Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación

Herramienta de Procesamiento de Publicaciones de Investigadores Latinoamericanos

Memoria para optar al Título de
Ingeniero Civil en Computación

Mario Antonio Liulión Sanhueza

Profesor Guía:
Sergio Ochoa Delorenzi

Miembros de la Comisión:
Jaime Hernán Sánchez Ilabaca
Benjamín Eugenio Bustos Cárdenas

Santiago, 21 de Octubre de 2009

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN
POR: MARIO ANTONIO LIULIÓN SANHUEZA
FECHA: 21/10/09
PROF. GUÍA: SR. SERGIO OCHOA

Herramienta de Procesamiento de Publicaciones de Investigadores Latinoamericanos

Desde hace mucho tiempo la globalización en la investigación es una realidad. Ya no es extraño ver conferencias en donde sus participantes vienen de distintas partes del planeta. Diferentes redes de colaboración se arman alrededor de temáticas particulares, con el objetivo de compartir resultados y reducir los esfuerzos de investigación. En este escenario las publicaciones juegan un rol protagónico, ya que no sólo dan cuenta de los resultados de una investigación particular, sino que también muestran cuáles son los grupos de investigación activos en una cierta región (o país) y cuáles son las áreas (o temas) sobre las que están trabajando. La visibilidad las publicaciones usualmente ayuda a aumentar y/o mejorar las redes de cooperación de sus autores.

En Latinoamérica estas instancias de colaboración suelen involucrar sólo a las principales universidades de un país o una región, pues su trabajo es un poco más visible que el de las universidades más pequeñas. Sin embargo, diversos reportes muestran que hay poca colaboración entre investigadores Latinoamericanos. Una posible causa es la poca de visibilidad que tiene el trabajo de estos investigadores, entre sus pares dentro de la región.

Como una forma de tratar de mejorar esta situación, este trabajo de memoria implementó una herramienta que captura, almacena y publicita las referencias a trabajos de investigadores Latinoamericanos. La fuente de datos utilizada fue la base de datos de DBLP. De esa manera se pretende ayudar a vincular a los investigadores de la región. Por el momento, este trabajo sólo consideró a investigadores del área de Ciencias de la Computación.

Como parte de esta memoria también se desarrolló un sitio Web, para el despliegue de los datos almacenados en el sistema. A través de este portal se pueden realizar diversas consultas, con el objeto de recuperar información de publicaciones de una persona o de una institución en particular.

El trabajo realizado es un punto de partida hacia la implementación de un observatorio de investigación de la región, en el área de Computación. Por esa razón el diseño del sistema soporta extensiones, como por ejemplo: la incorporación de información fidedigna de instituciones educativas y áreas de investigación, la búsqueda de información, la incorporación de indicadores estadísticos, el manejo de usuarios dentro del sitio web y la incorporación de otras fuentes de publicaciones.

Agradecimientos

Hay muchas personas importantes en mi vida a las cuales quisiera agradecer en este momento.

A mis padres, por todo el esfuerzo y dedicación hacia mí. Má, te amo infinito, eres la persona más importante en mi vida. Pá, donde quieras que estés, sé que estás orgulloso de mí en este momento, me gustaría recibir tu abrazo y que tú recibieras todo mi amor. Siento que han sido excelentes padres y les estoy infinitamente agradecido.

A mi hermana y a mi sobrino que viene en camino. Les agradezco por darme una oportunidad de ser un nuevo y mejor hermano y tío.

A mis abuelas. Han sido una parte muy importante de mi vida y gran parte de lo que soy ahora, es gracias a ustedes. Les estoy profundamente agradecido. A mi Akún, donde quiera que estés, quiero que compartas esta alegría conmigo.

A TODOS mis amigos y familiares. Han sido muy importantes para mí. A todos y a cada uno de mis compañeros de curso desde el Kinder C del 90 hasta el IVºC de la generación 2002 del San Gaspar. A mis amigos del colegio, en especial a Jama, Loewe y Abadal. A mis compañeros de Cruzaco, en especial a la Fabi. A mis compañeros de la sección 3 de mechones del 2003, en especial a Guatón, a Agustín, a la Jacque, a Felipe, a Gabriel y a Christian. A mis amigos del DCC, en especial a Dusan, la Viole, Renato, Feña, Jugoso, Alvaro Monares. A mis grandes amigos de Seres Naturales y en especial a mi equipo VIDA XV: Willy, Diego, Sophie y Juan Pablo. Los amo caleta.

A mis profesores, en especial a los de mi colegio. Siento que fueron muy importantes en mi formación. A mi profesor guía Sergio Ochoa, por su apoyo en este trabajo.

Y a todos los que no nombré, también les va mi agradecimiento.

Índice

Índice de Figuras

Índice de Tablas

1 Introducción

Los medios de comunicación han derribado diferentes fronteras físicas y han facilitado innumerables actividades del quehacer humano, siendo la investigación y el desarrollo de las ciencias una de ellas. Actualmente, los investigadores de diferentes partes del mundo colaboran entre sí para el desarrollo de todo tipo de actividades y estudios en conjunto. Las redes de investigación generadas de esta forma, han permitido abrir los horizontes de la ciencia hacia nuevos terrenos, e inclusive obtener resultados en forma más rápida y efectiva. Estos avances de la ciencia se reflejan a través de la publicación de dichos trabajos de investigación, muchos de los cuales se dejan en repositorios de información científica, a disposición de la comunidad.

Las ciencias de la computación no están ajenas a esta realidad, con sus redes de investigadores en áreas de conocimiento específicos de lenguajes, algoritmos, ingeniería de software, etc. Sin embargo, en Latinoamérica estas instancias de colaboración suelen involucrar solamente a las principales universidades, dejando postergadas a las pequeñas y medianas instituciones universitarias de la región.

En este escenario, el acceso a publicaciones científicas toma una gran importancia, pero a pesar de ello, existen algunas limitantes al respecto. Si bien cada país de la región tiene su propia conferencia nacional, rara vez estas publicaciones trascienden el escenario local, debido a que no existen repositorios públicos que almacenen estos artículos. De esa manera se inhibe (o se reduce) la posibilidad de colaboración entre estudiantes, docentes y/o investigadores latinoamericanos, especialmente porque se trata de publicaciones en español, que pueden ser fácilmente aprovechables por otros alumnos y docentes de universidades, que no necesariamente tienen dominio del idioma inglés.

Otra limitante importante que tienen las universidades latinoamericanas es la falta de visibilidad respecto a investigadores y laboratorios de investigación en la región. Si esa información estuviera disponible, probablemente habría más instancias de colaboración entre estas universidades, y definitivamente esto podría ayudar a las universidades más pequeñas para establecer vínculos con otras instituciones latinoamericanas en ámbitos específicos. Muchas veces instituciones vecinas no colaboran por falta de conocimiento mutuo respecto a lo que hace la otra.

Este trabajo de memoria propone enfrentar estos desafíos de acceso libre a la información -identificados en el Workshop CharLA08 [Cha08]- extendiendo la base de datos de investigadores latinoamericanos que posee LACCIR (Latin American and Caribbean Collaborative ICT Research Federation) [Lac08], e implementar un repositorio de artículos científicos de investigadores latinoamericanos. Esta información será de dominio público y podrá ser accedida a través de diversos criterios: por investigador, por país, por área de trabajo, etc.

1.1 Justificación

El trabajo propuesto maneja varios desafíos. Uno de estos desafíos es la integración con diversos sistemas ya existentes para extraer los registros de las publicaciones desde las distintas fuentes y centralizar la información en una sola herramienta. Estos sistemas no necesariamente mantienen un mismo estándar para publicar los trabajos. Es decir, la información disponible está distribuida, eventualmente duplicada y en distintos formatos. Por lo tanto, se requerían procedimientos de alimentación, clasificación, análisis de consistencia y almacenamiento de dicha información.

Por otra parte, existen diferentes comunidades de investigadores latinoamericanos que permanentemente realizan publicaciones en conjunto en diversas áreas de aplicación. El resultado del procesamiento de esta información, por área de investigación, por institución, etc., es de gran valor. Esto permite obtener una concepción global aproximada de las actividades de la región, indicadores de las actividades de los investigadores y las relaciones que se establecen producto de la colaboración.

Hoy en día no existe ninguna herramienta que permita resolver estos requerimientos a nivel de investigadores latinoamericanos. Parte de lo que existe a nivel global, es reutilizable para sacar ventaja de lo que ya está hecho. Específicamente para el trabajo de esta memoria, se saca provecho de la información publicada en el repositorio DBLP [Dbl08].

La herramienta desarrollada es de especial interés para las agencias de apoyo a la investigación científica (por ejemplo, CONICYT), universidades, estudiantes e incluso para los mismos investigadores, ya que facilita el procesamiento de la información de las publicaciones de los investigadores de la región y permite sacar indicadores que reflejan la actividad de los mismos.

1.2 Objetivo

A continuación se presentan el objetivo principal y los objetivos específicos asociados a este trabajo de memoria.

1.2.1 Objetivo General

El objetivo general de esta memoria es implementar un repositorio de información de artículos científicos de investigadores latinoamericanos. Dicho repositorio debe contar con mecanismos que faciliten su alimentación masiva, la actualización de artículos y también la consulta de los mismos a través de diferentes criterios.

1.2.2 Objetivos Específicos

Para el logro del objetivo general, se plantearon los siguientes objetivos específicos:

1. Extender la actual base de datos de investigadores que posee LACCIR, a fin de poder incluir en la herramienta el área de trabajo de las personas allí registradas.

2. Diseñar e implementar el repositorio de artículos científicos. Diseñar e implementar el subsistema encargado de la alimentación de documentos provenientes de DBLP y del CLEI [Cle08], de la conversión de formato[1], del chequeo de consistencia y de la administración del repositorio.

3. Diseñar e implementar un sistema de consulta que permita acceder a la información a través de diversos criterios e indicadores.

La información que se almacene acerca de cada publicación, deberá contener al menos la mínima recomendada por LaTeX y APA [APA08]. En base a estos datos, se podrá generar una lista de publicaciones y su descripción detallada.

Físicamente, el repositorio estará alojado en un servidor aún por definir del Departamento de Ciencias de la Computación de la Universidad de Chile. Este servidor, que probablemente sea el mismo que actualmente se utiliza para LACCIR, será administrado por personal contratado por esta entidad.

1.2.3 Posibles Indicadores de Interés

A continuación se listan algunos indicadores que eventualmente se podrían obtener del sistema implementado:

1. Cantidad de artículos por investigador en un determinado año.
2. Cantidad de artículos de conferencias y de journals de un investigador.
3. Cantidad de artículos por institución.
4. Cantidad de artículos, donde el investigador está como autor.
5. Cantidad de artículos, donde el investigador es el único autor.
6. Cantidad de publicaciones que involucran a investigadores de otro país.
7. Cantidad de publicaciones por país.

[1] La conversión se pretende llevar a cabo mediante la caracterización de campos de cada publicación para luego almacenarlas en un formato estándar, como lo es el de artículos de conferencia de LaTeX [Lam94].

1.3 Metodología

Dado los objetivos del trabajo de memoria, se realizó el desarrollo de la solución siguiendo las fases que se detallan a continuación.

1.3.1 Análisis del Problema a Resolver

En la primera fase se realizó una serie de reuniones con el profesor guía para una mejor concepción del problema, acotarlo y tener una noción de los requisitos para la herramienta a desarrollar. Se revisó algunos repositorios bibliográficos en Internet para analizar la información disponible y las ventajas y desventajas de trabajar con cada uno de ellos.

Una vez aclarados los requerimientos de la herramienta, se tomó la determinación de que con el fin de extraer información de publicaciones, lo más conveniente era abordar la tarea de trabajar con el repositorio DBLP. Dicha decisión se tomó dada la ventaja de que este sistema cuanta con gran cantidad de información y que la mayoría de la misma estaba condensada en un solo archivo y disponible en la Web.

Para la información de los investigadores se dispuso un archivo con algunas personas registradas en LACCIR. Con el fin de establecer un matching entre los autores dados y las publicaciones registradas en DBLP, se debió realizar una labor de obtención de datos, que consistió en recuperar el nombre estándar de publicación para un subconjunto de autores. Esta labor se realizó manualmente buscando en el sitio de DBLP por el nombre completo de algunos autores y almacenando el nombre con el cual dicha persona publica.

Finalmente, se realizó una revisión de las herramientas a utilizar para el desarrollo del sistema. Para esto, se exploró distintos sistemas de base de datos, entornos de desarrollo, lenguajes de programación, etc. El objetivo era conseguir las herramientas más adecuadas dados los requerimientos del problema a resolver.

1.3.2 Diseño de la Solución Propuesta

La segunda fase consistió en diseñar una solución que satisficiera los requerimientos. Se prestó especial énfasis al diseño del modelo de datos, ya que de dicho modelo se iba a generar automáticamente el código para la manipulación de la información del sistema. Esta generación automática corresponde a una de las ventajas del entorno de desarrollo escogido para llevar a cabo la implementación del sistema[2].

[2] Más adelante se profundizará en el ambiente de desarrollo usado para el trabajo de memoria.

1.3.3 Desarrollo de la Aplicación de Migración de Datos

En la tercera fase se desarrolló una aplicación de escritorio para la migración masiva de Datos.

La aplicación que se construyó, lee un archivo con los datos del repositorio DBLP y de él extrae la información correspondiente a las publicaciones de los investigadores latinoamericanos. Dicha información es persistida en la base de datos de manera local.

1.3.4 Desarrollo del Sitio Web de Reportes y Consultas

La cuarta fase consistió en desarrollar un sitio Web en el cual se desplegara la información solicitada. Para ello y usando otra funcionalidad del entorno de desarrollo, se generó automáticamente reportes de:

- Detalles de Persona
- Detalles de Publicación
- Detalles de Institución

Junto con lo anterior, se construyeron los siguientes módulos:

- Módulo para obtener las publicaciones para una cierta persona.
- Módulo para obtener las publicaciones para una determinada institución.
- Módulo para la búsqueda de personas.

1.4 Glosario

A continuación, en la Tabla 1: Glosario, se presenta un listado con algunos términos usados en este documento y que pueden ser desconocidos para el lector.

Tabla 1: Glosario

Término	Descripción
APA	American Psychological Association (Asociación Estadounidense de Psicología). Es una organización científica y profesional de psicólogos de EE. UU.
API	Application Programming Interface (Interfaz de Programación de Aplicaciones). Es el conjunto de funciones y procedimientos que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción.
BibTeX	Es una herramienta para dar formato a listas de referencias que se utiliza habitualmente con el sistema de preparación de documentos LaTeX.
CONICYT	Comisión Nacional de Investigación Científica y Tecnológica. Es la institución estatal chilena que coordina, promueve y fomenta la investigación científica y tecnológica en sus distintos campos. Depende del Ministerio de Educación.
DTD	Document Type Definition (Definición de Tipo de Documento). Es una descripción de estructura y sintaxis de un documento XML o SGML.
Entiy Bean	Es un tipo de Enterprise JavaBean que representa la data persistente mantenida en una base de datos.
GlassFish	Es un servidor de aplicaciones desarrollado por Sun Microsystems que permite ejecutar aplicaciones que siguen la especificación Java EE.
IDE	Integrated Development Environment (Entorno de Desarrollo Integrado). Es un programa compuesto por un conjunto de herramientas para un programador.
ISBN	International Standard Book Number (Número Estándar Internacional de Libro). Es un identificador único para libros, previsto para uso comercial.
JDBC	Java Database Connectivity. Es una API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java.
JPA	Java Persistence API. Es la API de persistencia desarrollada para la plataforma Java EE.
JSF	JavaServer Faces. Es una tecnología y framework para aplicaciones Java basadas en web. Simplifica el desarrollo de interfaces de usuario en aplicaciones Java EE.
JSP	JavaServer Pages. Es una tecnología Java que permite generar contenido dinámico para web, en forma de documentos HTML, XML o de otro tipo.
LaTeX	Es un sistema de composición de textos, orientado especialmente a la creación de libros, documentos científicos y técnicos que contengan fórmulas matemáticas.
NetBeans	Es un IDE para el desarrollo de aplicaciones para la red.
phpPgAdmin	Es una aplicación web, escrita en PHP, para administrar bases de datos PostgreSQL.
PostgreSQL	Es un sistema de gestión de base de datos relacional orientada a objetos.
SAX	Simple API for XML (API simple para XML) es una API para realizar un parsing de acceso serial en XML.
XML	Extensible Markup Language (Lenguaje de Marcas Extensible). Es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C).

2 Análisis del Problema a Resolver

En este capítulo, se profundiza en el problema a tratar. Específicamente, se describen en detalle el sitio DBLP y otros sitios de referencias bibliográficas. También se describen los requisitos de la herramienta a desarrollar al igual que la elección del ambiente de desarrollo con el cual se realizó este trabajo.

2.1 Revisión Bibliográfica

Como se dijo anteriormente en la Introducción, a nivel mundial se pueden encontrar diversas iniciativas de almacenamiento de publicaciones, como por ejemplo, Google Scholar [GSc09], CiteSeerX [Cit08], o el mismo DBLP ya mencionado. Si bien, estos sitios están diseñados para que el acceso a los registros se realice por medio de las interfaces gráficas, dichas interfaces no permiten el fácil procesamiento de los datos para recolectar otro tipo de información de interés. En el caso del problema aquí propuesto, no se pueden obtener inmediatamente estadísticas de colaboración de los investigadores latinoamericanos.

Sin embargo, los datos de las publicaciones alojadas en alguno de estos sistemas pueden ser recolectados para su posterior procesamiento. La recolección debe ser específica para cada sistema y depende del cómo se publica la información en ellos. A continuación se detalla alguna de las características de los sistemas de los cuales se pretende alimentar la herramienta a desarrollar.

2.1.1 DBLP

DBLP es la sigla de Digital Bibliography & Library Project y corresponde a un sitio Web especializado en bibliografía del área de las ciencias de la computación, alojado en la Universidad de Trier en Alemania (Ver Figura 1: Página inicial de DBLP).

