III Conferecia Bibliotecas y Repositorios Digitales de América Latina (BIREDIAL '13)
VIII Simposio Internacional de Bibliotecas Digitales (SIBD '13)
"ACCESO ABIERTO, PRESERVACIÓN DIGITAL Y DATOS CIENTÍFICOS"
Ciudad de la Investigación, Universidad de Costa Rica, del 15 al 17 de octubre de 2013.
Ponencia

La Representación de Recursos en los Repositorios Institucionales. El Caso de SEDICI, Repositorio Institucional de la Universidad Nacional de La Plata

Jose Texier
Universidad Nacional Experimental del Táchira (UNET), Venezuela
Servicio de Difusión de la Creación Intelectual, Universidad Nacional de La Plata (SEDICI), Argentina
CONICET, Argentina
jtexier@unet.edu.ve; dantexier@sedici.unlp.edu.ar

Marisa R. De Giusti, Nestor Oviedo, Gonzalo L. Villarreal y Ariel J. Lira
Servicio de Difusión de la Creación Intelectual, Universidad Nacional de La Plata (SEDICI), Argentina
{marisa.degiusti, nestor, gonzalo, ariel}@sedici.unlp.edu.ar

RESUMEN:
En el 2003, nace el Servicio de Difusión de la Creación Intelectual (SEDICI) como el repositorio institucional de la Universidad Nacional de La Plata (UNLP), con el objetivo prioritario de socializar el conocimiento generado en las diferentes áreas académicas de la Universidad, para devolver a la comunidad los esfuerzos destinados a la Universidad Pública con el pasar de los años la plataforma de software de SEDICI fue creciendo en funcionalidades por lo que el diseño, mantenimiento e implementación la convirtieron en una herramienta compleja. Por ello, se realizó una migración a la plataforma DSpace en el 2012, herramienta de software que mejor se adaptaba a las necesidades de SEDICI. DSpace, al igual que otras plataformas, presenta limitaciones en diferentes ámbitos tales como: procesos de depósitos, manejo de estadísticas y de las comunidades-colecciones-items, vocabularios controlados centrados en los autores, representación de los recursos, etc. De estas limitaciones, la representación de los recursos -proceso de registrar en forma persistente un conjunto de datos de los recursos- fue uno de los principales problemas presentados en la transición a DSpace, ya que se migraron muchos recursos en forma separada e incluso se realizaron adaptaciones por incompatibilidad de ambos sistemas porque la representación de distintos recursos era muy variada. Por tanto, este trabajo se centrará en describir el problema de la representación de recursos dentro de un repositorio institucional y en exponer la experiencia de SEDICI en este ámbito. El trabajo se limitará a la tipología de recursos: artículos de investigación, tesinas de grado y tesis de postgrado, libros, autores, instituciones, revistas y sus números, eventos y sus instancias.

Palabras clave: representación de recursos, repositorios institucionales, SEDICI, DSpace.

1. Introducción

En el 2003, nace el Servicio de Difusión de la Creación Intelectual (SEDICI) [1] como el repositorio institucional de la Universidad Nacional de La Plata (UNLP), con el objetivo prioritario de socializar el conocimiento generado en las diferentes áreas académicas de la Universidad, para devolver a la

comunidad los esfuerzos destinados a la Universidad Pública. SEDICI, como un portal de acceso libre, estuvo soportado por un desarrollo de software propio en PHP, MySQL y Java, llamado Celsius DL, que estaba adaptado a estándares internacionales como XML, OAI-PMH, SOAP, etc., y permitía el depósito y búsqueda de recursos (objetos físicos o digitales), por ejemplo: artículos de publicaciones periódicas, preprints, tesinas de grado y tesis de postgrado, producciones multimediales, libros electrónicos, autores, revistas, eventos, comunidades, colecciones, entre otros. Celsius DL fue creciendo en funcionalidades por lo que el diseño, mantenimiento e implementación convirtieron a la plataforma de software de SEDICI en una herramienta compleja, principalmente por dos razones: por una parte porque se requería de mayor tiempo y recurso humano y, por otra, porque se hacía necesaria una actualización de tecnologías de manera que no estuvieran en detrimento de nuevos desarrollos e investigaciones.

En razón de eso, para finales del 2011, se estudió la posibilidad de migrar a una plataforma que estuviera a la par de las nuevas tecnologías aplicadas al dominio y que fuera más amena para el usuario y para la gestión de recursos por parte del personal de SEDICI. Este análisis, realizado por el personal de SEDICI [2], tenía el objetivo de comparar diferentes plataformas de software: DSpace, EPrints, FEDORA y Greenstone, tomando en cuenta:

- licencias de uso libres y gratuitas,
- un alto nivel de aceptación por parte de la comunidad de repositorios digitales,
- contar con una sección de administración,
- proveer de mecanismos de personalización,
- mantener una actualización periódica de las versiones de la plataforma,
- soporte técnico para administradores y desarrolladores,
- manual de usuario y técnico,
- permitir selección de esquemas de metadatos,
- asegurar la performance para obtener tiempos de respuesta bajos,
- capacidades de escalabilidad, y,
- proveer de estándares que garanticen la interoperabilidad entre repositorios.

La conclusión del estudio determinó que DSpace era la herramienta de software que mejor se adaptaba a las necesidades de SEDICI [3], [4], [5], [6], [7]. El proceso de migración de Celsius DL a DSpace finalizó en el 2012. Esto permitió a SEDICI contar con nuevas funcionalidades como: búsquedas y faceting, proveedor de datos y de servicios de acuerdo con el protocolo OAI-PMH, autenticación de usuarios, facilidad en la indexación por parte de los buscadores Web, etc., En otras palabras, se logró una flexibilidad en los cambios estratégicos por parte de la gerencia del repositorio, modificaciones en la imagen institucional sin afectar el modelo de negocio del repositorio (procesos internos de software) y ampliar la gestión de las tipologías de los recursos aceptados, entre otros.

