PrEBi

SeDiCI SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

Bibliotecas y Repositorios Digitales

Tecnología y Aplicaciones

http://sedici.unlp.edu.ar

Participantes del dictado

- Marisa De Giusti
- Nestor Oviedo
- Silvia Peloche
- Matías Cánepa

Bibliotecas y repositorios digitales

Capítulo 4: Aspectos tecnológicos e informáticos. Software de gestión del repositorio. Requerimientos a nivel local. Preservación digital. Servicios: búsqueda, exploración, autoarchivo, DSI, citas, etc.

Contenido

- Software del repositorio
  - Características deseables
  - Alternativas libres
- Representación de recursos
  - Formatos planos vs. jerárquicos
  - Vocabularios controlados simples
  - Entidades abstractas
  - Representación física de los datos

Contenido

- Identificadores persistentes
  - Importancia
  - Algunas opciones disponibles
- Servicios de un repositorio digital
  - Búsqueda y recuperación
  - Exploración
  - Diseminación selectiva de la información
  - Autoarchivo

Contenido

- Estadísticas del repositorio
  - Objetivos
  - Estadísticas frecuentes
- Preservación de contenido
  - Digital obsolescence
  - Estrategias de solución

Contenido

- Repositorio semántico
  - Introducción
  - Problemas relativos a la representación
  - Recuperación de la información y navegación de las relaciones
  - Posibilidad de nuevas estadísticas

Software del repositorio

Software del repositorio

- Es uno de los pilares en la construcción de un repositorio digital.
- Tiene la capacidad de potenciar o limitar todos los aspectos del repositorio (servicios, tamaño, descripción de los recursos, etc.).
- Debe perdurar en el tiempo.

Software del repositorio

Aspectos a evaluar de un software de repositorio

Licencia: es un contrato entre el propietario de los derechos del software y los usuarios que lo utilizan. Este contrato especifica las condiciones bajo las cuales el primero cede derechos o permite actividades sobre el software a los segundos. Licencias conocidas son GPL, Creative Commons, BSD, LGPL, MIT, Apache, etc.

Nivel de impacto: nivel de uso del software por parte de la comunidad de repositorios digitales. Un nivel elevado proporciona confianza y promueve la constante actualización de la aplicación (reporte de errores y mejoras continuas).

Software del repositorio

Aspectos a evaluar de un software de repositorio

Nivel de personalización: medida de las posibilidades de adaptación, tanto de interfaz de usuario como de funcionalidad, para reflejar la identidad y las necesidades de la institución a la que representa. Esto incluye extensiones del software, logos y colores, estructura y organización de contenidos, etc.

Nivel de documentación: cantidad y calidad de la información de todos los aspectos relacionados al software. Desde la instalación y configuración hasta el uso del sistema por parte de usuarios finales y administradores.

Software del repositorio

Aspectos a evaluar de un software de repositorio

Frecuencia de actualizaciones: corrección de errores (de funcionamiento y seguridad) de forma continua, mejora en las funciones existentes e inclusión de nueva funcionalidad que amplíe las características del sistema.

Centros de soporte: listas de correo, wiki, foros, canal de chat y cualquier otro punto de contacto entre un usuario del sistema y los desarrolladores y/o la comunidad de usuarios del software, desde donde puede obtenerse asistencia ante dudas y problemas concretos.

Software del repositorio

Aspectos a evaluar de un software de repositorio

Facilidad de uso: medida referente a la curva de aprendizaje respecto del uso del sistema y todas sus funciones, tanto por usuarios como por administradores.

Formato de metadatos soportado: conjunto de elementos usado para almacenar los datos de cada recurso. Se destaca como un punto importante porque:

- propicia o limita parte de la funcionalidad
- influye en la precisión y completitud de la información
- es un factor de rechazo

Software del repositorio

Aspectos a evaluar de un software de repositorio

Performance: tiempos de respuesta del sistema ante cada solicitud, recursos físicos consumidos (disco, memoria, procesador, etc). La performance habla del balance entre velocidad de respuesta, consumo de recursos, costos, etc.

Escalabilidad: capacidad del software de mantener sus cualidades (performance, simplicidad, mantenibilidad, etc) en niveles aceptables aún cuando el volúmen de recursos, cantidad de usuarios, etc. aumenten considerablemente con el tiempo.

Software del repositorio

Aspectos a evaluar de un software de repositorio

Interoperabilidad: capacidad del sistema de comunicarse e interactuar con otros sistemas. En general los roles de un repositorio pueden ser:

- recolector de recursos/consumidor de servicios
- expositor de recursos/proveedor de servicios

Administración: sección del software de acceso restringido a usuarios con privilegios. Permite acceder a sectores privados del sistema para realizar principalmente acciones de control y mantenimiento.

Software del repositorio

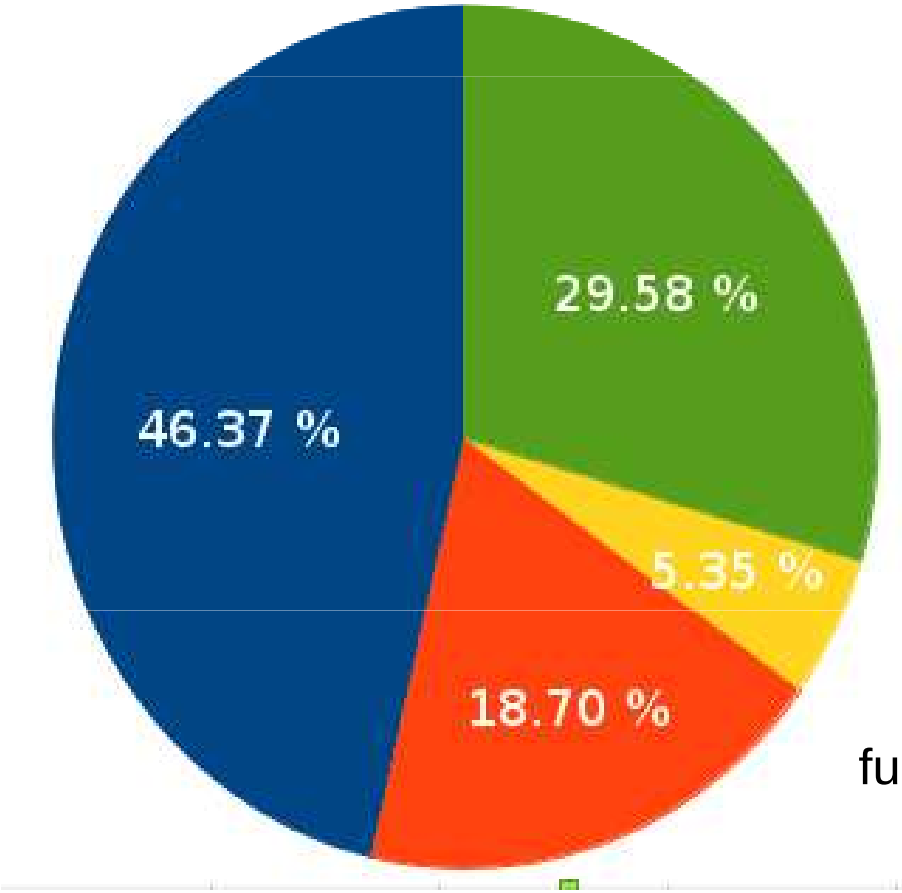
Aspectos a evaluar de un software de repositorio

