PrEBi

UNLP

SEDICI

REPOSITORIO INSTITUCIONAL DE LA UNLP

Bibliotecas y Repositorios Digitales

Tecnología y Aplicaciones

sedici.unlp.edu.ar

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

- Interoperabilidad
  - Niveles de interoperabilidad
  - Formas de interoperar
  - Formatos de metadatos
  - OAI-PMH
  - Recolección de recursos
  - Directrices de interoperabilidad

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

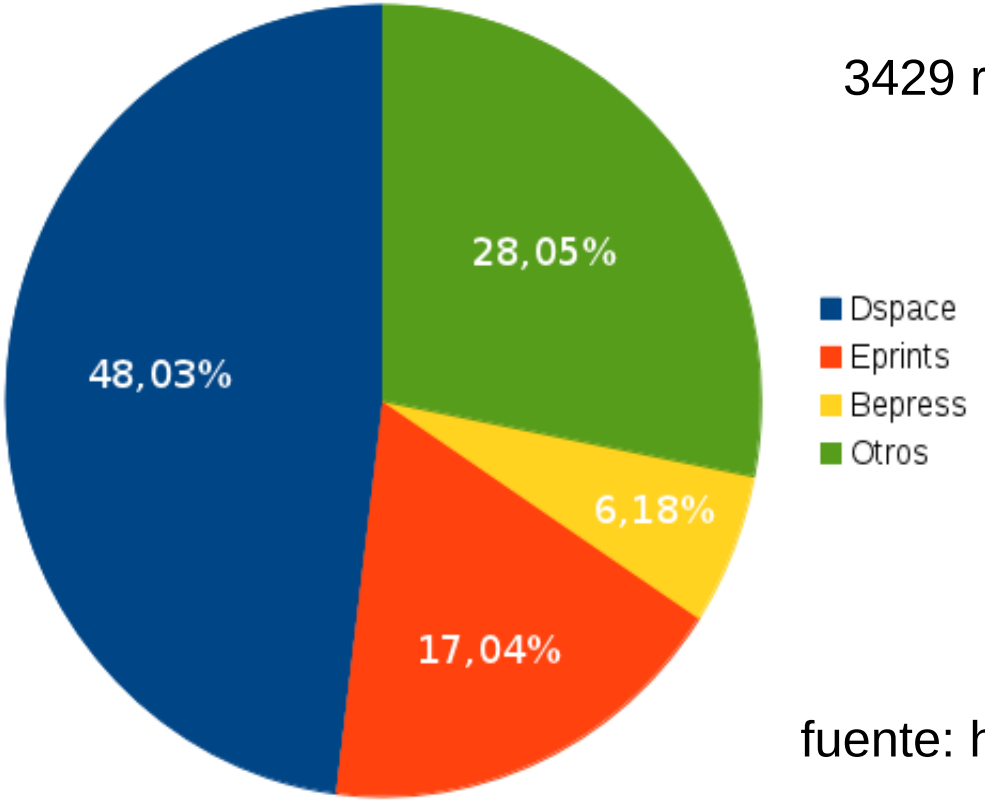
Aspectos a evaluar de un software de repositorio

