



UNLP



Proyecto de Enlace
de **Bibliotecas**



BIREDIAL 2011 Primer Conferencia sobre Bibliotecas y Repositorios Digitales

SeDiCI

SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

Desafíos y experiencias
en la vida de un Repositorio Digital

Ing. Marisa R. De Giusti :: Nestor F. Oviedo



Servicio de Difusión de la Creación Intelectual

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 primer posición en Argentina como repositorio institucional.



Desafíos y Experiencias

En este trabajo se presentan las principales problemáticas afrontadas por SeDiCI desde su creación, con la intención de que sirva como referencia para aquellas instituciones que se encuentren en vías de desarrollar sus propios Repositorios Institucionales.

En cada caso se describe el problema, su contexto y las vías de acción tomadas para superarlo



Tópicos cubiertos

Los temas a tratar son:

- Selección del software
- Representación de recursos
 - Formato de metadatos
- Catalogación
- Apoyo institucional
- Importación de recursos
- Servicios
 - Recuperación de la información
 - Diseminación Selectiva de la Información
 - Carpetas
 - Autoarchivo



Tópicos cubiertos

- Líneas de investigación actuales
 - Gestión de grandes volúmenes de información
 - Cosecha de recursos por diferentes medios
 - Procesos de transformación y mejora de la información
 - Ontologías y repositorios semánticos
- Problemática actual: cambio del software de soporte
- Comentarios finales



Selección del software

Decisión clave para las tareas administrativas y la exposición de recursos

Para SeDiCI se analizaron varias aplicaciones disponibles en aquel momento

- CyberThesis (Francia y Chile)
- Proyectos ETD de
 - UNICAMP (Brasil)
 - Virginia Tech (USA)
 - Montreal (Canadá)
 - Universidad de Valencia (España)



Selección del software

Características mínimas requeridas:

- de uso libre
- de código abierto
- con soporte de un formato de metadatos propio
- simple para la personalización
- con buena escalabilidad
- que permitiera actualizaciones frecuentes
- dotado de soporte y documentación

**No se encontró una aplicación que
reuniera estas características en los albores del 2003**

Se procedió al desarrollo de una aplicación propia



Selección del software

Celsius DL

- Aproximadamente 4 meses de desarrollo (análisis, implementación y pruebas)
- Administración: Java y Swing (Interfaz de usuario)
- Portal: PHP4
- Base de Datos: MySQL 4

Celsius DL ha sido el software de soporte del repositorio desde sus inicios, y lo sigue siendo en la actualidad



Selección del software

Celsius DL

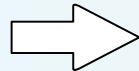
- Portal: funciona como una aplicación web accesible de forma pública. Cuenta con funciones de búsqueda y exploración, noticias, links y servicios adicionales para usuarios registrados.
- Administración: funciona como una aplicación Java Desktop. Además de catalogación provee registro de autores, tesauros, sistemas de clasificación, gestión del formato de metadatos propio, entre otras funciones.
- Funciona como OAI Data Provider, exponiendo todos los recursos catalogados, agrupados en conjuntos según su tipología documental.
- Provee un web-service de búsqueda, accesible de forma pública.



Representación de Recursos

- Los recursos son el elemento central en estos diseños.
- Existe una gran diversidad de Tipos de Recursos

Tesis
Libros
Artículos
Patentes



Diferentes modos de representación y tratamiento (estructura de metadatos, normalización, vocabularios controlados, etc.).

Almacenamiento cuidadoso en pos de recuperación eficiente, preservación e interoperabilidad.