¿Qué buscamos en cada aspecto a analizar?

- Licencia
- Nivel de impacto
- Nivel de personalización
- Nivel de documentación
- Frecuencia de actualizaciones
- Centros de soporte
- Facilidad de uso
- Formato de metadatos
- Performance
- Escalabilidad
- Interoperabilidad
- Administración

Software del repositorio

Software de repositorios mas usados

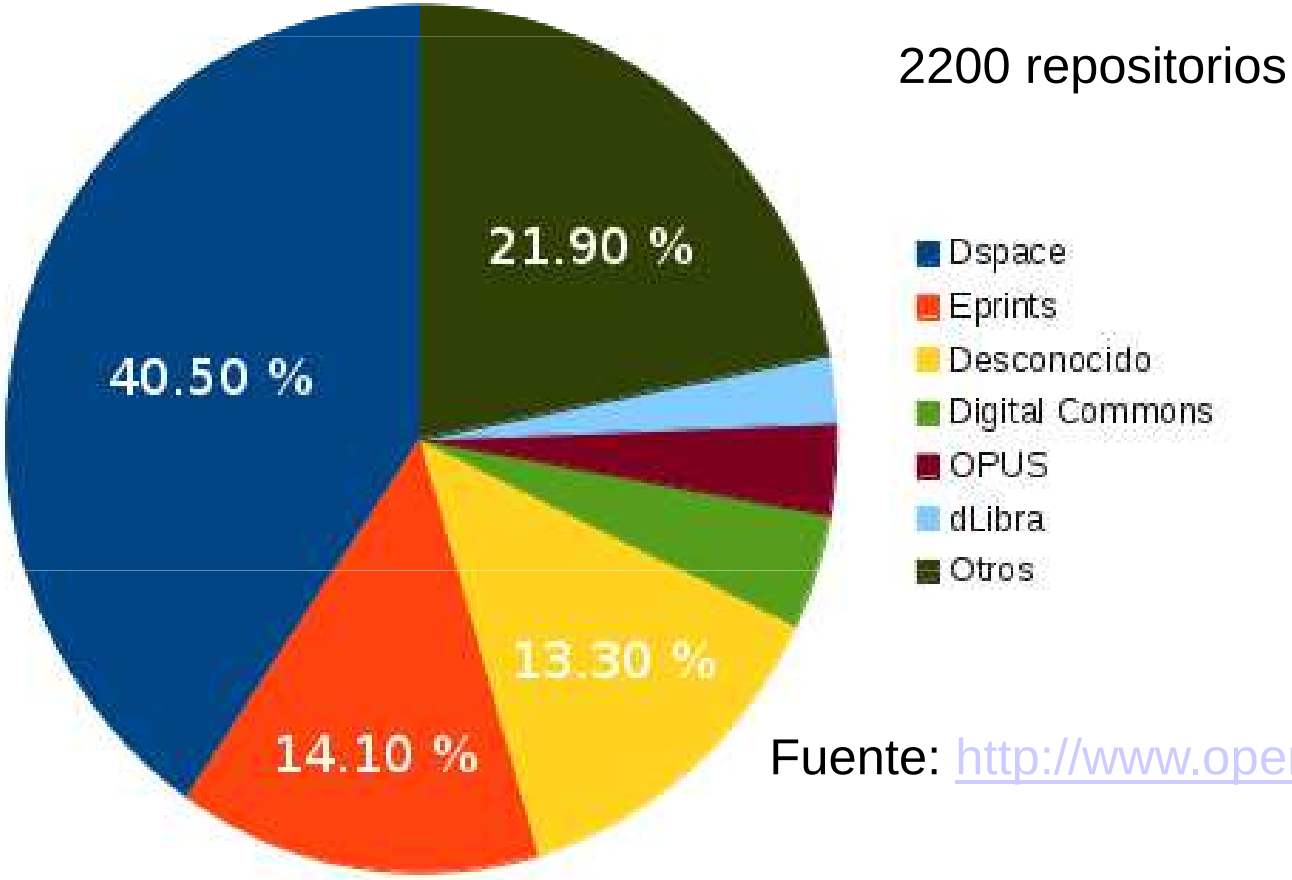


2950 repositorios

fuente: http://roar.eprints.org

Software del repositorio

Software de repositorios mas usados



2200 repositorios

Fuente: http://www.opendoar.org

Software del repositorio

Breve comparativa entre DSpace y EPrints

(más utilizados a nivel mundial)

	•DSpace	•EPrints
•Sitio web	•http://www.dspace.org	•http://www.eprints.org
•Creadores	•MIT (USA)	•University of Southampton (UK)
•Lenguaje	•Java	•Perl
•Plataforma	•Multiplataforma	•UNIX-like (portado a Windows)
•Base de Datos	•PostgreSQL	•MySQL
•Licencia	•BSD	•GPL v2
•Formato de Metadatos	•Qualified DC, formatos planos	•cualquier formato (incluso jerárquicos)
•Soporte para búsquedas	•Apache Solr (DSpace-Discovery)	•MySQL indexes

Representación de recursos

Representación de recursos

¿Qué se entiende por recurso?

Es todo objeto, físico o digital, que puede ser descripto a partir de la enumeración de un conjunto de datos específicos de dicho elemento, que lo distinguen entre otros objetos.

¿Qué significa representar un recurso?

Habla de registrar de forma persistente el conjunto de datos asociado a un recurso, usando este conjunto de datos como síntesis y reemplazo del objeto "real", permitiendo distribuir el recurso sin necesitar el objeto real (es decir, se usa su representación).

