SeDiCI – Desafíos y experiencias en la vida de un Repositorio Digital

Marisa R. De Giusti
Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CICPBA)
Proyecto de Enlace de Bibliotecas (PrEBi)
Servicio de Difusión de la Creación Intelectual (SeDiCI)
Universidad Nacional de La Plata
La Plata, Argentina
marisa.degiusti@sedici.unlp.edu.ar

Nestor F. Oviedo, Ariel J. Lira,
Ariel Sobrado, Juan P. Martínez, Analía Pinto
Proyecto de Enlace de Bibliotecas (PrEBi)
Servicio de Difusión de la Creación Intelectual (SeDiCI)
Universidad Nacional de La Plata
La Plata, Argentina
{nestor,alira,asobrado,dajuam,apinto}@sedici.unlp.edu.ar

***Abstract*—The Intellectual Creation Dissemination Service is the Institutional Digital Repository of La Plata National University. The project is intended to be the main distribution source of all the academic work produced inside UNLP. In view of main worldwide institutions's trend towards the publication of academic resources through open access digital repositories, SeDiCI has pointed to become a strategic tool to bring relevance to the University.**

Since its creation in the year 2003, SeDiCI has faced up many challenges and difficulties. Thus the initiative has been directly and indirectly affected during its development and growth. Despite these difficulties, SeDiCI hosts nowadays more than 14000 academic resources produced in the UNLP, which are freely accessible via a web portal. SeDiCI is currently one of the most important digital repositories at the national and regional (Latin America) level.

This work presents many of the experiencies and challenges that have shaped up SeDiCi, describing problems, context and solution approaches. Topics include the need for institutional support, many cataloguing issues, ways to improve services to users, resource digesting (import), among others. This document also describes some of the new and most recent challenges and the current research and development trends, which are oriented to improve and extend the services provided by SeDiCI (harvesting tools, text analysis tools, ontologies and semantic repositories, open access legislation, DSI, Self-archiving, and many others). The purpose of this document is to share the lived experiences with this project, which may be useful to other institutions working on their own digital repositories.

***Keywords*-digital repositores, open access, challenges, experiences**

***Resumen*—El Servicio de Difusión de la Creación Intelectual es un proyecto de Repositorio Digital Institucional creado dentro de la UNLP, orientado a funcionar como el punto central de difusión de toda la producción académica generada dentro de la institución. Dada la tendencia vista en las principales instituciones académicas del mundo de hacer pública su producción científica a través de repositorios digitales de acceso abierto, el SeDiCI pasa a ser una herramienta estratégica para la jerarquización de la institución.**

Desde su creación en el año 2003, el SeDiCI ha afrontado diversas dificultades que han influído directa o indirectamente en su desarrollo y crecimiento, pero aún a pesar de estas problemáticas, actualmente el SeDiCI cuenta con una base de datos documental que supera los 14000 recursos académicos propios (de la UNLP) expuestos bajo las políticas del Acceso Abierto. Esto convierte al SeDiCI en uno de los principales exponentes en su tipo, tanto a nivel nacional como regional (América Latina).

En este documento se presentan experiencias y desafíos que el SeDiCI ha enfrentado, describiendo en cada caso el problema, su contexto y las vías de acción tomadas para superarlo. Los principales tópicos son: necesidad de apoyo institucional, reglas y metodologías de catalogación, mejoras en los servicios provistos a los usuarios, importación de recursos, entre otros.

Adicionalmente se describen algunos de los desafíos actuales, diferentes líneas de investigación y desarrollo orientadas a resolver los retos y a expandir y mejorar los servicios proporcionados a la comunidad de usuarios. Entre estos se encuentran: mecanismos y herramientas de harvesting, gestión de grandes volúmenes de información, ontologías y repositorios semánticos, legislación relacionada al Acceso Abierto, Diseminación Selectiva de la Información, Autoarchivo, etc.

El principal objetivo de este documento es exponer la experiencia adquirida a partir de este proyecto, con la intención de que resulte de utilidad para aquellas instituciones que se encuentran en el proceso de creación de sus propios repositorios institucionales, o bien que se encuentren frente a problemáticas similares a las aquí expuestas.

Palabras Clave-repositorios digitales, acceso abierto, desafíos, experiencias

I. Introducción

La Universidad Nacional de La Plata (UNLP), atenta a la tendencia mundial de exponer la producción científica a través de repositorios digitales de acceso abierto, creó en el año 2003 un repositorio institucional que denominó Servicio de Difusión de la Creación Intelectual (SeDiCI). El principal objetivo de SeDiCI es el de preservar y dar visibilidad a todos los artículos, libros, tesis, reportes, obras pictóricas, entre otros, producidos por alumnos, docentes e investigadores de la UNLP.

SeDiCI ha adherido a las políticas del Acceso Abierto desde sus inicios, considerando que el acceso libre propende a una mayor visibilidad e impacto de los trabajos publicados, así como también representa una forma de retribución hacia la comunidad que deposita sus esfuerzos en la universidad pública.

La experiencia acumulada en el transcurso de la existencia del repositorio ha llevado a tomar decisiones vinculadas al desarrollo del mismo en todas sus áreas. La selección del software, formato de metadatos, personal, equipos de soporte, entre otros, han sido parte de los cambios.-Asimismo, SeDiCI ha servido como una herramienta estratégica para la jerarquización de la institución tanto a nivel nacional como internacional, por encontrarse posicionado (luego de más de 7 años de vida) en primer lugar en el ranking de repositorios nacionales y en octavo lugar en América Latina [1]. También cabe destacar que en el año 2010 SeDiCI fue reafirmado en su posición de Repositorio Institucional de la Universidad Nacional de La Plata a través de la resolución que obliga al depósito de las tesis de posgrado. SeDiCI cuenta hoy con una base documental de más de 14.000 recursos pertenecientes a las distintas áreas del conocimiento de la UNLP.