¿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

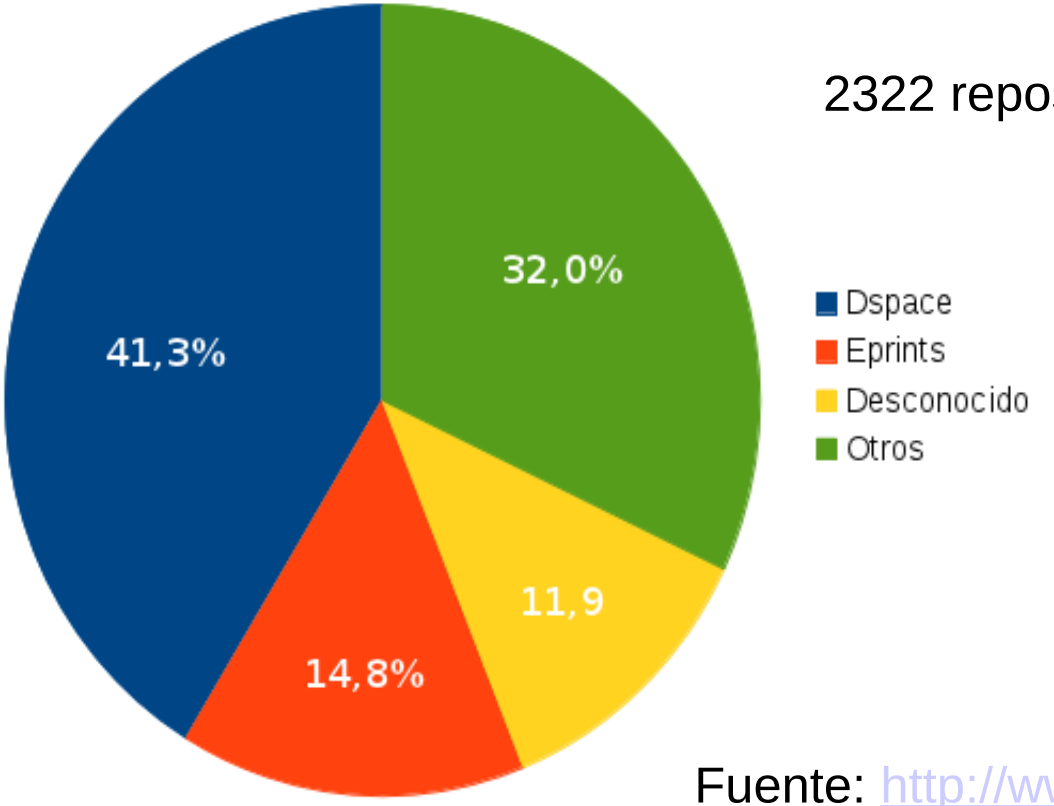


3429 repositorios

fuente: http://roar.eprints.org

Software del repositorio

Software de repositorios mas usados



2322 repositorios

Fuente: http://www.opendoar.org

Software del repositorio

Breve comparativa entre DSpace y EPrints

(más utilizados a nivel mundial)

	DSpace	EPrints
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	Qulified DC o cualquier formato plano	Cualquier formato (incluso jerárquicos)
Soporte para búsquedas	Apache Solr (Dspace-discovery)	MySQL indexes

PREBI

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).

PREBI

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.

PREBI

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.

PREBI

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.

PREBI

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).

PREBI

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

PREBI

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 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

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

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

PREBI

Identificadores persistentes

Importancia

Las URL cambian con el tiempo

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

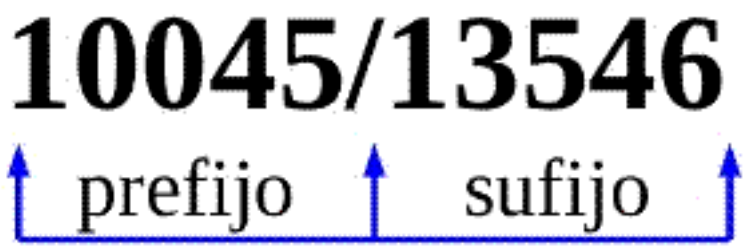
El servicio se basa en redireccionar la solicitud de una URL persistente a una URL (no persistente) real, la que efectivamente apunta hacia el recurso.

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.