Representación de recursos

La representación que se elija para los recursos del repositorio influye directamente en aspectos como:

- **complejidad del software**: una representación simple implica que los modelos de datos, los procesos de carga e incluso la interfaz de usuario, son más simples.
- **escalabilidad y performance**: cuando el número de recursos aumenta considerablemente, la representación de los recursos comienza a tomar un rol importante. Por ejemplo, en representaciones complejas basadas en bases de datos, la complejidad de las consultas aumenta considerablemente, y por lo tanto también aumentan los tiempos de respuesta.

Representación de recursos

- **Interoperabilidad**: para interoperar es necesario exponer los recursos propios en formatos entendibles por otros sistemas. La elección de la representación influirá en las capacidades del sistema para **derivar** otras representaciones (para su exposición) o bien generar recursos internos a partir de representaciones externas. Esto es, representaciones demasiado simples pueden llevar a transformaciones deficientes, mientras que representaciones muy complejas pueden llevar a procesos de transformación complicados.

Representación de recursos

Formatos de metadatos para la representación de recursos

Según estructura:

- Planos: no existe anidamiento de metadatos
- Jerárquicos: existe anidamiento de metadatos

Según especificidad:

- Simples: pocos elementos, más generales
- Complejos: muchos elementos, más específicos

Representación de recursos

Formatos de metadatos planos

```
<documento>
  <titulo>...</titulo>
  <autor>Gomez P.</autor>
  <filiacion>UNLP</filiacion>
  ...
</documento>
```

Parece adecuado, pero ¿qué sucede, por ejemplo, si se tiene más de un autor con disitintas filiaciones?

Representación de recursos

Formatos de metadatos planos

```
<documento>
  <titulo>...</titulo>
  <autor>Gomez P.</autor>
  <filiacion>UNLP</filiacion>
  <autor>Lopez R.</autor>
  <filiacion>UTN</filiacion>
  ...
</documento>
```

¿Cómo determinar de forma segura qué filiación corresponde a qué autor?

¿Qué pasa si el orden cambia en algún proceso de manipulación de metadatos?

Representación de recursos

Formatos de metadatos jerárquicos

```
<documento>
  <titulo>...</titulo>
  <autor>
    <nombre>Gomez P.</nombre>
    <filiacion>UNLP</filiacion>
  </autor>
  <autor>
    <nombre>Lopez R.</nombre>
    <filiacion>UTN</filiacion>
  </autor>
</documento>
```

Soluciona el problema planteado anteriormente, pero **complejiza el software** del repositorio, ya que la interpretación de estos datos para su validación, procesamiento y presentación ya no son tan simples.

Representación de recursos

La representación de un formato de metadatos plano es relativamente simple. Es decir, básicamente se trata de un listado de elementos con un nombre y un valor (sin considerar por el momento restricciones de tipos de datos, formatos, etc).

Su tratamiento y su representación son relativamente simples

Representación de recursos

Tratar con un formato de metadatos jerárquico dificulta considerablemente su representación. En bases de datos relacionales por ejemplo, debido a la naturaleza anidada de estos formatos, se tiende a crear consultas SQL demasiado complejas, con múltiples JOINS entre las mismas tablas, degradando la performance de forma considerable.

La opción mas viable para este tipo de formatos suele ser alguna forma de representación inherentemente anidada, como ser XML. Esto significaría la necesidad de contar con una Base de Datos XML (posiblemente solo para los documentos).

Representación de recursos

Formatos de metadatos simples frente a complejos

El caso ***simple*** se destaca por poseer poca cantidad de metadatos, cuya definición es amplia y, en general, poco restrictiva en cuanto a formatos.

En el caso ***complejo*** existe una mayor cantidad de metadatos, con contenidos mas explícitos y por lo tanto una definición mas restrictiva para cada uno.

Representación de recursos

Ejemplo: al catalogar una tésis con un formato simple como Dublin Core sin calificar, es probable que el director y co-director, junto con la institución de desarrollo, sean catalogados utilizando un mismo elemento: *dc:contributor,* ya que no existe una distinción para estos datos en la definición del formato.

Desde el punto de vista informático esto dificulta:
- presentación: no se puede distinguir de qué dato se trata
- validación: solo puede esperarse texto libre

Representación de recursos

Vocabularios controlados simples

Para determinados metadatos, se indica que su contenido se extrae de un vocabulario controlado, especificando además el vocabulario al que se hará referencia.

- Tesauros
- Sistemas de clasificación
- Idiomas
- Referencias geográficas
- Tipos de recursos
- Materias
- Frecuencias de entrega (mensual, bimestral, trimestral, etc)

Representación de recursos

Vocabularios controlados simples

Se necesita una forma de **Representación**

- Depende del tipo de vocabulario (lista simple de elementos o elementos relacionados).
- Puede ser una tabla en la base de datos, un archivo XML con un *schema* particular, un archivo de texto, etc.
- Debe permitir generar respuestas rápidas.
- Complejidad aportada por las **relaciones** entre elementos.

Representación de recursos

Vocabularios controlados simples

Se necesita **Referenciar** elementos

- Depende de la representación elegida para los recursos (XML, Bases de Datos, etc).
- Debe permitir distinguir de forma unívoca un elemento específico en un vocabulario determinado.
- Decisión entre:
  - Metadato vacío, con un dato adicional para la referencia
  - Metadato con valor del vocabulario replicado y un dato adicional para la referencia
  - Metadato con la referencia como valor

Representación de recursos

Vocabularios controlados simples

Se necesita una forma de **Presentación**

- Debe ser simple e intuitiva (suggest, select, search)
- Debe proporcionar respuestas rápidas
- De ser posible, debe ser **internacionalizable**
- Se debe utilizar desde un formulario de carga, desde una página de presentación de metadatos, desde la exportación de recursos, etc.

Representación de recursos

Entidades abstractas

¿A qué llamamos Entidades Abstractas?

Conjunto de elementos que poseen información descriptiva propia, utilizados en los procesos de catalogación de recursos como elementos de un vocabulario controlado.

Mismas consideraciones que para vocabularios controlados simples, adicionando algunos problemas.

Representación de recursos

Entidades abstractas

Ejemplos:

- Autores: apellido, nombres, email, institución de origen, etc.
- Instituciones: nombre, institución de la que depende, localidad, dirección, mail, responsables, etc.
- Revistas y sus números: nombre, ISSN, director, editor, staff, volúmen, tapa, etc.
- Eventos y sus instancias: nombre, año, ubicación, organizador, etc.