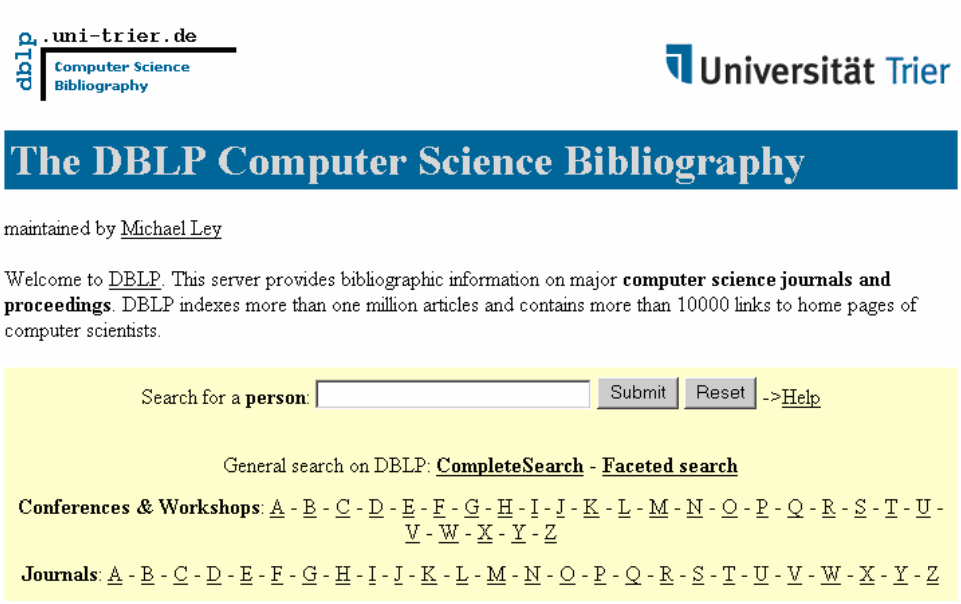


Figura 1: Página inicial de DBLP

Originalmente, esta herramienta se focalizaba a temas de Bases de Datos y Programación Lógica, pero gradualmente fue expandiéndose a otras áreas. Ha funcionado desde 1980 y actualmente contiene más de un millón de registros de publicaciones.

En este sitio Web sólo se pueden realizar búsquedas por persona y dichas consultas entregan un listado de sus publicaciones agrupadas por el año. Junto con este listado, se despliega un índice de coautores calculado a partir de los registros de dichas publicaciones (Ver Figura 2: Consulta en DBLP).

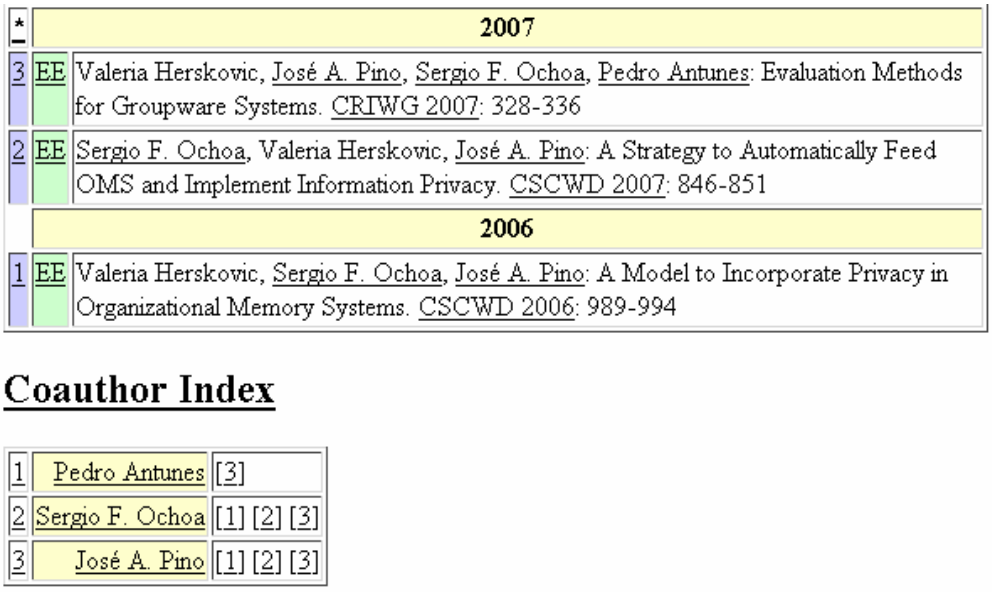


Figura 2: Consulta en DBLP

Además se despliegan una serie de links para la consulta de los coautores, para la eventual descarga de los documentos y para la generación de una referencia BibTeX a la publicación en cuestión.

Si bien, este índice de coautores entrega información de gran utilidad relativa a la colaboración de los investigadores, no entrega información sobre el país o institución de la cual proviene la persona. Tampoco se puede realizar una fácil comparación entre varios autores en cuanto a, por ejemplo, el número de publicaciones de cada uno.

Otra característica interesante de DBLP es que pone a disposición sus datos bibliográficos por medio de la descarga de un archivo en formato XML desde un repositorio. Dicho archivo, llamado “dblp.xml”, contiene todos los datos de las publicaciones ingresadas en el sistema y es continuamente actualizado.

El archivo dblp.xml es de gran importancia para el trabajo de esta memoria. Posteriormente se entrará en detalle sobre la estructura del archivo, ya que es fundamental para describir el funcionamiento de la aplicación de migración de datos.

2.1.2 Otros Sitios Bibliográficos

A continuación se describen otros sitios de referencia bibliográfica. Si bien no forman parte del desarrollo mismo de la herramienta de esta memoria, cabe destacarlos dada su popularidad en el ambiente académico.

2.1.2.1 CiteSeerX

CiteSeerX, desarrollado en la Universidad Estatal de Pensilvania, es una biblioteca digital de literatura científica, en especial de computación y de ciencias de la información (Ver Figura 3: Página inicial de CiteSeerX).

CiteSeerX es el sucesor de CiteSeer, que en 1997 fue el primer prototipo de biblioteca digital en tener un indexado y enlazamiento automático de citas. CiteSeerX surgió como necesidad de almacenar los más de 750 mil documentos y de satisfacer las más de 1,5 millones de visitas diarias que estaban exigiendo al límite a la primera versión de CiteSeer [Cit09].

Figura 3: Página inicial de CiteSeerX

En este sitio Web se pueden realizar búsquedas tanto por nombre del documento, como por autor. Para ambos casos, el resultado de cada consulta es un listado de publicaciones ordenadas por el número de citas a dicha publicación (Ver Figura 4: Consulta en CiteSeerX).

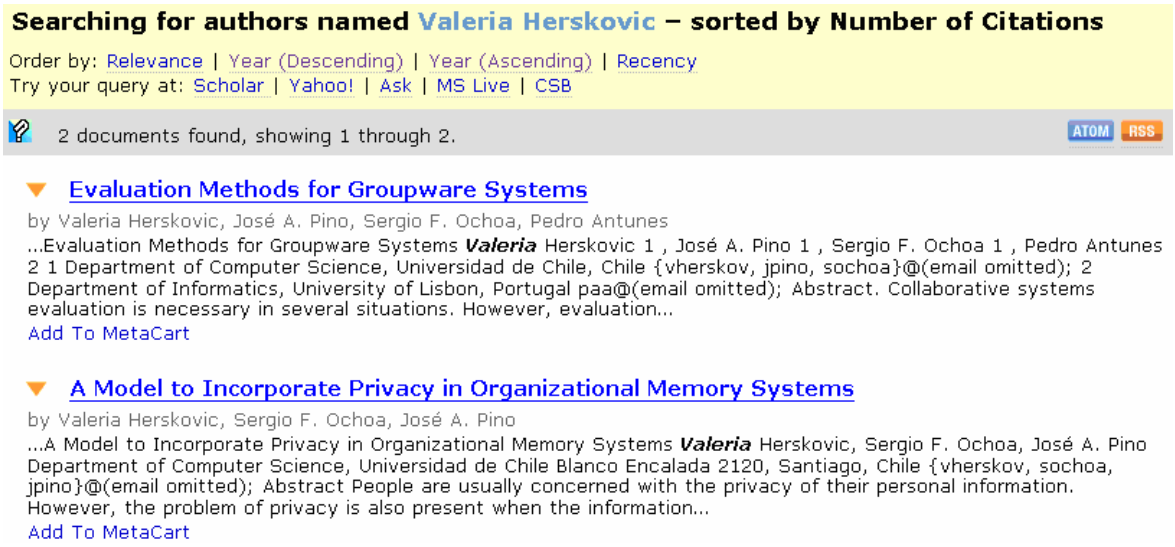


Figura 4: Consulta en CiteSeerX

Al poner el cursor sobre la viñeta de cada entrada, se puede tener acceso al abstract del documento. La descarga de las versiones completas está disponible en el link del título de cada publicación. Por otro lado, también se entregan links para ordenar el listado de publicaciones encontradas bajo otros criterios.

CiteSeerX, tiene la particularidad que el resultado de cada búsqueda se puede desplegar en forma de feeds, siguiendo un formato estándar de publicación. La ventaja de la información obtenida de esta forma, aparte de seguir un estándar de redifusión, se generan al consultar una URL, lo que potencialmente se puede utilizar para consultar automáticamente el sistema.

Otra ventaja de esta característica, es que la información generada es relativamente pequeñas (por ejemplo, en comparación con el archivo obtenido desde DBLP), lo que permite una fácil manipulación y un procesamiento más específico de la información.

CiteSeerX no maneja las instituciones a la cual pertenece cada investigador ni mantiene indicadores para cada autor.

2.1.2.2 Google Scholar

Este sistema, cuyo propietario es la compañía Google, es un buscador bibliográfico multidisciplinario de artículos y que presenta un enfoque en la literatura académica. Fue lanzado a público el 18 de Noviembre de 2004 y, hasta la realización de este trabajo de memoria, el sistema todavía se encuentra en versión Beta (Ver Figura 5: Página inicial de Google Scholar).

Figura 5: Página inicial de Google Scholar

Google Scholar encuentra publicaciones, extractos y citas, localizando la publicación completa en la web. Dentro de sus características más notables están:

- Ayuda a identificar la investigación más relevante en cualquier área de toda la investigación académica del mundo. Esto se logra mediante un ranking de los artículos.
- El índice de este sistema incluye las revistas más leídas en el mundo científico a excepción de Elsevier, la editorial más grande de libros de medicina y de literatura científica del mundo.

En este sitio, se pueden realizar todo tipo de búsquedas. Y los resultados son desplegados al usuario como se muestra en la Figura 6: Consulta en Google Scholar.

Figura 6: Consulta en Google Scholar

En los resultados se despliega un resumen de los datos de cada publicación, entre ellos el título, los autores y el año de publicación. El título es un link para ir a una descripción más detallada de la publicación.

Junto con los datos, se despliegan además opciones para cada artículo. Dentro de las más notables está el número de citas a cada publicación, las cuales se pueden ver usando el link asociado. De la misma forma se pueden desplegar los artículos relacionados y distintas versiones del documento.

Google Scholar, al igual que CiteSeerX, no maneja las instituciones a la cual pertenece cada investigador ni mantiene indicadores para cada autor.

2.1.2.3 Scopus

Scopus [Sco09] es una base de datos de extractos y citas multidisciplinaria de la Elsevier (Ver Figura 7: Página inicial de Scopus).

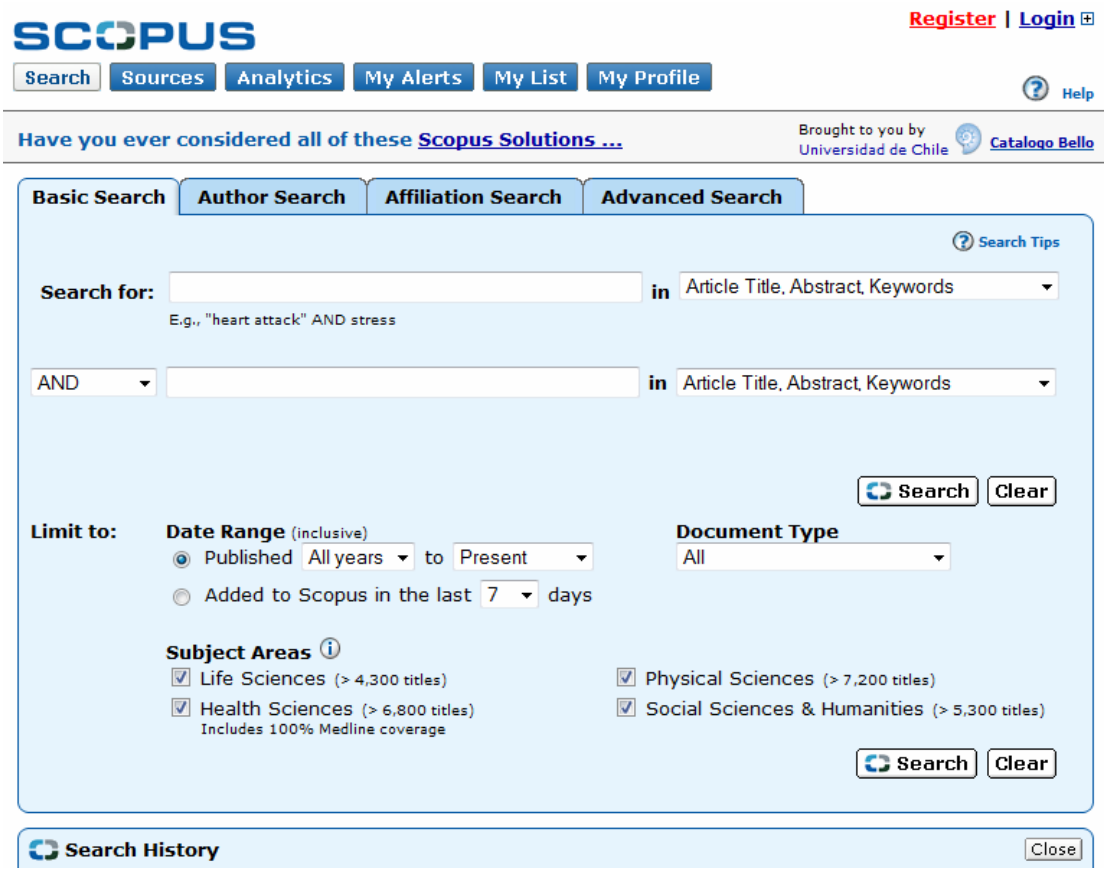


Figura 7: Página inicial de Scopus

Una desventaja importante de este sitio, es que el acceso a este sistema está restringido por suscripción. Algunas de sus características más notables son:

- Cuenta con artículos científicos, tecnológicos, médicos y sociales.
- Posee 38 millones de registros de publicaciones y 19 millones de ellos son posteriores a 1996 llegando a haber artículos que se remontan al año 1823. Cuenta con 3,6 millones de publicaciones de conferencias.
- Abarca alrededor de 18.000 títulos de más de 5.000 ediciones internacionales de los cuales 16.500 son de revisión por pares, 600 son publicaciones de negocio, 350 son series de libros.
- Cubre 435 millones páginas web científicas, incluyendo 23 millones de patentes.

El sistema ofrece campos para búsqueda por temas, por autor y por afiliación (por ejemplo, universidades). También existe un módulo de búsqueda avanzada. De no estar registrado, el usuario sólo puede consultar por autor.

El resultado de la búsqueda de autores, arroja listado de los autores que coinciden. En dicho resultado se presentan links para el despliegue de un perfil de cada investigador. En caso de estar registrado, se ofrecen filtros para acotar los resultados (Ver Figura 8: Búsqueda de Autores en Scopus).

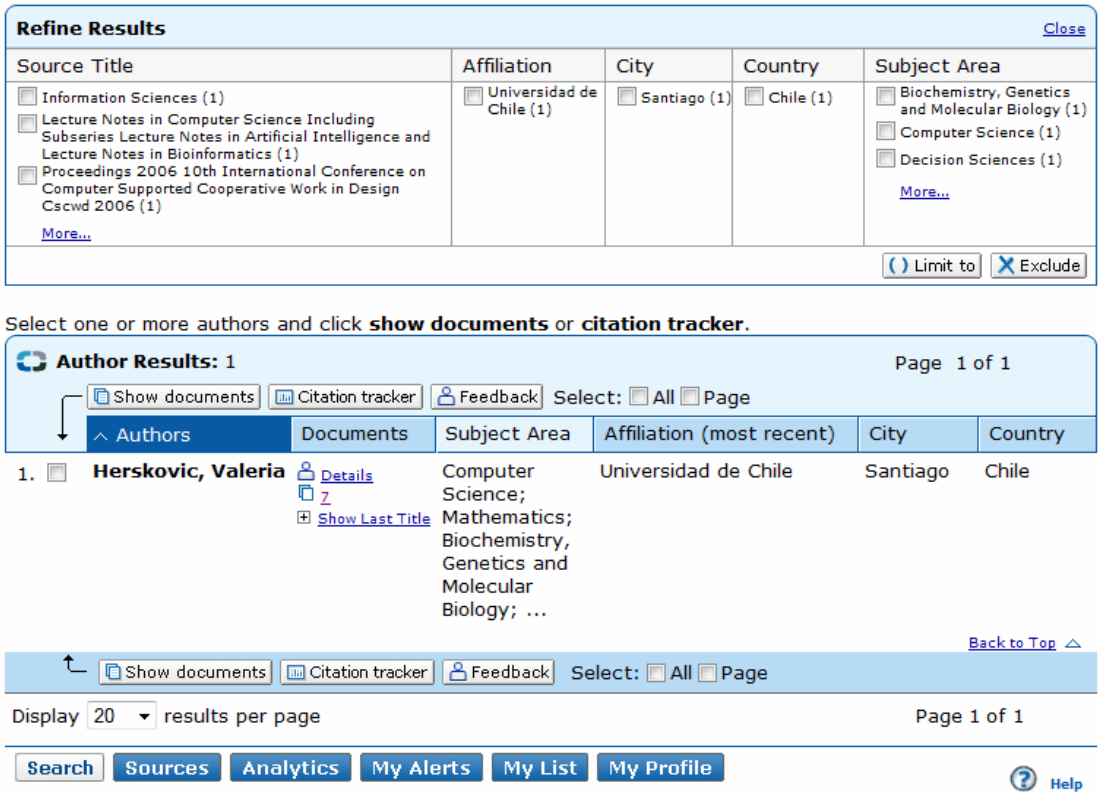


Figura 8: Búsqueda de Autores en Scopus

Como se dijo anteriormente, del listado de autores se puede acceder a un perfil del autor. En este perfil se despliega una información personal básica, junto con información relacionada a su actividad como investigador.

Si no se cuenta con una suscripción a Scopus, al momento de ver los detalles se muestra un perfil de cada autor con un resumen de la información almacenada en el sistema (Ver Figura 9: Perfil Personal en Scopus).