Celsius DL y SEDICI-DSpace, son plataformas de repositorios donde su principal proceso es registrar en forma persistente un conjunto de datos como síntesis y reemplazo del objeto "real" para poder identificarlo, recuperarlo y distribuirlo por parte de los usuarios. Este proceso, conocido como

la representación de recursos, por lo general, es llevado de diferentes formas en las mismas plataformas, por ejemplo, se observan diferentes personalizaciones en DSpace o otras plataformas usadas [8], [9]. Es importante destacar que DSpace, al igual que las demás plataformas, presenta limitaciones (procesos de depósitos, manejo de estadísticas y de las comunidades-colecciones-items, vocabularios controlados centrados en los autores) en cuanto al proceso de la representación de los recursos en los repositorios [6], [10], situación particular que generó un inconveniente en la transición de Celsius DL a DSpace en SEDICI, ya que se migraron muchos recursos en forma separada e incluso se realizaron adaptaciones por incompatibilidad de ambos sistemas porque la representación de distintos recursos en Celsius DL y DSpace era muy variada.

La representación de recursos es un problema recurrente que ha sido estudiado por algunos autores como Malizia [11], Paganelli [12], Gonçalves [13], Candela [14], entre otros. No obstante, sus trabajos abordan el tema mediante sistemas diseñados en forma general en los que no se toma en cuenta el recurso como el eje central.

2. Los Recursos

Los recursos son objetos físicos o digitales que se describen a partir de la enumeración de un conjunto de datos específicos (metadatos) que lo distinguen entre otros objetos y se pueden clasificar de la siguiente manera:

- Producción académica. Realizada en instituciones de educación e investigación, entre las que se encuentran: artículos de investigación, trabajos docentes, tesinas de grado y tesis de postgrado, disertaciones, libros electrónicos, presentaciones, objetos de aprendizaje.
- Producción multimedia. Por ejemplo: imágenes, música, audios, videos.
- Producción institucional y administrativa. Generada en instituciones privadas y públicas tales como: documentos generales, constancias, memorandos, ordenanzas, resoluciones, decretos, actas, minutas, notas, leyes.
- Entidades abstractas. Es el conjunto de elementos que poseen información descriptiva propia, utilizadas en los procesos de catalogación de recursos como elementos de un vocabulario controlado. Por ejemplo: autores, instituciones, revistas y sus números, eventos y sus instancias.
- Comunidades y colecciones. Las comunidades simbolizan a entidades administrativas de instituciones tales como departamentos, laboratorios, oficinas, centros de investigación, entre otros. En cambio, las colecciones simbolizan el lugar donde pertenecen los recursos y no pueden contener otras categorías, y deben encontrarse dentro de una comunidad.

El presente trabajo se limitará a describir la siguiente tipología de recursos presente en SEDICI, ya que existe una gran cantidad de tipos de recursos, los cuales con el paso del tiempo tienden a aumentar, de acuerdo con las necesidades de los usuarios y la evolución de las tecnologías. A continuación se van a describir cada uno de ellos y enumerar sus respectivas características:

1. Artículos de investigación. Es un documento donde se expresa, por parte del autor, un trabajo de investigación redactado en forma clara y precisa. Por lo general, siguen una estructura

sugerida con un aporte original a la comunidad científica y debe incluir referencias que permitan verificar y reproducir los aportes que se dan a conocer. La verificación es conocida como revisión por pares, por ello, son publicados en revistas arbitradas. Las principales características que identifican estos tipos de recursos son:

a. Título.
b. Subtítulo.
c. Fecha de publicación.
d. Idioma.
e. Autor*.
f. Autor institucional*.
g. Descripción física.
h. Palabras clave.
i. Descriptores.
j. Tipo de documento.
k. Identificadores geográficos.
l. Abstract.
m. URL de acceso.
n. Nota.

2. Tesis. Es un trabajo científico sometido a un sistema de reglas que dependen del grado académico que se esté optando, por ejemplo: pregrado, especialización, maestría, doctorado. Los tipos de tesis a tomar en cuenta en este trabajo son: tesina de grado, trabajo de especialización, tesis de maestría y tesis de doctorado. Las características que los engloban son:

    a. Título del documento.
    b. Subtítulo.
    c. Autor*.
    d. Autor institucional.
    e. Fecha de presentación.
    f. Director de la tesis.
    g. Co-director de la tesis*.
    h. Miembros del jurado*.
    i. Descriptores.
    j. Descripción del recurso.
    k. Grado alcanzado.
    l. Palabras clave.
    m. Lugar de desarrollo.
    n. Idioma.
    o. Tipo de documento.
    p. Abstract.
    q. Localización física.
    r. Localización electrónica.