Representación de recursos

Entidades abstractas

Desafíos: **Representación**

- Se debe definir un formato de metadatos (considerar los mismos problemas que para la representación de recursos)
- Opción de usar de WebServices como proveedor de entidades (hay que considerar qué información se incluye en la respuesta del servicio)

Representación de recursos

Entidades abstractas

Desafíos: **Referencia**

Una vez seleccionada una entidad abstracta, es necesario guardar la referencia.

Pueden suceder **problemas de compatibilidad** entre la representación elegida para la entidad abstracta y el o los metadatos del recurso a los cuales esa entidad se asocia.

Representación de recursos

Entidades abstractas

Ejemplo de problemas de compatibilidad

Entidad Autor:

- apellido
- nombre

Metadato autor:
(del formato de catalogación)

```
<author>
    <lastName/>
    <firstName/>
</author>
```

¿Cómo se indica que el campo *apellido* debe ir en el metadato /author/lastName y el campo *nombre* en /author/firstName?

Representación de recursos

Entidades abstractas

Desafíos: **Presentación**

Además de los elementos a tener en cuenta para los vocabularios simples, es necesario considerar los problemas de compatibilidad entre el formato de la entidad abstracta y el formato de catalogación utilizado.

Representación de recursos

Entidades abstractas

Alternativas de referencia que influyen en la presentación, según en qué momento se realiza la transformación de la entidad abstracta al metadato correspondiente

En ambos casos se asume que la referencia se guarda en un campo independiente

1. en el momento de catalogación

2. en el momento de presentación

Representación de recursos

Entidades abstractas

1. En el momento de la catalogación

- Una única transformación
- Problema de duplicidad de información
- Tiende a generar problemas de consistencia

Representación de recursos

Entidades abstractas

2. En el momento de la presentación

- Se requiere transformación cada vez que se muestra el recurso
- Mayor carga de procesamiento cada vez que se muestra el recurso
- Se evita la duplicidad de la información
- Se asegura la consistencia

Representación de recursos

Representación física de los datos

Es necesario analizar alternativas para el almacenamiento

- Performance
- Flexibilidad
- Escalabilidad

Algunas opciones:

- Base de datos XML (eXist)
- Base de datos relacional
- Base de datos orientada a objetos
- Base de datos RDF

Se pueden adoptar soluciones mixtas

Identificadores persistentes

Identificadores persistentes

¿Qué es un Identificador persistente?

Es un método de resolución de direcciones (URL) que busca garantizar el acceso a los objetos en internet, aún cuando éstos cambien su ubicación (URL de acceso).

Handle: *hdl.handle.net/123456789/1234*
DOI: *dx.doi.org/10.4100/jhse.2010.52.15*
PURL: *purl.org/net/example/purlName*

Identificadores persistentes

Importancia

Las URL cambian con el tiempo

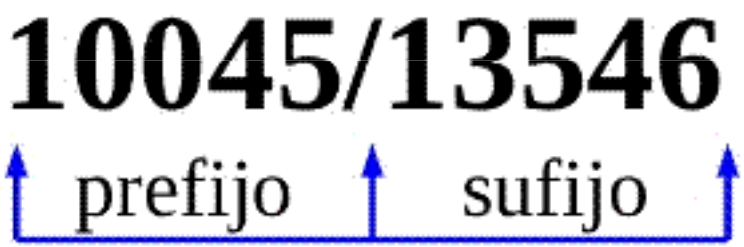
- Dominio: cambios poco frecuente
- Ruta: en general cambios frecuente

El servicio se basa en redireccionar la solicitud de una URL persistente a una URL (no persistente) real, la que efectivamente apunta hacia el recurso.

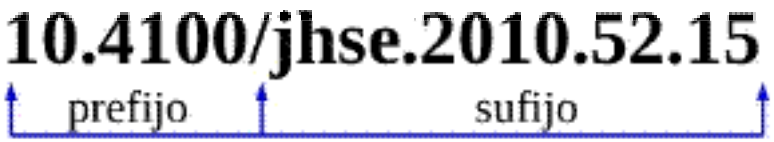
Cuando la URL real del recurso cambia, se informa de este cambio solo al manejador de identificadores persistentes contratado y este modifica las reglas de redirección.

Identificadores persistentes

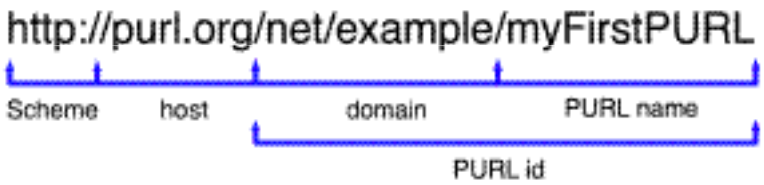
Algunas alternativas disponibles



Handle



DOI

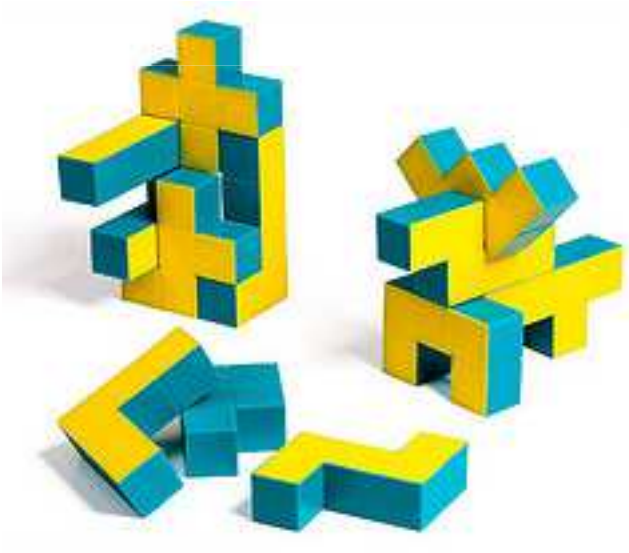


PURL

Servicios de un repositorio digital

Servicios de un Repositorio digital

- Búsqueda y Recuperación
- Exploración
- Diseminación Selectiva de Información
- Autoarchivo
- Servicios a otros sistemas



Servicios de un Repositorio digital

Búsqueda y Recuperación

- Un repositorio digital puede alojar cientos, miles o millones de recursos
- Es necesario proveer a los usuarios de mecanismos para buscar y recuperar estos recursos
- Los usuarios pueden estar buscando un recurso específico y único, o pueden requerir recursos con alguna característica en común (por ejemplo, artículos que traten sobre determinada área del conocimiento)
- A veces, los usuarios no saben bien que están buscando; suelen refinar los criterios de búsqueda una y otra vez hasta que localizan los recursos

Servicios de un Repositorio digital

Búsqueda y Recuperación

Un repositorio tiene que proveer un servicio de **búsqueda simple**, que permita ingresar algunos términos de búsqueda y retorne un conjunto de recursos como resultado

También debe proveer una **búsqueda avanzada**, que permita parametrizar los criterios de búsqueda y acotar así el conjunto resultante: por fecha de publicación de los recursos, por tipo de recurso, por idioma, por autor...

