SeDiCI (Servicio de Difusión de la Creación Intelectual): un recorrido de experiencias (2003-2011)

Marisa De Giusti[*]

Resumen

El Servicio de Difusión de la Creación Intelectual (SeDiCI) es el repositorio institucional de la Universidad Nacional de La Plata (UNLP), creado en el 2003 con el objetivo de dar visibilidad a la producción académica producida en esta casa de estudios considerando que el acceso libre posibilita un mayor número de citas y por tanto un mayor impacto, atendiendo al rol fundamental de una institución pública de socializar el conocimiento. Creado en el año 2003, actualmente SeDiCI se encuentra posicionado entre los primeros 10 principales repositorios digitales de América Latina según la Webometrics, y ocupa la primera posición en Argentina como repositorio institucional. En este trabajo se presentan algunas de las principales características y servicios ofrecidos por el portal, desde su fundación hasta la actualidad.

Palabras clave: acceso abierto; repositorios institucionales; visibilidad.

Presentación

El Servicio de Difusión de la Creación Intelectual es el Repositorio Digital Institucional de la Universidad Nacional de La Plata, creado para 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, SeDiCI es una herramienta estratégica para la jerarquización de la institución.

Desde su creación en el año 2003, SeDiCI ha afrontado diversas dificultades que han influido directa o indirectamente en su desarrollo y crecimiento, pero a pesar de estas problemáticas, actualmente cuenta con una base de datos documental que supera los 15000 recursos académicos propios (de la UNLP) expuestos bajo las políticas del Acceso Abierto. Esto convierte a SeDiCI en uno de los principales exponentes en su tipo, tanto a nivel nacional como regional (América Latina).

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, ya que esta práctica posibilita 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.

Del mismo modo, SeDiCI se ha convertido en una herramienta estratégica para la jerarquización de la universidad 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 el de América Latina (índice Webometrics, enero de 2011). El mandato de la Universidad Nacional de La Plata de febrero de 2011 que establece la obligatoriedad del depósito de las tesis de posgrado en SeDiCI lo reafirma como repositorio central.

[*] Directora del Proyecto de Enlace de Bibliotecas (PrEBi), del Servicio de Difusión de la Creación Intelectual (SeDiCI), y de la iniciativa LibLink del Iberoamerican Science and Technology Consortium (ISTEC)

A continuación se ofrece un panorama de las principales características del servicio.

Software

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) y 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 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 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. A lo largo de los años se ha ido actualizando la plataforma y compatibilizando el desarrollo con los estándares actuales.

Representación de los datos

Durante el desarrollo de SeDiCI se analizaron los 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, que permite así administrar de forma simple los metadatos disponibles (agregar, modificar y eliminar metadatos cuando sea necesario). Asimismo se establecieron normas de catalogación que indican al personal encargado de estas tareas qué metadatos deben utilizar (ya sea 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 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, lo que implica distintos tipos de consultas según lo que se desee obtener, 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.

Catalogación

A la hora de la catalogación en el repositorio se parte del supuesto de que cuantos más metadatos posea el recurso, más fácil y accesible será encontrarlo para los usuarios finales. Sin embargo, en ocasiones resulta 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 registros desde otros repositorios.

Como en todo sistema en continuo crecimiento, son numerosas las mejoras que deben realizarse para que el funcionamiento sea óptimo todo el tiempo. En este sentido, una de las estrategias implementadas fue la de 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 que abarcan el amplio rango de creaciones de las 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.

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 tenido 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, 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 central. Mandatos como el referido precedentemente aseguran que SeDiCI cuente con todos los datos requeridos para su correcta catalogación, a través del autoarchivo o bien que éste sea entregado personalmente por el autor.

Para optimizar el flujo de documentos 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 sus obras en SeDiCI.

Importación de recursos

Como ya se dijo, el principal objetivo de SeDiCI es reunir, preservar y publicar toda la producción de la Universidad Nacional de La Plata. Existen actualmente otros repositorios digitales que contienen y exponen documentos producidos en la universidad, lo que para SeDiCI representa una gran ventaja. Es decir, tener 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.

A continuación se citan algunas dificultades encontradas.