Representación de Recursos / Formato de metadatos

Aspectos a tener en cuenta:

- Diversidad de recursos
- Debe ser soportado por el software del repositorio: posibilidad de representar todos los metadatos necesarios para todos los tipos de recursos en/los formatos de metadatos y que el software los pueda gestionar
- Nivel de interoperabilidad deseado: puede estar limitado por las capacidades del software y/o por el formato de metadatos elegido
 - Necesidad de mapeos
 - Pérdida de información
- Es deseable que sea un formato completo (al menos para el tipo de recurso que se pretende representar)




Representación de Recursos / Formato de metadatos

En los inicios de SeDiCI, al momento de seleccionar el formato de metadatos, se observó:


- Incertidumbre en cuanto a los tipos de documento que se contemplarían en el futuro cercano aunque desde un principio se supo que no habría una tipología única
- El desarrollo de los estándares de formatos de metadatos no estaba tan avanzado como en la actualidad
- El software sería desarrollado por SeDiCI, lo cual aportaba libertad para la elección del formato de metadatos

Se optó por el uso de un formato de metadatos propio


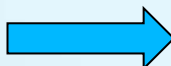


Representación de Recursos / Formato de metadatos

Ventajas

- 
- Alta flexibilidad para la catalogación
 - Soporte para múltiples tipos de documento
 - Creación de metadatos según la necesidad
 - Completo

Desventajas

- 
- Conjunto de tablas y relaciones complejo
 - Cada metadato puede ser un texto libre, una fecha, un término controlado, un código de un sistema de clasificación e incluso e incluso una referencia a otra tabla (un autor de la tabla de autores)
 -  consultas diversas
 - Con el crecimiento del repositorio se presentaron problemas de performance en la recuperación de registros



Catalogación

Contar con un formato de metadatos propio es una ventaja, ya que brinda la posibilidad de catalogar recursos no contemplados previamente, como vídeos, audio, etc., simplemente agregando los metadatos faltantes que se necesiten. Pero...la misma flexibilidad obliga a ser muy cuidadosos :

Se vio la necesidad de una guía de catalogación

- Describe metadatos de carácter obligatorio y optativo según el tipo de recurso
- Establece reglas de normalización a utilizar (fechas, nombres, vocabularios controlados, etc)
- Se aumenta la consistencia de las catalogaciones
- Se disminuyen los errores



Catalogación

Cuanto más metadatos posea el recurso es más fácil su ubicación, sin embargo existen problemas comunes:

- Falta de información básica (ej.: fecha de publicación)
- Información errónea (ej.: autores incorrectos, títulos incompletos)
- Falta del material físico, lo que dificulta
 - Comprobación de los datos
 - Obtener información adicional
- Dificultad para contactarse con los autores (ej.: falta de información de contacto)
- Necesidad de chequeos periódicos para verificar la existencia y accesibilidad a los documentos enlazados desde SeDiCI



Catalogación

Los procesos de catalogación en SeDiCI han mejorado considerablemente

- Aumento del número de recursos catalogados diarios
- Mejoras destacables en cuanto a la cantidad y calidad de los datos
- Designación de un grupo de personas especializadas en área
- Incorporación de nuevos tesauros



Apoyo institucional

SeDiCI no hubiera sido posible sin el apoyo de las autoridades de la Universidad Nacional de La Plata.

Gracias a esto:

- Aumento del personal
- Mejoras en equipamiento
- Mayor difusión de los servicios que SeDiCI brinda
- Resolución 78: depósito obligatorio de Tesis de Postgrado

La UNLP se ve beneficiada, ya que SeDiCI

- Amplía la visibilidad y el impacto de la institución en el mundo científico
- Es una herramienta estratégica para la jerarquización de la institución



Importación de recursos

Gracias al avance en el área de la interoperabilidad, particularmente el protocolo OAI-PMH, SeDiCI ve la posibilidad de recolectar recursos de la UNLP albergados en otros repositorios, buscando así agilizar las tareas de catalogación.

Sin embargo, se presentan diversos problemas:

- Identificación de los recursos de la UNLP
- Selección de la tipología deseada
- Necesidad de mapeos y transformación
 - Pérdida de información
 - Documentos incompletos como resultado
- La validación y corrección por parte de personal especializado es inevitable si se desea garantizar la calidad de los datos



Servicios

- Lo más importante: proveer a sus usuarios la mayor cantidad de posibilidades para la búsqueda y recuperación de documentos.
- Servicios adicionales:
 - Recuperación de información: búsqueda simple y avanzada. Mejoras en esta funcionalidad utilizando un indexador de texto que permite disminuir los tiempos de respuesta y brindar resultados más pertinentes.
 - Diseminación selectiva: desarrollo propio de perfiles de usuarios registrados.
 - Carpetas: se provee un mecanismo de organización dinámico para la lista de los recursos favoritos de los usuarios.