II. Software de soporte

La selección del software de soporte es una de las decisiones más importantes al momento de crear un repositorio digital, dado que constituye la base sobre la cual se llevarán a cabo las tareas administrativas diarias, así como la exposición de los recursos al mundo.

Para la elección del software de SeDiCI se analizaron distintas aplicaciones, buscando aquellas que reunieran las siguientes características: uso libre, código abierto, soporte de un formato de metadatos propio, simplicidad para la personalización, escalabilidad, actualizaciones frecuentes, soporte técnico eficiente, entre otros. Entre las aplicaciones analizadas se encontraban: CyberThesis (Francia y Chile) y los proyectos ETD de UNICAMP (Brasil), Virginia Tech (USA), Montreal (Canadá), la Universidad de Valencia (España).

Dado que no se logró encontrar una aplicación que reuniera todos los requisitos mencionados, se decidió llevar a cabo el desarrollo de una solución de software completamente nueva y a medida. El desarrollo de esta aplicación requirió alrededor de 4 meses entre análisis, implementación y pruebas. La misma consta de dos grandes partes: administración y portal.

Para la parte de administración se desarrolló una aplicación Java Desktop, ya que sería utilizada únicamente de forma local dentro de las oficinas de SeDiCI y se trataba de una tecnología simple, confiable y ampliamente difundida, en comparación con las tecnologías cliente-servidor del momento (PHP4, ASP, etc., en el año 2003).

Para la parte del portal se eligió PHP4 como lenguaje, ya que en este caso era necesario contar con una aplicación web simple, accesible de forma pública, que proveyera funciones de búsqueda y exploración de recursos, junto con noticias, links, eventos, etc.

Desde entonces, este desarrollo a medida es el software que da soporte a SeDiCI, funcionando por un lado como portal web, con búsquedas, exploración, noticias, links, etc., y servicios adicionales para usuarios registrados; mientras que por otro funciona como una administración desktop con registro de autores, tesauros, sistemas de clasificación, un formato de metadatos propio, etc.

Con el correr del tiempo, el desarrollo ha ido creciendo considerablemente: muchos desarrolladores han aportado a su código fuente y muchos servicios han sido agregados y otros mejorado. Pero debido al paso de los años y a los cambios tecnológicos que han surgido en estos siete años de vida (avances tecnológicos, cambios en el ámbito de las bibliotecas digitales, desarrollo de nuevas y mejores aplicaciones de código abierto por parte de grandes comunidades, etc.), el software de SeDiCI se ha convertido en una herramienta compleja, que requiere personal especializado dedicado a su mantenimiento, en detrimento de nuevos desarrollos e investigaciones. Esta realidad ha llevado a la decisión de realizar un reemplazo del software actual, para lo cual se ha iniciado un nuevo estudio con el fin de identificar aplicaciones candidatas para el reemplazo. Para esto se confeccionó una lista de características que la herramienta seleccionada debería reunir, entre las cuales se pueden mencionar:

- licencia de uso: el uso de la aplicación debe ser libre y gratuito, al menos para su uso en ámbitos académicos;
- *nivel de impacto*: es deseable que la herramienta cuente con un alto grado de aceptación por la comunidad global de repositorios digitales, ya que esto ayuda a mantener el proyecto en actividad y en constante actualización;
- *administración:* la aplicación debe contar con una sección de administración accesible sólo por usuarios con privilegios, para llevar a cabo las tareas de catalogación y administración en general (colecciones, backups, configuración, etc.);
- *personalización:* debe proveer mecanismos de personalización, tanto de interfaz de usuario, como de funcionalidad (ya sea por medio de configuración o por el agregado de extensiones);
- *documentación:* debe proveer documentación para desarrolladores (manuales, instructivos, ejemplos, etc);
- actualizaciones: se debe tener acceso a las actualizaciones y nuevas versiones de la aplicación;
- *soporte:* se debe contar con soporte técnico eficiente para los administradores y para los desarrolladores encargados de su instalación, configuración, personalización y extensión;
- *facilidad de uso:* es deseable contar con un manual de usuario; además, es importante que el diseño de la interfaz de usuario sea simple, con los elementos ubicados de forma intuitiva;
- *formato de metadatos:* debe permitir la selección del formato de metadatos a usar, según las necesidades del repositorio. Es deseable que se permita la definición de un formato de metadatos propio;
- *performance:* las respuestas del sistema deben ser en tiempo y forma;

- *escalabilidad:* las capacidades del software no deben limitar el crecimiento del repositorio;
- *interoperabilidad:* se deben proveer características que permitan la interacción entre varios repositorios digitales (importación, exportación, web-services de búsquedas, OAI-PMH, etc.).

III. Representación de recursos

La representación de recursos es una problemática recurrente en el diseño de software para repositorios digitales. Esto se debe principalmente a que los recursos son el elemento central en estos diseños, al tiempo que son muy variados en cuanto a su tipología (tesis, libros, artículos, disertaciones, presentaciones, audio, vídeo, imágenes, etc.), lo que modifica considerablemente su representación y tratamiento (estructura de metadatos, normalización, vocabularios controlados, etc). Además, su almacenamiento físico debe realizarse cuidadosamente, ya que debe asegurarse la recuperación en forma eficiente, la preservación en el tiempo y las capacidades de interoperabilidad con otros repositorios.

A. Formato de metadatos

Existen varios aspectos a tener en cuenta para la selección del formato de metadatos a utilizar. Uno de estos es la diversidad de tipos de recursos que serán tratados en el repositorio, contemplando además las capacidades del software de soporte. Es decir, es necesario asegurarse de que todos los metadatos necesarios para todos los tipos de recursos que serán catalogados puedan ser representados en el/los formatos de metadatos que se seleccionen, y que además el software de soporte esté capacitado para gestionar estos formatos.