    s. Nota.
3. Libro. Es una obra científica, literaria o de cualquier otra índole que puede aparecer impresa o en cualquier otro soporte. Actualmente los libros pueden ser escaneados o producidos en algún formato electrónico. Las propiedades que se destacan de ellos son:
    a. Título del documento.
    b. Subtítulo.
    c. Autor*.
    d. Fecha de publicación.
    e. Idioma.
    f. Editor*.
    g. ISBN.
    h. Editorial.
    i. Descripción física.
    j. Palabras clave.
    k. Descriptores.
    l. Tipo de documento.
    m. Localización electrónica.
    n. Nota.
4. Entidad abstracta autor. Representa la información del creador de algún tipo de recurso, del director de una tesis, del jurado, etc.
    a. Nombres.
    b. Apellidos.
    c. Correo electrónico*.
    d. Ciudad.
    e. Pais.
    f. Institución*.
    g. Miembro de grupos.
    h. Fecha de nacimiento.
    i. Fecha de defunción.
    j. Profesión*.
    k. Ocupación.
    l. Sexo.
    m. Idiomas.
    n. Alias*.
    o. Nota.
5. Entidad abstracta institución. Identifica la entidad a la que pertenece el autor y/o el recurso.
    a. Nombre.
    b. Ciudad.
    c. Pais.
    d. Dirección.
    e. Alias*.

    f. Nota.
6. Entidad abstracta revista. Es una publicación periódica que agrupa un conjunto de artículos científicos de diversas disciplinas y se asegura su calidad a través de la revisión por pares (arbitraje).
    a. Título de la serie.
    b. Subtítulo.
    c. ISSN.
    d. Frecuencia.
    e. Editor*.
    f. Materias*.
    g. Indizada*.
    h. Director*.
    i. Subdirector*.
    j. Cómite de referato*.
    k. Secretaría de redacción*.
    l. Volumen.
    m. Número.
    n. Tipo de documento.
    o. Localización física.

2.1. Esquemas de Metadatos

Las características enumeradas para cada uno de los recursos descritos anteriormente representan los metadatos de cada uno de ellos. El concepto de metadatos no es algo nuevo, antes de la aparición de Internet ya se había usado en la catalogación de libros y revistas para normalizar la información de manera que se pueda recuperar de una forma organizada. Pero en el ámbito de las ciencias de la información, los metadatos se emplean para referirse a registros de recursos de información disponibles [15]. Por tanto, se puede entender por metadatos como datos que describen otros datos, es decir, es información estructurada que describe, explica y/o localiza un recurso de información para poder identificarlo, recuperarlo, utilizarlo, administrarlo o preservarlo de una manera más clara y sistemática. Para la representación de metadatos se han desarrollado distintos modelos, esquemas, formatos o estándares, que si bien comparten una sintaxis y estructura de la información en XML, difieren atendiendo a los propósitos de la información que describen [16]. Entre estos esquemas de metadatos más utilizados en el mundo de la ciencia de la información se encuentran: DC (Dublin Core Metadata Initiative), MARC (Machine Readable Cataloging), MODS (Metadata Object Description Schema), MADS (Metadata Authority Description Schema), METS (Metadata Encoding and Transmission Standard), PREMIS (Preservation Metadata: Implementation Strategies).

- Dublin Core (DC) [17]. Es un simple pero eficaz conjunto de elementos para describir recursos, cada elemento es opcional y puede repetirse. La norma del Dublin Core conlleva dos niveles: Simple y Cualificado. El Dublin Core Simple presenta quince elementos (Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Subject, Source, Title y Type) y el Dublin Core Cualificado presenta otros

elementos adicionales al simple. La semántica del Dublin Core ha sido establecida por un grupo internacional e interdisciplinario de profesionales de la biblioteconomía, la informática, la codificación textual, y otros campos teórico-prácticos relacionados.

- MARC 21 (MAchine-Readable Cataloging - Century 21st) [18]. Permite estructurar e identificar los datos bibliográficos (tal como títulos, nombres, temas, notas, información sobre publicación, y descripción físicas de ítems) de tal forma que puedan ser reconocidos y manipulados por computadora. Este formato fue creado en 1999 como un resultado de la combinación de los formatos MARC de Estados Unidos y Canadá. Por tanto, es un formato redefinido de MARC y por ende, también tiene cinco clases: *Bibliographic Format, Authority Format, Holdings Format, Community Format, and Classification Data Format*.
- Metadata Object Description Schema (MODS) [19]. Esquema para la representación de registros derivados del formato bibliográfico del MARC 21 pero usando etiquetas basadas en denominaciones textuales. Según Tránsito Ferreras [20], MODS es más completo y está más orientados a bibliotecas que DC, está más orientado al usuario MARC-XML.
- Metadata Authority Description Schema (MADS) [21]. Relacionado con MODS, representa el formato de autoridad según MARC 21. De este modo, permite incluir información sobre agentes (personas y organizaciones), eventos y términos (conceptos, géneros, etc.).
- Metadata Encoding and Transmission Standard (METS). [22] Está desarrollado por Network Development and MARC Standards Office de la Library of Congress. Es un formato que registra la estructura jerárquica y contenedora de un objeto digital: nombre, archivos, ubicación, estructura y metadatos asociados. Un documento METS posee un formato estandarizado para transmisión de metadatos que se estructuran en XML. Un documento METS consta de siete secciones: cabecera METS, metadatos descriptivos, metadatos administrativos, archivo, mapa estructural, enlaces estructurales y comportamientos.
- Preservation Metadata Implementation Strategies (PREMIS) [23]. Se enfoca en estrategias de implementación de metadatos de preservación de recursos digitales. PREMIS esta conformado por un grupo de trabajo internacional patrocinado por Online Computer Library Center (OCLC) y Research Libraries Group (RLG) que en el 2008, elaboró el diccionario de datos PREMIS para metadatos de preservación [24], el cual define los metadatos de preservación como "la información que utiliza un repositorio para dar soporte al proceso de preservación digital". El diccionario define un conjunto de unidades semánticas, de propiedades y de información que la mayoría de los repositorios necesita conocer de sus entidades para asegurar la preservación. El modelo de datos PREMIS define cinco entidades: entidades intelectuales, objetos, agentes, acontecimientos y derechos.
- Finalmente, existen otros estándares de metadatos que se han desarrollado por distintos usuarios e instituciones, y, que han ayudado a consolidar repositorios de datos en diferentes áreas. Algunos de los estándares más conocidos se nombran a continuación pero no serán parte de esta propuesta:
    - Darwin Core.
    - DDI (Data Documentation Initiative for Social and Behavioral Sciences Data).
    - DIF(Directory Interchange Format for Scientific Data).