Servicios

- **Autoarchivo:** desde los inicios se brindó la posibilidad a los usuarios registrados de cargar los datos básicos y subir archivos. Constatación posterior por parte de los administradores.
 - Discusiones generadas: cesión de derechos y licencias de uso.
 - Actualmente se ha completado el circuito de autoarchivo:
 - Registro del usuario
 - Ingreso al sitio personal
 - Acceso al Acuerdo de cesión no exclusiva de derechos patrimoniales
 - Ingreso de datos
 - Subida de archivos
 - Posibilidad de elegir una licencia de uso



Líneas de investigación

SeDiCI se encuentra en continuo desarrollo de distintas líneas de investigación, siempre con el primordial objetivo de proveer más y mejores servicios.

Principales líneas de investigación actuales

- Gestión de grandes volúmenes de información
- Cosecha de recursos por diferentes medios
- Procesos de transformación y mejora de la información
- Ontologías y repositorios semánticos



Líneas de investigación

Gestión de grandes volúmenes de información

Una problemática recurrente en el área de los repositorios digitales.
Más de 16 millones de registros obtenidos por harvesting OAI.

¿Cual es la mejor forma para buscar y recuperar un documento dentro de esta gran cantidad de registros?

- BD XML
 - Consultas complicadas (XQuery)
 - Funciones administrativas (ABM) optimizadas
 - Matching exacto (poca utilidad)
- BD Relacional
 - Representación poco clara (Tablas contra jerarquías en XML)
 - Generación de tuplas de datos a partir de registros XML
 - Consultas potencialmente complicadas (SQL y JOINS)
 - Indices fulltext en algunos casos



Líneas de investigación

Gestión de grandes volúmenes de información

- Indexador de Texto (Apache Solr)
 - Consultas simples
 - Matching por aproximación
 - Faceting
 - Calculo de y ordenamiento por relevancia
 - Gran número de funciones adicionales (faceting, dismax, boosting, stopwords, stemming, etc)
 - Tiempos de respuesta en el orden de los milisegundos
 - Fácil configuración

Se desarrolló un portal de búsquedas que explota la funcionalidad provista por este motor de indexación (aún en etapas de prueba)



Líneas de investigación

Cosecha de recursos por diferentes medios

OAI-PMH es el protocolo de interoperabilidad mas difundido

Otras potenciales fuentes de datos

- Web-services
- Bases de Datos
- Web (por medio de crawling)

Cada tipo de fuente de datos establece sus propias reglas de comunicación y transferencia

Se buscó explotar estas fuentes alternativas de información, intentando extender la recolección de recursos para alcanzar a aquellos repositorios que no cuentan con un OAI Data Provider, al tiempo que se amplía la cobertura sobre los repositorios que exponen recursos por otros medios



Líneas de investigación

Cosecha de recursos por diferentes medios

Se realizó una extensión al software de recolección de metadatos desarrollado en SeDiCI, denominado **Celsius Harvester**.

Estas extensiones transformaron a Celsius Harvester en una implementación de la arquitectura Extract, Transform and Load (ETL, comunmente utilizada en el área de data mining)



Líneas de investigación

Cosecha de recursos por diferentes medios

Extract

soporte (extensible) de múltiples protocolos de comunicación y transferencia para la recolección de recursos

Transform

Transformación de los recursos a un formato intermedio, sobre el cual se pueden aplicar un conjunto de transformaciones, en busca de mejorar la calidad de los datos. Ejemplos son el reemplazo de términos por vocabularios controlados, normalización de fechas y códigos de idioma, extracción de información derivada de los datos, etc. Los componentes de transformación pueden ser conectables.

Load

Soporte (extensible) de múltiples medios de almacenamiento (o destino) para la información transformada. Ejemplos de esto son Apache Solr, archivos, web-services, etc.