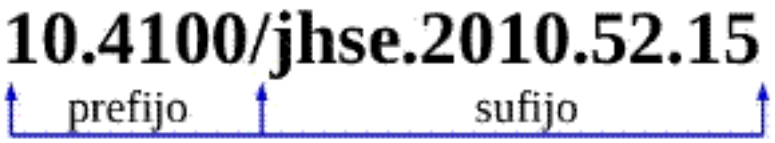
PREBI

Identificadores persistentes

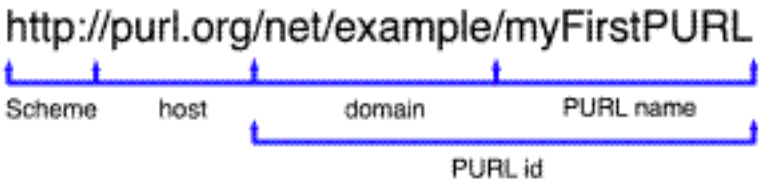
Algunas alternativas disponibles



Handle



DOI

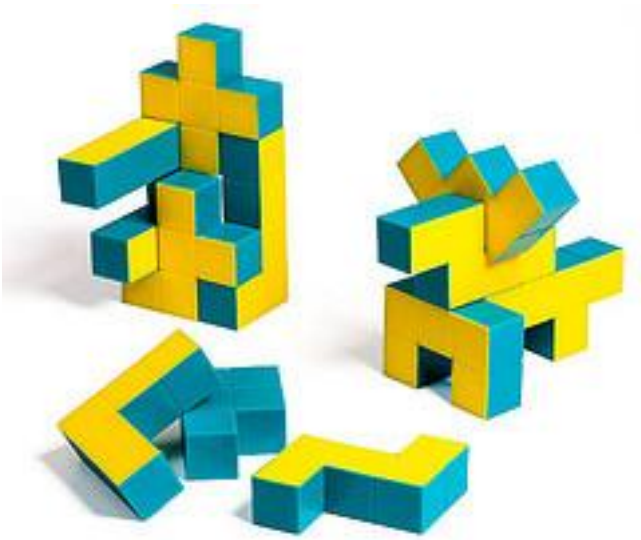


PURL

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:

PREBI

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

Refine su busqueda

Tipo de documento

- Articulo (2521)
- Audio (4)
- Capitulo de libro (9)
- Comunicacion (131)
- Contribucion a revista (197)
- Documento de trabajo (40)
- Informe tecnico (40)
- Libro (90)
- Musica (1)
- Objeto de conferencia (5942)
- Ordenanza (21)
- Preprint (4)
- Revision (334)
- Tesis de doctorado (739)
- Tesis de grado (214)
- Tesis de maestria (270)
- Trabajo de especializacion (76)

Fecha de publicación

- 2000 - 2013 (9452)
- 1900 - 1999 (1071)

Materia

- Ciencias Informaticas (5537)
- Educacion (1442)

Resultados de su búsqueda...

desarrollo de software

Mostrando ítems 1-10 de 10633

Objeto de conferencia XIV Congreso Argentino de Ciencias de la Computación 2008

Herramientas para el desarrollo de software embebido para aplicaciones aeronáuticas

Arias, Silvia; Montes, Alfredo Miguel; Banchio, Enrique; Gonzalez Kriegel, Bernardo; Muñoz, Gabriel J.

Libro Facultad de Informática 2010

Desarrollo de software dirigido por modelos . *Conceptos teóricos y su aplicación práctica*

Pons, Claudia; Giandini, Roxana Silvia; Pérez, Gabriela

Objeto de conferencia V Workshop de Investigadores en Ciencias de la Computación 2003

Proceso ágil para desarrollo de software

Servetto, Arturo Carlos; García Martínez, Ramón; Perichinsky, Gregorio

Objeto de conferencia XIII Workshop de Investigadores en Ciencias de la Computación 2011

Desarrollo de software sensible al contexto

Quincoces, Viviana Elizabet; Gálvez Díaz, María del Pilar; Cáceres, N. R.; Vega, Ariel; Brouchy, C. V.; Velázquez, E. C.; González, O. M.; Guzmán, A. N.

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.

PREBI

Servicios de un Repositorio Digital

Exploración. Ejemplos.

Navegar por autor

Todos 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

O introducir las primeras letras:

Mostrando ítems 1-60 de 26990

AA, Jiye
Aamir, Muhammad F.
Aamir, Muhammad N.
Abad-Grau, María M.
Abadía, Anselmo
Abadía, David
Abadie, Diego Gustavo Edwin
Abadi, Florencia
Abajo, Rosaura
Aballay, Alicia
Aballay, Laura
Aballay, Laura N.

Abarca Cedeño, Mireya Sarahí
Abásolo Guerrero, María José
Abásolo, María José
Abatedaga, Nidia
Abate, Stella Maris
Abba, Agustín Manuel
Abba, Martin Carlos
Abbas, Alaa M.
Abbas, Ash Mohammad
Abbas, Ghulam
Abbas, Khizar
Abbas, Mateen

Abdala, Cristian
Abdala, Cristian Simón
Abdala de Aredez, Virginia
Abdala, Fernando
Abdala, G.
Abdala, Juan
Abdala, Lidia R.
Abdala, Virginia
Abdalla, Dulcinéia S. P.
Abdelahad, Corina
Abdel-Hamid, Magdi
Abdel Masih, S.
Abdel Masih, Samira