- EML (Ecological Metadata Language).
- FGDC/CSDGM (Content Standard for Digital Geospatial Metadata).
- NBII (National Biological Information Infrastructure).
- MIAME (Minimum Information About a Microarray Experiment).
- MINSEQE (Minimum Information about a high-throughput SeQuencing Experiment).
- TEI (Text Encoding Initiative).
- EAD (Encoded Archival Description).
- ETD (Electronic Theses and Dissertations).
- MPEG-7 (Multimedia Content Description Interface).

Estos esquemas de metadatos junto con los recursos se convierten en el elemento central del diseño de software para repositorios. Se ha observado como los recursos son muy variados en cuanto a su tipología, lo que modifica considerablemente su representación y tratamiento (esquema de metadatos, vocabularios controlados, etc.). El 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 de los recursos a través de sus metadatos. Por tanto, la representación de recursos dentro de un repositorio institucional es compleja para el manejo de los recursos de una forma clara y transparente. A continuación se presentan unas características básicas que deben tomar en cuenta para una representación básica de los recursos en repositorios:

- Los recursos deben estar catalogados en más de un esquema de metadatos a fin de evitar pérdida de información cuando se almacenen y se realicen mapeos entre los esquemas de metadatos deseados.
- Debido a la gran diversidad en las estructuras (planas y jerárquicas) y restricciones (simples y complejos) de cada esquema de metadatos (formatos), es necesario plantear una solución de representación clara, flexible, factible, escalable, interoperable, mantenible y que no represente un desafío de configuración para los administradores, así como tampoco sea complicada su utilización por parte de los usuarios.
- Muchos campos presentes en los esquemas de metadatos necesitan ser estandarizados (por ejemplo: aplicación de vocabularios controlados o tesauros) para garantizar una identificación y recuperación correcta de acuerdo con las características de los recursos usados, a fin de evitar la redundancia y garantizar la integridad de la información.
- En estos sistemas, frecuentemente, se encuentran elementos que poseen información descriptiva propia, conocidos como entidades abstractas_que son representadas de forma separada. Por ello, es necesario permitir su reutilización en los diferentes procesos del repositorio y garantizar su representación.
- Las entidades abstractas pueden tener distintas representaciones según el esquema de metadatos en el que se represente el recurso y la semántica deseada en dicha representación.
- Todos los metadatos de los recursos de un repositorio institucional debe almacenarse de forma persistente, por ello, este almacenamiento debe ser configurable a cualquier tipo de paradigma de base de datos e independiente del modelo de representación de los recursos definido.

- Para una representación de los recursos básica, se deben tomar en cuenta los siguientes módulos: almacenamiento, catalogación, indexación, infraestructura de la plataforma de software y preservación.

3. Los Repositorios Institucionales

Los diferentes conceptos y descripciones de Repositorios Institucionales (RI) han tenido su entrada en el mundo científico desde principios de los años 2000 con autores como Clifford Lynch [14] y Van de Sompel [26]. Por ello, se puede decir que los RI son estructuras web interoperables de servicios informáticos, dedicadas a difundir los recursos científicos y académicos (físicos o digitales) de las universidades a partir de la enumeración de un conjunto de datos específicos (metadatos), para que se puedan recopilar, catalogar, acceder, gestionar, difundir y preservar [27]. Las actividades de catalogación, acceso, gestión y difusión de los contenidos son las más consolidadas con el crecimiento de los repositorios, por el contrario, la recopilación de materiales y la preservación todavía se encuentran en sus primeros pasos. Las funciones básicas de los RI son: búsqueda y recuperación de recursos, exploración, diseminación selectiva de información, autoarchivo y servicios a otros sistemas [28].

El concepto descrito anteriormente junto a otros trabajos [11], [12], [13], [14], pone en el contexto del mundo científico el problema de la representación de recursos dentro de los RI, problema que se ha convertido en complejo poco a poco principalmente por la diversidad de plataformas de software y esquemas de metadatos existentes para representar los recursos de diferentes tipologías en los repositorios. Entre las plataformas de software más usadas se tiene a DSpace, EPrints y Digital Commons [8], [9], pero se encuentran plataformas consolidadas de desarrollos propios que abarcan un gran porcentaje de los recursos esparcidos en el mundo como: arXiv, CiteSeerX, Social Science Research Network, entre otros [29]. Se observa que en estas diferentes plataformas se encuentran los recursos representados en diversos esquemas de metadatos, situación que seguirá en expansión. De hecho, muchos esquemas de metadatos se desarrollan de acuerdo con una variedad de usuarios y de disciplinas para facilitar la identificación, recuperación, utilización y/o gestión de recursos en los repositorios [30]. Por tanto, los esquemas de metadatos a través de su conjunto de elementos diseñados en una estructura formal y relacionados con la semántica, la sintaxis y la obligatoriedad en sus valores, pasan a ser un elemento importante en la representación de recursos dentro de un RI. Adicionalmente, se observa en los últimos años, que la tendencia ha sido que los esquemas estén codificados en XML, dado que XML es un estándar del W3C (World Wide Web Consortium) predominante para codificación e intercambio de datos, que están agrupados en elementos delimitados por etiquetas [31].