Líneas de investigación

Procesos de transformación y mejora de la información

Ligado al problema de la gestión de grandes volúmenes de información. El problema es la **heterogeneidad** de los datos recolectados.

| | Tipo | Fecha | Idioma |
|----------------------|----------|--------------------------|---------|
| Valores recolectados | Article | 25/01/2011 (dd/mm/yyyy) | spa |
| | Art | 25-01-2011 (dd-mm-yyyy) | spanish |
| | Artículo | 01-25-2011 (mm-dd-yyyy) | español |
| | Arti | 25-ene-2011 (dd-mm-yyyy) | es |
| Valor normalizado | ARTÍCULO | 25-01-2011 (dd-mm-yyyy) | ESPAÑOL |

Estos y otros tipos de normalización han sido incluidos dentro de la etapa de Transformación de Celsius Harvester.

Estas transformaciones representan grandes mejoras para la búsqueda y recuperación de información. También permitirían continuar con investigaciones en áreas relacionadas a la extracción de conocimiento.



Líneas de investigación

Ontologías y repositorios semánticos

Una de las mayores ambiciones de SeDiCI

Representan nuevas y mejores formas de navegar dentro de la base documental.

Amplían el espectro de posibilidades en el área de recuperación de la información, proporcionando un marco propicio para nuevos desarrollos

Se trabaja en:

- La definición de ontologías correctas y extensibles
- Extracción de relaciones a partir de los recursos catalogados, haciendo hincapié en la automatización, para la población de las ontologías definidas
- Mejorar el portal web para que refleje las nuevas funciones derivadas de estas extensiones sobre la información



Problemática actual: cambio del software de soporte

Con el paso de los años, Celsius DL se vio afectado por

- múltiples cambios de requerimientos, políticos y estratégicos, que se reflejaron en modificaciones en el software.
- nuevas y mejores versiones de las tecnologías utilizadas en el desarrollo original.
- Aumentó la necesidad de personal dedicado a su mantenimiento.
- La aplicación se volvió compleja y difícil de mantener.
- Se incrementó notablemente la cantidad y calidad de desarrollos Open Source dedicados a las bibliotecas y repositorios digitales.
- Algunas aplicaciones evolucionaron y tuvieron una gran diseminación.



Cambio del software de soporte

Se resolvió reemplazar Celsius DL

Se confeccionó una lista de características deseables para la aplicación que reemplazaría a Celsius DL

- licencia de uso libre y gratuita
- alto grado de aceptación en la comunidad de repositorios digitales
- sección de administración solo para usuarios con privilegios
- capacidades de personalización y extensión (interfaz de usuario y funcionalidad)
- buena y completa documentación
- actualizaciones continuas (proyecto activo)
- soporte de usuarios administradores y desarrolladores
- facilidad de uso, tanto para usuarios como para administradores
- posibilidad de elegir el formato de metadatos a utilizar
- posibilidad de especificar un formato de metadatos propio
- buena performance en los servicios
- escalabilidad
- funciones para la interoperabilidad: importación, exportación, web-services de búsqueda, OAI-PMH, etc.
- de simple instalación y configuración



Cambio del software de soporte

Principal Candidato



- ✓ Flexible
- ✓ Extensible con plugins
- ✓ Instalación simple
- ✓ Código abierto
- ✓ Formato de metadatos configurable
- ✓ Respaldo de una comunidad de desarrolladores

El equipo de SeDiCI se encuentra evaluando todos los aspectos de DSpace para determinar si es el reemplazo adecuado para Celsius DL



Comentarios finales

Al menos cinco elementos pilares con los que se debe contar para lograr un repositorio digital

- Software de soporte
- Formato de Metadatos
- Catalogación
- Apoyo Institucional
- Mejora continua

Se ha presentado la experiencia de SeDiCI acerca de estos y otros elementos, los caminos tomados y sus consecuencias.



Agradecimientos

GRACIAS!!

A la Universidad Nacional de La Plata

A los organizadores de BIREDIAL por invitarnos a este evento

A los asistentes

A Lucas Folegotto, diseñador de PrEBi-SeDiCI