PREBI

Servicios de un Repositorio Digital

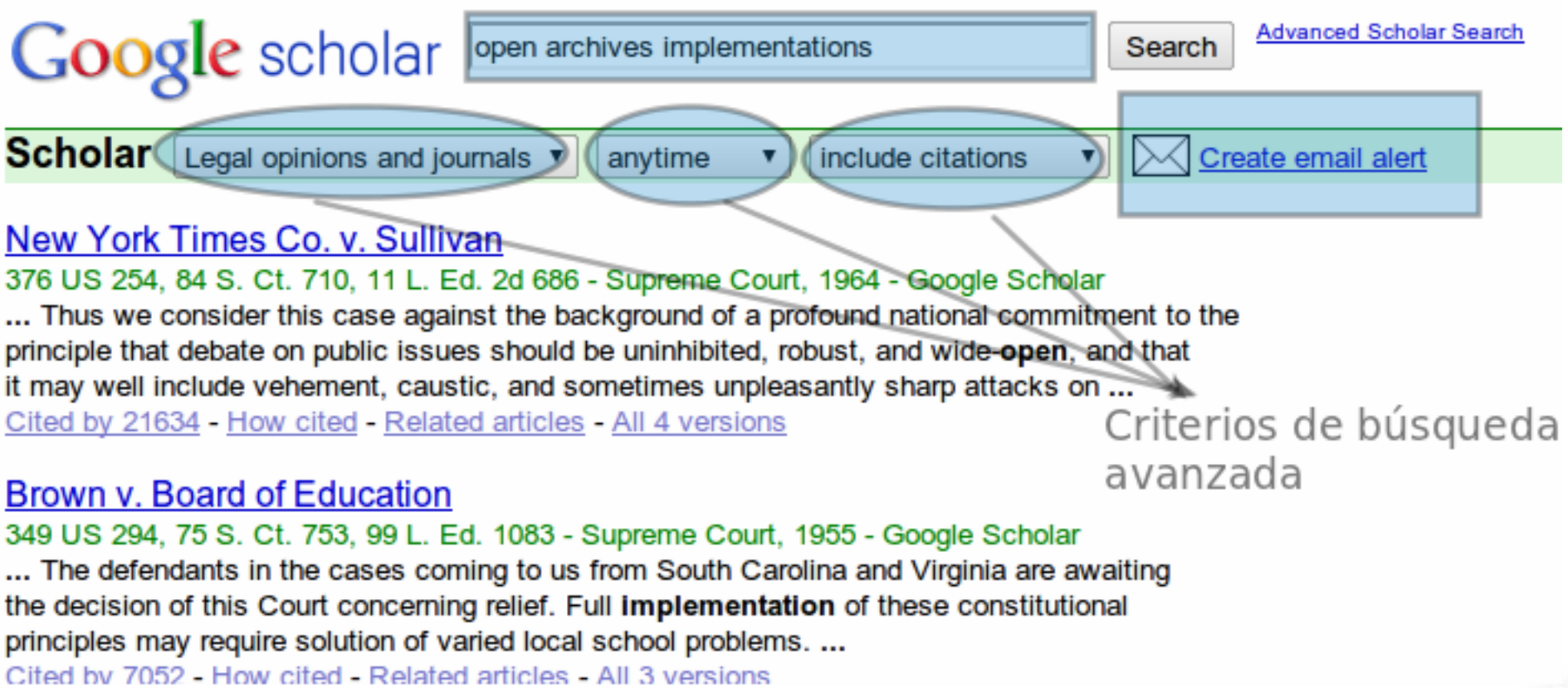
Diseminación selectiva de información

- 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
- 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/