Los esquemas han sido diseñados e implementados principalmente en los RI a partir de unos principios de los metadatos que permiten reusos de los registros de metadatos e integrar los mismos. Los principios son: modularidad, extensibilidad, refinamiento y multilingüismo [32]. El principio de modularidad se considera como el principio clave de organización para caracterizar los entornos de las diferentes fuentes de contenidos de los recursos, de estilos de gestión de contenidos y de enfoques

de descripción de los recursos. Este entorno modular permite combinar a los diferentes tipos de elementos de metadatos de diferentes esquemas, vocabularios y bloques de construcción para hacerlos interoperables. Es importante destacar que el contexto de un elemento en particular esta limitado gracias a los *namespaces* en los esquemas de metadatos. El principio de extensibilidad indica la posibilidad de poderse ajustar a las necesidades de una aplicación en particular. El principio de refinamiento se observa por la aplicación en diferentes dominios de acuerdo con el nivel de detalle definido por los cualificadores o por el conjunto de valores posibles de los elementos obtenidos por vocabularios controlados u algoritmos particulares. Finalmente el multilingüismo permite la diversidad lingüística y cultural de estos esquemas. Estos principios de metadatos junto con la representación de recursos que se elija para la plataforma de software de los repositorios influyen directamente en los siguientes aspectos de un RI [2]:

- Complejidad del software. El software debe ser perdurable, es decir, debe estar en continua corrección de errores y generación de actualizaciones. Por tanto, entre más simple sea la representación más simple serán los modelos de datos, los procesos de carga e incluso la interfaz de usuario.
- Escalabilidad y performance. Directamente proporcional con el número de recursos, por ello, cuando ellos aumentan considerablemente afectan la escalabilidad y la perfomance. Por ejemplo, en representaciones complejas basadas en bases de datos, la complejidad de las consultas aumenta considerablemente al igual que los tiempos de respuesta.
- Interoperabilidad. Definida como la capacidad que tienen algunos sistemas para intercambiar y utilizar información procedente de otro sistema diferente. La elección de la representación influirá en las capacidades del sistema para derivar otras representaciones para su exposición o para generar recursos internos a partir de representaciones externas. Por tanto, representaciones demasiado simples pueden llevar a un proceso de transformación deficiente, mientras que representaciones muy complejas pueden llevar a un proceso de transformación complicada.

Finalmente, algunos de los problemas de la representación de recursos dentro de los RI que se desean solucionar en la presente propuesta basados en los siguientes aspectos [2]:

1. Formatos de metadatos:
    I. Debido a la gran diversidad de formatos, se pueden clasificar por su estructura y por su especificidad. Según su estructura pueden ser planos donde no existe anidamientos de metadatos y jerárquicos donde sí existe anidamientos de metadatos. En cambio según su especificidad pueden ser simples (pocos elementos y por ende más generales) o complejos (muchos elementos significa que se tiene que ser más específicos).
    II. El problema presentado en los formatos se complejiza en formatos jerárquicos porque si se esta usando base de datos relacionales, las consultas SQL serían muy complejas degradando considerablemente su performance, por ello, una opción viable es el uso de formatos inherentemente anidados como XML con una base de datos de ese tipo. Por el contrario, si se usan formatos simples, su tratamiento estaría solamente limitado a elementos con un nombre (clave) y un valor.

2. Vocabularios controlados:
    I. Muchos campos presentes en los esquemas de metadatos necesitan ser estandarizados y puedan ser identificados, recuperados y usados de forma clara y precisa. Por ejemplo tesauros (listas de palabras o términos empleados para representar conceptos) para definir las temáticas de los recursos (Eurovoc o DeCS), sistemas de clasificación del conocimiento como el CDU (Clasificación Decimal Universal), idiomas, referencias geográficas, tipos de recursos, materias como los LCHS (Library of Congress Subject Heading), frecuencia de entrega (mensual, bimensual, trimestral), entre otros.
    II. Para evitar la redundancia y garantizar la integridad de la información en los vocabularios, se pueden considerar tres puntos de vista: representación, referencia y presentación.
        - Forma de representación. Puede estar en una lista simple de elementos o en elementos relacionados, que se encuentran en una tabla de una base de datos, en un archivo XML con un esquema particular, en un archivo de texto, etc.
        - Forma de referenciar. Se necesita una relación para distinguir de manera unívoca un elemento en un vocabulario determinado, por ello, se debe tomar la decisión entre: un metadato vacío con un valor adicional para la referencia, un metadato con valor del vocabulario replicado junto con un dato adicional para la referencia y un metadato con la referencia como valor.
        - Forma de presentación. Es como el usuario estará observando y utilizando los metadatos en el portal, es decir, en formularios de carga, en la página de presentación de metadatos, en la exportación de recursos, etc.. La forma puede ser simple, intuitiva (suggest, select o search), e internacionalizable.
3. Entidades abstractas:
    I. Entendidas como un conjunto de elementos con información descriptiva propia son utilizadas en los procesos de catalogación. Algunos ejemplos de entidades abstractas son los autores, instituciones, revistas y sus números, eventos y sus instancias. Se pueden considerar los mismos puntos de vista que los vocabularios controlados (representación, referencia y presentación) pero con algunos problemas adicionales.
        - Forma de representación. Depende de esquema de metadatos seleccionados con los mismos problemas que para la representación de recursos. Adicionalmente, se puede pensar en el uso de servicios web como proveedor de entidades.
        - Forma de referenciar. Una vez seleccionada una entidad abstracta es necesario guardar la referencia, surgiendo problemas de compatibilidad entre la representación elegida para la entidad abstracta y los metadatos del recurso a los cuales se asocia esa entidad.
        - Forma de presentación. Además de las consideraciones presentes en los vocabularios controlados, es necesario considerar los problemas generados en el formato de catalogación usado. Entonces se tienen dos alternativas: en el momento de catalogación debe realizarse una transformación única que elimine el problema de duplicidad y de consistencia; y en el momento de presentación que se requiere de una

transformación cada vez que se muestre el recurso y por ende, mayor carga de procesamiento, pero se evita la duplicidad y se asegura consistencia.

