Experiencia en la utilización y desarrollo de software libre en la Implementación de Gestión de Jornadas, Biblioteca Digital y la aplicación de Wrappers.

Luis Alejandro Vargas[1], Alberto Laender[2], German Montejano[3],
Nilda Perez Otero[4] Miguel Arturo Bernechea[5], Angel Valdez [6]

[1, 4, 5, 6] Departamento de Sistemas – Facultad de Ingeniería – Universidad Nacional de Jujuy
[2]Departamento de Ciencia da Computacão - Universidade Federal de Minas Gerais - Brasil
[3]Facultad de Ciencias Físico-Matemáticas y Naturales – Universidad Nacional de San Luis

[1]avargas@arnetbiz.com.ar, [2]laender@dcc.ufmg.br, [3]gmonte@unsl.edu.ar
[4]nilperez@fi.unju.edu.ar, [5]macber10@hotmail.com, [6]betovar282@hotmail.com

Resumen

El presente trabajo resume la experiencia de desarrollo e implantación de Software en entorno Libre, que consiste en el envío, recepción y evaluación de paper's, que docentes y/o investigadores elevan a las Jornadas de Ciencia y Tecnología de las Ingenierías del NOA. Dichas jornadas se desarrollan anualmente en una de las universidades del noroeste argentino, que conforman el CODINOA[1]. Una vez finalizadas las Jornadas, los artículos ACEPTADOS formar parte de la Biblioteca Digital, software con características de repositorio, desarrollado para las Jornadas, que permite el acceso libre e irrestricto a la producción científica llevada a cabo por los docentes y/o investigadores de las universidades del NOA.

Debido a la poca información del repositorio, pero no a su calidad, implementamos varios wrappers, para la extracción de datos de otras Librerias Digitales, y así poder disponer de toda esa información en el instante que los usuarios la precisan. Concentrar la información proveniente de diferentes fuentes de datos, obviamente, relativas a una misma área de interés, beneficiaría en gran medida la búsqueda de información que el usuario está interesado y permitiría estar posicionado en un mismo portal de búsqueda.

1 .Consejo de Decanos de Ingeniería del NOA. www.codinoa.edu.ar.

Palabras Claves

Software Libre, Recuperación de Información. Librería Digital. Jornada de Ciencia y Tecnología del las Ingenierías del Noa. Wrappers. Repositorio Digital.

Contexto

El presente trabajo se implementó en la Facultad de Ingeniería de la Universidad Nacional de Jujuy (UNJu), para la organización de las VI Jornadas de Ciencia y Tecnología de Facultades de Ingeniería del NOA, que se llevaron a cabo durante los dias 4 y 5 de octubre del 2010 en la facultad de Ingeniería. El conjunto de base de datos de los artículos evaluados, y ACEPTADOS pasaron a conformar un repositorio de artículos de las Jornadas.

Las VI Jornadas de Ciencia y Tecnología de las Facultades de Ingeniería del NOA son promovidas por el CODINOA y en la edición VI del año 2010, fueron organizadas por la Facultad de Ingeniería de la Universidad Nacional de Jujuy, cuyo propósito es difundir las actividades científico tecnológico que se desarrollan en las Facultades de Ingeniería de las Universidades Nacionales del NOA. En la presente edición, 2011, la organización de las Jornadas es llevada a cabo por la Universidad Nacional de Catamarca. (http://jctnoa.unca.edu.ar/)

Objetivos Generales que persigue la organización de las Jornadas

Facilitar la vinculación entre profesionales que están abordando problemáticas regionales y con otros actores del desarrollo regional.

Integrar los docentes investigadores del NOA con el propósito de trabajar en conjunto en temáticas similares para generar programas estratégicos de resolución de problemas regionales comunes.

Objetivos Particulares:

Establecer cuáles son las principales líneas de investigación en cada área temática.

Detectar el grado de vinculación e interacción que cada investigador / grupo de trabajo tiene con otros grupos.

Proponer modos de implementación de políticas que viabilicen nuevas y/o mayores interacciones.

Desde el año 2005, se vienen desarrollando dichas Jornadas, en diferentes universidades del NOA. Ver Gráfico 1. La primera edición se llevó a cabo en la Facultad de Ingeniería de la Universidad Nacional de Jujuy, y en cada edición, las jornadas tienden a mejorar en todo aspecto, con mayor repercusión y compromiso en la comunidad universitaria de las universidades correspondientes al Noroeste argentino, y extendiéndose a otras universidades allegadas a la región. Ver Cuadro 1.