Otro aspecto importante a tener en cuenta es el nivel de interoperabilidad que se desea lograr, lo cual suele estar constreñido tanto por las capacidades del software de soporte, así como por las limitaciones del formato de metadatos elegido. Esto es, existen casos en los que el formato de metadatos utilizado es el mismo que aquel utilizado por el repositorio con el cual se desea interoperar, pero en otros casos se necesita de mapeos para generar nuevos formatos de metadatos a partir del propio. Esto trae como consecuencia la inevitable pérdida de información, llevando a la necesidad de analizar qué tan graves son esas pérdidas. Un ejemplo simple se puede observar cuando un formato de metadatos de origen mantiene el nombre de un autor en campos separados: campo *nombre* y campo *apellido*, y el formato de metadatos de destino contiene un único campo *nombre* en el cual se concatenan estos dos valores, perdiendo la capacidad de distinción.

Durante el desarrollo de SeDiCI se analizaron los distintos formatos de metadatos más utilizados, y dado que ninguno llegaba a cubrir todas las necesidades planteadas se optó por un formato propio, buscando principalmente flexibilidad en la definición del mismo. Actualmente, la estructura de metadatos que se utiliza está basada en un conjunto de tablas relacionales, permitiendo así administrar los metadatos disponibles (agregar, modificar y eliminar metadatos cuando sea necesario) de forma simple. Asimismo se establecieron normas de catalogación que instruyen al personal encargado de estas tareas sobre qué metadatos deben utilizar (de forma obligatoria, recomendada u opcional) para catalogar cada tipo de recurso (tesis, libros, artículos, etc.).

Si bien esta representación del formato de metadatos es flexible, uno de sus puntos débiles es la complejidad, ya que la estructura de tablas necesaria para la representación de los metadatos, las restricciones de contenido y atributos, entre otros, dificultan su comprensión y propician la pérdida de claridad acerca de cómo se relacionan las tablas. El problema precedente se debe principalmente a que cada metadato puede ser un texto libre, una fecha con determinado formato, un término de un vocabulario controlado, un código de un sistema de clasificación, o incluso una referencia a otra tabla de la base de datos (ejemplo: un autor del registro de autores), implicando distintos tipos de consultas según lo que se desee obtener y afectando la performance en la recuperación de los registros por el gran número de uniones entre tablas que es necesario realizar. Finalmente, cabe destacar que los dos aspectos negativos mencionados anteriormente también perjudican la escalabilidad del software.

IV. Catalogación: problemas y desafíos

Como se dijo más arriba, el formato de metadatos utilizado en SeDiCI es un desarrollo propio, con todas las ventajas y desventajas que esto conlleva. Entre las ventajas, a la hora de la catalogación, debe señalarse la extrema flexibilidad del mismo, que permite, por ejemplo, la rápida catalogación de recursos tan dispares como una imagen, un vídeo o una tesis. Sin embargo, la misma flexibilidad del sistema obliga a que los administradores sean sumamente cuidadosos a la hora de catalogar recursos no textuales o que no cuentan con toda la información requerida.

Con el objeto de minimizar toda posibilidad de error y de maximizar la inmediata recuperación por el sistema del recurso buscado, se ha consensuado, para uso interno, un conjunto de metadatos obligatorios y un conjunto de metadatos recomendados u opcionales. Entre los obligatorios se encuentran aquellos que hacen a la identificación unívoca del recurso: nombre y apellido de los autores, título del documento, descriptores y año de publicación, entre otros. Entre los recomendados u opcionales se encuentran aquellos metadatos que pueden aportar información adicional sobre el recurso (por ejemplo, identificación geográfica) pero que no hacen a su identificación.

Se parte del supuesto de que cuantos más metadatos posea el recurso, más fácil y accesible será para los usuarios finales. Sin embargo, en ocasiones resulta muy difícil contar incluso con los datos más básicos de un recurso, por diversas razones, que van desde la falta de datos en el propio documento (cuando se trata de recursos textuales), la información escasa o errónea y las transformaciones que se operan cuando los recursos se obtienen a través de la cosecha de datos desde otros repositorios. En este sentido, el objetivo es, a futuro, consensuar un marco mínimo de interoperabilidad con las bibliotecas de las diferentes unidades académicas que poseen sus propios repositorios para minimizar errores o inexactitudes a la hora de cosechar recursos tanto de uno como de otro lado.

Si bien en SeDiCI se prioriza la carga de recursos a texto completo, no siempre es posible contar con ello y en aras de mantener un registro lo más fiel posible de la producción académica de la universidad se han cargado (y se cargan)

recursos de los que sólo se cuenta con sus respectivos metadatos, con nula o escasa posibilidad de confrontar adecuadamente la fuente original de dichos datos. Para enfrentar esta dificultad, el personal de SeDiCI realiza periódicos chequeos de los recursos catalogados y siempre que sea posible procura confrontar los mismos con otras fuentes (o bien, con los propios autores) para asegurar en todo momento la veracidad y corrección de los datos. En este sentido, SeDiCI opera como un "megacatálogo" de todas las producciones académicas y por ello se procura que los metadatos expuestos sean los más adecuados y pertinentes.

Con el objeto de mejorar la descripción adecuada de los recursos se han planteado varias estrategias, entre las que pueden mencionarse:

- agregado de nuevos metadatos,
- división entre materias y descriptores, y
- la inclusión de nuevos tesauros.

El agregado de nuevos metadatos obedeció no sólo al crecimiento del repositorio en general sino al objetivo de lograr una catalogación más pertinente de los recursos. Entre los metadatos agregados se puede citar "Evento", para distinguir si un recurso pertenece o ha sido presentado en un congreso/simposio. Otros metadatos fueron cambiados, como el caso de "Localización electrónica" y "Localización física": ambos son, actualmente, metadatos con formato de links que llevan, respectivamente, a la ubicación en la red y a la ubicación en un determinado catálogo. Con esta distinción se discrimina también cuándo un recurso es a texto completo (y se encuentra alojado externamente en la red) y cuándo se cuenta solamente con su catalogación, pero se conoce también su ubicación en una biblioteca u otro repositorio dentro de la universidad. Con el mismo objeto de discriminar cuándo se cuenta con el texto y cuando no, y a los fines de elaborar estadísticas pertinentes, se ha implementado otro metadato, no visible para el usuario final, denominado "Full-text", en el que los administradores deben indicar, mediante sí/no, si se cuenta con el texto.