4. Representación física de los datos:
   I. Todos los metadatos de los recursos de un repositorio institucional debe almacenarse de forma persistente, por ello, este almacenamiento debe ser configurable a cualquier tipo de paradigma de base de datos e independiente del modelo de representación de los recursos definido, para ello, es necesario analizar alternativas desde el punto de vista de la performance (tiempos de respuestas y consumo de recursos), flexibilidad y escalabilidad (mantener sus cualidades aún aumentando sus recursos, usuarios, etc).
   II. Algunas opciones para la persistencia de los datos pueden ser: base de datos en XML (eXist), relacionales (postgreSQL), orientadas a objetos, RDF, o soluciones mixtas.

Por todo lo expuesto, esta propuesta se enfoca a diseñar un modelo de datos que mejore la representación de los recursos dentro del repositorio bajo una metodología de desarrollo software dirigida por modelos que tome en cuenta las características necesarias ya descritas y se adapte a las diversas tecnologías.

4. Casos de Prácticos: Celsius DL y SEDICI-DSpace

4.1. Celsius DL

Desde sus inicios en el 2003 hasta la migración a DSpace en el 2012, fue adquiriendo una experiencia en el transcurso de la existencia del repositorio, lo que llevó a tomar decisiones vinculadas al desarrollo del mismo en todas sus áreas, tales como: selección del software, formato de metadatos, personal, equipos de soporte, entre otros [2]. Asimismo, SEDICI (en sus dos versiones) ha servido como una herramienta estratégica para la jerarquización tanto a nivel nacional como internacional, por encontrarse posicionado (luego de más de 9 años de vida) en primer lugar en el ranking de repositorios nacionales, en noveno lugar en América Latina y en el centésimo quincuagésimo lugar en el mundo [29]. Para el momento de la migración, Celsius DL contaba con una base documental de más de 15.000 recursos pertenecientes a las distintas áreas del conocimiento de la UNLP. A continuación la representación de recursos en Celsius DL a partir de los cuatro principales aspectos explicados anteriormente [2]:

4.1.1. Formato de metadatos. Durante el desarrollo de Celsius DL 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, donde su estructura no presentaba anidamientos y era simple. Asimismo, se establecieron normas de catalogación para instruir al personal encargado de estas tareas sobre qué metadatos deben utilizar (de forma obligatoria, recomendada u opcional) para catalogar cada tipo de recurso existente (tesis, libros, artículos, etc.).

4.1.2. Vocabularios controlados. La necesidad de normalizar algunos campos en los esquemas de metadatos para los recursos puedan ser identificados, recuperados y usados de forma clara, pone en contexto los vocabularios controlados, es decir, estos vocabularios son usados para representar,

referenciar y presentar los valores de los campos de una forma normalizada para los usuarios y administradores del sistema. 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. Celsius DL había implementado el uso de un listado de términos controlados al que se denominó "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 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, Celsius DL utilizó el tesauro de la UNESCO y un tesauro propio, elaborado con base en el tesauro de la 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 y el DECs; 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. Por ejemplo, en Celsius DL si se quería identificar la temática de un libro sobre "*Edificios e instalaciones oficiales de enseñanza media*" según el lenguaje documental usado, sería de la siguiente manera (tabla 1):

Lenguaje Documental Usado	Traducción al Lenguaje Documental
Sistema de clasificación decimal	371.6
Lista de encabezamientos de materias	ARQUITECTURA ESCOLAR
Tesauro	ESCUELAS ESPACIOS EDUCATIVOS INSTALACIONES NORMAS DE CONSTRUCCIÓN DISEÑO ARQUITECTONICO
Lista de términos libres	ESCUELAS-RANCHO

Tabla 1. Identificación de un libro

4.1.3. 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. Al igual que los vocabularios controlados, las entidades abstractas son usadas para representar, referenciar y presentar valores de metadatos en forma normalizada. Por ejemplo, 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. Celsius DL contaba con 17 tipos de documentos, 4 de los cuales son entidades abstractas: serie documental o publicación periódica, entrega documental o número de publicación periódica, congresos y objeto de conferencia. Además de estas entidades abstractas presentes en las tipologías documentales, en Celsius también se encuentran las entidades abstractas autores e instituciones.

4.1.4. Representación física de los datos. La estructura de metadatos que se utiliza está basada en un conjunto de tablas relacionales (en MySQL), para administrar los metadatos disponibles (agregar, modificar y eliminar metadatos cuando sea necesario) de forma simple. 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: la entidad abstracta autor del registro de autores), lo que implica distintos tipos de consultas según lo que se desee obtener, además afecta la performance y escalabilidad del software en la recuperación de los registros por el gran número de uniones entre tablas que es necesario realizar.