Las Universidades que integran el CODINOA son:

Universidad	Página WEB	Edición y Año	Trabajos
Universidad Nacional de Jujuy	www.unju.edu.ar	I, - 2005 VI - 2010	88 136
Universidad Nacional de Catamarca	www.unca.edu.ar	II - 2006 VII - 2011	119 -
Universidad Nacional de Tucuman	www.unt.edu.ar	III - 2007	141
Universidad de Santiago del Estero	www.unse.edu.ar	IV - 2008	154
Universidad Nacional de Salta	www.unsa.edu.ar	V - 2009	166

Cuadro 1

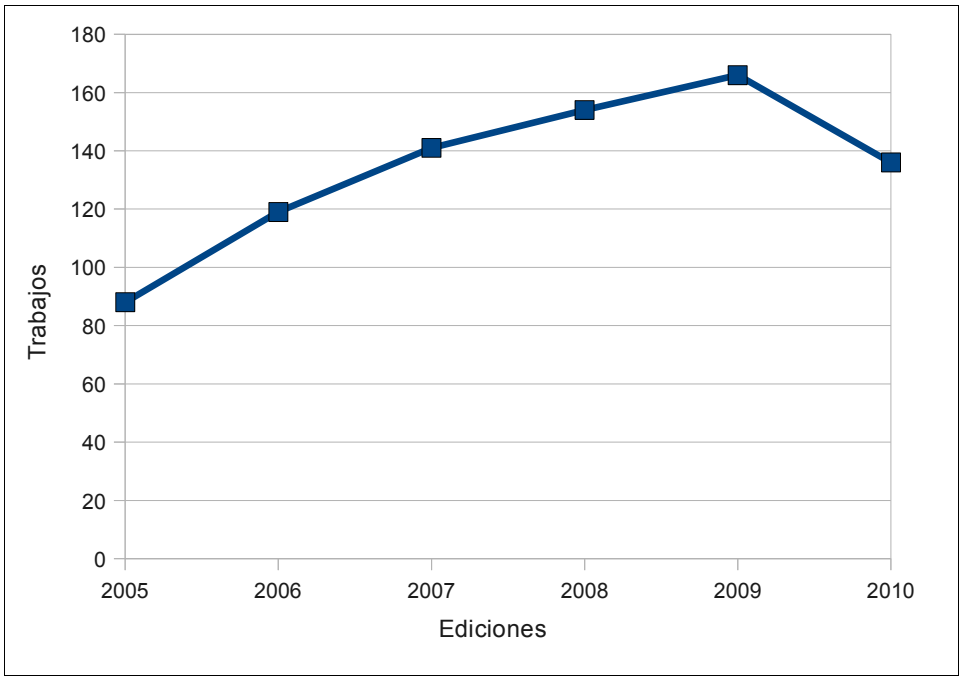

Gráfico 1 Cantidad de Trabajos

Plataforma de Trabajo

El trabajo cumple con el diseño, desarrollo, e implementación bajo una plataforma web que reune las siguientes características:

Lenguaje de Programación PHP y javascript

Motor de Base de Datos MySQL

Servidor web APACHE, siendo el más utilizado, cubre cerca del 44% de los dominios web. [8]

La contratación de un HOSTING, de Buenos Aires, que cumplía con las exigencias y características, para albergar el sitio de las jornadas, donde se instaló las correspondientes Bases de Datos, y la implementación del sistema de las Jornadas (www.jctnoajujuy.com.ar).

Motivos de elección de plataforma WEB y Software Libre.

Distribución geográfica de los potenciales usuarios del sistema, distribuidos en otras universidades de provincias pertenecientes a la región NOA de Argentina.

En cada edición de las Jornadas, se utilizaban herramientas aisladas, planillas de cálculos, o pequeños sistemas, que no se utilizadas en la siguientes ediciones.

No había una continuidad de los trabajos que se realizaban.

Los organizadores, autores de trabajos, y evaluadores, tienen acceso a los trabajos desde cualquier computadora con acceso a INTERNET.

Los organizadores de cada edición de las Jornadas, puedan instalar y adaptar el Sistema a sus nuevos requerimientos.

Acceso a los artículos de las jornadas, de todas las ediciones, por parte del público en general desde cualquier lugar con acceso a INTERNET.

Los recursos físicos para instalar el software son mínimos y están disponibles en la mayoría de los proveedores de hosting.

Eliminar Software propietario.

Tendencia de mudar hacia la web aplicaciones y sistemas.

Reducción de costos, en la adquisición de software.

No existía una Biblioteca Digital de Trabajos Aceptados de las Jornadas, para que investigadores, docentes, alumnos y público en general accedieron a esa fuente de información.

Diseño del Sistema a Implementar

La propuesta de desarrollar e implementar el Sistemas con Software Libre, va a permitir en primer instancia adquirir herramientas, Gestores de Base de Datos y otros software sin costos. Software y herramientas necesarias para el desarrollo y funcionamiento del Sistema.