PREBI

Servicios de un Repositorio Digital

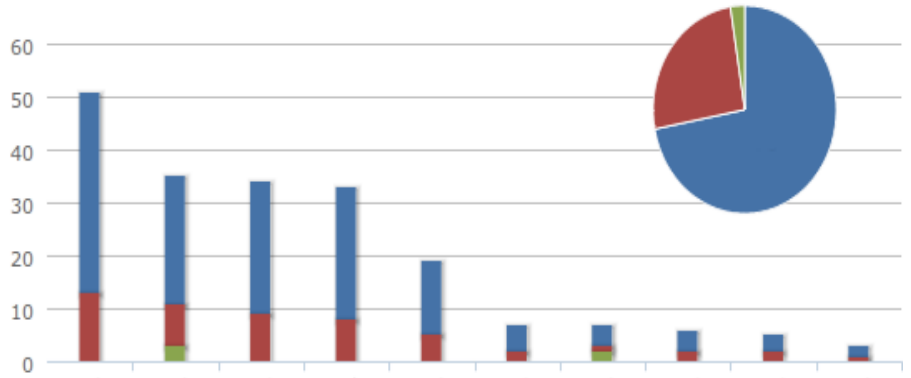
Servicios a otros sistemas

- 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

Estadísticas del repositorio

Necesidad e importancia



Clasificación de estadísticas

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

Estadísticas del repositorio

Necesidad de 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 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 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 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 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 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

PREBI

Estadísticas del repositorio

Necesidad de estadísticas. Nivel de impacto

- Nivel de Impacto: debemos medir el alcance global y local del repositorio
    - 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 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 estadísticas a partir de dos grandes criterios:

- a partir del tipo de información que nos brindan
  - información interna
  - información del entorno
- a partir del encargado de recolectarlas y generarlas
  - el software que sustenta al repositorio
  - herramientas integradas al repositorio
  - servicios de terceros

Estadísticas del repositorio

Clasificación de estadísticas

- 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

- Información del entorno:
  - está muy relacionado con el nivel de impacto
  - **este entorno no es controlado por nosotros**
  - 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
    - 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
  - Podemos tercerizar la recolección de algunos datos
  - 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-DSpace

http://sedici.unlp.edu.ar/handle/10915/15920/statistics

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

PREBI

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

PREBI

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 contínua

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 internacional

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.

PREBI

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

PREBI

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

Interoperabilidad

Interoperabilidad

¿Qué es la interoperabilidad?

Capacidad de los sistemas informáticos de interactuar a través del intercambio de información y servicios, para lograr un objetivo.

Interoperabilidad

¿Por qué es importante interoperar?

El intercambio de servicios y recursos ayuda a cumplir parte de los objetivos de un repositorio digital:

- Mayor visibilidad e impacto de los recursos propios
- Mayor cantidad de recursos ofrecidos a los usuarios
- Mayor cantidad y diversidad de servicios para ofrecer

Introducción

El contexto del Open Access

Los movimientos de Acceso Abierto y la tendencia mundial hacia estas políticas plantea un marco altamente propicio para la interoperabilidad entre repositorios digitales.

Interoperabilidad

Agregadores de recursos

Existen repositorios que se dedican exclusivamente a la recolección y exposición de recursos de terceros. Esto significa que no cuentan con producción propia.

Hispana: más de 3 millones de registros recolectados de entre más de 150 repositorios de España. http://hispana.mcu.es

Europeana: más de 15 millones de registros recolectados de entre más de 1500 repositorios de Europa (específicamente de la Unión Europea). http://www.europeana.eu

OAIster: más de 23 millones de recursos recolectados de entre más de 1100 repositorios de acceso abierto de todo el mundo. http://www.oclc.org/oaister

Interoperabilidad

Directrices de interoperabilidad

Son un conjunto de reglas y recomendaciones que buscan establecer un marco de trabajo a fin de que dos sistemas puedan interactuar de forma exitosa y confiable.

Niveles de interoperabilidad

Niveles de interoperabilidad

Dado que *interoperabilidad* es un término muy amplio (aplicable en muchas disciplinas), existen múltiples clasificaciones del mismo.

En lo que respecta a los repositorios digitales, interesa analizar una perspectiva mas bien tecnológica y acotada:

- Interoperabilidad Sintáctica
- Interoperabilidad Semántica

Niveles de interoperabilidad

Sintáctica

Hace referencia a todo lo necesario para que dos sistemas sean capaces de establecer una comunicación e intercambiar información.

Esto incluye:

- protocolos de comunicación y transferencia
- codificación de caracteres
- formatos de datos

Niveles de interoperabilidad

Sintáctica

Elementos que corresponden a la interoperabilidad sintáctica pueden ser, por ejemplo:

- protocolo TCP/IP
- protocolo HTTP
- protocolo OAI-PMH
- formato XML y esquemas XML (XSD)
- Directrices de interoperabilidad

Niveles de interoperabilidad

Semántica

Hace referencia a todo lo necesario para que el sistema receptor haga una correcta interpretación de la información recibida, de forma automática.