4.2. SEDICI-DSpace

La actual versión de la plataforma de software que soporta al repositorio institucional de la UNLP es el DSpace 1.8. A continuación la representación de recursos en SEDICI-DSpace a partir de los cuatro principales aspectos expuestos en secciones anteriores:

4.2.1. Formato de metadatos. DSpace no soporta esquemas de metadatos jerárquicos, por ello, se tiene que limitarse a esquemas existentes o propios de estructuras no anidadas. Actualmente, se cuenta con un esquema propio y combinado con dos esquemas conocidos como DC y MODS. Adicionalmente, se conoce que DSpace sólo puede tratar con formatos de metadatos planos y una forma de "simular" esta jerarquización es estableciendo una relación entre metadatos de detalle (o hijo) y un metadato principal (o padre). Por ejemplo, los metadatos `<autor>` y `<filiacion>` en

un formato plano no tienen relación entre sí, por lo que no hay certeza sobre qué filiación corresponde a cada autor.

```
<autor>Oviedo, Nestor F.</autor>
<autor>Texier, Jose</autor>
<filiacion>Universidad Nacional de La Plata (UNLP)</filiacion>
<filiacion>Universidad del Táchira (UNET)</filiacion>
```

En el caso de contar con un formato de metadatos jerárquicos, estas ambigüedades no existirían. Por tanto, la estructura anidada sería:

```
<autor>
      <nombre>Oviedo Nestor F.</nombre>
      <filiacion>Universidad Nacional de La Plata (UNLP)</filiacion>
</autor>
<autor>
      <nombre>Texier, Jose</nombre>
      <filiacion>Universidad del Táchira (UNET)</filiacion>
</autor>
```

En ejemplo, si se desea adaptarlo a DSpace, el metadato principal sería <autor>, mientras que el metadato de detalle sería <filiacion>. Considerando que existe un ID único para cada metadato, esto quedaría de la siguiente manera.

```
<autor id="1">Oviedo Nestor F.</autor>
<autor id="2">Texier, Jose</autor>
<filiacion id="3" padre="1">Universidad de La Plata (UNLP)</filiacion>
<filiacion id="4" padre="2">Universidad del Táchira (UNET)</filiacion>
```

En SEDICI aún no se ha definido la utilidad de definir campos hijos en más de 1 nivel. La simplicidad o no de permitir estas relaciones estará en parte de dada por las alternativas de implementación que se planteen.

4.2.2. Vocabularios controlados y entidades abstractas. Al igual que en Celsius DL, los vocabularios y entidades abstractas usadas se siguen manteniendo salvo una particularidad, gracias a la libertad que ofrece DSpace para establecer un vínculo con los autores, en SEDICI-DSpace se creo una tabla de referencia que mantiene una relación con los autores que se usaban en Celsius DL. Se entiende que es un proceso no conveniente pero dada la libertad para este caso de DSpace y el control de autoridades de DSpace, el personal de SEDICI prefiere ofrecer este servicio de administración interna que es transparente para el usuario de la plataforma.

4.2.3. Representación física de los datos. En DSpace la persistencia de los datos es asegurada bajo un modelo de base de datos relacional conocido como PostgreSQL 8.1. En el caso de garantizar una especie de simulación de un esquema de metadatos jerárquicos, se deben realizar pequeños ajustes en tablas particulares de la base de datos de DSpace. Por ejemplo, la tabla metadatavalue es donde se almacenan los metadatos, y tiene los siguientes campos:

- metadata_value_id: id del metadato.
- item_id: id del recurso.
- metadata_field_id: id del tipo de metadato (autor, title, etc).
- text_value: valor textual del metadato.
- text_lang: idioma del valor textual.
- authority: id del elemento correspondiente a un vocabulario controlado.

Para llevar a cabo esta pseudo-jerarquía, sería necesario agregar un campo a esta tabla (por ejemplo parent_metadata_value_id), en el cual se indique el metadato de nivel superior en la jerarquía.

5. Trabajos Futuros

- Analizar y comparar los modelos de los repositorios de los autores referentes que permita establecer los pro y los contra que tienen cada uno de ellos, para generar un modelo general que permita una representación de recursos de manera indistinta a la plataforma que se use, por ejemplo, EPrints y/o Greenstone, que son otras plataformas en software libre para repositorios.
- El concepto de los repositorios ha evolucionado y se ha relacionado con el concepto la biblioteca digital, donde se destaca el gran auge que los RI han tenido con un incremento sostenido en los últimos años de acuerdo con los diferentes directorios de repositorios existentes, además del afianzamiento de la filosofía del acceso abierto en la comunidad de investigadores y académicos. Todas estas realidades incentivan el estudio del dominio de los RI e incentiva el desarrollo de aplicaciones y componentes de software en este dominio.

6. Referencias

[1] SEDICI, "SeDiCI - Repositorio de la Universidad Nacional de La Plata," 2013. [Online]. Available: http://sedici.unlp.edu.ar/. [Accessed: 14-Mar-2012].

[2] M. De Giusti, N. Oviedo, A. Lira, A. Sobrado, J. Martínez, and A. Pinto, "SeDiCI – Desafíos y experiencias en la vida de un repositorio digital," *RENATA*, vol. 1, no. 2, pp. 16–33, Aug. 2011.

[3] J. Tramullas Saz and P. Garrido Picazo, "Software libre para repositorios institucionales: propuestas para un modelo de evaluación de prestaciones," *El Prof. Inf.*, vol. 15, no. 3, pp. 171–181, 2006.

[4] D. P. Madalli, S. Barve, and S. Amin, "Digital Preservation in Open-Source Digital Library Software," *J. Acad. Librariansh.*, vol. 38, no. 3, pp. 161–164, May 2012.

[5] S. K. Singh, M. Witt, and S. Dorothea, "A Comparative Analysis of Institutional Repository Software," presented at the Fifth International Conference on Open Repositories, Madrid, 2010.

[6] G. Pyrounakis and M. Nikolaidou, "Comparing Open Source Digital Library Software," *Handb. Res. Digit. Libr. Des. Dev. Impact*, pp. 51–60, Feb. 2009.