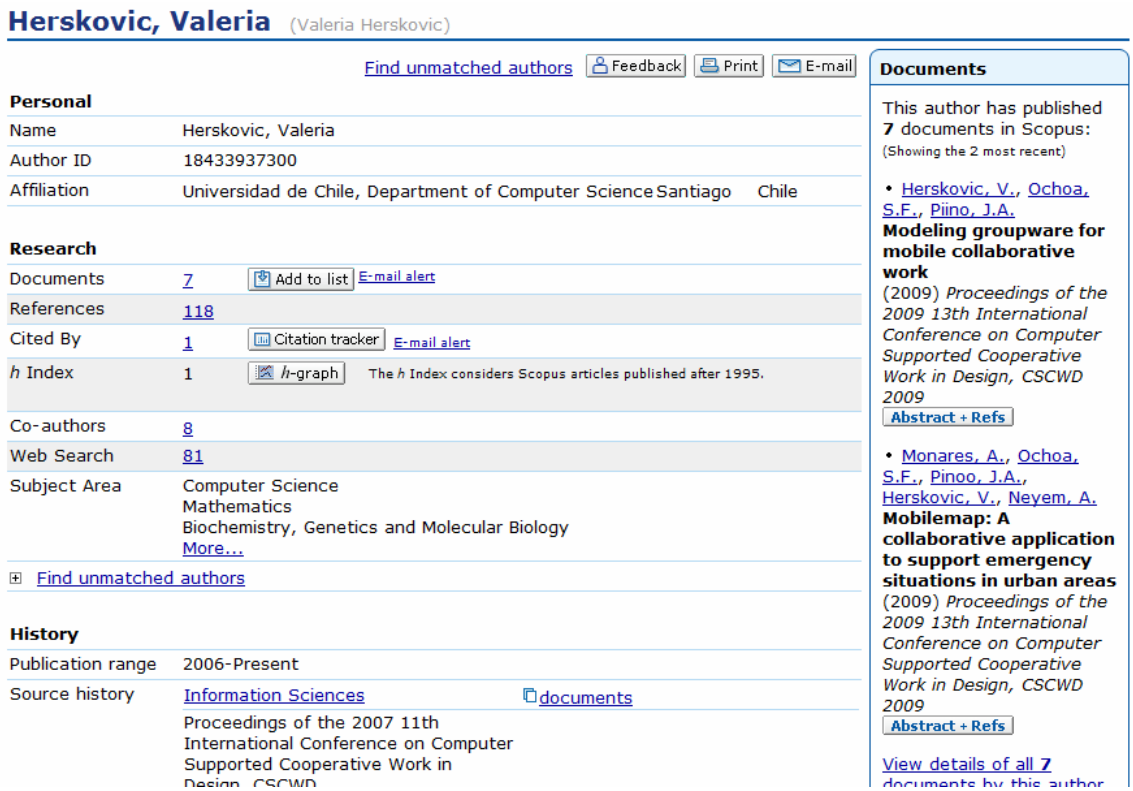


Figura 9: Perfil Personal en Scopus

Del resultado de la consulta temática se devuelve un listado de las publicaciones que hacen coincidencia. De dicho listado se puede acceder a extractos o a artículos completos. También se presenta un sub-módulo para filtrar la búsqueda.

Análogamente, para el resultado de la búsqueda por afiliación, se arroja un listado de las distintas instituciones que coincidan y el sub-módulo correspondiente para filtrar los resultados. Del listado de afiliaciones se puede acceder a un detalle en donde salen algunas estadísticas de la misma.

2.1.2.4 ISI Web of Knowledge

ISI Web of Knowledge [ISI09] es el sistema bibliográfico del Institute for Scientific Information (ISI) de la Thomson Scientific (Ver Figura 10: Página inicial de ISI Web of Knowledge).

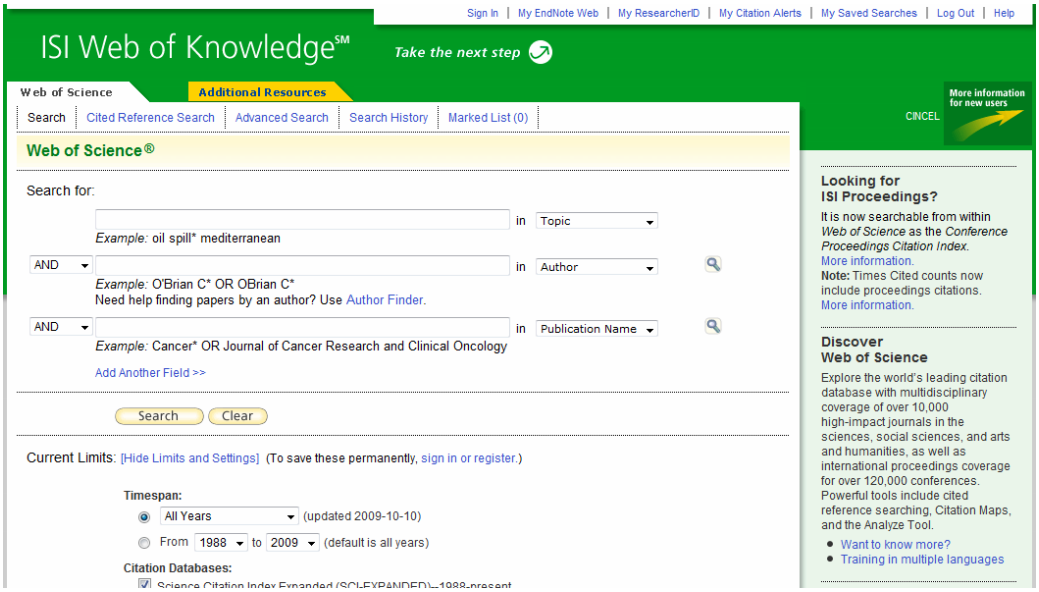


Figura 10: Página inicial de ISI Web of Knowledge

Una desventaja del sitio es que el uso de este sistema está restringido en su totalidad. Solamente instituciones como universidades y a áreas de investigación de grandes compañías pueden acceder a esta herramienta.

La pantalla principal permite una búsqueda por diferentes criterios, los más importantes son por tema, por autor (siguiendo la convención ISI) o por nombre de la publicación. El resultado de esta búsqueda arroja un listado de publicaciones que coinciden con los criterios. Junto con el listado se muestra un sub-módulo para refinar los resultados (Ver Figura 11: Consulta ISI Web of Knowledge).

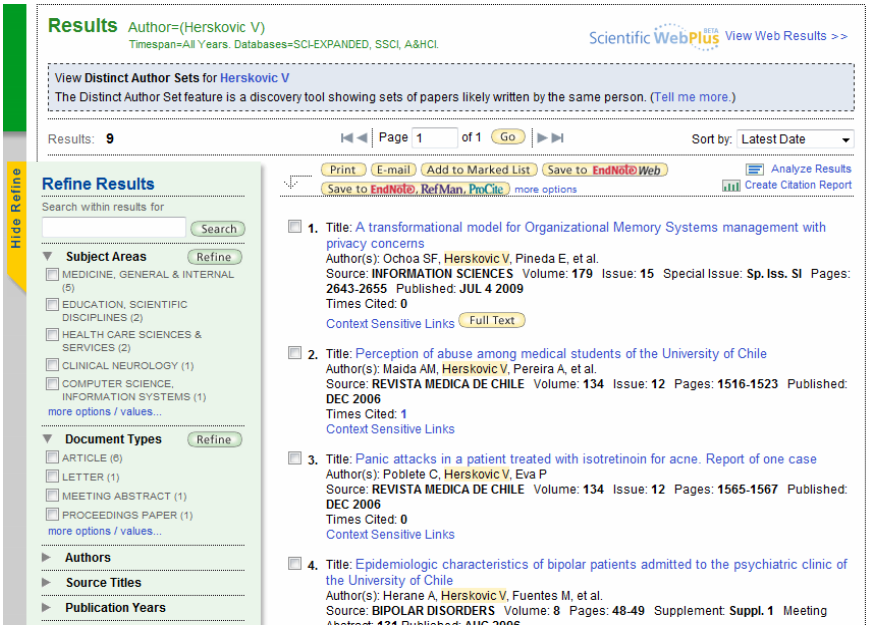


Figura 11: Consulta ISI Web of Knowledge

Del listado de resultados se puede acceder a una descripción detallada de cada publicación (Ver Figura 12: Detalles de Publicación en ISI Web of Knowledge).

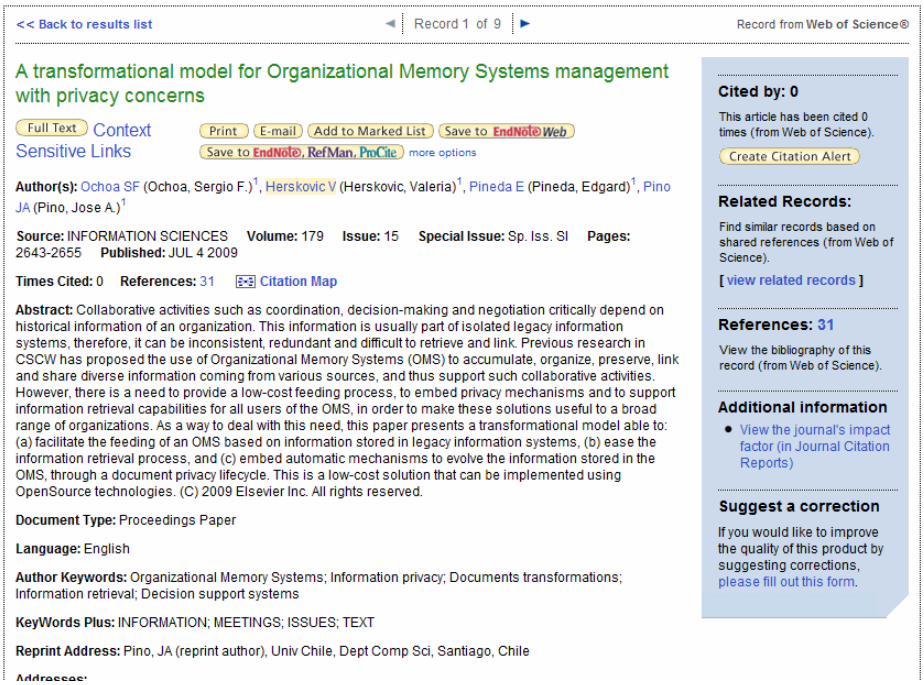


Figura 12: Detalles de Publicación en ISI Web of Knowledge

2.1.2.5 Latindex

Latindex [Lat09] es un sistema gestado en 1995 en la Universidad Nacional Autónoma de México (Ver Figura 13: Página inicial de Latindex).

Figura 13: Página inicial de Latindex

Si bien este repositorio multidisciplinario está focalizado en América Latina, el Caribe, España y Portugal, sólo cuenta con datos que caracterizan a cada journal académico. En la página principal se encuentra un buscador que arroja como resultado un listado de journals que coinciden con la palabra dada. Figura 14: Consulta en Latindex.

Figura 14: Consulta en Latindex

A partir de dicho listado, se puede acceder a los detalles del journal en cuestión.
Figura 15: Detalles de Publicación en Latindex.

Figura 15: Detalles de Publicación en Latindex

2.2 Requisitos de la Solución

Con el fin de cumplir los objetivos mencionados, se plantearon una serie de requerimientos que se describen a continuación.

2.2.1 Información a Considerar

El sistema debe manejar información relativa a los investigadores. Esto es, sus datos personales, su grado académico, sus áreas de investigación y aplicación. También se debe considerar los datos de la institución a la cual pertenece cada persona. Para esto último son de interés la universidad detrás de cada institución y la región geográfica a la cual pertenece.

El sistema debe manejar información de publicaciones. Son de interés el tipo de publicación, el título de la publicación, el año de publicación, si es una publicación ISI o no, la cantidad de páginas, el journal, el publisher.

2.2.2 Sitio Web de Reportes y Consultas

La herramienta debe contar con un sitio web para la generación de reportes construidos a partir de los datos almacenados en el sistema. Las consultas principales a satisfacer son las publicaciones de una persona y de una institución. El reporte que se genere, debe venir ordenado por tipo de publicación y por año.

2.2.3 Migración Masiva desde DBLP

La herramienta debe contar con un procedimiento de carga masiva de datos provenientes del repositorio DBLP. Esta información permitirá tener un estado inicial de datos en el sistema.

2.2.4 Considerar a Autores Latinoamericanos de LACCIR

El sistema a desarrollar debe considerar a los autores latinoamericanos. Para esto, desde LACCIR será provista una lista de investigadores de los cuales se necesita recuperar la información de sus publicaciones.

2.3 Análisis Detallado de DBLP

Para la carga masiva desde DBLP se sacó ventaja de que este sistema pone a disposición un archivo dblp.xml con sus datos.

Al comienzo de este trabajo de memoria, dicho archivo tenía un tamaño de 483MB y su contenido es básicamente una gran lista de registros de distintos tipos de publicaciones (artículos, libros, tesis, entre otros) con los correspondientes datos asociados. Ver Figura 16: Screenshot del archivo dblp.xml.

Figura 16: Screenshot del archivo dblp.xml

Este archivo XML presenta la particularidad de su gran tamaño, que dificulta su manipulación, lectura y parsing. La mayoría de las herramientas para el manejo XML están diseñadas para archivos más pequeños. Si bien del sitio se pueden descargar unas clases Java de ejemplo especialmente construidas para el parsing del archivo dblp.xml, las eventuales consultas para sacar información de este archivo, pueden a su vez, derivar en resultados de gran tamaño que también se vuelven difíciles de manejar.

Este archivo XML tiene una estructura definida por un archivo DTD llamado "dblp.dtd"[3]. A continuación en la Figura 17: Screenshot del archivo dblp.dtd, se presenta un extracto del archivo que define la estructura del archivo dblp.xml.

[3] En el anexo se detalla el procedimiento de obtención para el archivo dblp.xml y dblp.dtd.

```
<!ELEMENT dblp (article|inproceedings|proceedings|book|incollection|
                phdthesis|mastersthesis|www)*>
<!ENTITY % field "author|editor|title|booktitle|pages|year|address|journal|volume|
number|month|url|ee|cdrom|cite|publisher|note|crossref|isbn|series|school|chapter"
>

<!ELEMENT article       (%field;)*>
<!ATTLIST article
                        key CDATA #REQUIRED
                        reviewid CDATA #IMPLIED
                        rating CDATA #IMPLIED
                        mdate CDATA #IMPLIED
>

<!ELEMENT inproceedings (%field;)*>
<!ATTLIST inproceedings key CDATA #REQUIRED
                        mdate CDATA #IMPLIED
>

<!ELEMENT proceedings   (%field;)*>
<!ATTLIST proceedings   key CDATA #REQUIRED
                        mdate CDATA #IMPLIED
>

<!ELEMENT book          (%field;)*>
<!ATTLIST book          key CDATA #REQUIRED
                        mdate CDATA #IMPLIED
>

<!ELEMENT incollection  (%field;)*>
<!ATTLIST incollection  key CDATA #REQUIRED
                        mdate CDATA #IMPLIED
>

<!ELEMENT phdthesis     (%field;)*>
<!ATTLIST phdthesis     key CDATA #REQUIRED
                        mdate CDATA #IMPLIED
>

<!ELEMENT mastersthesis (%field;)*>
<!ATTLIST mastersthesis key CDATA #REQUIRED
                        mdate CDATA #IMPLIED
```

Figura 17: Screenshot del archivo dblp.dtd

De la definición se puede concluir que el archivo dblp.xml tiene una estructura muy flexible y básicamente consiste en una larga lista de registros de publicaciones, los cuales se dividen en alguno de los tipos descritos en la Tabla 2: Descripción de los Tipos de Publicaciones.

Tabla 2: Descripción de los Tipos de Publicaciones

Tipo	Descripción
article	Artículo de un journal o revista.
book	Libro con una editorial explícita.
incollection	Parte de un libro que tiene su propio título.
inproceedings	Artículo en las actas de sesiones (proceedings) de una conferencia.
mastersthesis	Tesis de maestría o proyecto fin de carrera.
phdthesis	Tesis de doctorado.
proceedings	Actas de sesiones (proceedings) de una conferencia.
www	Referencias a páginas Web o Homepages de algunos autores.

Cada registro contiene por lo menos un atributo "key", el cual es un identificador único. Estos registros pueden contener 0 o más de alguno de los campos mencionados en la Tabla 3: Descripción de los Campos de Publicaciones[4].

Tabla 3: Descripción de los Campos de Publicaciones

Campo	Descripción	Usado
address	Usualmente la dirección asociada al publisher.	Sí
author	Autor de la publicación.	Sí
booktitle	Título del libro al cual pertenece la publicación.	Sí
cdrom	Referencia local a un archivo con contenidos.	
cite	Cita a un registro (key de otra publicación)	
crossref	Referencia cruzada (key de otra publicación).	
chapter	Número del capítulo al cual pertenece la publicación.	Sí
editor	Editor (persona) de la publicación.	Sí
ee	Link a un abstract o a un texto completo de la publicación.	Sí
isbn	Código ISBN de la publicación.	Sí
journal	Journal al cual pertenece la publicación.	Sí
month	Mes de la publicación.	Sí
note	Información adicional de ayuda al lector.	Sí
number	Número de la publicación.	Sí
pages	Páginas de la publicación.	Sí
publisher	Publisher de la publicación.	Sí
school	Institución asociada a la publicación (para tesis).	Sí
series	Serie a la cual pertenece la publicación.	Sí
title	Título de la publicación.	Sí
url	Link a una tabla de contenidos de la publicación.	Sí
volume	Volumen del Journal o del Libro.	Sí
year	Año de la publicación.	Sí

En dicha tabla además se muestra los campos que son almacenados en la base de datos de la herramienta desarrollada. Estos atributos abarcan el recomendado para describir una publicación según LaTeX satisfaciendo los requerimientos.

2.4 Ambiente de Desarrollo

Con el fin de desarrollar el sistema, ser revisó una serie de tecnologías y herramientas que pudiesen formar parte de la solución. El ambiente de desarrollo final que se ocupó lo componen los siguientes elementos.

[4] Parte de la descripción se realizó en base a los datos del archivo dblp.xml, por lo que eventualmente pueden discrepar de la semántica del campo.

2.4.1 Hardware

El hardware final se fue conseguido en forma particular en la segunda mitad del trabajo de memoria y consistió en un equipo portátil Dell Studio XPS 1340, cuyas características relevantes se describen a continuación en la Tabla 4: Ambiente de Desarrollo (Hardware).

Tabla 4: Ambiente de Desarrollo (Hardware)

Elemento	Descripción
Procesador	Intel Core 2 Duo P8600 (2.4GHz/1066Mhz FSB 3M L2 Cache)
Memoria	4GB, DDR3, 1067 MHz 2 Dimm
Disco Duro	160GB 5400 RPM SATA Hard Drive

2.4.2 Software

El software finalmente utilizado en el desarrollo de la memoria se detalla a continuación en la Tabla 5: Ambiente de Desarrollo (Software).