Que el nuevo producto permitiría la distribución, modificación, e incorporación de mejoras y módulos por parte de los organizadores de las próximas ediciones de las Jornadas.

Describimos a los Organizadores de la Jornadas las características de catalogar el sistema como Software Libre con su correspondiente licencia GNU GPL[9], siendo un ejemplo típico de trabajo colaborativo, donde grupos distribuidos y dispersos geográficamente coordinan esfuerzos. Su principal ventaja es el acceso y uso del **código fuente del SISTEMA,** que van a obtener los organizadores de las siguientes Jornadas, les va a permitir:

1. Libertad de instalar y utilizar el programa, no solo en la edición 2010, sino en las próximas ediciones de las organizaciones de las jornadas.
2. libertad de estudiar el funcionamiento del programa y modificarlo, adaptándolo a sus nuevas necesidades y requerimientos. Acelerando los tiempos de desarrollo del sistema, para nuevos módulos
3. Libertad de distribuir copias del programa, incorporando las mejoras, para otras ediciones de las Jornadas.
4. Libertad de incorporar mejoras en el programa y hacer públicas dichas mejoras a los demás, de tal manera que toda la comunidad se beneficie.

El desarrollo del Sistemas consistía en el diseño e implementación de dos grandes módulos Fig. 1

a) Gestión de Trabajos

b) Gestión de Biblioteca Digital.

a) Sistema de Gestión de Jornadas.

Inconvenientes y problemas detectados:

A nivel de Organización.

Recepción de los paper's de los trabajos que se realizaban via mail. Revisión diaria de mail, inclusive posterior a la fecha de cierre de presentación de Trabajos.

Impresión de los trabajos y envío de la documentación a los EVALUADORES. Via correo, y en algunas oportunidades se tuvo que llevar la documentación a las universidades de origen de los evaluadores.

Recepción via mail, y por correo de los resultados de las evaluaciones por parte de los evaluadores.

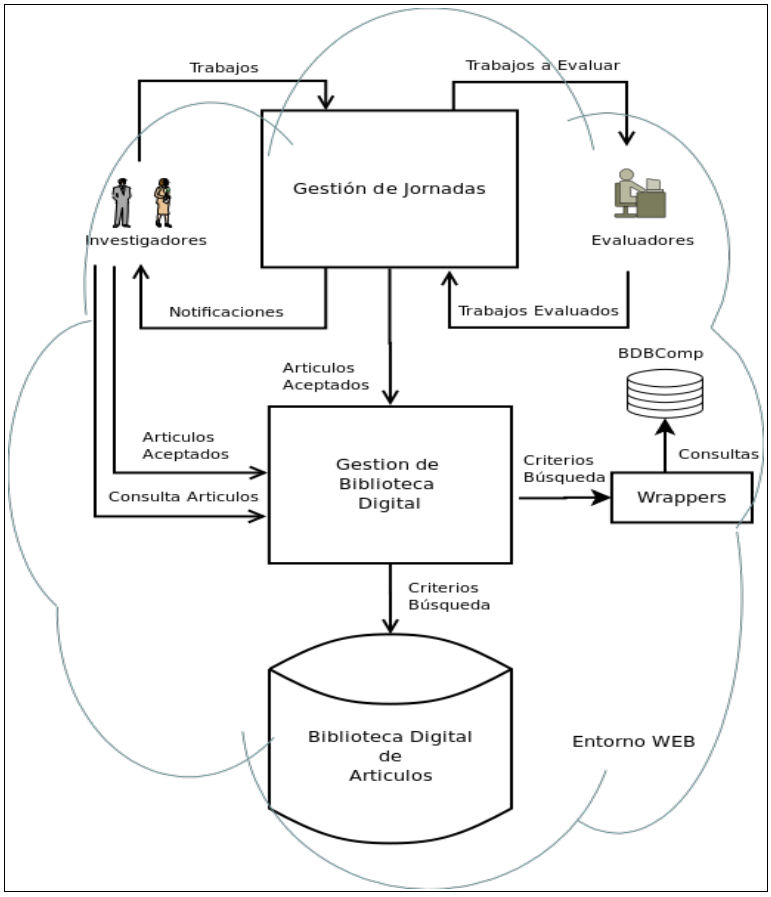

Fig. 1 Sistema propuesto

Notificación via mail de los trabajos ACEPTADOS, NO ACEPTADOS y de los trabajos que se deben realizar CORRECCIÓN. Evaluar un trabajo en calidad de NO ACEPTADO, llevaba nuevamente a asignar y enviar el trabajo a un nuevo EVALUADOR, dicho evaluador notificaba via mail el resultado del trabajo, .y según su calificación, si era necesario enviar a un TERCER EVALUADOR, y realizar nuevamente el circuito, o se notificaba al investigador que su trabajo NO FUE ACEPTADO. Posteriormente llegaban via correo las planillas o grillas de evaluación.