Como en todo sistema en continuo crecimiento, son numerosas las mejoras que deben realizarse de forma constante para que el funcionamiento sea óptimo todo el tiempo. En este sentido, otra de las estrategias implementadas fue la de comenzar a utilizar tres tipos de términos para catalogar temáticamente el material disponible. Durante una buena cantidad de años los recursos se catalogaban temáticamente mediante el uso de descriptores (términos controlados) y palabras-clave (términos no controlados). Por descriptores se entiende un vocabulario finito y controlado de términos mientras que las palabras-clave surgen de los propios textos, proporcionadas por los autores de los mismos. En la actualidad, SeDiCI ha implementado el uso de otro listado de términos controlados al que se ha denominado "Materias", en el cual se incluye un conjunto restringido de términos controlados, seleccionados por administradores idóneos, que hacen referencia a las grandes áreas temáticas del conocimiento de que se compone el amplio rango de unidades académicas en que se divide la universidad. De este modo, los recursos son catalogados en primer término mediante una "macrocatalogación" (materias), luego una catalogación temática más restringida (descriptores) y finalmente, si las hay, mediante las palabras-clave proporcionadas en el texto por su autor. Puede decirse, sintéticamente, que se parte de lo general para llegar a lo particular de cada recurso.

En el mismo sentido, el tema de los descriptores a utilizar ha sido largamente discutido. Como se sabe, los descriptores están contenidos en una estructura jerárquica que establece las relaciones entre ellos, denominada "tesauro". Existe gran cantidad y variedad de tesauros. Durante mucho tiempo, SeDiCI utilizó el tesauro de la UNESCO [2] y un tesauro propio, elaborado en base al de UNESCO, que incorporaba términos que no se encontraban allí y que eran necesarios para los recursos existentes en el repositorio. Con el crecimiento del repositorio fue evidente que dichos tesauros no alcanzaban a cubrir todas las necesidades de los recursos y, tras un adecuado estudio de los tesauros disponibles, se decidió incorporar dos nuevos tesauros: el Eurovoc [3] y el DECs [4]; éste último, a pesar de estar orientado mayormente hacia las ciencias de la salud incluye también numerosos términos de otras áreas, con lo cual su inclusión ha resultado más beneficiosa de lo que se esperaba. No se descarta que a futuro se incorporen otros tesauros.

Otro aspecto de la catalogación que está sometido a discusión actualmente es el de las llamadas "entidades abstractas". Las entidades abstractas son formas de agrupar recursos que, por una u otra causa, deben presentarse visualmente juntos, como en el caso de los artículos de un número determinado de una revista. Para evitar que los artículos se presenten en forma desordenada o apartada es que se los incluye dentro de estas "entidades" que tienen la función de presentar la información ordenadamente. Sin embargo, esto supone un doble trabajo para los administradores: si, por caso, desean cargar un artículo de una revista, por un lado, deben generar una primera entidad abstracta, la Serie Documental, que comprende sólo los datos generales de la revista (nombre, director, frecuencia), y, por otro, una segunda entidad abstracta, la Entrega Documental, que hace referencia al número o volumen específico de la dicha revista donde fue publicado ese artículo. Sólo cuando estas dos primeras entidades están generadas, es posible que los administradores puedan cargar los artículos en cuestión. Este es, claramente, uno de los desafíos pendientes a futuro.

No menos problemático es otro aspecto vinculado a la catalogación: el de las tipologías documentales. En estos momentos, en los que se analizan todos los aspectos vinculados a la integración con el Sistema Nacional de Repositorios Digitales en Ciencia y Técnica propulsado por el Ministerio de Ciencia, Tecnología e Innovación Productiva de la Nación, uno de los aspectos más críticos ha sido precisamente este, ya que cada repositorio cuenta con una tipología propia y SeDiCI no es la excepción. Más allá de lo que finalmente resulte del consenso del comité de expertos convocado por el ministerio, en la actualidad SeDiCI cuenta con 17 tipos de documentos, 4 de los cuales son entidades abstractas, como se señaló más arriba. Vale decir que hay por lo menos 13 tipos de documentos disponibles para cubrir las necesidades básicas de un repositorio institucional. Claramente, a la luz del avance de las tecnologías, es necesaria una revisión a fondo de cada tipo documental, ya que los usados en estos momentos fueron pensados y diseñados en un momento histórico muy diferente del actual en todo sentido.

Del mismo modo, en los últimos tiempos se han agregado nuevos tipos documentales que anteriormente no existían en el

repositorio, como los documentos legales y las patentes. La misma dinámica del repositorio, su constante flujo de recursos, es la que dictamina, como en los casos precedentes, la creación de nuevos tipos documentales para clasificar adecuadamente sus existencias. Esto, que por supuesto constituye una ventaja del software propio y de su ya mencionada flexibilidad, puede resultar un aspecto problemático en lo que hace a la interoperabilidad y a la adecuación de las normativas que finalmente se expidan por parte del ministerio para el Sistema Nacional de Repositorios, del que SeDiCI será, por supuesto, uno de los más importantes y caudalosos.

Un último aspecto a considerar dentro de los desafíos que plantea la catalogación de recursos tan heterogéneos como los que alberga SeDiCI tiene que ver directamente con el diseño de su software y con el trabajo que deben realizar sus administradores. Como en un principio la cantidad de tipos de documentos y los metadatos para dar cuenta de ellos no eran tan numerosos como en la actualidad, no se tuvo en cuenta la posibilidad de que al momento de cargar un recurso, al administrador le aparecieran solamente los metadatos correspondientes al recurso en cuestión. Por ejemplo, si el recurso a cargar es una tesis, el metadato "director de la tesis" será obligatorio, mientras que no aplicará para el caso de un artículo.