En cualquier caso, las búsquedas deben cumplir ciertos criterios mínimos:

Servicios de un Repositorio digital

Búsqueda y Recuperación

- Simpleza: el formulario de búsqueda debe ser simple, y mostrar campos de búsqueda avanzada si el usuario lo requiere. De todos modos, la búsqueda avanzada también debe permanecer simple

- Eficiencia: las búsquedas deben resolverse casi inmediatamente, en cuestión de milisegundos, o muy pocos segundos a lo sumo

- Relevancia: Todos los resultados de una búsqueda tendrán un valor de relevancia. Cuanto más relevante, más arriba deberá mostrarse entre los resultados

Servicios de un Repositorio digital

Búsqueda y Recuperación

- Filtrado: la búsqueda avanzada permite definir ciertos criterios a aplicarse durante la búsqueda
- En ocasiones, es deseable aplicar **filtros** una vez realizada la búsqueda
- Para ello, es necesario definir criterios de agrupamiento de resultados, y permitir al usuario agregar o eliminar criterios
- Una técnica muy utilizada es el *faceting (faceted search, faceted navigation o faceted browsing),* que permite a los usuarios explorar filtrando la información disponible en los resultados de la búsqueda

Servicios de un Repositorio digital

Búsqueda y Recuperación . Faceting

SeDiCI SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

PrEBi

Servicios de un Repositorio digital

Exploración

- Mediante la exploración, los usuarios pueden acceder a los recursos a partir de *un orden* preestablecido
- Este *orden* puede variar de repositorio en repositorio: colecciones, temas, fechas, etc.
- La exploración permite obtener un pantallazo general del repositorio

Servicios de un Repositorio digital

Exploración. Ejemplos

SeDiCI SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

PrEBi

Servicios de un Repositorio digital

Exploración. Ejemplos

Navegar por autor

0-9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Mostrando ítems 1-60 de 20939 Página siguiente ▶

Nombre de los autores	Nombre de los autores	Nombre de los autores
AA, Jiye	Abatedaga, Nidia	Abdala, Juan
Aamir, Muhammad F.	Abate, Stella Maris	Abdala, Lidia R.
Aamir, Muhammad N.	Abba, Agustín Manuel	Abdala, Virginia
Abad-Grau, María M.	Abba, Martin Carlos	Abdalla, Dulcinéia S. P.
Abadía, Anselmo	Abbas, Alaa M.	Abdelahad, Corina
Abadi, Florencia	Abbas, Ash Mohammad	Abdel-Hamid, Magdi
Abajo, Rosaura	Abbas, Ghulam	Abdel Masih, S.
Aballay, Alicia	Abbas, Khizar	Abdel Masih, Samira
Aballay, Laura	Abbas, Mateen	Abdo, Diego
Aballay, Laura N.	Abbas, Tanveer	Abdo Ferez, María Cecilia
Aballay, P.	Abbate, Florencia	Abdulameer, Shaymaa A.
Aballay, P.	Abbate, Horacio Antonio	Abdul Rasool, Bazigha K.
Abalo, Facundo	Abbate, Sandro Giuseppe	Abedini, Walter

SeDiCI SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

PrEBi

Servicios de un Repositorio digital

Diseminación Selectiva de Información

- DSI es una técnica de envío de información de interés a los usuarios
- En un servicio DSI, los usuarios **solicitan** que se les envíe información
- Esta solicitud debe estar acompañada de algunos criterios de selección de información: temas, idiomas, tipos de recursos, períodos...
- En algunos casos, los usuarios pueden *suscribirse a búsquedas;* el software del repositorio ejecutará la misma búsqueda periódicamente, y enviará al usuario aquellos recursos que aparecen como nuevos entre los resultados

Servicios de un Repositorio digital

Diseminación Selectiva de Información

Google Scholar: Alertas por correo

Servicios de un Repositorio digital

Autoarchivo

- Es importante que todos los miembros de la organización se involucren con el repositorio. Una forma de hacerlo es que ellos mismos aporten su propia producción
- El servicio de *autoarchivo* permite a los miembros de la organización cargar sus propios recursos al repositorio
- De este modo, los autores se aseguran la publicación y difusión de sus trabajos en forma rápida y sencilla
- Este servicio implica la carga de un archivo, y una pre-catalogación del recurso por parte de quién realiza el autoarchivo
- La interfaz de catalogación debe ser muy simple, y se presenta un subconjunto de metadatos al usuario

Servicios de un Repositorio digital

Autoarchivo

- Existen restricciones en cuanto al tipo de archivo a enviar, y también en cuanto al tamaño de los mismos
- Los recursos enviados mediante autoarchivo quedan en un estado *pendiente de revisión*: debe hacerse un control de calidad sobre los recursos subidos, especialmente sobre aquellos subidos por personas no especializadas en catalogación
- Los autores deben seleccionar una licencia CC para su obra
- Los autores deben aceptar una licencia de difusión para SeDiCI