Tabla 5: Ambiente de Desarrollo (Software)

Elemento	Descripción
Sistema Operativo	Microsoft® Windows Vista™ Home Premium Service Pack 2
Base de Datos	PostgreSQL 8.4
IDE	NetBeans 6.7.1
Servidor de Aplicaciones	GlassFish 2.1
Diseñador de Base de Datos	DBDesigner 4

El sistema operativo fue elegido debido a que venía por defecto con el equipo adquirido y en él se podían instalar todas las herramientas necesarias. El resto de las herramientas fueron elegidas principalmente porque, a parte de ser gratuitos, facilitan tremendamente el trabajo de implementación y que en conjunto existe un alto nivel de integración.

Para la elección de la base de datos primaron las siguientes ventajas: es capaz de soportar las exigencias de la herramienta a desarrollar, es de fácil administración ya sea vía la aplicación cliente que viene por defecto o vía web, usando la herramienta phpPgAdmin que se puede instalar junto con la base de datos. También se integra con NetBeans.

El diseñador de bases de datos fue elegido debido a que, aparte de ser fácil de usar, existe un procedimiento para generar un script compatible con PostgreSQL que crea la estructura de la base de datos a partir del modelo diseñado.

La principal ventaja del IDE seleccionado es que posee procedimientos de generación automática de código a partir de la estructura de la base de datos. Entre

otros, de esta forma se generan clases que encapsulan la estructura de la base de datos, clases encargadas de los procedimientos de recuperación, creación, actualización y eliminación de datos. También se puede generar el código correspondiente a un sitio web en donde presentar la información almacenada. Otra ventaja notable es la integración con el servidor de aplicaciones[5] y la base de datos, lo que facilita la labor de implementación.

2.4.3 Tecnologías y Librerías

El lenguaje de programación utilizado fue Java. Las principales ventajas que primaron en la elección fueron: permite desarrollar la solución al problema, se contaba con experiencia, en la web hay muchas herramientas e información que soportan el desarrollo en este lenguaje.

Para el sitio web, se ocupó JSP junto con librerías JSF para la presentación. En la elección primó la ventaja de su integración con el lenguaje de programación y que el IDE de desarrollo presentaba varias funcionalidades de soporte. Además de las librerías estándar, en la herramienta desarrollada se ocuparon las librerías que se muestran en la Tabla 6: Ambiente de Desarrollo (Librerías Externas).

Tabla 6: Ambiente de Desarrollo (Librerías Externas)

Elemento	Descripción
Swing Application Framework	Framework para la Interfaz gráfica
Swing Layout Extensions	Extensiones para la Interfaz gráfica
JSF Extensions (Ajax)	Presentación Web
TopLink Essentials	Comunicación con la Base de Datos
postgresql-8.3-604.jdbc4	Comunicación con la Base de Datos
log4j-1.2.15	Manejo del registro.

Estas librerías utilizadas vienen con el IDE elegido para el desarrollo, a excepción de las 2 últimas que son descargables de la web.

2.4.4 Otras Herramientas y Tecnologías Revisadas

A continuación en la Tabla 7: Herramientas y Tecnologías Descartadas, se mencionan algunas herramientas y tecnologías que fueron revisadas, pero que fueron descartadas y no formaron parte del ambiente de desarrollo final.

[5] El instalador del IDE elegido viene con el servidor de aplicaciones GlassFish.

Tabla 7: Herramientas y Tecnologías Descartadas

Elemento	Descripción
eXist DB	Con el fin de aprovechar el formato XML en el cual venía el archivo con los registros de DBLP, se exploró el sistema de base de datos XML eXist. Si bien, se logró cargar en éste el archivo dblp.xml, el procedimiento de carga era extremadamente lento. Otro inconveniente que se detectó, es que se podían realizar algunas consultas específicas, pero cuando se trataba de consultas cuyo resultado era una gran cantidad de elementos, había problemas de memoria y no se obtenía el resultado. Dentro de la información disponible hasta ese entonces, no se encontró alguna configuración que reparara en ese hecho.
SQLite	Este sistema de base de datos es bastante simple y conveniente de usar para aplicaciones pequeñas, sin embargo, se descartó porque el sistema requería garantizar un buen funcionamiento para grandes cantidades de datos y contar con una buena mantenibilidad.
Microsoft SQL Server	Si bien, este sistema de base de datos funcionó para los requerimientos de volumen de datos, la versión disponible en el Repositorio del DCC no era compatible con Windows Vista (sistema operativo utilizado por el alumno).
lxml (Python)	Dentro de las herramientas exploradas, se llegó a lxml de Python, una librería para la lectura de archivos XML. Esta librería carga todo el árbol XML en memoria, por lo que fue descartada dada el volumen de datos que se requería manipular. En general se descartó Python, porque se contaba con más experiencia en Java.

3 Diseño de la Solución Propuesta

A continuación se presenta la arquitectura de la solución, el modelo de clases y el modelo de datos.

3.1 Arquitectura General

La herramienta desarrollada está compuesta por 3 partes fundamentales: una base de datos, una aplicación de migración de datos y un sitio web de reportes y consultas. La interacción de dichas partes se muestra en la Figura 18: Arquitectura General.

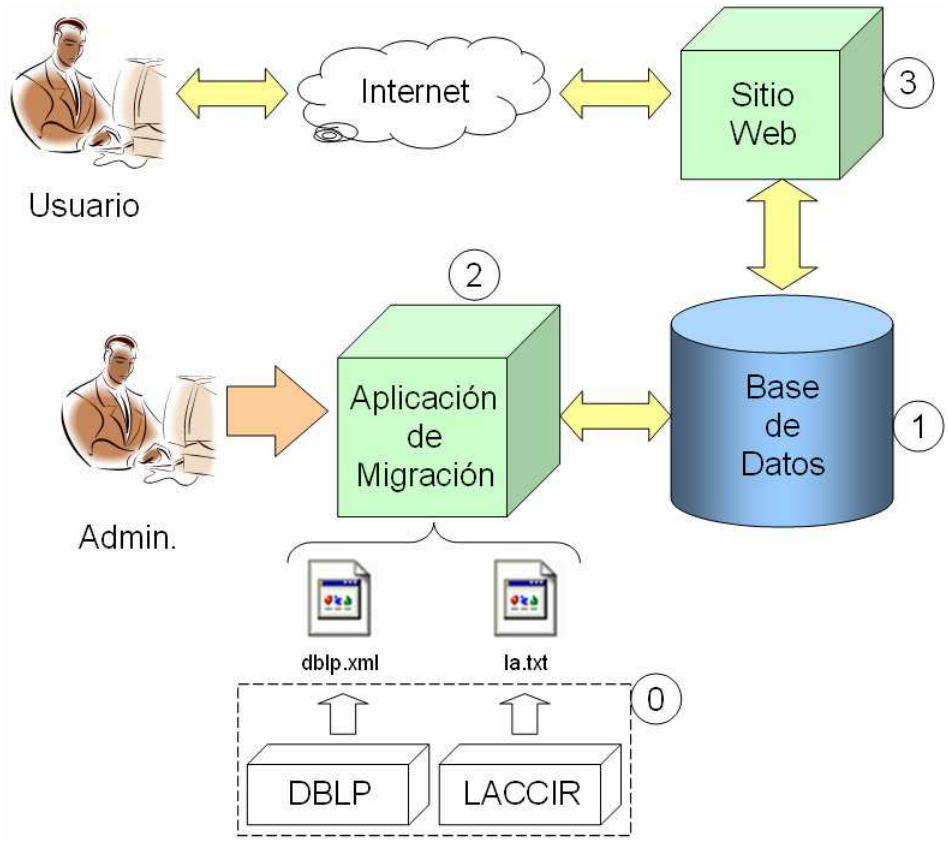


Figura 18: Arquitectura General

La primera parte corresponde a la base de datos, para la cual se utilizó PostgreSQL. En esta componente del sistema, se almacenan todos los datos que maneja la herramienta.

La segunda componente refiere a la aplicación de migración de datos. Esta parte del sistema interactúa con sistemas externos a la herramienta (corresponde a la parte 0 en la Figura 18: Arquitectura General). La aplicación toma datos provenientes del

sistema DBLP vía el archivo dblp.xml disponible en la web y migra los registros de publicaciones a la base de datos del sistema. Además, el procedimiento toma información proveniente desde LACCIR a través de una lista con los nombres que los autores a considerar en la migración. Este listado corresponde al nombre estándar que los investigadores usan para publicar sus trabajos.

En esta fase inicial de la herramienta y para el desarrollo de la misma, este listado se generó manualmente y sólo abarcó una pequeña muestra de investigadores. Se espera que en el futuro LACCIR cuente con este dato adicional para cada persona ahí registrada.

La tercera componente es el sitio web de reportes y consultas. Es la parte del sistema encargada de mostrar la información que se puede extraer desde la base de datos.

3.2 Modelo de Datos

El modelo de datos del sistema se presenta en la Figura 19: Modelo de Datos. Dicho modelo fue construido a partir de la información que debiese almacenar y procesar la herramienta según los requerimientos y de acuerdo a las indicaciones del profesor guía.

Este modelo se inspira en el esquema de la base de datos de investigadores de LACCIR, el cual se destaca por incluir las áreas de investigación de las personas y la institución de la que provienen (Ver en el Anexo el Modelo de Datos de LACCIR). Dicho concepto fue modificado y expandido, con el objetivo de que el nuevo modelo incorporara la estructura necesaria para incluir los datos recolectados desde DBLP y así considerar las publicaciones de cada investigador.

Dentro del modelo de datos propuesto, las tablas más importantes se muestran en la Tabla 8: Descripción de las Tablas Principales del Modelo de Datos, junto con una breve descripción. En el Anexo se encuentra el Diccionario de Datos del modelo con una explicación detallada de cada campo.

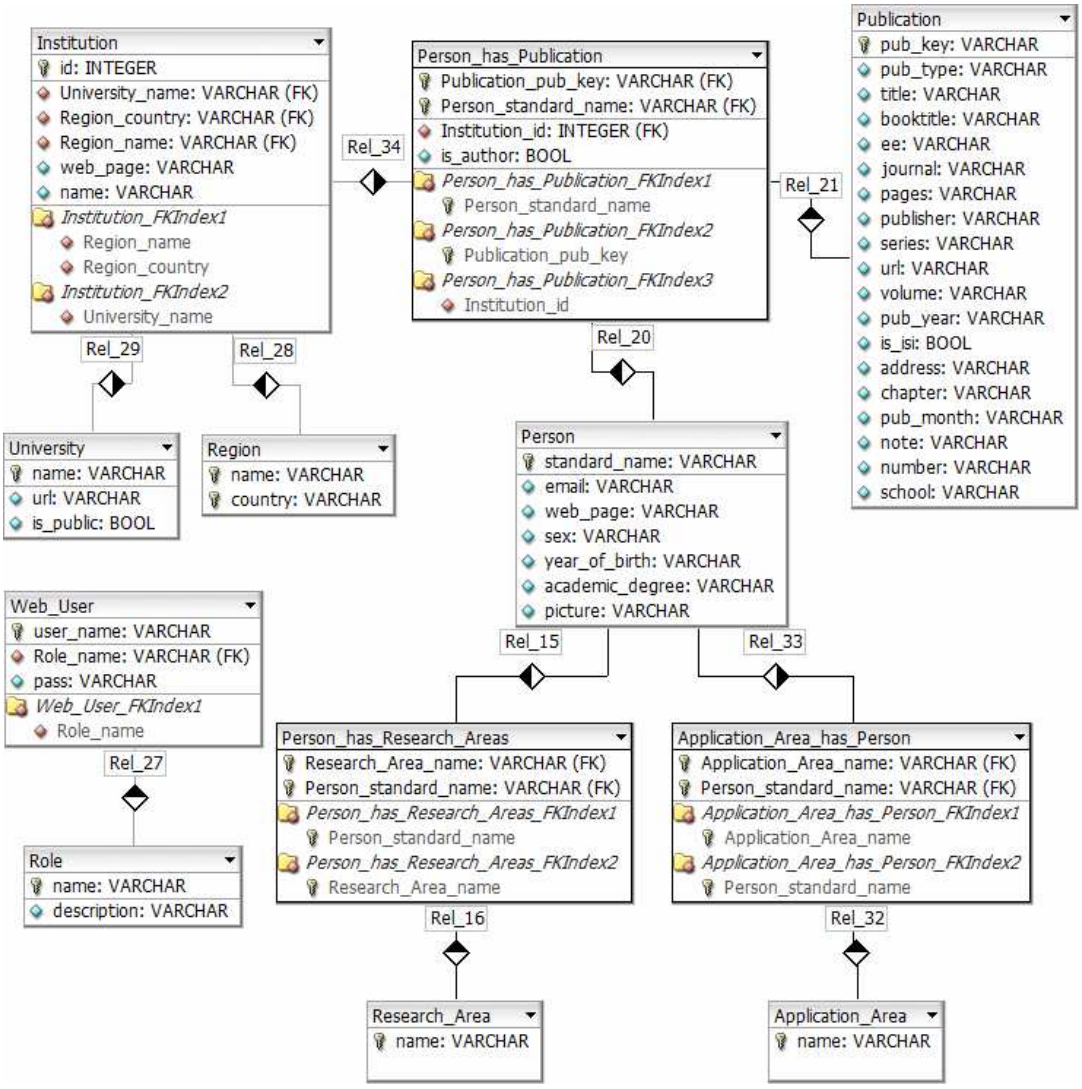


Figura 19: Modelo de Datos

Tabla 8: Descripción de las Tablas Principales del Modelo de Datos

Tabla	Descripción
Application_Area	Tabla con las áreas que un investigador dice trabajar.
Institution	Tabla con las instituciones registradas.
Person	Tabla con los autores y editores de las publicaciones.
Person_has_Publication	Tabla que almacena las relaciones (persona, publicación, institución).
Publication	Tabla con las publicaciones registradas en el sistema.
Region	Tabla con las regiones de cada institución.
Research_Area	Tabla con áreas de investigación predeterminadas.
University	Tabla con las universidades a las cuales pertenecen las instituciones.
WebUser	Tabla con los usuarios del sistema.
Role	Tabla con los roles de usuario para el sitio web

Una modificación importante al modelo original de LACCIR, es que el nombre del autor pasa a ser identificador de la persona y es exactamente igual al que dicho investigador usa para publicar sus trabajos. Aquí se consideró válido suponer que el nombre para publicar es de responsabilidad de cada persona. De esta forma si se genera alguna incompatibilidad o conflicto de nombres, cada investigador debiese hacer las gestiones para modificar dicho nombre en los registros bibliográficos y así indicar cuáles son efectivamente sus trabajos.

La segunda modificación importante es el modelado de la institución. En el nuevo modelo, se le otorga más importancia haciendo que participe en la relación de publicar. Con esta nueva perspectiva, el investigador se independiza de la institución quedando cubierto el hecho de que una persona pueda cambiarse y publicar trabajos en distintos lugares. En esta misma línea, la región pasa a ser exclusiva de la institución y no de la persona. De esta forma, la institución pasa a ser "la institución a la cual pertenecía el autor al momento de publicar".

Otra modificación al modelo original de LACCIR es el área de aplicación y de investigación de cada persona. En el nuevo modelo, ambas son independientes del investigador y eventualmente común entre varias personas.

En cuanto a las publicaciones, aparte de los datos básico para describirlas bibliográficamente, se le incorporó un flag para indicar si se trata de una publicación ISI o no. De esta forma el modelo puede distinguir este tipo de publicaciones.

La migración masiva de datos implementada en este trabajo de memoria escribe en la tabla Publication (a excepción del flag is_isi), la tabla Person_has_Publication (a excepción de la institución) y la tabla Person (sólo el campo standard_name). El resto de los datos y tablas del modelo, son ingresados a mano a través de la herramienta web.

3.3 Modelo de Clases

Para desarrollar la solución al problema se requirió implementar clases que modelaran la información y que la manipularan, clases que realizaran la migración masiva y clases que realizaran consultas a la base de datos y que mostraran resultados.

Para el modelo de la información y la manipulación de datos, se usó la interfaz JPA[6] la cual facilita las labores de interacción con la base de datos. Siguiendo dicho esquema de desarrollo, se generaron Entity Beans y clases controladoras encargadas de la persistencia.

En la implementación de la aplicación cliente encargada de la migración masiva de datos se utilizó Swing Application Framework[7], un Framework de desarrollo de

[6] La interfaz JPA corresponde al paquete *javax.persistence*
[7] El Swing Application Framework corresponde al paquete *org.jdesktop.application*

aplicaciones de escritorio que viene con NetBeans. Gracias a este Framework extendiendo las clases correspondientes, se abstrae la presentación de la lógica y los distintos procedimientos pueden ser ejecutados en background.

Para el proceso de migración masiva propiamente tal, se utilizó la interfaz estándar SAX[8] la cual facilita la manipulación de archivos XML. La gran ventaja de esta interfaz es que permite realizar un parsing por eventos, lo cual es idóneo para el archivo dblp.xml debido a su gran tamaño.

Para la lógica de presentación del sitio web se utilizó JSF, un Framework de desarrollo web que usa JSP. En este Framework, se generan clases que controlan el comportamiento las páginas, es así como se generó clases controladoras que utilizan las distintas clases de interacción con la base de datos y que preparan los datos para ser presentados al usuario.

A continuación se muestra una serie de diagramas de las principales clases que componen todo el sistema y su descripción detallada.

3.3.1 Entity Beans

Estas clases son las que representan las entidades que interactúan en el sistema. Ver Figura 20: Diagrama de Clases: Entity Beans. Estas entidades están bajo el paquete *cc69f.jpa.entity*.