Verificar la inscripción de los expositores de los TRABAJOS que solo tienen calificación de ACEPTADOS.

Organización y clasificación de los trabajos, para preveer los recursos necesarios

para que los autores de los trabajos puedan realizar sus correspondientes exposiciones.

Listados para los certificados, por áreas temáticas, por autores de Trabajos, por expositores, asistentes a la Jornada, y otros reportes necesarios para llevar a cabo una mejor organización de las Jornadas.

La comunicación con los Autores de Trabajos y/o Evaluadores se llevaba a cabo via mail, por la cual se tenia que llevar un registro detallado de la recepción y envio de mails a cada autor y/o evaluador.

El medio de control era en forma manual, utilizando planillas electrónicas.

La presentaciones que no se realizaban en tiempo y en forma, tanto de los trabajos, como de las evaluaciones conducía a:

- demoras en la organización de las Jornadas,
- informes incorrectos,
- reimpresión de trabajos.
- envio del trabajo a otro evaluador.
- pedido de datos incompletos a los autores del trabajo.
- imposibilidad de efectuar una estadística.
- dedicación de recursos humanos
- comunicación via telefónica y via email con los autores o evaluadores.
- incorrecta asignación de recursos físicos.

Nivel de Evaluadores.

Esperar la información de los trabajos para iniciar la evaluación, via mail o por correo.

Enviar resultados de las evaluaciones, via mail.

Enviar la documentación de las evaluaciones via correo.

Nivel de autor de Trabajo:

Hasta último momento no precisaban con la información si su correspondiente trabajo había sido ACEPTADO o NO.

Demoras en recibir contestación por parte de la organización, ya que la comunicación se realizaba via mail.

No podian acceder a la grilla de evaluación.

La actualización de datos, se tornaba dificultoso.

Solicitud de reimpresión de Certificados, debido a datos incorrectos.

No podían realizar un seguimiento de los trabajos presentados.

La confirmación de la recepción del trabajo a los autores no era en el momento.

El control de la inscripción de un autor en calidad de expositor, en las Jornadas, se presentaba dificultosa, en la mayoría de las veces el pago se debía realizar en el lugar del evento o por transferencia bancaria.

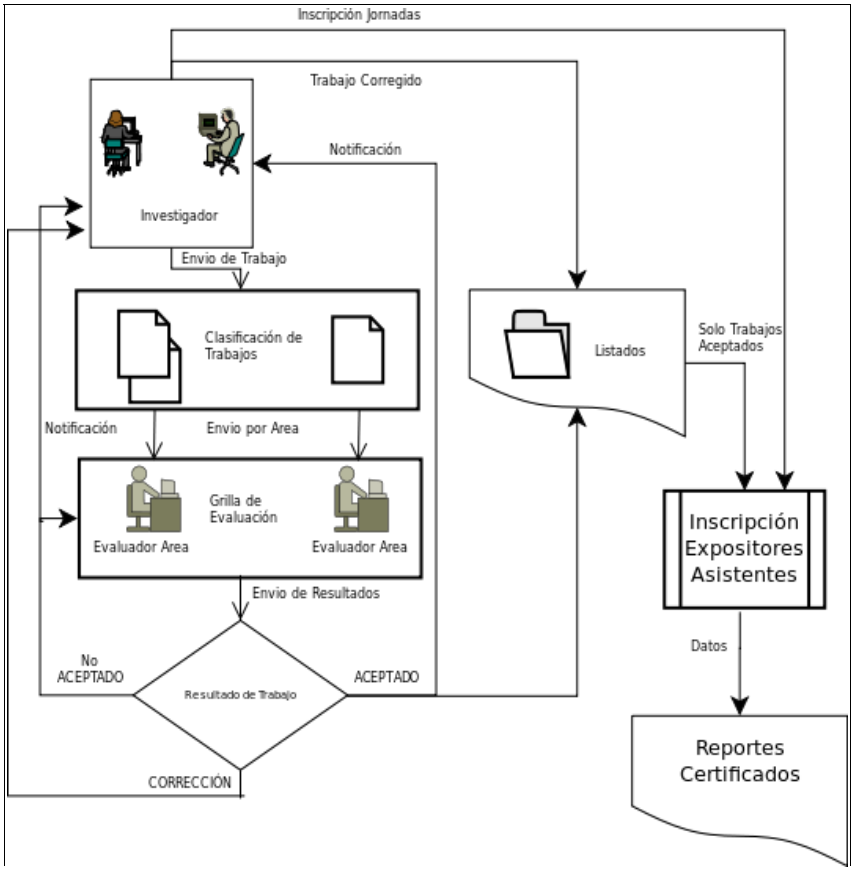

Fig. 2 Gestión de Trabajos

Ventajas en la implementación del Sistema: Gestión de Jornadas

Acceso restringido a opciones del sistema, tanto para el autor, evaluador y organizadores.