En la actualidad, los administradores acceden al conjunto total de los metadatos en una lista desplegable y allí deben seleccionar el que corresponda según la naturaleza del recurso: esto no sólo conlleva un mayor tiempo de carga (ya que cada metadato debe seleccionarse de una lista en crecimiento) sino también un margen de error más elevado, ya sea por distracción o por la similitud entre los metadatos (ejemplo: "título del documento" y "título de la serie" pueden confundirse fácilmente). Con el fin de evitar esto se han renombrado algunos metadatos pero aún subsiste el desafío de generar automáticamente el conjunto de metadatos correspondientes para cada tipo documental, de modo que el administrador sólo deba completar los campos con la información requerida sin tener que elegirlos de una lista (y eventualmente agregar un metadato si lo considera necesario).

V. Apoyo institucional

Un servicio de estas características, como puede suponerse, no hubiera sido posible ni sustentable sin el apoyo firme y decidido de las autoridades de la Universidad Nacional de La Plata. La gestión precedente y la actual de la UNLP han adquirido un conocimiento profundo del valor del repositorio institucional en relación a la visibilidad de la institución y de las obras de sus actores. Este conocimiento ha ido formalizando los caminos para el aporte de material al repositorio.

En el mismo sentido, uno de los mayores logros ha sido la resolución 78 de febrero de 2011 [5], en la que se instituye que todas las tesis de posgrado deben ser depositadas en SeDiCI para su preservación, como contraparte digital del depósito de una copia en la biblioteca de su respectiva unidad académica. De esta manera se asegura no sólo un ingreso constante de recursos al servicio sino también poder cumplir de este modo con la responsabilidad de curatela del recurso digital que es el compromiso del repositorio institucional. Mandatos como el referido precedentemente aseguran que SeDiCI cuente con todos los datos requeridos para su correcta catalogación, ya sea que el autor decida subir su trabajo mediante autoarchivo o bien éste sea entregado personalmente. En ambos casos siempre será posible confrontar los datos incompletos o erróneos con la fuente original, lo que reducirá sensiblemente toda posibilidad de error en la catalogación.

De todos modos, y aunque esto no podría ser más auspicioso, hay que tener en cuenta que el correcto funcionamiento de esta nueva etapa dependerá en buena medida del conocimiento y comportamiento de los alumnos e investigadores. A tales efectos se están diseñando campañas de difusión y publicidad, para que todos los actores involucrados en este proceso (alumnos/investigadores, secretarías de posgrado, bibliotecas, etc.) estén al tanto de la resolución así como de los pasos a seguir para depositar correctamente sus tesis en SeDiCI. En el pasado ha sido palpable un gran desconocimiento, por las razones que fuera, de los principales interesados acerca de las modalidades y beneficios de difundir sus obras mediante SeDiCI. Es, sin ninguna duda, otro de los desafíos a vencer en lo inmediato, tanto el desconocimiento como cierto grado de desconfianza, que aún subsiste en muchos, acerca de la difusión por medios electrónicos. En este último sentido, SeDiCI toma todos los recaudos necesarios para resguardar los derechos inalienables de cada autor: siempre se ha trabajado codo a codo con la Dirección de Propiedad Intelectual de la universidad para proteger los derechos de todos.

VI. Importación de recursos

Como se mencionó previamente, el principal objetivo de SeDiCI es reunir, preservar y publicar toda la producción de la Universidad Nacional de La Plata. Existen además muchos repositorios digitales que actualmente contienen y exponen documentos producidos en esta universidad, lo que para SeDiCI podría representar una gran ventaja. Es decir, poseer la capacidad de importar estos recursos directamente a SeDiCI, permitiría agilizar los procesos de catalogación de dichos documentos, ya que la información necesaria se encuentra públicamente accesible, evitando así la necesidad de recopilar todos los datos para cada documento desde cero. Además, visto el gran avance en cuanto a tecnologías dedicadas a mejorar la interoperabilidad entre aplicaciones, particularmente el protocolo OAI-PMH para intercambio de metadatos, esto no debería demandar un esfuerzo significativo.

Sin embargo, luego de distintos análisis acerca de cómo concretar estas importaciones, se descubrieron algunos problemas que debían ser solucionados.

El primer gran problema fue la necesidad de separación de los recursos propios de la UNLP de entre el conjunto de documentos expuestos por el repositorio a través del protocolo OAI-PMH. A esto se agrega el hecho de que SeDiCI, al presente, incorpora sólo ciertos tipos de recursos, desestimando por ejemplo programas de materias, planes de estudio, documentos administrativos, entre otros, los que tal vez sí formen parten de sus contenidos a futuro.

Otro de los problemas encontrados parte de que el formato más difundido para el intercambio de metadatos bajo el protocolo OAI-PMH es Dublin Core. Este formato, en contraste con el formato de metadatos de SeDiCI (mucho más completo y descriptivo), plantea la necesidad de realizar

mapeos y transformaciones a la información importada, lo cual implica pérdida de información o incluso generación de documentos incompletos. Una forma de evitar la generación de documentos con información incorrecta o incompleta producto de los mapeos, es la intervención del personal especializado, encargado de la revisión y corrección de los recursos importados. Asimismo, si se desea garantizar que toda la información importada sea correcta, completa y respete todas las reglas de catalogación de SeDiCI (normalización de términos, uso de vocabularios controlados, etc.), la obligación de destinar personal especializado para esta tarea es inevitable, aumentando así el costo de las importaciones.