[8] La interfaz SAX corresponde al paquete *org.xml.sax*

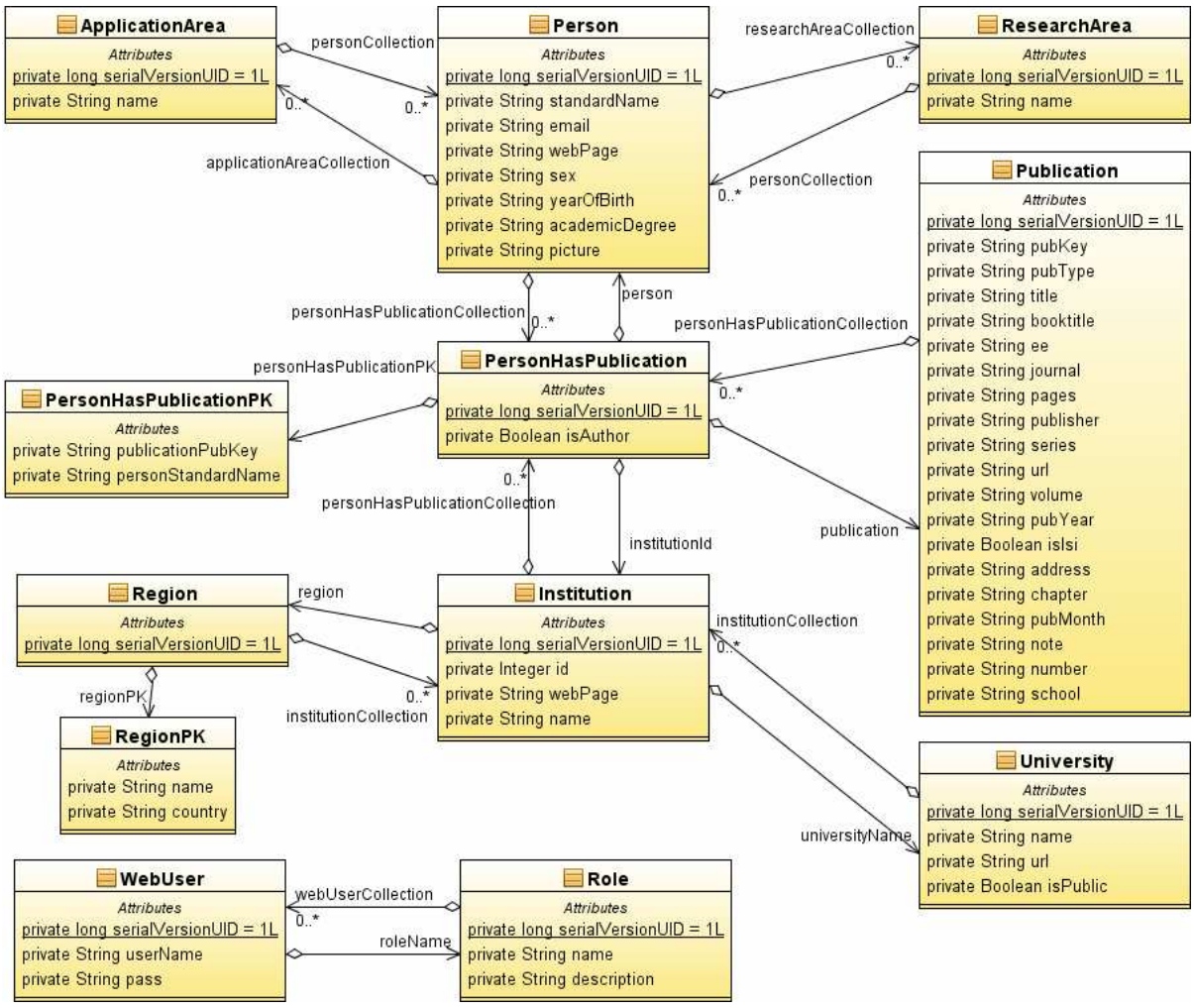


Figura 20: Diagrama de Clases: Entity Beans

Estas clases fueron generadas automáticamente a partir del modelo de datos del sistema usando una funcionalidad de NetBeans. El procedimiento consiste en que el IDE se conecta a la base de datos y, a partir del esquema de la misma, genera clases que representan a cada entidad del sistema. Es así como las tablas principales se convierten en clases con variables de instancia iguales a los atributos de las tablas. También se establecen las asociaciones respectivas gracias a las llaves foráneas de la base de datos. De esta forma, se encapsulan los datos y las relaciones entre las tablas.

Las entidades generadas de esta forma son usadas para la manipulación de los datos en todas las partes que componen la herramienta desarrollada. A estas clases generadas automáticamente, se les añadió los constructores vacíos y se les modificó el método “toString()” en cada clase. Esto último con fines netamente estéticos. La persistencia de estos datos está a cargo de unas clases controladoras que se describen a continuación.

3.3.2 Controladores JPA

Como se dijo anteriormente, estas clases están encargadas de la persistencia de las entidades. Estas clases también son generadas por NetBeans a partir de las clases entidades. Ver Figura 21: Diagrama de Clases: Controladores JPA (local).

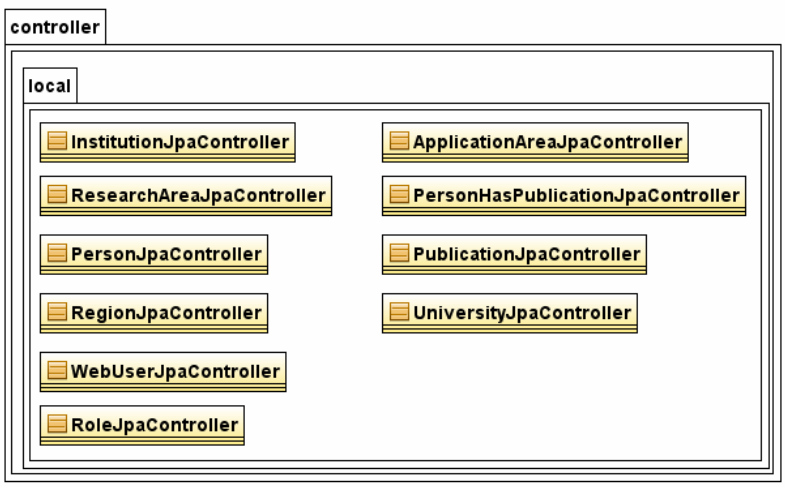


Figura 21: Diagrama de Clases: Controladores JPA (local)

Como se puede ver, se cuenta con un controlador para cada entidad. Entre otros, cada uno de ellos tiene métodos para la Creación, Edición, Destrucción, Búsqueda para las entidades en las que se especializan. Es en estos métodos en donde se mantiene la consistencia de los datos para cada una de las entidades, uno de los requisitos para la herramienta.

Dentro del sistema se encuentran 2 tipos de controladores:

- Los controladores locales, correspondientes al paquete *cc69f.jpa.controller.local*. Manejan la persistencia localmente, es decir, se conectan directamente a la base de datos para realizar las consultas. Estos controladores son utilizados por la aplicación de migración de datos.

- Los controladores del lado del servidor, correspondientes al paquete *cc69f.jpa.controller.server*. Interactúan con la base de datos usando el pool de conexiones del servidor de aplicaciones. Estos controladores, son utilizados por el sitio web de reportes y consulta.

En el caso del sitio web, para la interacción con la base de datos se extendieron clases generadas automáticamente. Esto se hizo con el fin de obtener más funcionalidades en cuanto a la búsqueda de personas, publicaciones e instituciones. Ver Figura 22: Diagrama de Clases: Controladores JPA (server).

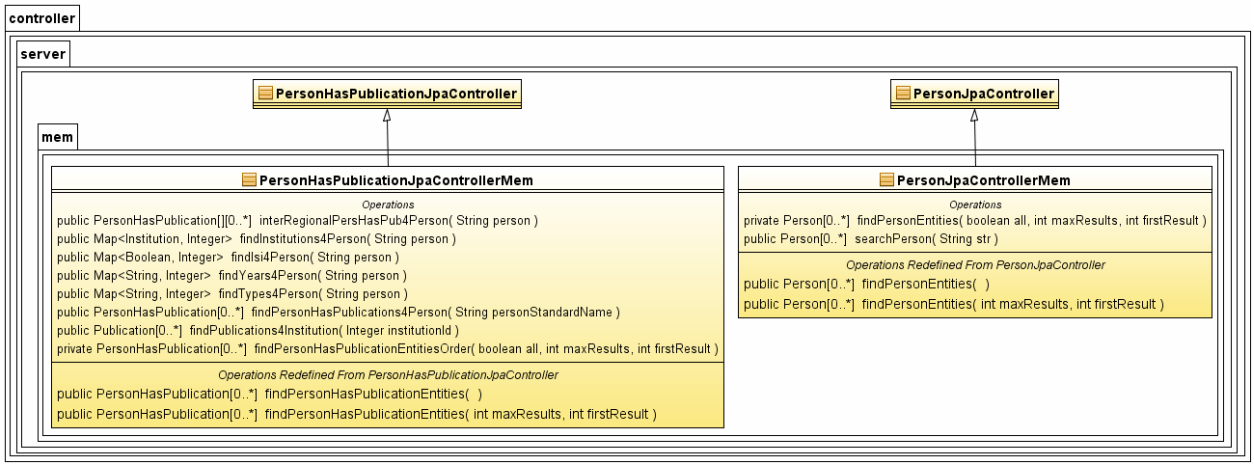


Figura 22: Diagrama de Clases: Controladores JPA (server)

Para satisfacer las consultas requeridas se extendió 2 clases:

- La clase *PersonJpaControlerMem*, fue especializada con la funcionalidad de buscar a personas dado un *String* cualquiera.

- La clase *PersonHasPublicationJpaControlerMem*, se especializó para la búsqueda publicaciones por persona o por institución. También es la encargada de buscar las instituciones a la cual ha pertenecido una persona y en general recolecta toda la información para generar estadísticas personales.

A ambas clases se les modificó el método por defecto para buscar entidades, con el fin de que los resultados arrojados viniesen ordenados. Este par de clases conforman el paquete *cc69f.jpa.controller.server.mem* y son utilizadas por las clases que controlan el comportamiento en el sitio web para realizar consultas a la base de datos. Estas últimas clases del comportamiento del sitio, se detallarán más adelante.

3.3.3 Presentación de la Aplicación de Migración

En la Figura 23: Diagrama de Clases: Presentación de la Aplicación de Migración, se muestran las clases encargadas de la presentación al usuario y del funcionamiento general de la aplicación cliente.

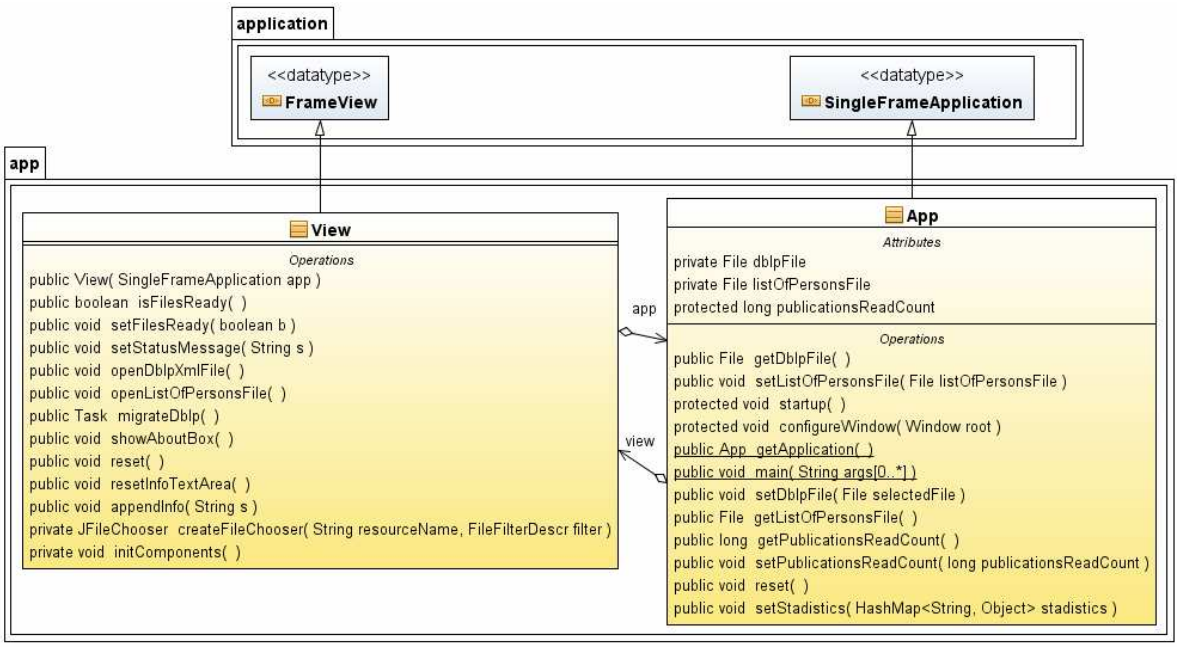


Figura 23: Diagrama de Clases: Presentación de la Aplicación de Migración

Del Swing Application Framework se consideró las clases *SingleFrameApplication* y *FrameView* de las cuales extienden las clases A*pp* y *View* respectivamente:

- La clase *View*, como su nombre lo indica, es la encargada de la interfaz de usuario: la instanciación de los componentes y del comportamiento de los mismos.

- La clase *App*, es la encargada de mantener el estado de la aplicación y de ir modificando interfaz de usuario para el despliegue de los datos correspondientes.

En esta aplicación conformada por estas 2 clases, el usuario indica la ruta al archivo dblp.xml y la ruta a una lista con los investigadores de interés para migrar sus publicaciones. Desde aquí se ejecuta la tarea de migración de datos que realiza el parsing del XML.

3.3.4 Parsing del Archivo dblp.xml

En la Figura 24: Diagrama de Clases: Parsing del Archivo dblp.xml, se muestran todas las clases encargadas del procedimiento de lectura del archivo dblp.xml, el procesamiento de la información y de la persistencia de los datos de publicaciones.

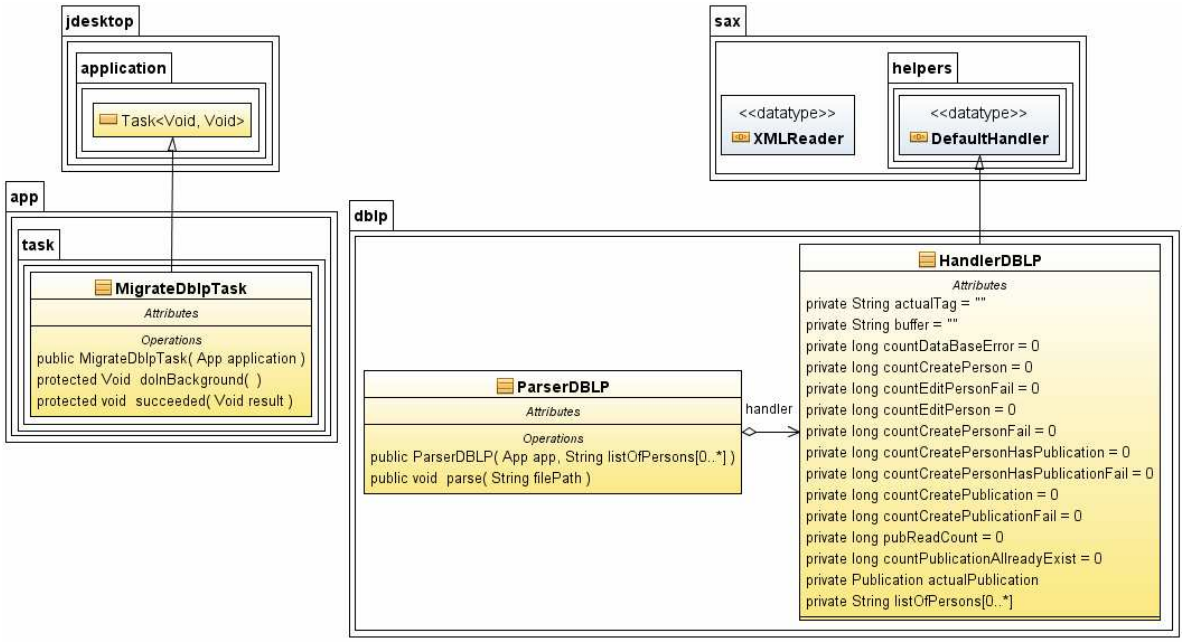


Figura 24: Diagrama de Clases: Parsing del Archivo dblp.xml

La clase *MigrateDblpTask* extiende a la clase abstracta *Task* del Swing Application Framework, y es la responsable de realizar la tarea de la migración de datos desde el archivo dblp.xml hacia la base de datos considerando las personas de interés.

Lo primero que se hace en dicha tarea, es cargar en memoria el listado de personas a considerar en el parsing mediante la lectura del archivo indicado por el usuario.

Junto con esto, inicializa un objeto *ParserDBLP* el cual es el encargado del parsing del archivo dblp.xml y la persistencia de los datos de interés. Las personas que se consideran en el procedimiento son pasadas como parámetros en un arreglo de *String* al momento de crear el objeto parser.

Para el parsing del archivo dblp.xml, el objeto *ParserDBLP* crea 2 objetos especializados:

- Un handler de la clase *HandlerDBLP* extendido de la clase *DefaultHandler*, que almacena la lógica detrás del esquema del archivo dblp.xml
- Un reader de la clase *XMLReader*[9] encargado de la lectura del XML.

El reader va detectando los elementos XML y genera eventos para cada ocurrencia. Dichos eventos son capturados por el handler y según el tipo de evento y los datos leídos, va persistiendo los registros en el sistema para las personas indicadas en los parámetros.

[9] Ambas clases, *XMLReader* y *DefaultHandler*, pertenecen a la interfaz de SAX

3.3.5 Presentación del Sitio Web de Reportes y Consulta

Para la Presentación del Sitio Web, se utilizó otra funcionalidad el IDE NetBeans, la cual genera páginas JSP para la lectura, creación, actualización y eliminación de datos a partir de los Entity Beans que maneja el sistema. Estas páginas, fueron la base para la construcción de los reportes que genera la herramienta desarrollada.

Junto con las páginas JSP, Netbeans genera automáticamente clases *Controller* para cada entidad, las cuales están encargadas del comportamiento de las distintas páginas. Cada página JSP invoca métodos de estas clases los cuales dan funcionalidad a cada página.

Además, existe una clase *Converter* asociado a cada *Controller*. Estas clases se encargan de la manipulación de información a través de las distintas páginas que componen al sitio. Cada *Converter* se especializa en una entidad en particular.

Todas estas clases, encargadas de la lógica de presentación del sitio web, conforman el paquete *cc69f.jsf.classes* y se muestran en la Figura 25: Diagrama de Clases: Presentación del Sitio Web.