Envio de trabajos, por parte de los investigadores, bajo el entorno web, donde se establecen fechas de cierres de presentación para los autores; de evaluación para los evaluadores; fechas de inscripción, que afectan a la totalidad de los usuarios del sistema.

Los autores, pueden modificar su trabajo, en todo momento, y/o completar información, via online de los autores que comprenden la elaboración de dicho trabajo.

Los evaluadores, llevan a cabo la evaluación, y una vez completada la grilla de evaluación, los organizadores tienen conocimiento online de tal situación, donde ellos

deben comunicar a los autores del trabajo, o habilitar a otro evaluador. La condición de Evaluación del trabajo está presente via online, para que el responsable del trabajo pueda visualizar en todo momento, la Aceptación, corrección, o No Aceptación de su trabajo.

Los autores pueden visualizar y completar la grilla de evaluación de su trabajo pertinente.

Los organizadores pueden visualizar los trabajos a medida que vayan ingresando al Servidor WEB.

Se pueden realizar consultas, listados, estadísticas sobre datos verdaderos.

b) Biblioteca Digital de los trabajos de las Jornadas.

Introducción

En la Web encontramos que la información crece constantemente y parte de ella está disponible a través de servicios especializados de Bibliotecas Digitales.

Una gran fuente de información generaba cada edición de las Jornadas, donde año tras año, toda esa información de Trabajos Aceptados, solo se podía consultar en los anales impresos correspondientes.

En la edición de las Jornadas 2010, toda la información de Articulos se encontraban disponibles en una base de datos, eso nos dio lugar a que desarrollemos el software BTC[2] para el acceso a dicha información. Concentrar la información proveniente de diferentes fuentes de datos, obviamente, relativas a una misma área de interés, beneficiaria en la búsqueda de información en la que el usuario está interesado.

Propusimos el desarrollo de

1.- Construcción de un Repositorio Digital

. 2.- Creación de wrappers.

1.- Desarrollo e Implementación del Respositorio Digital

Con la colaboración del Laboratorio de Base de Datos, de la UFMG, Brasil, se ha desarrollado un repositorio de Articulos, presentados y aprobados en la Jornadas.

La búsqueda de artículos se realiza por TITULOS, RESUMEN, EVENTOS, TRABAJOS y por AUTORES.

Además de la difusión de la información, uno de los objetivos principales del desarrollo del repositorio es la preservación de sus contenidos.

2 BTC Biblioteca de Trabajos Científicos. Portal desarrollado por la Facultad de Ingeniería-Unju. Argentina.

2.- Desarrollo de Wrappers.

Introducción

El portal BTC (Biblioteca de trabajos científicos), portal que almacena artículos publicados en la VI Jornada de Ciencia y Tecnologia de las Ingenierías del Noa, llevadas a cabo en la provincia de Jujuy – Argentina, tiene un total de 136 trabajos en sus correspondientes bases de Datos. El portal tiene muy pocos artículos incorporados a su Base de Datos, y por tal motivo propusimo realizar la búsqueda de información no solo en dicho portal, sino efectuar la búsqueda de información en otras bibliotecas digitales, como ser la BDBComp, en su servicio de búsqueda de artículos por Títulos.

Propusimos desarrollar extractores de datos o también denominados wrappers que apliquen los criterios de búsqueda que se han introducido en el portal BTC, en el portal BDBComp[3] , y así lograr obtener los datos-resultados de diferentes páginas. Los criterios de búsqueda son ingresados en cualquier de los siguientes idiomas: español, portugués e inglés, en el portal BTC, y la traducción se efectúa via ONLINE a otros idiomas mediante Google Translator, donde también aplicamos el concepto de extracción de datos. Dichos procesos son llevados a cabo en forma transparente para el usuario que efectúa la consulta. Los resultados son formateados, clasificados según el idioma de escritura y visualizados mediante archivos XML dentro del portal BTC, sitio donde se va a concentrar la información.

El portal BDB Comp (Biblioteca Digital Brasileira de Computacão), portal que ofrece servicios de búsqueda de documentos científicos en sus bases de datos. Comprende 6060 trabajos publicados en diferentes eventos y periódicos de computación realizado en Brasil. También incluye trabajos publicados en los siguientes periódicos : JBCS, RITA, IP e INFOCOMP.