VII. Servicios

Si bien el servicio más importante que un repositorio digital debe proveer a sus usuarios es la búsqueda y recuperación de recursos, existen muchos otros servicios adicionales que aportan gran valor al sitio, con la constante intención de simplificar el trabajo de los usuarios y brindar funciones útiles para los mismos.

A continuación se mencionan algunos de los servicios que se ofrecen desde el portal de SeDiCI.

A. Recuperación de la información

Desde su creación, el portal de SeDiCI ha proporcionado la búsqueda y recuperación de documentos, brindando la posibilidad de realizar una búsqueda simple, otra avanzada y distintos tipos de exploración del repositorio. Estos son los servicios más importantes y más utilizados del portal, por lo que son pasibles de mejoras continuas, en el intento de estar al día con las nuevas metodologías y tecnologías para la recuperación de la información de la forma más eficiente.

Si bien el tiempo de respuesta ante una consulta de un usuario en el portal es un factor de alta importancia, existe otro factor que resulta más importante aún: la relevancia de los resultados. Esto es, qué tan acertados sean los resultados devueltos según lo que el usuario desee encontrar, o bien, qué tan cercanos sean los resultados retornados según el criterio de búsqueda especificado por el usuario. Actualmente, parte de las mejoras propuestas para esta funcionalidad se basa en la utilización de un motor de indexación de texto denominado Apache Solr, el cual se destaca por proveer tiempos de búsqueda del orden de los milisegundos, aportando funciones que permiten optimizar los resultados obtenidos en cuanto a su relevancia.

B. Diseminación Selectiva de la Información

La Diseminación Selectiva de la Información es un procedimiento mediante el cual se distribuyen periódicamente referencias a recursos según los intereses de cada usuario que se suscribe al servicio. Para esto, los usuarios crean perfiles que determinan los criterios de filtrado a aplicarse.

El servicio de DSI surgió de la necesidad de mantener informados a los usuarios de cualquier recurso nuevo que se agregue al repositorio y que pueda llegar a ser de su interés (según la configuración de los perfiles).

Previo al desarrollo de este servicio se analizó un amplio espectro de herramientas de DSI disponibles de forma pública, y aunque cada una de ellas cumplía sus funciones adecuadamente, ninguna se adaptaba a las características estructurales propias de SeDiCI, lo que era imprescindible dada la necesidad de crear perfiles basados en dichas estructuras. Por esto, se inició el desarrollo del servicio de DSI, integrado al portal de SeDiCI y disponible inicialmente sólo para los usuarios registrados.

Durante el desarrollo, además de la tarea de diseñar los perfiles, uno de los principales puntos de discusión trató sobre la inclusión de perfiles para usuarios no registrados en el portal, lo que implicaba la necesidad de un desarrollo más complejo y extenso. Por cuestiones de prioridades en el conjunto de tareas a llevar a cabo dentro de SeDiCI, se decidió realizar una implementación que no contemplase a este tipo de usuarios, dejando esta extensión para un futuro no muy lejano.

C. Carpetas

Cuando se realiza una búsqueda en cualquier buscador web y se encuentran sitios de interés, siempre es importante contar con algún mecanismo que permita registrar esos sitios para un acceso posterior. Esto mismo sucede en el contexto de los repositorios digitales. Cuando se descubre un recurso que resulta de interés, es deseable contar con un mecanismo de marcado de dicho recurso, de modo que pueda ser accedido en el futuro sin la necesidad de realizar una nueva búsqueda. Para esto SeDiCI cuenta con la posibilidad de marcar un recurso como favorito, dejándolo accesible directamente desde la sección de usuarios registrados.

Por otro lado, con el paso del tiempo, la lista de recursos favoritos puede incrementarse y llegar a ser extremadamente larga, afectando considerablemente su legibilidad. Esto puede transformar la tarea de ubicar un recurso específico en una labor aún más difícil que realizar la búsqueda nuevamente. Para atacar esta problemática, se incluyeron las carpetas de usuarios: así se provee un mecanismo de organización dinámico de estas listas, permitiendo agrupar recursos según el criterio y las necesidades de cada usuario.

D. Autoarchivo

Desde sus inicios y durante varios años, el portal de SeDiCI contó con una sección especial accesible sólo por usuarios registrados, en la que se les permitía sugerir la inclusión de un recurso en la base de datos de SeDiCI. Los usuarios debían proporcionar sólo algunos datos básicos sobre el recurso (autor y título, entre otros) y la carga del archivo correspondiente. Estos datos eran posteriormente revisados y completados por personal especializado dentro de SeDiCI, y en caso de aceptarse el trabajo propuesto, éste quedaba accesible desde el portal web.

Esta metodología simple de colaboración por parte de los usuarios implicaba numerosas discusiones. Entre ellas, se destacan: la necesidad de una autorización para el resguardo y publicación del trabajo en cuestión por parte de SeDiCI, firmada por al menos uno de los autores del trabajo, y la necesidad de definir una licencia que garantice y proteja los derechos de los autores sobre su obra.

En la actualidad, esta herramienta para aportar recursos ha sido formalizada, alineando a SeDiCI a otras grandes instituciones del mundo que implementan el Autoarchivo.

VIII. Líneas de investigación actuales

Con el constante objetivo de mejorar y descubrir nuevos servicios y herramientas dentro del contexto de los repositorios digitales, SeDiCI se encuentra en continuo desarrollo y despliegue de diferentes líneas de investigación. A continuación, se mencionan algunas de las principales áreas de investigación a las que SeDiCI se dedica.

A. Gestión de grandes volúmenes de información

Una problemática recurrente en el área de los repositorios digitales es la gestión de millones de registros de forma eficiente. En SeDiCI, esto se presenta a medida que avanzan las tareas de cosecha OAI sobre diversos repositorios mundiales. Actualmente se llevan recolectados más de 16 millones de recursos (sólo metadatos) en formato Dublin Core (XML). Por supuesto, no es útil poseer esta gran cantidad de documentos si no se provee de algún mecanismo eficiente de búsqueda y recuperación, y aquí es donde comienzan los desafíos.