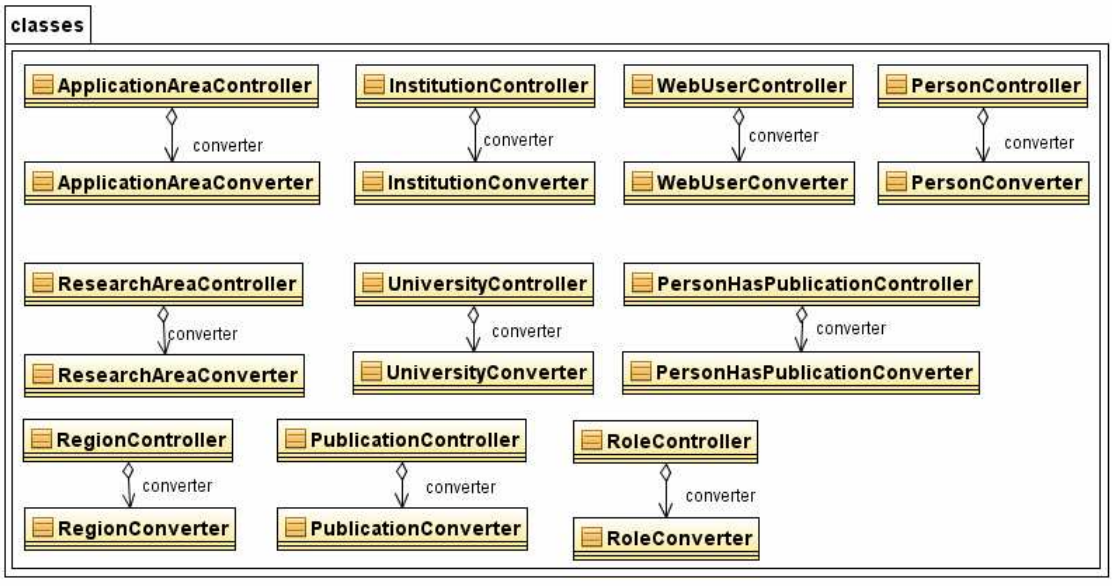


Figura 25: Diagrama de Clases: Presentación del Sitio Web

Para satisfacer el requerimiento de las consultas a realizarse en el sistema, se extendió algunas de las clases generadas automáticamente con el fin de agregar comportamiento a lo ya hecho por defecto. Estas clases conforman parte de la extensión "*mem*" del paquete original y se muestran en la Figura 26: Diagrama de Clases: Controladores de Presentación Modificados.

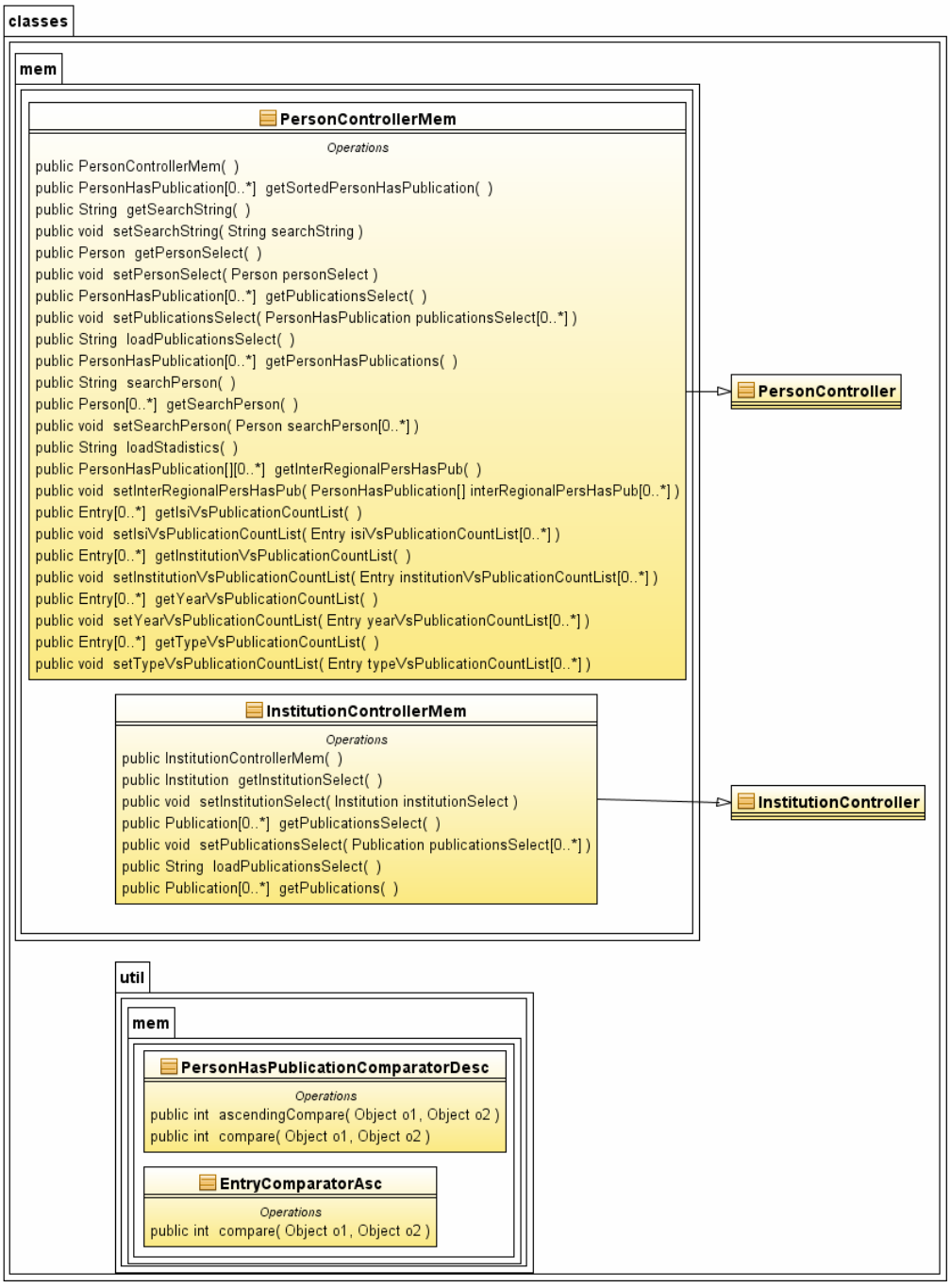


Figura 26: Diagrama de Clases: Controladores de Presentación Modificados

Específicamente se extendió 2 clases *Controller*:

- La clase *PersonControllerMem* extiende al controlador original, para incorporar la lógica de presentación de la búsqueda de personas y la presentación de la consulta de publicaciones por persona.

- La clase *InstitutionControllerMem* incorpora la lógica para la consulta de publicaciones por institución.

También se implementó clases para ordenar los resultados obtenidos y así poder presentarlos mejor al navegante, estas clases se encuentran en la extensión "*util.mem*".

Por otro lado, también se implementó una clase para administrar información para la aplicación. Ver Figura 27: Diagrama de Clases: Session.

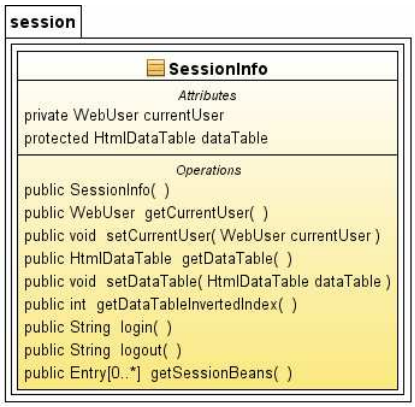


Figura 27: Diagrama de Clases: Session

La clase *SessionInfo*, está diseñada para manejar la información general de navegación por el sitio. Tiene implementado una funcionalidad de login, que toma en cuenta a los usuarios registrados en la base de datos, pero que no está funcional dentro de la aplicación.

4 Descripción de la Herramienta Desarrollada

A continuación se describe a las distintas partes que conforman la herramienta desarrollada y de cómo el usuario debe interactuar.

4.1 Instalación y Puesta en Marcha

Para probar el sistema se sugiere seguir los siguientes pasos:

1. Instalar la base de datos y el IDE ocupados en este trabajo de memoria. Ver el ambiente de desarrollo descrito anteriormente.
    a. Al momento de instalar el IDE, asegurarse de que en el instalador venga incluido el servidor de aplicaciones.

    b. Al momento de instalar la base de datos, asegurarse de que exista la siguiente cuenta de usuario:
    username: postgres
    pasword pass.

2. Una vez instalada la base de datos, crear una base de datos de nombre “cc69f” y con codificación “utf-8”.

3. Bajar el script de creación de la base de datos desde la siguiente dirección.
    http://www.dcc.uchile.cl/~mliulion/cc69f/create_scrip.sql

4. Ejecutar el script para esa base de datos. Para esto se recomienda usar la plataforma phpPgAdmin que viene con la base de datos.

5. Crear un conjunto de conexiones hacia la base de datos, en el servidor de aplicaciones. Esto se puede hacer en la consola de administración[10] en la sección Recursos> JDBC> Conjuntos de conexiones.
    a. El tipo de conexión debe ser “javax.sql.DataSource” y el proveedor “PostgreSQL”.
    b. Asegurarse de que el “Nivel de aislamiento” esté “Garantizado” y configurar las “Propiedades Adicionales”, como se muestra en la Figura 28: Configuración de GlassFish (Propiedades Adicionales)

[10] La consola de administración de GlassFish por defecto se encuentra en http://localhost:4848. La cuenta por defecto es: username “admin” y password “adminadmin”.

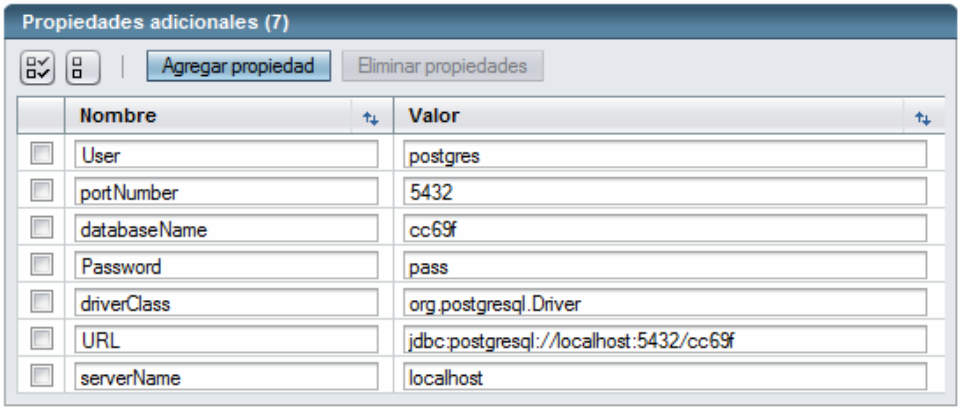


Figura 28: Configuración de GlassFish (Propiedades Adicionales)

6. Crear un data source “jdbc/cc69f” en el servidor de aplicaciones asociado al conjunto de conexiones del paso anterior. Esto se puede hacer en la consola de administración en la sección “Recursos> JDBC> Recursos JDBC”.

7. Bajar el archivo comprimido con el entorno de trabajo NetBeans desde:
   http://alumnos.dcc.uchile.cl/~mliulion/cc69f/NetBeansProjects.zip

8. Reemplazar el entorno de trabajo de NetBeans. Se sugiere respaldar el directorio original, en caso de haber proyectos anteriores.

9. Bajar un listado de autores, un dblp.xml y un dblp.dtd. Estos dos últimos archivos son descargables desde DBLP (Ver Anexo la Obtencion del Archivo dblp.xml). La lista de autores utilizada, se encuentra en la siguiente dirección:
   http://www.dcc.uchile.cl/~mliulion/cc69f/la36.txt

10. Al abrir NetBeans se podrán ver y ejecutar los proyectos El proyecto de nombre “cc69f-app” corresponde a la Aplicación de Migración. El proyecto “cc69-war” corresponde al Sitio Web VerFigura 29: Proyectos en NetBeans

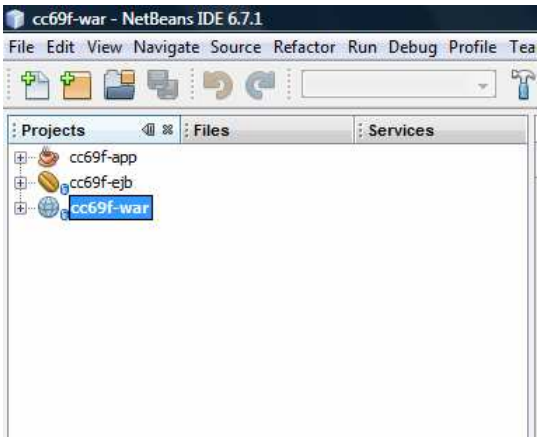


Figura 29: Proyectos en NetBeans

4.2 Aplicación de Migración de Datos

Como se mencionó en la Arquitectura General, el procedimiento de migración desarrollado se realiza mediante una aplicación de escritorio que toma datos provenientes del repositorio DBLP y de ellos extrae información para los autores latinoamericanos y la almacena en la base de datos del sistema.

Dado esto, los inputs del procedimiento son 2 archivos: un archivo dblp.xml[11] y otro archivo con la lista de personas de interés para el sistema. Este listado de autores consiste en un archivo de texto plano con el nombre de una persona por línea. Dichos nombres corresponde al nombre estandarizado registrado en DBLP[12].

Ambos archivos son indicados mediante un recuadro de exploración de archivos usando los botones correspondientes. Cuando el usuario elige los archivos, la aplicación responde mostrando una información básica de los archivos seleccionados de esta forma Elegidos ambos archivos se habilita el botón para ejecutar la migración.

Durante la ejecución de la migración, al usuario se le muestra una información dinámica indicándole que se está realizando el parsing del archivo dblp.xml. Específicamente se activa una barra de proceso junto con un icono móvil “busy” y se mantiene actualizado un contador con las publicaciones leídas hasta el momento. Ver Figura 30: Aplicación de Carga de Datos durante la carga.

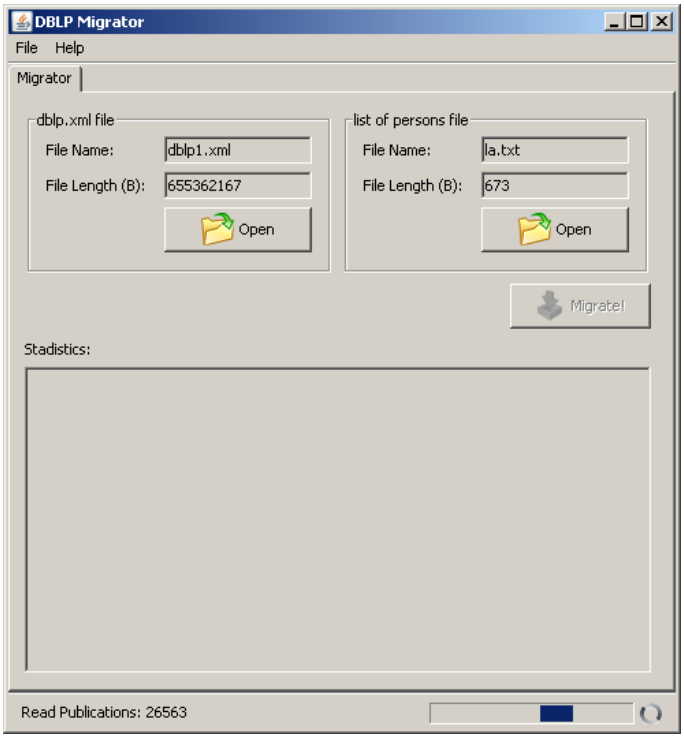


Figura 30: Aplicación de Carga de Datos durante la carga

[11] Junto con el archivo dblp.xml debe estar el archivo dblp.dtd en el mismo directorio para que se pueda realizar el parsing.
[12] Para el presente trabajo, este último archivo fue confeccionado manualmente.

Terminado el proceso, en la base de datos se habrán cargado los datos de las nuevas publicaciones, las personas asociadas a las mismas y se habrán creado las relaciones necesarias que indican si una persona es autor o editor de dichas publicaciones. Al finalizar, en la aplicación se le muestra al usuario un pequeño resumen de lo realizado en el proceso.

4.3 Sitio Web de Reportes y Consultas

Esta es la tercera componente de la Arquitectura General y, como se mencionó anteriormente, es la parte del sistema encargada de mostrar la información posible de recabar desde la base de datos. A continuación se describen las distintas secciones de la herramienta web.

4.3.1 Pantalla Principal

El sitio web tiene una pantalla principal que cumple el rol de mostrar al navegante las distintas partes del sistema (Ver Figura 31: Pantalla Principal).

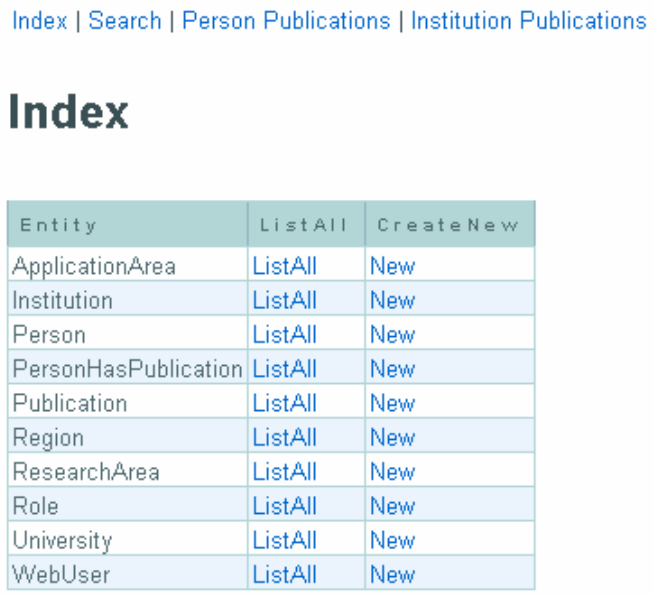


Figura 31: Pantalla Principal

En la parte superior, se ve un menú el cual permanece constante en toda la aplicación. En este menú, se puede ir a la Pantalla Principal (Index), a una Búsqueda de Personas (Search) y a la consulta de Publicaciones de una Persona (Person Publications) o bien de una Institución (Institution Publications). Estas secciones se describirán más adelante.

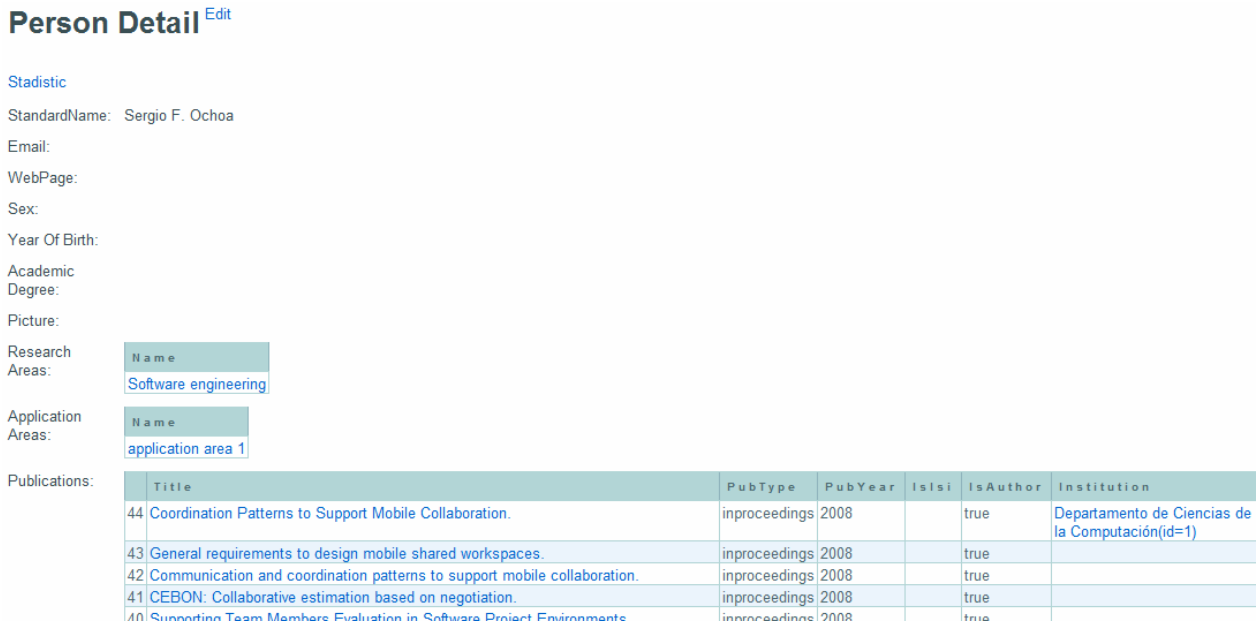
En la parte central se despliega un menú con los tipos de entidades que conforman el sistema, links para acceder a un listado de cada uno y links para la creación de una nueva entidad.

