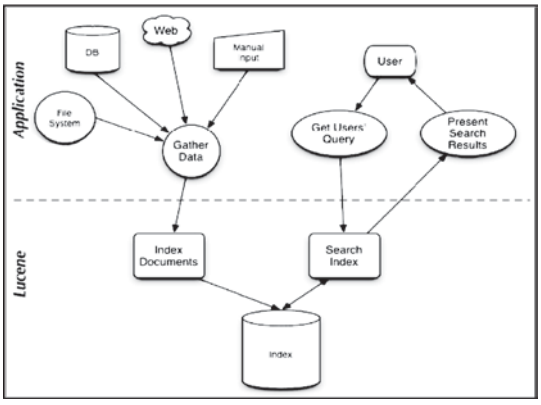

Figura N.° 3. Arquitectura de Lucene [4].

3. REQUERIMIENTOS Y DISEÑO DE LA APLICACIÓN

En esta sección se presenta y diseña una propuesta de aplicación para la construcción de un repositorio digital que utilice alguna de las herramientas libres para su implementación.

3.1. Área de aplicación y requerimientos

El instituto de Investigación de la Facultad de Ingeniería de Sistemas e Informática de la UNMSM, recibe y registra documentos impresos y digitales, de sus investigadores como:

- Propuestas de proyectos de investigación,
- Informes técnicos de proyectos de investigación, estos proyectos pueden ser con financiamiento o sin financiamiento por parte del Vicerrectorado de investigación.
- Asimismo, el área produce revistas de investigación al menos dos revistas al año, los cuales contienen artículos elaborados por los investigadores.
- Registra Proyectos de tesis de pregrado y postgrado que se van a sustentar
- Entre otros documentos de investigación, como informes técnicos de grupos de investigación, etc.

Contar con un repositorio digital de estos documentos de investigación que maneja el Instituto de Investigación permitiría un control, manejo y documentación eficiente de los mismos. Cada año se generan en el Instituto decenas de estos documentos. Permitiendo además compartir esta información con otros investigadores nacionales e internacionales. De esta manera, el Instituto, así como cientos de organizaciones, utilizarían este tipo de tecnologías de la información.

Los requerimientos funcionales para la construcción del repositorio serían los siguientes:

- Registro y actualización de los documentos de investigación: propuestas, informes técnicos, proyectos de tesis, pre y postgrado, artículos de revistas, etc.
- Consultas por tipos de líneas de investigación, autores, fechas, etc.
- Información estadística de la producción científica de la Facultad.
- Administración de usuarios que registran documentos.
- Los requerimientos No funcionales básicos para la construcción del repositorio serían:
- Utilizar DSpace para la construcción del repositorio
- Lenguaje de programación Java para el desarrollo de las interfaces
- Sistema operativo Windows para las interfaces de usuario
- Se requiere de un computador de escritorio para el desarrollo de las aplicaciones.

3.2. Funcionalidades de la aplicación del repositorio

En este punto se presentan los Casos de uso del Sistema prioritarios, que representan las funcionalidades que el sistema debe realizar, y se presentan con sus respectivos diagramas, los que se muestran en la Figura 4:

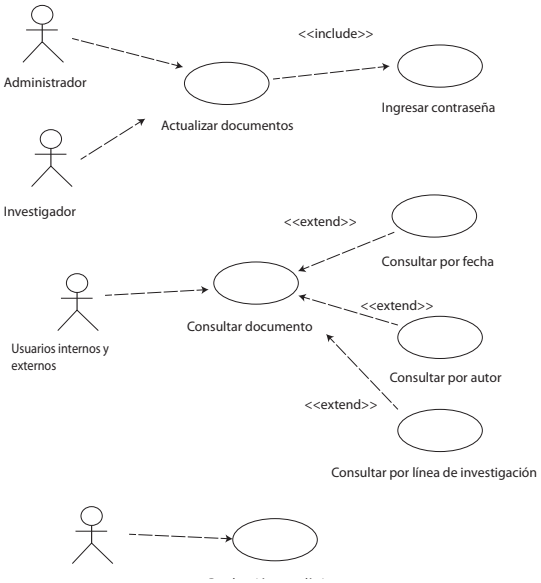

Figura N.° 4. Diagrama de casos de uso del repositorio propuesto [6].

Actualizar documentos permite registrar, actualizar y eliminar documentos del repositorio. El usuario es un administrador del sistema o un investigador.

Consultar documentos, los usuarios de esta funcionalidad pueden ser del Instituto o investigadores interesados en la información registrada, de otros centros nacionales o internacionales.

Emitir de información estadística, los usuarios de este caso de uso serían los directores y asesores del Instituto interesados en la producción científica de la Facultad.

3.3. Selección de la tecnología utilizada

Debido a una evaluación realizada de las tecnologías y herramientas presentadas en la sección 2 de este artículo; así como basados en la experiencia realizada en el curso de Taller de Proyectos, en donde se realizaron dos implementaciones de prototipos de repositorios de proyectos de tesis de pregrado de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos. Se puede concluir que las herramientas Lucene y DSpace presentan las características más adecuadas y viables para ser seleccionadas en el presente trabajo.

De las dos herramientas preseleccionadas podemos argumentar que Dspace es una de las más utilizadas en la construcción de repositorios; sin embargo, no es fácil su implementación, y es rígida en cuanto al diseño de interfaces, limitando algunas funcionalidades que una aplicación de repositorio podría querer considerar.