[7] E. Tzoc, "A Mobile Interface for DSpace," *-Lib Mag.*, vol. 19, no. 3/4, Mar. 2013.

[8] OpenDOAR, "OpenDOAR - Home Page - Directory of Open Access Repositories," 2013. [Online]. Available: http://www.opendoar.org/. [Accessed: 21-Mar-2013].

[9] ROAR, "Registry of Open Access Repositories (ROAR)," 2013. [Online]. Available: http://roar.eprints.org/. [Accessed: 21-Mar-2013].

[10] M. Kökörvcený and A. Bodnárová, "Comparison of digital libraries systems," in *Proceedings of the 9th WSEAS international conference on Data networks, communications, computers*, Stevens Point, Wisconsin, USA, 2010, pp. 97–100.

[11] A. Malizia, P. Bottoni, and S. Levialdi, "Generating Collaborative Systems for Digital Libraries: a Model-Driven Approach," *Inf. Technol. Libr.*, vol. 29, Dec. 2010.

[12] F. Paganelli and M. C. Pettenati, "A Model-driven Method for the Design and Deployment of Web-based Document Management Systems," *J. Digit. Inf.*, vol. 6, no. 3, Jan. 2006.

[13] M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp, "Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries," *Acm Trans Inf Syst*, vol. 22, no. 2, pp. 270–312, Apr. 2004.

[14] Leonardo Candela, Castelli, Y. Ioannidis, S. Ross, C. Thanos, P. Pagano, G. Koutrika, H.-J. Schek, and H. Schuldt, "Setting the Foundations of Digital Libraries," *-Lib*, vol. 13, no. 3/4, Mar. 2007.

[15] R. Heery, "Review of metadata formats," *Program Electron. Libr. Inf. Syst.*, vol. 30, no. 4, pp. 345–373, Dec. 1996.

[16] E. Méndez, "Tratamiento de los objetos de información en los archivos : retos y estándares para la descripción basada en metadatos," 2003. [Online]. Available: http://eprints.rclis.org/handle/10760/12691#.UAAZFuEzfgM. [Accessed: 13-Jul-2012].

[17] DCMI, “Dublin Core Metadata Element Set, Version 1.1,” 2012. [Online]. Available: http://www.dublincore.org/documents/dces/. [Accessed: 13-Jul-2012].
[18] MARC, “MARC STANDARDS (Network Development and MARC Standards Office, Library of Congress).” [Online]. Available: http://www.loc.gov/marc/. [Accessed: 11-Jun-2013].
[19] MODS, “Metadata Object Description Schema: MODS (Library of Congress),” 2012. [Online]. Available: http://www.loc.gov/standards/mods/. [Accessed: 13-Jul-2012].
[20] T. Ferreras Fernández, “Preservación digital en repositorios institucionales: GREDOS,” 2010. [Online]. Available: http://gredos.usal.es/jspui/handle/10366/83130. [Accessed: 13-Jul-2012].
[21] MADS, “Metadata Authority Description Schema (MADS) - (Library of Congress).” [Online]. Available: http://www.loc.gov/standards/mads/. [Accessed: 11-Jun-2013].
[22] METS, “Metadata Encoding and Transmission Standard (METS) Official Web Site,” 2012. [Online]. Available: http://www.loc.gov/standards/mets/. [Accessed: 13-Jul-2012].
[23] PREMIS, “PREMIS: Preservation Metadata Maintenance Activity (Library of Congress),” 2012. [Online]. Available: http://www.loc.gov/standards/premis/. [Accessed: 13-Jul-2012].
[24] PREMIS, *PREMIS Data Dictionary for Preservation Metadata -- version 2.0*. 2008.
[25] C. A. Lynch, “Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age,” *ARL:*, Feb-2003. [Online]. Available: http://www.arl.org/resources/pubs/br/br226/br226ir.shtml. [Accessed: 28-Jan-2013].
[26] H. Van de Sompel, S. Payette, J. Erickson, C. Lagoze, and S. Warner, “Rethinking Scholarly Communication,” -*Lib Mag.*, vol. 10, no. 9, Sep. 2004.
[27] J. Texier, M. R. De Giusti, N. Oviedo, G. L. Villarreal, and A. J. Lira, “Los beneficios del desarrollo dirigido por modelos en los repositorios institucionales,” presented at the BIREDIAL - Conferencia Internacional Acceso Abierto, Comunicación Científica y Preservación Digital, 2012.
[28] J. Texier, M. R. De Giusti, N. Oviedo, G. L. Villarreal, and A. J. Lira, “El uso de repositorios y su importancia para la educación en Ingeniería,” presented at the World Engineering Education Forum (WEEF 2012) “Educación en Ingeniería para el Desarrollo Sostenible y la inclusión social,” 2012.
[29] Webometrics, “Ranking Web of Repositories,” 2013. [Online]. Available: http://repositories.webometrics.info/. [Accessed: 08-Mar-2013].
[30] L. M. Chan and M. L. Zeng, “Metadata Interoperability and Standardization - A Study of Methodology, Part I,” -*Lib Mag.*, vol. 12, no. 6, p. 3–, 2006.
[31] E. Méndez, “La descripción de documentos electrónicos a través de metadatos : una visión para la Archivística desde la nueva e- Administración,” *Revista d’Arxius*, 2003. [Online]. Available: http://eprints.rclis.org/12684/. [Accessed: 05-Jun-2013].
[32] E. Duval, W. Hodgins, S. Sutton, and S. L. Weibel, “Metadata Principles and Practicalities,” -*Lib Mag.*, vol. 8, no. 4, Apr. 2002.