En la experiencia de SeDiCI, luego de analizar y probar varias alternativas (archivos en un file system, bases de datos relacionales, bases de datos XML, entre otros), los mejores resultados fueron obtenidos utilizando un motor de indexación de texto denominado Apache Solr. Contando con una configuración finamente adaptada, este motor permite realizar búsquedas en el orden de los milisegundos, al tiempo que provee gran cantidad de funcionalidad adicional muy valiosa (como facets, ordenamiento por relevancia, etc.). Cabe destacar que los principales factores que intervienen en el nivel de desempeño de este poderoso motor de indexación son:

- *hardware*: debe ser un servidor relativamente poderoso. SeDiCI cuenta con servidor equipado con un procesador Xeon de 4 núcleos, 16GB de RAM y discos SAS de 15000RPM organizados en RAID para performance;
- *schema*: la lista de campos que serán almacenados, sus tipos de datos y sus atributos específicos, son elementos muy importantes en la creación del índice. Muchos de estos parámetros determinan la performance del motor de indexación;
- *configuración*: Apache Solr es altamente configurable, pero de forma similar a lo que sucede con la definición del schema, estos parámetros deben ser modificados cuidadosamente. De esto depende la performance general del motor de indexación (valores de cache, tamaño de archivos, componentes de software adicionales, etc), así como también puede determinar los límites en su desempeño.

B. Cosecha de recursos por diferentes medios

En el contexto de los repositorios digitales, la forma de interoperabilidad más difundida es el protocolo OAI-PMH, ya que es relativamente simple de implementar y utilizar. Además de este protocolo, existen muchas otras fuentes de información desde las cuales se podrían obtener recursos relevantes. Ejemplos de esto son la web, web-services, bases de datos, etc. Es claro que existen grandes diferencias tanto en la forma de obtención de los datos, como en su organización y procesamiento. La potencialidad de estas nuevas fuentes de información no tradicionales es inmensa y, al contar con mecanismos automáticos para la recolección, se evitan grandes esfuerzos.

SeDiCI, en su búsqueda de fuentes de información alternativas a las tradicionales, se enfocó en generar una herramienta que permitiera simplificar las tareas de recolección de recursos desde estas nuevas fuentes. De esto surgió un software de recolección configurable y extensible, que se ajusta a la arquitectura ETL (Extract, Transform and Load), e incluye un gran número de características y capacidades de procesamiento que suman un valor agregado importante a la información recolectada, siempre considerando su posterior uso. Entre estas características se destacan: la aplicación de transformaciones simples sobre los datos y la posibilidad de especificar cualquier tipo de almacenamiento (archivos, base de datos, motor de indexación, etc). Este software se encuentra actualmente en etapa de pruebas.

C. Procesos de transformación y mejora de la información

Como se mencionó anteriormente, la gestión de grandes volúmenes de información obliga a lidiar con el problema de la búsqueda y recuperación de forma eficiente. Sumado a esto existen otros tipos de problemas, no menos importantes, derivados de la heterogeneidad de los datos. Por ejemplo, el metadato dc:type de Dublin Core es completado por cada repositorio según sus criterios particulares, con lo cual puede suceder que muchos valores distintos encontrados en este metadato en realidad representen el mismo valor: valores como *Article*, *Artículo*, *ART* significan que el tipo de recurso es un artículo científico. El metadato de idioma es otro ejemplo de este caso. Asimismo, los campos de fechas también son problemáticos, ya que cada repositorio provee las fechas en el formato que considera más apropiado. Según los requisitos del repositorio que realiza la agregación de recursos, estos problemas pueden llegar a ser muy complejos o incluso imposibles de solventar de forma automática.

La herramienta de software comentada en el punto anterior, dedicada a la recolección de recursos desde distintos orígenes, incluye entre sus funcionalidades la capacidad para realizar transformaciones simples a los metadatos descargados. Entre estas se encuentran el reemplazo de términos según un diccionario predefinido, la normalización de fechas a un formato común, la eliminación de metadatos con datos inválidos, la definición de valores por defecto para metadatos que no están presentes, entre otros.

De esta forma, SeDiCI logró mejorar ampliamente la información recolectada, permitiendo así búsquedas más exactas y mayor navegabilidad en los resultados. Además, esto permite obtener información estadística más precisa.

D. Ontologías y repositorios semánticos

Contar con un repositorio semántico es una de las ambiciones más grandes de SeDiCI. Este gran paso implicaría enormes capacidades para la búsqueda y navegación del repositorio, determinando un avance significativo en la búsqueda de soluciones que simplifiquen las tareas de los usuarios al momento de buscar información relevante dentro del repositorio.

El diseño de ontologías adecuadas, flexibles y extensibles es uno de los principales objetivos, seguido de la adecuación del portal de forma tal que se refleje esta potencia y presente una visión simplificada del complejo mundo de relaciones subyacente. A partir de aquí se plantea la posibilidad de investigar sobre nuevas formas de interacción con los usuarios para el descubrimiento de información, como servicios de DSI avanzados, entre otras cosas.

Para esto se requiere inicialmente un análisis en profundidad de todos los aspectos que rodean a SeDiCI, como los formatos de metadatos, los medios de almacenamiento, los servicios que se ofrecen desde el portal, el software de administración, etc. para luego proponer alternativas de implementación y realizar la población de las ontologías correspondiente.

También se ha planteado la posibilidad de realizar parte de la población de las ontologías de forma automática, a partir de relaciones inferidas según la información existente en el repositorio actual, como relaciones de jerarquía (revistas y artículos), por áreas temáticas, por palabras-clave, por autores, etc. Esto mismo se ha planteado para ser aplicado sobre los recursos obtenidos por medio de la recolección desde repositorios externos, considerando las diferencias en complejidad, ya que en este caso en particular se cuenta con un volumen de información mucho mayor.