Se busca que el sistema receptor "**entienda**" los datos tal como los "**entiende**" el emisor.

Para contar con interoperabilidad semántica, primero debe asegurarse la interoperabilidad sintáctica

Niveles de interoperabilidad

Semántica

Entran en juego:

- Formatos de metadatos
- Vocabularios controlados:
  - Tesauros
  - Sistemas de clasificación
- Ontologías
- Directrices de interoperabilidad

Niveles de interoperabilidad

Estándares internacionales

La adopción de estándares internacionales aumenta las capacidades de interoperabilidad del repositorio.

Protocolos de transferencia: REST, Z39.50, etc

Formatos de archivos: XML, etc

Formatos de metadatos: DC, MODS, MARCXML, etc

Directrices: DRIVER, Lucis MODS, OpenAIRE, etc

Formas de interoperar

Formas de interoperar

En general, en el contexto de los repositorios digitales se habla de:

- Búsqueda remota
- Recolección de recursos
- Depósito remoto

Formas de interoperar

Búsqueda remota: Z39.50

- Definido en los estándares internacionales ANSI/NISO z39.50 e ISO 23950
- Protocolo cliente-servidor de búsqueda y recuperación desde bases de datos remotas.
- Ampliamente utilizado en sistemas integrados de bibliotecas (ILS - *Integrated Library Systems*) para la búsqueda remota y la gestión de préstamos interbibliotecarios (*Interlibrary Loan*).
- Sintaxis de consulta específica: PQF (*Prefix Query Format*)

Formas de interoperar

Búsqueda remota: Z39.50

```
Z> find @attr 1=1003 software
Sent searchRequest.
Received SearchResponse.
Search was a success.
Number of hits: 66, setno 1
records returned: 0
Elapsed: 0.267659
```

```
Z> show 1
Sent presentRequest (1+1).
Records: 1
[INNOPAC]Record type: USmarc
00770nam  2200193I  4500
001 547843
008 730130s1970    enkm          a100 0 eng u
040    $c MIA $d m.c. $d IQU
049    $a IQUU
099    $a QA $a 76.6 $a S64 $a 1970
111 2  $a Software 70 Conference $d (1970 : $c University...)
245 10 $a Software 70: $b proceedings of a conference ...
260    $a Princeton, N. J., $b Auerbach, $c 1970.
300    $a 197 p. $b illus. $c 29 cm.
500    $a Includes bibliographical references.
650  0 $a Computer programming $v Congresses.
650  0 $a Programming languages (Electronic computers) $v Congresses.
700 1  $a Evans, David J.
710 2  $a Software World (Firm)

nextResultSetPosition = 2
Elapsed: 0.296679
Z>
```

Formas de interoperar

Búsqueda remota: Z39.50

Ventajas y desventajas

- Las consultas son abstractas respecto de la estructura de la base de datos que se está consultando
- Los mapeos de campos de búsqueda dependen de la implementación de cada servidor
- No aprovecha las ventajas de la web actual (protocolo REST)

Formas de interoperar

Búsqueda remota: SRU/SRW

SRU (*Search / Retrieve via URL*) y SRW (*Search / Retrieve via Web*) nacen como los sucesores del protocolo Z39.50, y se apoyan sobre tecnologías actuales y muy difundidas (HTTP, XML).

Al igual que Z39.50, la agencia responsable del mantenimiento de estos dos estándares es la Library of Congress

Ambos son considerados muy simples de entender e implementar

Formas de interoperar

Búsqueda remota: SRU

Se caracteriza por enviar la expresión de búsqueda (y cualquier otra indicación) dentro de una URL.

Esto es, todos los comandos necesarios para que el servidor entienda una petición y lleve a cabo las acciones pertinentes, se envían dentro de la URL misma de la petición.

http://fedora.dlib.indiana.edu:8080/SRW/search/GSearch?query=dc.title=road

Formas de interoperar

Búsqueda remota: SRW

Al igual que su *mellizo* SRU, trabaja sobre tecnologías actuales y muy difundidas: XML y HTTP, pero presenta una importante diferencia: el envío de la petición se realiza mediante un POST al servidor, en el que se envía un documento XML que contiene todas las instrucciones y datos correspondientes.