4.3.2 Reportes

A continuación se describen las pantallas que corresponden a los principales reportes, los cuales muestran los detalles con la información almacenadas en el sistema.

4.3.2.1 Detalles de una Persona

En el reporte de los detalles de una persona se muestra toda la información relativa a una persona. Ver Figura 32: Detalles de una Persona.



Person Detail Edit

Stadistic

StandardName: Sergio F. Ochoa

Email:

WebPage:

Sex:

Year Of Birth:

Academic Degree:

Picture:

Research Areas:

Name
Software engineering

Application Areas:

Name
application area 1

Publications:

	Title	PubType	PubYear	IsIsi	IsAuthor	Institution
44	Coordination Patterns to Support Mobile Collaboration.	inproceedings	2008		true	Departamento de Ciencias de la Computación(id=1)
43	General requirements to design mobile shared workspaces.	inproceedings	2008		true	
42	Communication and coordination patterns to support mobile collaboration.	inproceedings	2008		true	
41	CEBON: Collaborative estimation based on negotiation.	inproceedings	2008		true	
40	Supporting Team Members Evaluation in Software Project Environments	inproceedings	2008		true	

Figura 32: Detalles de una Persona

Este reporte es generado cada vez que el navegante hace clic en el link del nombre de una persona. Los datos de este reporte se describen en la Tabla 9: Datos del Reporte de Detalles de una Persona.

Tabla 9: Datos del Reporte de Detalles de una Persona

Dato	Descripción
StandardName	Nombre estándar de la persona.
Email	Email de la persona.
WebPage	Página web de la persona.
Sex	Sexo de la persona.
YearOfBirth	Año de nacimiento de la persona.
AcademicDegree	Grado académico de la persona.
Picture	Path a una imagen de la persona.

Después de los datos básicos y en el caso de que esté registrado en la base de datos, se muestran tablas con las áreas de investigación y las áreas de aplicación la persona. El nombre de cada área es un link para acceder a un listado de las personas las cuales comparten dicha área.

Finalmente se muestra un listado de las publicaciones la persona en cuestión. Los datos que se muestran son los descritos en la Tabla 10: Datos del Reporte de Detalles de una Persona (Publicaciones).

Tabla 10: Datos del Reporte de Detalles de una Persona (Publicaciones)

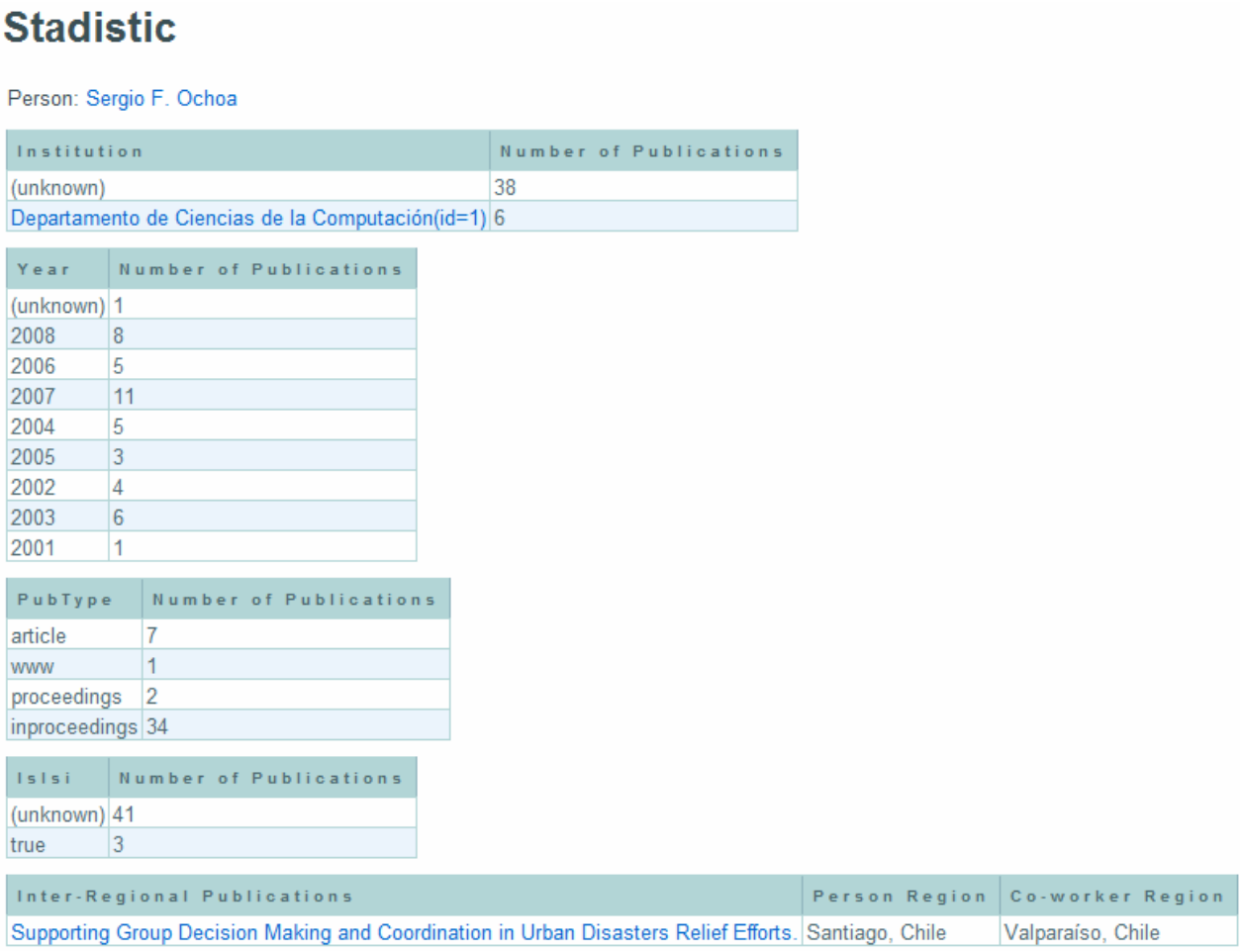
Dato	Descripción
Title	Título con el cual está registrada la publicación.
PubType	Tipo de publicación.
PubYear	Año de la publicación.
IsIsi	Flag que indica si la publicación es ISI o no. Si el valor es true, la publicación es ISI; si es false, no es ISI.
IsAuthor	Flag que indica si la persona es autor o editor. Si el valor es true, es autor; si es false, es editor.
Institution	Nombre de la Institución y su código.

Siguiendo los respectivos links, en este reporte el navegante puede desplegar la información detallada de una publicación o bien la de la institución a la cual estaba asociado al momento de publicar cada trabajo.

Tanto en este reporte, como en todos los demás, los listados de publicaciones vienen ordenados por año y por tipo de publicación.

4.3.2.2 Detalles de una Persona (Estadísticas)

A partir del módulo de Detalles de una Persona, el navegante puede acceder a un módulo que calcula algunas estadísticas para el investigador en cuestión. Ver Figura 33: Detalles de una Persona (Estadísticas).



Stadistic

Person: Sergio F. Ochoa

Institution	Number of Publications
(unknown)	38
Departamento de Ciencias de la Computación(id=1)	6

Year	Number of Publications
(unknown)	1
2008	8
2006	5
2007	11
2004	5
2005	3
2002	4
2003	6
2001	1

PubType	Number of Publications
article	7
www	1
proceedings	2
inproceedings	34

IsIsi	Number of Publications
(unknown)	41
true	3

Inter-Regional Publications	Person Region	Co-worker Region
Supporting Group Decision Making and Coordination in Urban Disasters Relief Efforts.	Santiago, Chile	Valparaíso, Chile

Figura 33: Detalles de una Persona (Estadísticas)

Dentro de los indicadores podemos encontrar:

- Cantidad de publicaciones bajo una Institución determinada
- Cantidad de publicaciones por año
- Cantidad de publicaciones por tipo
- Cantidad de publicaciones ISI
- Publicaciones en donde se involucra a una persona de otra región

4.3.2.3 Detalles de una Publicación

Este reporte condensa toda la información de la base de datos relativa a una publicación. Ver Figura 34: Detalle de una Publicación.

Publication Detail Edit

PubKey: conf/criwg/NeyemOP08
PubType: inproceedings
Title: Coordination Patterns to Support Mobile Collaboration.
Booktitle: CRIWG
Ee: http://dx.doi.org/10.1007/978-3-540-92831-7_21
Journal:
Pages: 248-265
Publisher:
Series:
Url: db/conf/criwg/criwg2008.html#NeyemOP08
Volume:
PubYear: 2008
Isisi:
Address:
Chapter:
PubMonth:
Note: .
Number:
School:
Persons:

Person	IsAuthor	Institution (Id)	
Sergio F. Ochoa	true	Departamento de Ciencias de la Computación(id=1)	Edit

Figura 34: Detalle de una Publicación

La descripción de cada elemento del reporte se encuentra en la Tabla 11: Datos del Reporte de Detalles de una Publicación.

Tabla 11: Datos del Reporte de Detalles de una Publicación

Dato	Descripción
PubKey	Identificador de la publicación dentro del sistema de DBLP.
PubType	Tipo de publicación.[13]
Title	Título con el cual está registrada la publicación.
Booktitle	Título del libro en donde se publicó el trabajo.
Ee	Link a un abstract o a la publicación completa.
Journal	Journal de la publicación.
Pages	Cantidad de páginas de la publicación.
Publisher	Publisher de la publicación.
Series	Serie de la publicación.
Url	Link (relativo al sistema DBLP) a una tabla con los contenidos de la publicación.
Volume	Volumen de la publicación.
PubYear	Año de la publicación.
IsIsi	Flag que indica si la publicación es ISI o no. Si el valor es true, la publicación es ISI; si es false, no es ISI.
Address	Dirección asociada a la publicación.
Chapter	Capítulo de la publicación.
PubMonth	Mes de la publicación.
Note	Nota asociada a la publicación.
Number	El número de la publicación.
School	Escuela asociada a la publicación.

Cabe destacar que todos estos datos corresponden a los migrados desde el archivo dblp.xml, a excepción del flag “IsIsi”, que no figura en la definición de dicho archivo.

Junto con los datos anteriores, se despliega un listado de los autores y editores (si es que los hubiese) de dicha publicación. Para cada persona asociada a la publicación en cuestión, se muestran los datos descritos en la Tabla 12: Datos del Reporte de Detalles de una Publicación (Autores y Editores).

Tabla 12: Datos del Reporte de Detalles de una Publicación (Autores y Editores)

Dato	Descripción
Person	Nombre estándar de la persona.
Institution	Nombre de la Institución y su código.
IsAuthor	Flag que indica si la persona es autor o editor. Si el valor es true, es autor; si es false, es editor.

[13] Ver Tabla 2: Descripción de los Tipos de Publicaciones

4.3.2.4 Detalles de una Institución

En este reporte, al navegante se le presenta condensada toda la información relativa a una institución. Ver Figura 35: Detalles de una Institución.

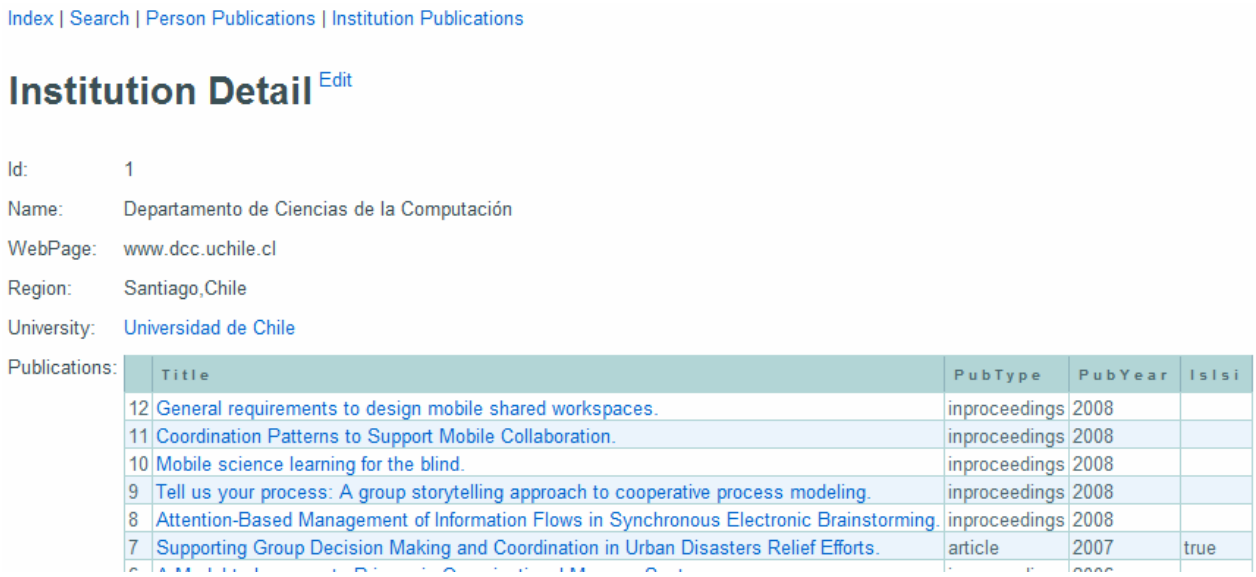


Figura 35: Detalles de una Institución

Los datos que se muestran en pantalla se describen en la Tabla 13: Datos del Reporte de Detalles de una Institución.

Tabla 13: Datos del Reporte de Detalles de una Institución

Dato	Descripción
Id	Identificador único de la institución.
University	Nombre de la universidad de la institución.
Region	Nombre y país de la región de la institución.
WebPage	Página web de la institución.
Name	Nombre de la institución.

Junto con los datos anteriores, también se muestra un listado de las publicaciones asociadas a esa institución. Dichos datos se describen en la Tabla 14: Datos del Reporte de Detalles de una Institución (Publicaciones).

Tabla 14: Datos del Reporte de Detalles de una Institución (Publicaciones)

Dato	Descripción
Title	Título con el cual está registrada la publicación.
PubType	Tipo de publicación.
PubYear	Año de la publicación.
IsIsi	Flag que indica si la publicación es ISI o no. Si el valor es true, la publicación es ISI; si es false, no es ISI.

En el reporte el navegante puede desplegar los detalles de la universidad a la cual pertenece la institución o bien los detalles de una determinada publicación usando los respectivos links.

4.3.3 Consultas

A continuación se describen las consultas que la herramienta web puede responder.

4.3.3.1 Buscador de Personas

En este módulo el navegante puede buscar a un determinado investigador. La página se muestra en la Figura 36: Buscador de Personas.

Este buscador es capaz de encontrar a un autor por su nombre o apellido. Las coincidencias son listadas al usuario y éste puede hacer clic en el nombre para desplegar el reporte de los detalles de la persona en cuestión.

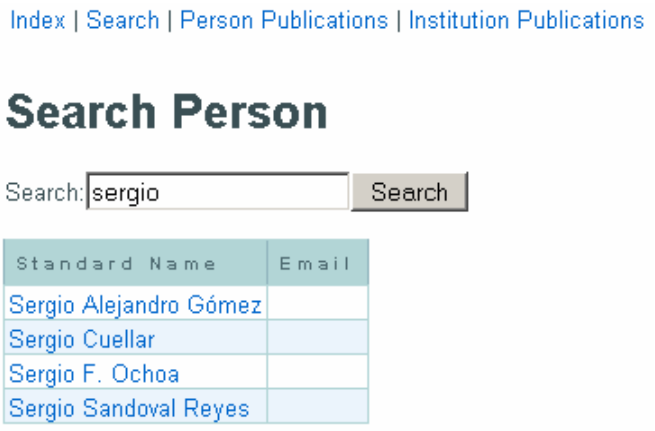


Figura 36: Buscador de Personas

4.3.3.2 Publicaciones de una Persona

En esta consulta se le presenta al navegante un listado de las publicaciones de una determinada persona.

El usuario debe seleccionar a la persona en cuestión mediante un DropBox que lista a todas las personas registradas en la base de datos. Una vez hecho esto, se despliega una lista con las publicaciones del investigador seleccionado. Ver Figura 37: Publicaciones de una Persona.

Figura 37: Publicaciones de una Persona

En la parte superior del resultado, al navegante se le presenta un link para ver los detalles de la persona seleccionada. Luego, se le presentan una serie de links para desplegar los detalles de una determinada publicación o bien un detalle de la institución asociada a la publicación.

4.3.3.3 Publicaciones de una Institución

En esta consulta, de forma análoga al de las publicaciones por persona, se presenta al usuario un DropBox con el listado de las instituciones registradas en la base de datos. Dado que el nombre de varias instituciones puede coincidir, el DropBox despliega el nombre junto con el identificador de cada institución.

El usuario debe seleccionar una institución y presionar el botón para recuperar los registros. Una vez hecho esto presentará el listado de publicaciones en donde alguno de sus autores o editores haya estado asociado a la institución seleccionada al momento de registrar la publicación. Ver Figura 38: Publicaciones de una Institución.

Figura 38: Publicaciones de una Institución

En el resultado arrojado, se despliega un link para acceder al detalle de la institución, seguido por un listado con las publicaciones asociadas a la misma. En cada una de estas publicaciones, hay un link para acceder al detalle correspondiente.

4.4 Mantenimiento del Sistema

El mantenimiento del sistema es básicamente incremental y se pueden usar tanto la aplicación cliente, como la aplicación web, según la operación que se requiera realizar:

- Para la migración masiva de nuevos datos, se debe usar el cliente ya que es la pieza del sistema que realiza el parsing el archivo dblp.xml.
- Para la edición manual de datos, se debe utilizar la aplicación web.