Servicios de un Repositorio digital

Autoarchivo

Servicios de un Repositorio digital

Autoarchivo

http://e-archivo.uc3m.es/

http://riuma.uma.es/

SeDiCI SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

PrEBi

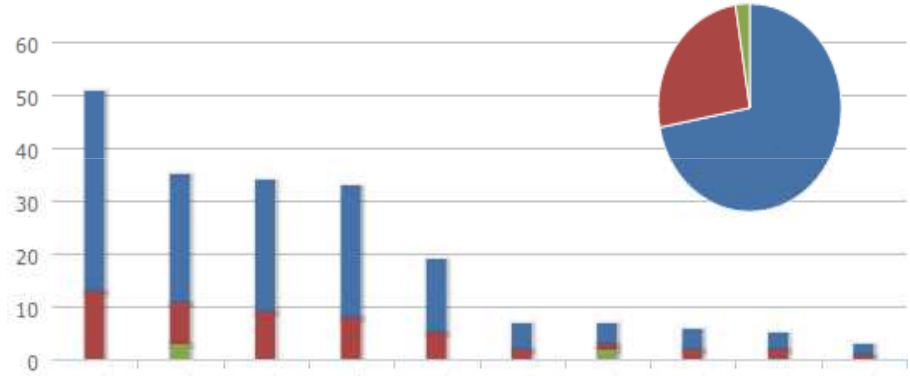
Servicios de un Repositorio digital

Servicios a otros sistemas

- Un Repositorio Institucional no está aislado en el mundo: debe ser capaz de interactuar con otros sistemas y otros repositorios, de compartir recursos y de recuperar recursos remotos
- Esto aumentará la visibilidad del repositorio en la web y maximizará la difusión de los recursos
- El repositorio podrá también aumentar la cantidad de recursos disponibles para sus usuarios
- Algunos servicios comunes: OAI PMH, SRU/SRW, RSS

Estadísticas del repositorio

Necesidad e importancia



Clasificación de estadísticas

- a partir de la información que nos brindan
- a partir de quién las genera

Ejemplos

Estadísticas del repositorio

Necesidad de las estadísticas

- Las estadísticas son una herramienta clave a la hora de medir nuestro repositorio
  - Tamaño y Tasa de Crecimiento
  - Nivel de Impacto

- Obtener tablas y gráficos estadísticos avanzados, y no aprovechar esta información es casi lo mismo a no tener estadísticas

- El repositorio debe *retroalimentarse* con estos datos y utilizarlos bajo una política de *expansión y mejora continua*

Estadísticas del repositorio

Necesidad de las estadísticas

- Los datos obtenidos sirven como control de calidad, para saber dónde estamos parados como repositorio
- La interpretación de estos datos permitirá la toma de decisiones en varios niveles:
  - político/estratégico: cooperar con otros grupos, interactuar más con determinados actores...
  - táctico: cambiamos la forma de agrupar cierto tipo de recurso, incorporamos un nuevo tipo de recurso, implementamos una nueva metodología de carga
  - tecnológico: necesitaremos más hardware y mejor conectividad, debemos ampliar nuestro software para integrar cierta tecnología, será mejor revisar los índices de la base de datos...

Estadísticas del repositorio

Necesidad de las estadísticas. Tamaño y Tasa de Crecimiento

- Estadísticas de Tamaño y Tasa de Crecimiento
  - Necesitamos conocer cuántos recursos aloja nuestro repositorio
  - Es importante saber cómo han crecido estos recursos en el tiempo
    - de este modo, podemos detectar mesetas en las curvas de crecimiento y apuntalar donde sea necesario
    - podemos también predecir tendencias, como períodos de mayor o menor actividad, y prepararnos con antelación

Estadísticas del repositorio

Necesidad de las estadísticas. Tamaño y Tasa de Crecimiento

- El concepto de "tamaño" es muy amplio
  - cantidad de recursos locales
  - cantidad de recursos en full-text
  - cantidad de usuarios registrados
- Tasa de crecimiento también puede interpretarse de diferentes maneras
  - recursos incorporados año tras año
  - usuarios registrados cada semana
  - alertas por correo creadas mes a mes

Estadísticas del repositorio

Necesidad de las estadísticas. Tamaño y Tasa de Crecimiento

- Además de las cantidades mencionadas, tenemos otras "cantidades" de interés
  - Cantidad de Recursos locales
    - Tesis de grado, de posgrado
    - Artículos de revista, en congresos
    - Libros, e-books
  - Recursos a partir del origen
    - por dependencia, por departamento, área...
  - Por área temática
    - informática, ingeniería, literatura y letras, ciencias jurídicas...

Estadísticas del repositorio

Necesidad de las estadísticas. Tamaño y Tasa de Crecimiento

- Las clasificaciones nos permiten detectar *desequilibrios*
- Algunos desequilibrios son normales y esperables
  - *"en el último año, se sumaron más de 2000 tesis de grado y solamente 50 libros"* ⇨ natural, considerando la cantidad de alumnos que se recibe anualmente
- Otros desequilibrios puede ser indeseables y podrían corregirse si se detectan a tiempo
  - "El 70% de los recursos proviene del 35% de las dependencias" ⇨ quizás debamos promocionar el uso del repositorio en el 65% restante, o quizás debamos adaptar el repositorio para que les sea de mayor utilidad
- Nuevamente, las estadísticas serán de utilidad si brindan información **precisa**, y si dicha información es **utilizada** apropiadamente

Estadísticas del repositorio

Necesidad de las estadísticas. Nivel de Impacto

- Nivel de Impacto: debemos medir el alcance global y local del repositorio
  - quiénes lo utilizan y para qué
  - desde dónde acceden los usuarios (países, regiones, instituciones)
  - cómo se posiciona en rankings y en buscadores
  - qué se busca y qué no se busca
  - con qué dispositivos y plataformas se accede (computadoras, tablets, sistemas operativos, navegadores)
  - a partir de cuáles servicios llegamos a nuestros usuarios (web, feeds, SRU/SWR, DSI, e-mail...)

Estadísticas del repositorio

Necesidad de las estadísticas. Nivel de Impacto

- Aquí también podremos tomar decisiones en niveles muy diversos:
  - Incorporar nuevos idiomas, a partir del origen de los usuarios
  - Optimizar las páginas web para maximizar su visibilidad en los buscadores
  - Reorganizar los contenidos para darles mayor relevancia a aquellos menos utilizados
  - Promocionar servicios con bajo nivel de uso
  - Desarrollar servicios, herramientas y estrategias para aumentar el acceso desde ciertos dispositivos
  - Mejorar las herramientas de búsqueda

Estadísticas del repositorio

Clasificación de estadísticas

Podemos clasificar las estadísicas a partir de dos grandes criterios:

- a partir del tipo de información que nos brindan
  - información sobre recursos, usuarios, servicios del repositorio, búsquedas realizadas, descargas ...
  - información del entorno o contexto: visitas, visibilidad en la web, navegadores utilizados, hardware de acceso...
- a partir del encargado de recolectarlas y generarlas
  - el software que sustenta al repositorio
  - otras herramientas integradas al repositorio
  - servicios de terceros