De esta manera, la herramienta que estamos seleccionando es el Lucene, debido a la facilidad de implementación, y al volumen de documentos que maneja el presente proyecto, así como un diseño flexible de las interfaces de la aplicación del repositorio que se plantea implementar.

3.4. Arquitectura del diseño

En la Figura 5 se muestra un diseño de Arquitectura del repositorio propuesto. Se observan los procesos que corresponden al área de aplicación y al área de Lucene. En la parte izquierda de la imagen se observan el conjunto de documentos digitales que después del proceso de registro se van a almacenar en el repositorio digital. En la parte derecha de la imagen, un usuario del sistema realiza una consulta (pregunta por un documento), y el sistema le devuelve un resultado de la búsqueda.

3.5. Diseño de interfaces

En la Figura 6 se muestra un diseño estándar de interface para repositorios digitales, en el ejemplo corresponde al repositorio de la Universidad de Alcalá, modelo que se podría adoptar para el repositorio de II-FISI. Los aspectos principales que se pueden destacar, en el recuadro de la izquierda, la etiqueta “Navegar”, aquí se presenta un menú para realizar búsquedas por alguna de las opciones: comunidades

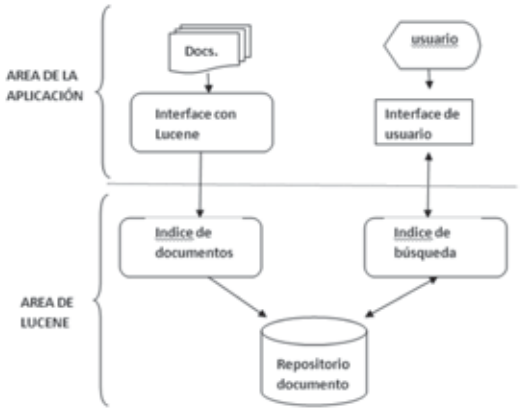

Figura N.° 5. Arquitectura de repositorio propuesto [6].

(áreas de la institución), fechas, autor, títulos o materias. En la parte central de la imagen, primero se presenta al repositorio, hacia el medio se presenta la opción “Buscar”, a partir de textos que se ingresan, y luego se exhiben los enlaces de las comunidades definidas en el repositorio.

4. EVALUACIÓN Y CONCLUSIONES DEL TRABAJO DESARROLLADO

El trabajo que se presenta en este artículo se desarrolla en el marco de los Repositorios virtuales para contenidos digitales. En particular, se presenta una propuesta de requerimiento y diseño de implementación de un repositorio digital para el Instituto de Investigación de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos. El repositorio permitirá organizar, mantener y consultar documentos relacionados a la producción de la investigación en el Instituto. Se ha hecho un estudio y selección de la herramienta libre que permitirá la construcción del repositorio, esta técnica es el Luecene. En este trabajo, además se presentan los requerimientos y un diseño de la aplicación.

Se ha cumplido en un 100% de los objetivos propuestos en este Trabajo, los cuales son:

1. Estudio detallado de las herramientas libres más utilizadas.

2. Elaborar los requerimientos de un área usuaria de la universidad para la parte de aplicación.

3. Desarrollar el análisis y diseño para el repositorio di-

Figura N.° 6. Diseño de interface de la aplicación propuesta.

gital de documentos de investigación de la II-FISI.

Si bien se propone el análisis y diseño del repositorio para documentos de investigación de la FISI UNMSM. Resta por realizar la implementación. Dos de las herramientas libres que son candidatas para ser seleccionadas para la construcción del repositorio serían DSpace y Lucene.

Asimismo la implementación del repositorio podría replicarse en otros Institutos de Investigación de la Universidad.

A continuación se presentan los resultados obtenidos en la realización del presente trabajo. Como trabajo futuro se propone la implementación, es decir la codificación y pruebas, del repositorio digital.

Los productos obtenidos como resultado de la investigación son:

1. 100 % de los objetivos propuestos del proyecto.

2. Análisis y diseño de la Propuesta de desarrollo de un repositorio digital para documentos de investigación de la FISI UNMSM.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] dspace.mit.edu Repositorio institucional, consultado el 05-11-2010.

[2] www.dspace.org Página oficial Dspace, consultado el 05-11-2010.

[3] www.eprints.org Página oficial de Eprints, consultado el 05-03-2010.

[4] Lucene.apache.org Página oficial Lucene, consultado el 05-03-2010.

[5] www.openarchives.org/OAI/ The Open Archives Initiative Protocol for Metadata Harvestin. Consultado el 05-03-2010.

[6] Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval. Maryland: Addison-Wesley-Longman Publishing co, 1999.

[7] Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. Computer Netsorks and ISDN Systems, 30, 1998. p. 107-117

[8] Chu, H. and Rosenthal, M. “Search engines for the WWW: A comparative study and evaluation methodology” En http://www.asis.org/annual-96/ElectronicProceedings/chu.html

[9] Delgado Domínguez “Mecanismos de recuperación de Información en la www”, Universidad de Islas Baliares, España. 1998. http://dmi.uib.es/people/adelaida/tice/modul6/memfin.pdf

[10] Frakes W.B. y Baeza Yates R. “Information Retrieval: data structures and algorithms”. Prentice Hall 1998.

[11] La Serna P. N. y grupo, Diseño del Sistema de Recuperación de Información para la biblioteca FISI. Vol 2. Revista RISI 2005.

[12] Manning, C . Prabhakar R., and Hinrich S. Introduction to Information Retrival. Cambridge University Press, 2008.

[13] The Directory of Open Access Repositories – OpenDOAR. www.opendoar.org.