La primera gran dificultad fue la necesidad de discriminación de los recursos propios preexistentes dentro de SeDiCI de los documentos recolectados a través del protocolo OAI-PMH. Por otra parte, SeDiCI 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.

Otra de las dificultades encontradas es 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 (más completo y descriptivo), plantea la necesidad de realizar mapeos y transformaciones a la información importada, lo que implica pérdida de información o incluso generación de registros incompletos. Una forma de evitar esto es la intervención del personal especializado, encargado de la revisión y corrección de los recursos importados, pero esto aumenta notablemente el costo de las importaciones.

Servicios

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

A. Recuperación de la información

Si bien el menor tiempo de respuesta posible a una consulta de un usuario en el portal es un factor de extrema importancia, existe otro factor que resulta más crucial: 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 basan 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 (DSI)

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

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 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. Hasta el momento el servicio sólo está disponible para usuarios registrados en el sitio; no se descarta que en un futuro próximo el servicio se haga extensible a cualquiera que desee usarlo.

C. Carpetas

Cuando el usuario descubre un recurso que resulta de su interés, es deseable que cuente con un mecanismo de marcado de dicho recurso, de modo que pueda ser accedido en el futuro sin necesidad de realizar una nueva búsqueda. Para esto, SeDiCI cuenta con la posibilidad de marcar un recurso como favorito, dejándolo accesible desde el sitio de usuario.

Por otro lado, con el paso del tiempo, la lista de recursos favoritos puede incrementarse y llegar a ser demasiado 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 podían incluir 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 cargar el archivo correspondiente. Estos datos eran posteriormente revisados y completados por personal especializado de SeDiCI, y en caso de aceptarse el trabajo propuesto, éste quedaba accesible desde el portal web.

Esta metodología de colaboración por parte de los usuarios implica numerosas discusiones. Entre ellas: la necesidad de una autorización para el resguardo y publicación del trabajo 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, y los usuarios comienzan a usarla y disfrutar de sus ventajas, alineando así a SeDiCI a otras grandes instituciones del mundo que implementan el autoarchivo.

Líneas de investigación actuales

SeDiCI se encuentra en continuo desarrollo 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 (es decir, sólo sus 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.

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. Con una configuración debidamente adaptada, este motor permite realizar búsquedas en el orden de los milisegundos, al tiempo que provee gran cantidad de funcionalidad adicional muy valiosa.

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 simple de implementar y utilizar. Además de este protocolo, existen otras fuentes de información desde las cuales se pueden obtener recursos relevantes, como 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 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 gran número de características y capacidades de procesamiento que suman valor agregado a la información recolectada. 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, 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 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: Article, Artículo, ART significan todos que el tipo de recurso es un artículo científico. 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 la capacidad para realizar transformaciones simples a los metadatos descargados, como 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, que determinará un avance significativo en la búsqueda de soluciones que simplifiquen las tareas de los usuarios al buscar información relevante en el 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.

Para esto se requiere 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, por áreas temáticas, por palabras-clave, por autores, etc. Esto también 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.

Cambio del software de soporte

El software de SeDiCI lleva muchos años en funcionamiento, y en la actualidad cumple con la funcionalidad básica de un repositorio digital. Sin embargo, por las razones que se comentan a continuación, se ha resuelto realizar la migración a 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, sus capacidades de ampliación son limitadas, al tiempo que se dificulta el mantenimiento del sistema por no contar con el soporte adecuado (tecnología discontinuada).

Como se mencionó, 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. Con el correr de los años, fue notable también el avance de las aplicaciones de código abierto en el área de las bibliotecas digitales, área en la que se cuenta hoy con varias opciones muy desarrolladas, con gran soporte y en continua ampliación.

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

Para esto se evaluaron 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 de SeDiCI, debido principalmente a su flexibilidad en la personalización de la herramienta: apariencia estética, formato de metadatos, extensibilidad con plugins, etc.

Actualmente, personal informático de SeDiCI se encuentra abocado al análisis detallado de esta herramienta, para determinar el costo total de migrar desde la plataforma actual a DSpace.

Conclusiones

Hasta aquí 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 el repositorio ha atravesado, llegando a ser hoy día uno de los 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 en la búsqueda de nuevas instancias de superación.