Estadísticas del repositorio

Clasificación de estadísticas

- A partir de Tipo de información
  - Información interna:
    - es específica para el repositorio
    - dependiente del software en uso
    - qué datos se almacenan
    - con cuánta granularidad
    - qué estadísticas se generan a partir de estos datos
    - podemos incorporar nuevas estadísticas y obtener datos mucho más precisos
    - recursos almacenados, usuarios registrados, accesos, servicios del repositorio, búsquedas realizadas, descargas

Estadísticas del repositorio

Clasificación de estadísticas

- Entorno o contexto:
  - obtenemos información acerca del entorno del repositorio
  - está muy relacionado con el nivel de impacto
  - **este entorno no es controlado por nosotros**
  - por lo general, no debemos preocuparnos por registrar estos datos

  - Incluye cantidad de visitas al portal, visibilidad del portal en la web, tipos de navegadores utilizados, dispositivos desde los que acceden los usuarios

Estadísticas del repositorio

Clasificación de estadísticas

- Recolectadas y generadas por el mismo software
  - La recolección de datos debe estar en todos los rincones del software
  - Podremos controlar por completo las estadísticas, generar versiones más simples y más avanzadas, derivar nuevas estadísticas, etc...
  - Software más complejo
    - mayor dificultad de desarrollo y mantenimiento
      - importancia del diseño en capas
    - podría degradar la performance
    - diseñar un módulo de generación estadísticas no es una tarea simple

Estadísticas del repositorio

Clasificación de estadísticas

- Recolectadas por herramientas integradas al repositorio
    - El software que sustenta nuestro repositorio requiere otros programas para funcionar. Como mínimo, tendremos:
        - un sistema operativo, ej. Linux, Windows
        - un servidor web, ej. Apache, IIS, Tomcat, Jetty
        - una base de datos, ej. MySQL, Oracle
        - un servidor de correos, ej. Postfix,Exim

Estadísticas del repositorio

Clasificación de estadísticas

- Todos estos programas generan registros de acceso, de errores, de potenciales problemas (slow-log)... No nos preocupamos por guardar la información
- El desafío es cómo explotarla: debemos interpretarla, procesarla y mostrarla de manera útil (análisis de logs, minería de datos...)
- Afortunadamente, hay programas que realizan esto por nosotros
- Desafortunadamente, si bien podemos controlar parcialmente qué datos se registran, no tendremos la misma flexibilidad comparado con las estadísticas recolectadas por el software del repositorio

Estadísticas del repositorio

Clasificación de estadísticas

- Servicios de terceros
  - Como tercer alternativa, podemos tercerizar la recolección de estadísticas
  - Existen varios servicios externos capaces de recolectar y generar estadísticas
  - Puede requerir mínimos cambios en nuestro software, aunque a veces los sistemas están preparados para integrarse con algunos servicios populares
  - Aquí tendremos estadísticas de acceso, visibilidad, crecimiento del repositorio...
  - Algunos servicios son gratuitos, otros poseen una parte gratuita y otra paga, otros son solamente pagos

Estadísticas del repositorio

Ejemplos

Estadísticas de SeDiCI

Aplicaciones instalables
- Awstats

Servicios on line
- Google Analytics
- StatCounter
- Yahoo! Site Explorer
- Rankings y registros globales
  - Webometrics
  - Roar http://roar.eprints.org/1193/

Preservación de contenido

Preservación de contenido

Hay una muy importante necesidad de preservar el contenido digital en el tiempo, con el objetivo de conservarlo accesible frente a riesgos como

- Incendios, Inundaciones, etc
- Robos
- Problemas de hardware (rotura de discos, etc.)
- Cambios tecnológicos constantes

Es un proceso continuo

Preservación de contenido

Digital obsolescence

Es el resultado de la evolución de las tecnologías: a medida que surgen nuevas tecnologías, las viejas van quedando en desuso y se vuelven obsoletas.

Mantener tecnologías obsoletas en funcionamiento puede ser justificado en casos particulares, pero no en la mayoría.

Cornell University Library creó la "Cámara de los horrores"
http://www.dpworkshop.org/dpm-eng/oldmedia/chamber.html

Preservación de contenido

Digital obsolescence

Mantener tecnologías obsoletas requiere conservar

- Hardware
- Software (aplicaciones, librerías, sistema operativo, etc)
- Documentación (manuales, instructivos, etc)
- Personal con la capacitación y las habilidades necesarias para trabajar en ese entorno obsoleto

Suelen ser opciones muy difíciles de mantener y muy costosas.

En general no suele ser la mejor opción

Preservación de contenido

Estrategias

Las formas de atacar los problemas de preservación, y en particular los problemas de obsolescencia, son:

- Migración continua
- Adhesión a estándares internacionales
- Emulación
- Encapsulamiento
- Metadatos de preservación
- Políticas de backup

Preservación de contenido

Migración continua

Migrar la información de una tecnología a la siguiente de forma continua, evitando así la obsolescencia.

- Es una de las opciones de mayor uso
- Asegura el acceso en todo momento (los datos son siempre accesibles mediante una tecnología actual)

- Requiere transformación de los datos originales
- Decisiones sobre qué se desea preservar

Preservación de contenido

Adhesión a estándares internacionales

Es una estrategia que busca apoyarse en la afirmación de que los estándares internacionales son relativamente estables en el tiempo.

- En la actualidad, los estándares evolucionan casi tan rápido como las tecnologías
- Es una estrategia que debería usarse en combinación con otras
- Según la National Initiative for Networked Cultural Heritage, los formatos que no serán declarados obsoletos (al menos en un futuro cercano) son: TIFF y PDF sin compresión, y ASCII y RTF sin compresión, para imágenes y texto respectivamente.

Preservación de contenido

Emulación

Se trata de imitar las características y capacidades de un software y/o hardware, de modo que los procesos "crean" que están funcionando en la plataforma original.

- No hay necesidad de modificar los datos originales (como en la migración), manteniendo la integridad de la información.
- Una vez que se archivaron los datos, solo hay que asegurarse que el soporte físico utilizado siga siendo accesible
- Se puede usar un mismo emulador para múltiples objetos del mismo tipo.

Preservación de contenido

Encapsulamiento

Se basa en agrupar cada objeto a preservar junto con todos los elementos (incluso software) necesarios para asegurar su acceso en el tiempo.