IX. Cambio del software de soporte

El software de SeDiCI lleva muchos años en funcionamiento, y en la actualidad cumple con toda la funcionalidad básica de un repositorio digital. Sin embargo, por razones que se comentan a continuación, se ha resuelto realizar una migración hacia una nueva aplicación.

El primer punto a destacar es la desactualización de las tecnologías sobre las que el software de SeDiCI está implementado. Al tratarse de tecnologías con más de 7 años de antigüedad, las capacidades de ampliación se ven limitadas, al tiempo que se dificulta el mantenimiento del sistema por no contar con el soporte adecuado (tecnología discontinuada). Asociado a esto, debe sumarse el hecho de que el personal pierde motivación al estar dedicado a un mismo proyecto durante mucho tiempo, y como consecuencia se disminuye su productividad.

Como se mencionó anteriormente, el software de SeDiCI fue desarrollado y puesto en marcha en el año 2003. En el transcurso de estos años, la generación de nuevos requerimientos, los cambios en los planes estratégicos, las modificaciones de imagen, la ampliación en cuanto a la tipología de documentos aceptados, entre otros, fueron generando la necesidad de modificaciones en el código fuente y la estructura del sistema. Es así que, a medida que se realizaban cambios en el software, éste se fue volviendo cada vez más complejo y difícil de mantener. Paralelo a esto, con el correr de los años, se evidenció también el avance de las aplicaciones de código abierto en el área de las bibliotecas digitales, contando hoy con varias opciones muy desarrolladas, con gran soporte y en continua ampliación.

Es por esto que se decidió proceder al reemplazo del software de SeDiCI por una nueva aplicación, de código abierto, relativamente simple de instalar y configurar, así como de adaptar para reflejar la imagen de la institución, que cuente con soporte adecuado y que se encuentre en continuo progreso.

Para esto se evaluaron principalmente dos de las aplicaciones más difundidas de la actualidad: DSpace y ePrints. De entre estas se vio que DSpace era la más cercana a los requisitos planteados para esta nueva etapa del proyecto, debido principalmente a su flexibilidad en la personalización de la herramienta: apariencia estética, formato de metadatos, extensibilidad con plugins, etc.

Actualmente, parte del personal informático de SeDiCI se encuentra abocado al análisis y evaluación detallado de esta herramienta, para intentar determinar el costo de migrar desde la plataforma actual a la aplicación DSpace.

X. Comentarios finales

En este trabajo se han presentado los aspectos más relevantes de la evolución de SeDiCI, desde el momento de su creación hasta la actualidad. Entre estos se destaca la gran evolución por la que SeDiCI ha atravesado, llegando a ser hoy en día uno de los repositorios más importantes de América Latina. Junto a esto se destaca la situación del software de SeDiCI, en vías de ser reemplazado por una aplicación de código abierto, desarrollada y mantenida por una comunidad de desarrolladores y utilizada en todo el mundo.

La experiencia adquirida con los años, junto con los continuos procesos de mejora, y las líneas de investigación en desarrollo, hablan del compromiso de SeDiCI para con la institución que lo alberga, así como para con la comunidad científica que lo rodea, siempre buscando nuevas instancias de superación.

Este documento busca servir de apoyo para aquellas instituciones que se encuentren en vías de desarrollo de su propio repositorio digital. Se han intentado delinear las principales y más comunes problemáticas en esta área, posibles medidas que viabilizan las soluciones estudiadas.

Referencias

[1] Ranking web of world repositories. http://repositories.webometrics.info. Ultima visita: 15 de Abril de 2011.

[2] Tesauro de la UNESCO. http://databases.unesco.org/thessp. Ultima visita: 15 de Abril de 2011.

[3] EuroVoc. http://eurovoc.europa.eu/drupal/?q=es. Ultima visita: 15 de Abril de 2011.

[4] DeCS – Descriptores en Ciencias de la salud. http://decs.bvs.br/E/homepagee.htm. Ultima visita: 15 de Abril de 2011.

[5] Las Tesis de Maestría y Doctorado serán preservadas en formato digital, a través del Servicio de Difusión de la Creación Intelectual (SeDiCI). http://www.sedici.unlp.edu.ar/noticias/mostrarnews.php?id_noticia=71. Ultima visita: 15 de Abril de 2011.

[6] M. R. De Giusti, A. Sobrado, A. Vosou, G. Villarreal, “Plataforma de recolección en fuentes heterogéneas de la web y su aplicación a la organización de un repositorio semántico en SeDiCI: preliminares”, III Conferencia Internacional de Biblioteca Digital y Educación a Distancia, 2009.

[7] M. R. De Giusti, A. Sobrado, A. J. Lira, M. M. Vila, G. Villarreal, “SeDiCI | Servicio de Difusión de la Creación Intelectual – UNLP”, Revista Iberoamericana de Bibliotecología, vol. 31, no. 2, 2008.

[8] M. R. De Giusti, “Aspectos técnicos de SeDiCI”, Seminario-taller "Rumbo a la Biblioteca Digital", 2004.

[9] M. R. De Giusti, G. Villarreal, A. Sobrado, A. Vosou, “Recuperación y clasificación automática de información, resultados actuales y perspectivas futuras”, V Simposio Internacional de Bibliotecas Digitales, 2009.

[10] M. R. De Giusti, G. Villarreal, A. Vosou, J. P. Martinez, “An ontology-based context aware system for Selective Dissemination of Information in a digital library”, Journal of Computing, vol. 2, no. 5, 2010.
[11] M. R. De Giusti, N. F. Oviedo, A. J. Lira, “Service Cloud for information retrieval from multiple origins”, International Conference on Engineering Education (ICEE 2010), Gliwice, Polonia, 2010.
[12] M. R. De Giusti, N. F. Oviedo, A. J. Lira, “Extract, Transform and Load architecture for metadata collection”, unpublished.