Los criterios de búsqueda que el usuario introduzca, en el portal BTC (A), puede aplicarlos en el portal BDBComp (B), para efectuar la búsqueda de artículos científicos. Los resultados del portal B son extraídos, analizados y presentados dentro del portal A. Figura 3.

Para migrar la información de un portal a otro, se crea por cada consulta un archivo con formato XML. En internet hay un gran incremento de la cantidad de datos en formato XML [2], siendo considerando el formato de elección para el intercambio de datos, ya que es flexible para representar diferentes tipos de información, sobre todo datos semi-estructurados[3].

3 BDBComp Biblioteca Digital Brasileira de Computacão desarrollado por el Departamento de Ciencia da Computacão de la UFGM, Brasil.

Tenemos tres clases de tareas relacionadas al manejo de información en la Web [1], donde nos detalla:

i) Modelado y consultas en la web,

ii) extracción de Información e integración y

iii) construcción y reconstrucción de sitios en la web.

En base a esto, determinamos extraer una representación estructurada de los datos de páginas HTML, que nos brinda la BDB Comp, después de realizar una consulta, a través de su motor de búsqueda, y aplicamos el concepto de crear nuevos sitios.

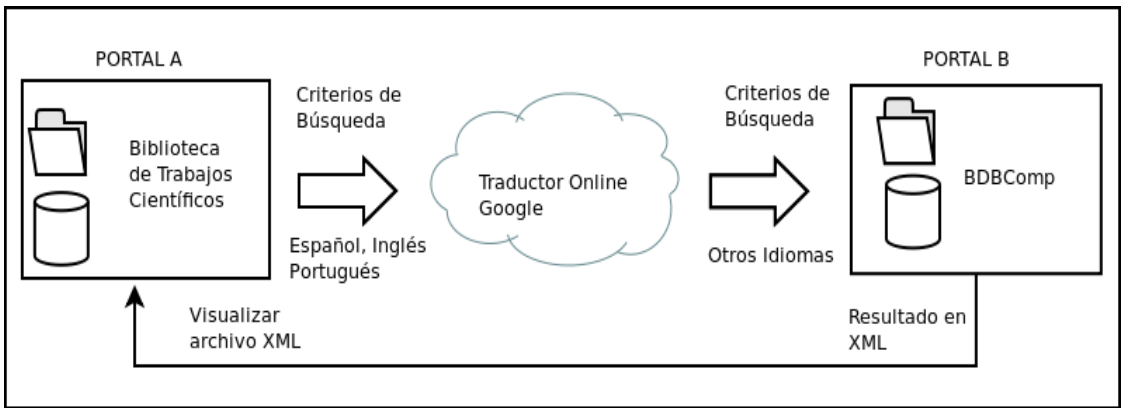

Figura 3. Aplicación de wrappers

Abordaje de la Propuesta: Extracción de datos.

Extraer la información correcta, es de suma importancia, por la cual en su mayoría las aplicaciones utilizan wrappers.

Los wrapper son programas capaces de reconocer y extraer objetos de interés dentro de páginas fuentes de la web. En [4] se efectúa una caracterización de los diferentes tipos de extractores de datos, según la técnica principal utilizada para la generación de wrappers. En [5] establece tres enfoques: la dificultad de un extractor de información, la técnica utilizada en la extracción, y por último el esfuerzo del usuario en el proceso de llevar dicho extractor a otro dominio.

El conjunto de páginas generadas del sitio BDBComp son dinámicas, donde la diferencias de las paginas devueltas sólo se puede apreciar en el contenido, como lo muestra en la Figura 4.

Por cada paginación que se realiza de información, representa ejecutar una nueva consulta al motor de Bases de Datos, llevando a que cada vez que se realiza la extracción de datos, para una nueva página de información, se efectúa la consulta correspondiente al servidor para rescatar todos los datos de dicha página.

Figura 4 Registros devueltos en la BDB Comp

Los artículos almacenados en la BDBComp, como el campo TÍTULO, donde se aplican los criterios de búsqueda están en lenguaje portugués, español e inglés. Los criterios de búsqueda ingresados en la BTC, se han traducido a los idiomas PORTUGUÉS, INGLES y ESPAÑOL, mediante Google Translator, y la extracción de datos, en forma transparente para el usuario, y asi poder recuperar los artículos en esos idiomas.

Se efectuaron los siguientes pasos, para la extracción de información.