Como elementos a encapsular podemos tener:

- Especificaciones del formato de archivo
- Instructivos relacionados a la emulación necesaria
- Información de configuración de alguna herramienta en particular
- Software de emulación
- Especificaciones de hardware

Preservación de contenido

Metadatos de preservación

Generalmente considerados como metadatos administrativos

Buscan registrar información relativa a la evolución de los recursos en el tiempo según las acciones de preservación aplicadas, incluyendo información sobre formatos, usos, actividades de preservación realizadas, responsables de dichas actividades en el tiempo, etc.

Varias iniciativas:
- PREMIS: PREservation Metadata: Implementation Strategies
- OAIS: Open Archival Information System
- NEDLIB: Networked European Deposit Library

Preservación de contenido

Políticas de backup

Los riesgos de pérdida de datos por eventos desafortunados siempre son posibles:

- Incendios
- Inundaciones
- Robos
- Fallas de hardware

Para disminuir esos riesgos es necesario contar con un sistema de backups (datos, configuración, documentación, etc)

- Incremental
- Espejo

Repositorio semántico

Repositorio semántico

Web semántica

Si bien en general se afirma que la web es una base de datos gigante, colaborativa, distribuida, en continuo crecimiento, etc, también existe consenso respecto de que esa base de datos tiene algunos problemas:

- Información mayormente semi-estructurada o completamente desestructurada
- Mucha información desactualizada
- Información redundante
- Información íntimamente relacionada, aunque sin ningún vínculo

Repositorio semántico

Web semántica

¿Por qué todos estos problemas?

Carga descentralizada
Flexibilidad (texto, imágenes, videos, etc)
Libertad de expresión

Repositorio semántico

Web semántica

El objetivo de la web semántica es aportar "**significado**" a toda la información disponible, de forma que sea "interpretable" por máquinas a través de agentes inteligentes, para así proveer información coherente, completa, competente, etc., de forma automática o semi-automática.

De esta forma se logra

- Aumentar la interoperabilidad entre sistemas
- Generar nuevos tipos de servicios de búsqueda y recuperación

Repositorio semántico

Los repositorios digitales cuentan con:

- Información estructurada (metadatos)
- Carga controlada (reglas de catalogación)
- Vocabularios controlados (tesauros, entidades abstractas)
- Base de datos centralizada (en general)

Repositorio semántico

Un repositorio semántico se caracteriza por la existencia de relaciones entre sus componentes (documentos, entidades, etc).

Algunas relaciones posibles son:

- Composición
- Traducciones
- Misma temática
- Autores relacionados
- Instituciones relacionadas

Repositorio semántico

Problemas relativos a la representación

Es necesario encontrar una forma flexible y eficiente para representar estas relaciones.

- Por inferencia, a través de relaciones establecidas en el modelo de datos
- De forma explícita, por ejemplo con **Ontologías**

Repositorio semántico

Problemas relativos a la representación

Por inferencia en base al modelo

Título: Función endotelial en el embarazo
Autor: Ros, Natalia
Descriptores: Cardiología; Embarazo

Título: Donantes de tejidos valvulares cardíacos: modelo de selección
Autor: Olano, Ricardo Daniel
Descriptores: Cardiología; Cultivo de tejidos

Título: Diferencias en la forma de presentación y diagnóstico de la enfermedad coronaria en la mujer
Autor: Corneli, Mariana
Descriptores: Cardiología; Enfermedades cardiovasculares

Autores relacionados con *Cardiología*

- *Ros, Natalia*
- *Olano, Ricardo Daniel*
- *Corneli, Mariana*

Repositorio semántico

Problemas relativos a la representación

Ontologías

Representa conceptos/objetos y las relaciones entre ellos

Las componentes mas importantes son:

- **Clases**: tipos de objetos
- **Instancias**: objetos concretos que pertenecen a una clase particular
- **Atributos**: características de una Clase (y por lo tanto de todas las instancias de esa clase)
- **Relaciones**: formas en las que los objetos se conectan entre sí

Repositorio semántico

Problemas relativos a la representación

Ontologías del dominio

Este tipo particular de ontologías se limita a representar elementos de un dominio particular, aportando un *contexto* a los conceptos/objetos que representa.

Ej.: el concepto *Ratón*

- En el dominio *Informática* hace referencia a un periférico de entrada
- En el dominio *Animales* hace referencia a un roedor

Repositorio semántico

Problemas relativos a la representación

Las ontologías se construyen usando un **lenguaje**.

Los más destacados son:

- RDFSchema: Resource Description Framework Schema

Provee un conjunto de clases base, utilizando RDF como lenguaje de base.

- OWL: Ontology Web Language

Es una familia de lenguajes (OWL Lite, OWL DL y OWL Full), con varias sintaxis alternativas (una es una extensión al vocabulario de RDFS)

Repositorio semántico

Problemas relativos a la representación

Ejemplo de RDFS

```
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xml:base="http://www.animals.fake/animals#">

        <rdfs:Class rdf:ID="animal" />

        <rdfs:Class rdf:ID="horse">
                <rdfs:subClassOf rdf:resource="#animal"/>
        </rdfs:Class>
</rdf:RDF>
```

Repositorio semántico

Problemas relativos a la representación

Ejemplo de OWL (usando RDFS como sintaxis)

```
<rdf:RDF xmlns:owl ="http://www.w3.org/2002/07/owl#" ...
  xmlns:rdfs ="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd  ="http://www.w3.org/2001/XMLSchema#">

<owl:Class rdf:ID="Animal">
      <rdfs:label>Animal</rdfs:label>
      <owl:Class rdf:ID="Male">
            <rdfs:label>Male</rdfs:label>
            <rdfs:subClassOf rdf:resource="#Animal"/>
      </owl:Class>
      <owl:Class rdf:ID="Female">
            <rdfs:label>Female</rdfs:label>
            <rdfs:subClassOf rdf:resource="Animal"/>
            <owl:disjointWith rdf:resource="Male"/>
      </owl:Class>
</owl:Class>
</rdf:RDF>
```

Repositorio semántico

Recuperación de la información y navegación de relaciones

Se necesita una forma de almacenamiento y recuperación eficiente

Existen:

- Bases de datos RDF (openRDF, Mulgara)
- Lenguajes de consulta RDF (SPARQL)

Además es necesario adaptar la interfaz de usuario para proveer elementos de navegación pertinentes, contextuales, eficientes, simples, etc.

Repositorio semántico

Posibilidad de nuevas estadísticas

El agregado de relaciones al repositorio permite generar un nuevo conjunto de estadísticas mas complejas, pero más interesantes:

- Tendencias en cuanto a temáticas
- Relación entre autores e instituciones
- Instituciones y autores mas productivas