A continuación se detalla la forma de uso para realizar algunas tareas típicas.

4.4.1 Agregar Personas Masivamente

1. Seleccionar el archivo dblp.xml.
2. Seleccionar un archivo de personas que contenga el nombre de las personas a agregar.
3. Correr el proceso.

La aplicación cargará las nuevas publicaciones, junto con las personas que no estén registradas y que estén asociadas a dichas publicaciones. En caso de que una publicación ya exista, la aplicación no modificará el contenido de la base de datos.

4.4.2 Agregar Nuevas Publicaciones Masivamente

1. Seleccionar un archivo dblp.xml actualizado (que contenga las nuevas publicaciones).
2. Seleccionar el archivo de las personas ingresadas en el sistema.
3. Correr el proceso

La aplicación omitirá las publicaciones ya registradas en la base de datos y persistirá las nuevas que no lo estén.

4.4.3 Edición Manual de Datos

En cada reporte de Detalles, aparece un link para la edición de los datos que se muestran en pantalla. Cuando se presiona el link “Edit”, el reporte se convierte en un formulario para editar los campos.

Para realizar los cambios, el usuario debe presionar “Save”. Si no desea realizar cambios, debe presionar “Cancel”.

4.4.4 Creación Manual una Nueva Entidad

Para agregar manualmente una entidad se debe ir al link correspondiente dentro de la pantalla principal. Ver Figura 39: Creación manual de una nueva entidad.

Figura 39: Creación manual de una nueva entidad

Al presionar el link, se despliega un formulario para insertar los campos manualmente.

Para guardar los cambios el usuario debe presionar el link “Create”. Esto creará una nueva entidad en la base de datos. Si el usuario se arrepiente, puede presionar el link “Cancel” para volver a la Pantalla Principal.

5 Conclusiones y Trabajo a Futuro

El trabajo realizado consistió en el desarrollo de una herramienta para el procesamiento de publicaciones de autores latinoamericanos, en la cual se pudiese alimentar masivamente de fuentes externas, actualizar datos de publicaciones y en donde se pudiese acceder a dicha información.

En esta herramienta, a diferencia de las existentes hoy en día, se pueden establecer asociaciones entre investigadores, publicaciones, instituciones, regiones y áreas de investigación. Esto le otorga a la herramienta una funcionalidad y potencialidad valiosa para obtener una concepción global aproximada de las actividades de investigación latinoamericana.

Específicamente, se trabajó en la implementación de un procedimiento para la migración masiva de datos provenientes del sistema DBLP y la implementación de un sitio web para el despliegue de información almacenada en la base de datos y la consulta de dicha información bajo ciertos criterios.

El procedimiento de la migración masiva extrae información desde un archivo XML publicado en DBLP y lo procesa con el fin de persistir registros para un conjunto de autores determinado por el usuario.

El sitio web, despliega reportes con la información almacenada en la base de datos. En él se pueden consultar por las publicaciones de una determinada persona o institución. También se puede editar manualmente la información almacenada.

Dado que la herramienta está en su fase inicial, es posible expandirla en varios aspectos. Uno de ellos es, la incorporación de información fidedigna relativa a las instituciones y áreas de aplicación e investigación de las personas. Actualmente hay implementado reportes para ello, de contar con datos reales, la herramienta tendría gran valor agregado.

Otro aspecto mejorable de la herramienta es la elaboración de más y mejores indicadores y estadísticas. La herramienta podría contar con información gráfica que facilite el análisis de la información almacenada en el sistema.

La usabilidad de la interfaz gráfica y el manejo de usuarios para el sitio web, es otro posible aspecto para mejorar la herramienta. Es posible que se requiera que cada investigador administre sus datos, para ello el sitio debe ser fácil de usar para los navegantes. También puede ser deseable restringir el acceso a ciertas partes del sistema.

Finalmente, la incorporación de más fuentes de publicaciones es otro aspecto con posibles mejoras. Actualmente la herramienta sólo procesa los datos extraídos desde el repositorio DBLP. Para que la herramienta cuente con una buena aproximación a la totalidad de publicaciones latinoamericanas de computación, sería interesante poder estimar la información restante y averiguar de dónde obtenerla.

6 Bibliografía y Referencias

[Apa08] American Psychological Association (APA). URL: http://www.apa.org/. Última visita: Sep. 2008.

[Cha08] Proceedings del Workshop CharLA'08: Grand Challenges in Computer Science Research in Latin America. Buenos Aires, Argentina. Sep. 5-6, 2008.

[Cit08] CiteSeerX. URL: http://citeseerx.ist.psu.edu. Última visita: Nov. 2008.

[Cit09] CiteSeerX About. URL: http://citeseerx.ist.psu.edu/about/site. Última visita: Oct. 2009.

[Cle08] CLEI: Centro Latinoamericano de Estudios en Informática. URL: http://www.clei.cl. Última visita: Sep. 2008.

[Dbl08] DBLP: Digital Bibliography & Library Project. URL: http://dblp.uni-trier.de/. Última visita: Nov. 2008.

[GSc09] Google Scholar. URL: http://scholar.google.com/. Última visita: Oct. 2009.

[ISI09] ISI Web of Knowledge. URL: http://www.isiknowledge.com/. Última visita: Oct. 2009.

[Lac08] LACCIR: Latin American and Caribbean Collaborative ICT Research Federation. URL: http://www.laccir.org. Última visita: Sep. 2008.

[Lam94] Lamport, L. LaTeX: A Document Preparation System (2nd Edition). Addison-Wesley, 1994.

[Lat09] Latindex. URL: http://www.latindex.unam.mx/. Última visita: Oct. 2009.

[Sco09] Scopus. URL: http://www.scopus.com/. Última visita: Oct. 2009.

7 Anexos

En el presente Anexo se muestra información complementaria al resto del documento que puede ser de interés para el lector.

7.1 Modelo de Datos LACCIR

A continuación se ilustra el modelo de datos de LACCIR. Este modelo fue conseguido por el profesor guía, para tomarlo como base en el desarrollo del modelo de datos la herramienta desarrollada en este trabajo de memoria.

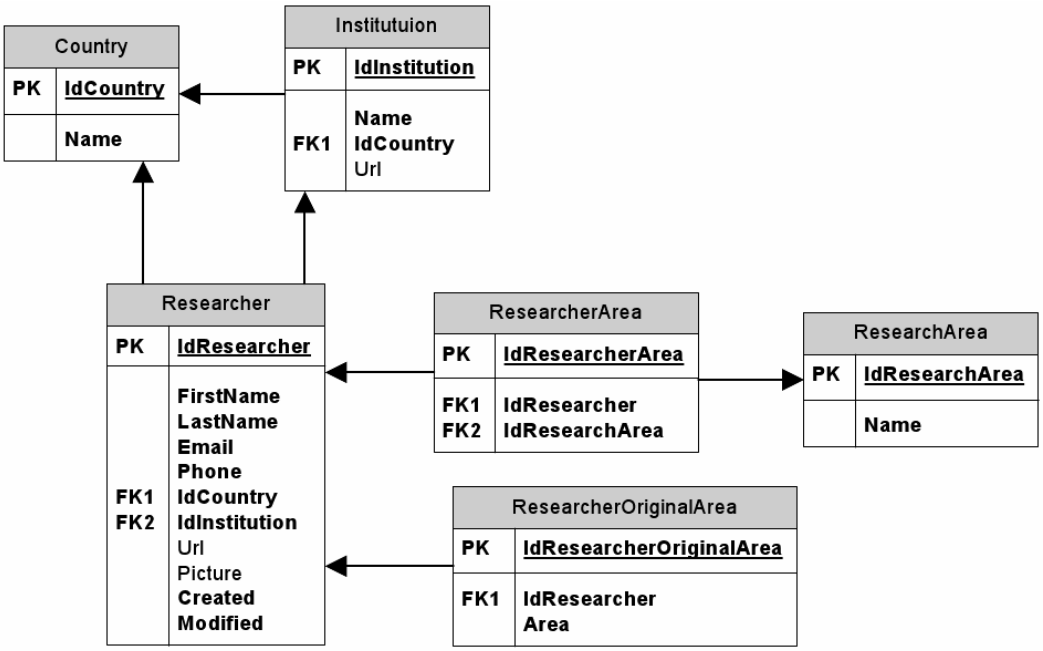


Figura 40: Modelo de Datos de LACCIR

7.2 Obtención del Archivo dblp.xml

El archivo dblp.xml con el cual trabaja la aplicación de migración de datos, está disponible en Internet en el sitio web de DBLP. Junto con este archivo, también está disponible el archivo dblp.dtd para que se pueda manipular el archivo XML de forma correcta, manteniendo el esquema original.

A continuación se describe la forma para descargar ambos archivos de forma directa mediante una URL y también la navegación necesaria para llegar a dichos archivos a partir de la página inicial de DBLP.

Procedimiento directo por URL

El archivo XML, dado su gran tamaño, conviene que sea obtenido a partir de su versión comprimida en formato .gz, ahorrando así en tiempo de descarga. Dicho archivo se encuentra disponible en la siguiente URL:

http://dblp.uni-trier.de/xml/dblp.xml.gz

A su vez, el archivo DTD se encuentra en la siguiente URL:

http://dblp.uni-trier.de/xml/dblp.dtd

Procedimiento desde la página inicial de DBLP

Para la obtención de los archivos se deben seguir los siguientes pasos:

1. Desde la página inicial de DBLP se debe ir a las FAQ Pages, mediante el link indicado en la Figura 41: Link FAQ en la página inicial de DBLP.

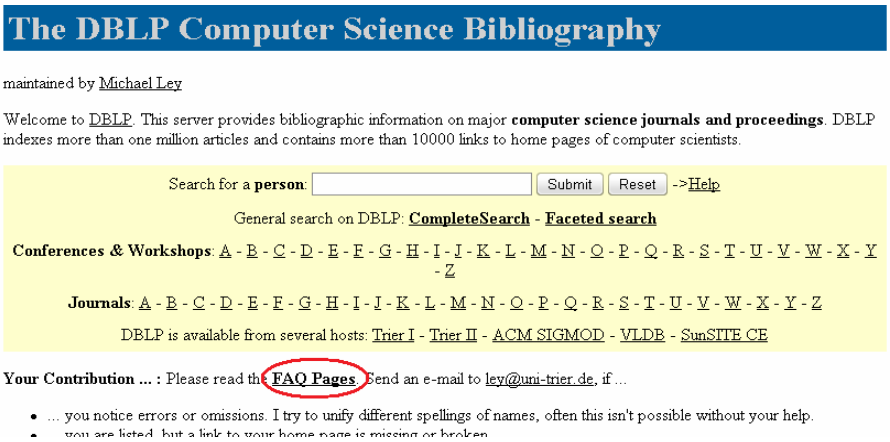


Figura 41: Link FAQ en la página inicial de DBLP

2. Dentro de la lista de preguntas de las FAQs, hay una que se refiere al parsing del archivo dblp.xml. Ver Figura 42: Link FAQ sobre el parsing del archivo dblp.xml.

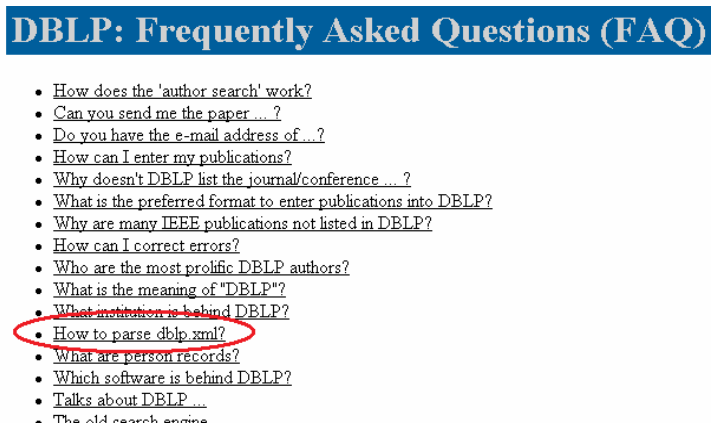


Figura 42: Link FAQ sobre el parsing del archivo dblp.xml

3. Dentro del desarrollo del tema, se encuentra un link en donde están disponibles una serie de archivos con los datos de DBLP. Ver Figura 43: Link al repositorio de archivos de DBLP.

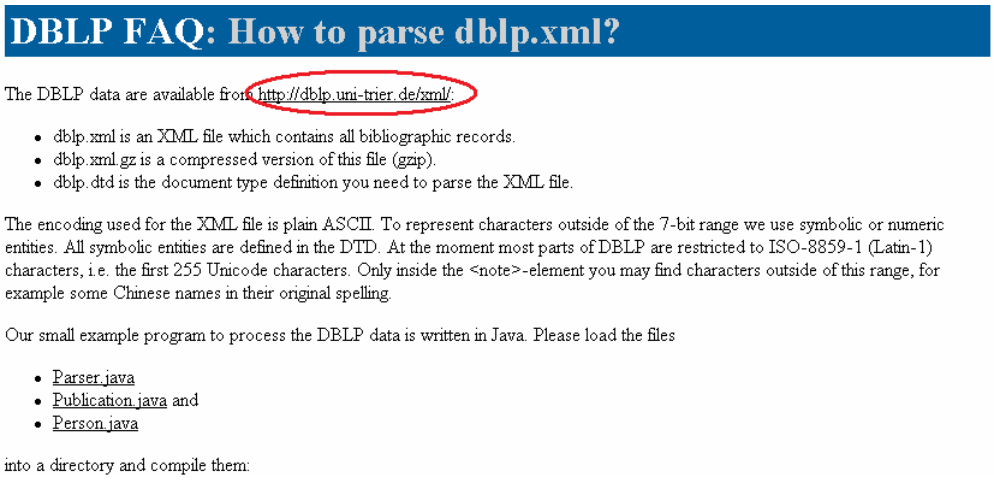


Figura 43: Link al repositorio de archivos de DBLP

4. Desde esa ubicación se pueden descargar varios archivos. En este caso solo interesan el archivo dblp.xml.gz y dblp.dtd. Ver Figura 44: Links a archivos dblp.xml.gz y dblp.dtd.

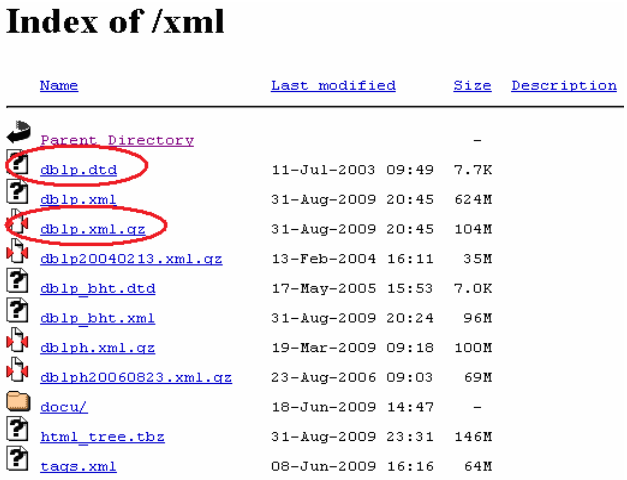


Figura 44: Links a archivos dblp.xml.gz y dblp.dtd

7.3 Diccionario de Datos

A continuación se muestra un detalle de las tablas del sistema y una descripción de cada atributo y sus restricciones.

Application_Area						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
name	VARCHAR	PK	NN		nombre del área	

Application_Area_has_Person						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
Application_Area_name	VARCHAR	PK	NN		Nombre del área	
Person_standard_name	VARCHAR	PK	NN		nombre de la persona	

Institution						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
id	INTEGER	PK	NN	UNSIGNED	identificador único de la institución	AI
University_name	VARCHAR		NN		nombre de la universidad de la institución	
Region_country	VARCHAR		NN		país de la región de la institución	
Region_name	VARCHAR		NN		nombre de la región de la institución	
web_page	VARCHAR				página web de la institución	
name	VARCHAR				nombre de la institución	

Person						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
standard_name	VARCHAR	PK	NN		nombre estándar de la persona	
email	VARCHAR				email de la persona	
web_page	VARCHAR				página web de la persona	
sex	VARCHAR				sexo de la persona	
year_of_birth	VARCHAR				año de nacimiento de la persona	
academic_degree	VARCHAR				grado académico de la persona	
picture	VARCHAR				path a una imagen de la persona	

Person_has_Publication						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
Publication_pub_key	VARCHAR	PK	NN		identificador de la publicación	
Person_standard_name	VARCHAR	PK	NN		nombre estándar de la persona	
Institution_id	INTEGER			UNSIGNED	identificador de la institución	
is_author	BOOL				true, la persona es autor; false, la persona es editor	

Person_has_Research_Areas						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
Research_Area_name	VARCHAR	PK	NN		área de investigación	
Person_standard_name	VARCHAR	PK	NN		nombre de la persona	

Publication						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
pub_key	VARCHAR	PK	NN		identificador de la publicación	
pub_type	VARCHAR		NN		tipo de publicación (article, inproceedings, etc)	
title	VARCHAR				título de la publicación	
booktitle	VARCHAR				título del libro de la publicación	
ee	VARCHAR				link a un abstract	
journal	VARCHAR				journal de la publicación	
pages	VARCHAR				cantidad de páginas de la publicación	
publisher	VARCHAR				publisher de la publicación	
series	VARCHAR				serie de la publicación	
url	VARCHAR				link a una tabla de contenidos de la publicación	
volume	VARCHAR				volumen de la publicación	
pub_year	VARCHAR				año de la publicación	
is_isi	BOOL				true, la publicación es ISI; false, no es ISI	
address	VARCHAR				Dirección asociada a la publicación	

chapter	VARCHAR				Capítulo de la publicación	
pub_month	VARCHAR				Mes de la publicación	
note	VARCHAR				Nota asociada al registro bibliográfico	
number	VARCHAR				Número de la publicación	
school	VARCHAR				Escuela asociada a la publicación	

Region						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
name	VARCHAR	PK	NN		nombre de la región	
country	VARCHAR	PK	NN		país de la región	

Research_Area						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
name	VARCHAR	PK	NN		nombre del área de investigación	

Role						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
name	VARCHAR	PK	NN		nombre del rol	
description	VARCHAR				descripción del rol	

University						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
name	VARCHAR	PK	NN		nombre de la universidad	
url	VARCHAR				página web de la publicación	
is_public	BOOL				true, la universidad es publica; false, no es publica	

Web_User						
ColumnName	DataType	Key	Not Null	Flags	Comment	Auto Inc
user_name	VARCHAR	PK	NN		nombre de usuario	
Role_name	VARCHAR		NN		rol del usuario	
pass	VARCHAR		NN		password del usuario	