1.- Identificar los objetos-información a extraer de la BDBComp, figura 5

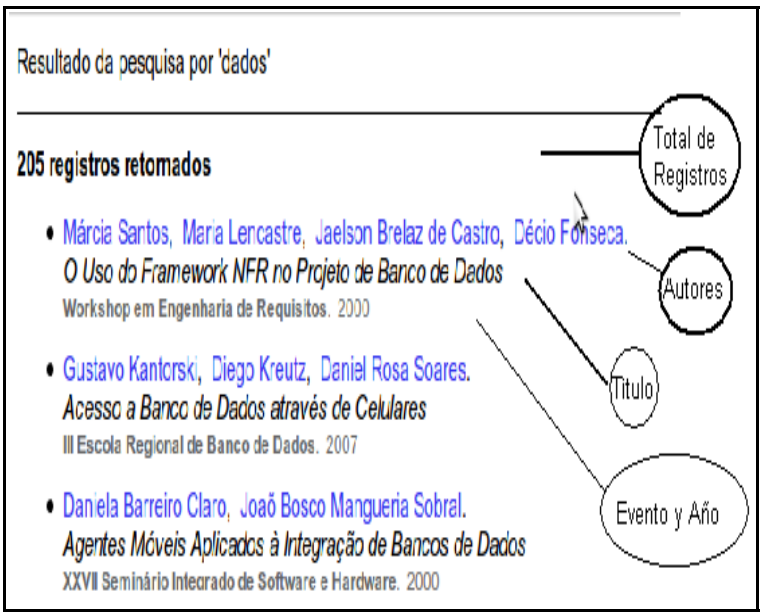

Figura 5. Elementos a Extraer

Los Objetos a Extraer son:

Total de Registros (R) , Autores (A) , título (T) , Evento (E), Año (N) y Link del trabajo (L)

El Conjunto de Objetos $O = \{ R, \{A_1 \ldots A_n\}, T, E, N, L\}$ So , donde

L, tiene como valores {Null, enlace del trabajo}

$A1 \ldots An \subset A$, siendo A un conjunto de Autores.

$So \subset S$, siendo S conjunto de página resultados de una consulta a la BDBComp

$S \subset B$, siendo B Base de datos de BDBComp.

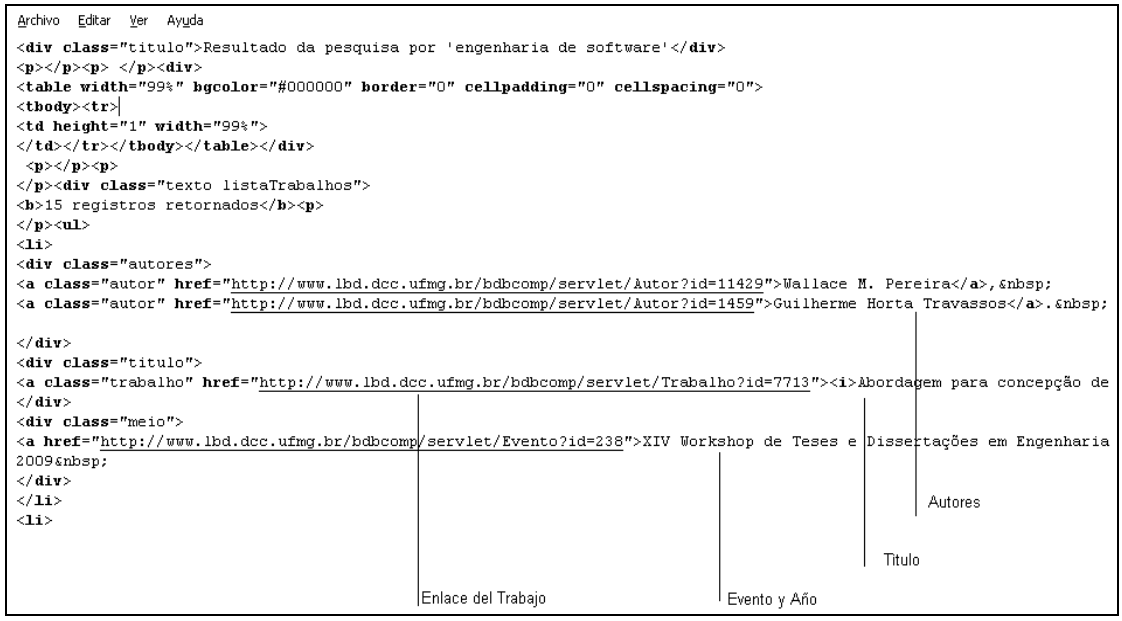

Figura 6: Identificación de Patrones

2) Identificar en S entre qué patrones se almacena la información So, y así poder realizar la extracción de los objetos O. La extracción de información se llevó a cabo mediante el lenguaje de programación PHP. Ver figura 6.

3) Recordemos que estamos en presencia de datos semi-estructurados, en la cual los objetos $O_1..O_n$ pueden tener o no valores. Los datos son almacenados en una estructura XML, como lo muestra la figura 7. Dicho archivo es utilizado para el traspaso de información de un portal a otro.

4) Extracción 100% de los datos, donde el usuario ha introducido una o más palabras como criterio de búsqueda. .

5) Google Translator realiza la traducción, donde la extracción de dichos resultados es 100 % efectiva, pero en algunos casos los resultados no son los esperados, debido a que en varios idiomas se utilizan varias palabras en inglés que no requerien una traducción, y por tal efectuar la traducción, la

recuperación de información no es la deseada por el usuario.

Figura 7 Ej. de archivo XML generado por el wrappers

Conclusión y Trabajos Futuros

El uso de Software Libre, permitió la instalación de Base de Datos, herramientas de programación, y en gran medida acelerar los tiempos de programación, para el desarrollo del Sistema.

La tecnología libre permitió liberar código, para desarrollar e incorporar mejoras en el Sistema, adecuándose a nuevas características y/o funcionalidades, logrando un software de mayor calidad.

El software desarrollado al ser sistema modular, permite mayor oportunidad de contribuciones por parte de desarrolladores y colaboradores del proyecto.

Caso de éxito: Liberar el Software con las características de licencia GNU GPL, permitió la incorporación de mejoras al sistema. La organización de las VII Jornadas, edición 2011, llevadas a cabo por la Universidad Nacional de Catamarca, pudieron efectuar sin inconvenientes, la instalación e implementación del sistema, y teniendo al alcance el **codigo fuente** se están realizando mejoras e incorporando nuevas funcionalidades y módulos al Sistema. Adecuaron módulos y funciones existentes en el sistema a los requerimientos que la organización necesitaba.

La colaboración y asesoramiento por parte del Laboratorio de Banco de Dados, del Departamento de Ciencia da Computacão, de la Universidad Federal de Minas Gerais, Brasil, dirigida por el Dr. Alberto Laender, ha permitido aplicar técnicas en el desarrollo de la Biblioteca Digital, utilizado para almacenar los artículos científicos que se presentaron en la VI Jornadas de Ciencia y Tecnología de las Ingenierías del NOA.

Se desarrolló diferentes extractores de datos, que extraen información de las páginas del portal B y lo muestra en el portal A, mediante un archivo con formato XML.

Logramos que el proceso de la extracción de información sea **RESISTENTE**, ya que con nuevas páginas de consultas So, tomadas de la misma fuente Web B, donde el formato HTML ha cambiado, pero el contenido de las páginas siguen siendo el mismo [6].

No podemos decir que el proceso sea **ADAPTABLE**, ya que en el presente trabajo solamente nos limitamos a extraer los datos de una sola fuente B.

Sería de sumo interés aplicar una verificación de la calidad de los datos extraidos, mediante cálculo de medidas de similitud probabilísticas, utilizando posicionamiento y una estructura de los datos en las páginas de origen [7] y para tal el archivo XML que se ha generado, almacena los valores de las posiciones de los datos que se han extraido.

Sería interesante la colección y prestación de información mediante un repositorio de código, funciones y módulos con características de código libre y abierto para la comunidad universitaria de la Universidad Nacional de Jujuy.

Bibliografía

[1] Daniela Florescu, Alon Levy, Alberto Mendelzon. Databases Techniques for the World -Wide Web: A Survey. SIGMOD Record 27 (3): 59-74, 1998.

[2] Dongwon Lee, Wesley W. Chu, Comparative Analysis of Six XML Schema Languages. ACM SIGMOD Record Volume 29 (3): 76-87, 2000

[3] D. Chamberlin, Xquery: An XML query language. IBM Systems Journal, vol 41, N° 4, 2002

[4] A.H.Laender, B. Ribeiro-Neto, A.S. Da Silva, and J.S. Teixeira. A Brief Survey of Web Data Extraction Tools. SIGMOD Record, 31(2): 84-93, 2002.

[5] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled Shaalan. A Survey of Web Information Extraction Systems. Journal IEEE Transactions on Knowledge and Data Engineering. Volume 18 (10):1411-1428, 2006

[6] P. B. Golgher, A. S. da Silva, A. H. F. Laender, and B. A. Ribeiro-Neto. Bootstrapping for Example-Based Data Extraction. In Proceedings of the Tenth ACM International Conference on Infomration and Knowledge Management, pages 371-378, Atlanta, Georgia, 2001.

[7] Olga Regina Fradico de Oliveira . Uma Abordagem para Verificação Automática da Qualidade de Dados Extraídos da Web. Tesis de PosGraducación- 2003 UFMG, Belo Horizonte. Brasil

[8] Netcraft, http://www.netcraft.com . Acceso en: 05/2011

[9] GNU Operting System, http://www.gnu.org . Acceso en: 05/2011