Esto es, la consulta al servidor se "empaqueta" en XML y se envía, recibiendo XML como respuesta (al igual que en el caso de SRU)

Formas de interoperar

Búsqueda remota: SRW

Las reglas y restricciones utilizadas para armar e interpretar el paquete XML están dadas por el protocolo **SOAP**.

SOAP fue creado y es mantenido por la W3C, en el área de los Web Services.

SOAP es un protocolo estándar y muy difundido.

Casi cualquier lenguaje de programación moderno tiene librerías para trabajar con SOAP.

Formas de interoperar

Búsqueda remota: SRW

Petición SRW

```
<SOAP:Envelope xmlns:SOAP="http://schemas.xmlsoap.org/soap/envelope/">
  <SOAP:Body>
    <SRW:searchRetrieveRequest xmlns:SRW="http://www.loc.gov/zing/srw/">
      <SRW:version>1.1</SRW:version>
      <SRW:query>(dc.author exact "jones" and  dc.title >= "smith")</SRW:query>
      <SRW:startRecord>1</SRW:startRecord>
      <SRW:maximumRecords>10</SRW:maximumRecords>
      <SRW:recordSchema>info:srw/schema/1/mods-v3.0</SRW:recordsSchema>
    </SRW:searchRetrieveRequest>
  </SOAP:Body>
</SOAP:Envelope>
```

Formas de interoperar

Búsqueda remota: SRW

Respuesta

```
<SOAP:Envelope xmlns:SOAP="http://schemas.xmlsoap.org/soap/envelope/">
  <SOAP:Body>
    <SRW:searchRetrieveResponse xmlns:SRW="http://www.loc.gov/zing/srw/"
      <SRW:version>1.1</SRW:version>
      <SRW:numberOfRecords>2</SRW:numberOfRecords>
      <SRW:resultSetId>8c527d60-c3b4-4cec-a1de-1ff80a5932df</SRW:resultSetId>
      <SRW:resultSetIdleTime>600</SRW:resultSetIdleTime>
      <SRW:records>
        <SRW:record>
          <SRW:recordSchema>info:srw/schema/1/mods-v3.0</SRW:recordSchema>
          <SRW:recordPacking>string</SRW:recordPacking>
          <SRW:recordData> DATOS </SRW:recordData>
          <SRW:recordPosition>1</SRW:recordPosition>
        </SRW:record>
      </SRW:records>
    </SRW:searchRetrieveResponse>
  </SOAP:Body>
</SOAP:Envelope>
```

Formas de interoperar

Búsqueda remota: OpenSearch

Es un protocolo que extiende otros formatos para agregar la búsqueda remota.

Las peticiones se realizan vía GET (los parámetros van en la URL)

Proporciona **Autodiscovery**: permite que los navegadores detecten que el sitio soporta OpenSearch y así el sitio podrá seleccionarse como motor de búsquedas del navegador

La respuestas pueden ser:

- la página de resultados del sitio en cuestión
- en RSS o ATOM, extendidos con elementos OpenSearch que agregan información sobre la búsqueda

Ejemplos: Youtube, SEDICI, Facultad de Informática

PREBI

Formas de interoperar

Recolección de recursos: OAI-PMH

Open Archives Initiative - Protocol for Metadata Harvesting

Establece un conjunto de reglas a partir de las cuales puede realizarse el intercambio de recursos de forma exitosa.

Se centra en la **transferencia** de metadatos de un extremo a otro, sin establecer restricciones en cuanto a los datos que se transfieren.

Formas de interoperar

Recolección de recursos: OAI-PMH

Define dos perfiles de trabajo

Data Provider: es aquél repositorio que ofrece sus recursos bajo el protocolo OAI-PMH, para que otros los recolecten mediante cosechas.

Service Provider: es aquél que recolecta recursos desde distintos Data Providers y brinda un servicio a una comunidad de usuarios en base a los recursos recolectados y el valor agregado aportado sobre los mismos (deduplicación, normalización, ordenamiento, búsquedas, etc).

Formas de interoperar

Depósito remoto: SWORD

Simple Web service Offering Repository Deposit

Protocolo basado en APP (Atom Publishing Protocol, a.k.a ATOMPUB)

Permite realizar el depósito de documentos de forma remota: desde otros sistemas.

Es un protocolo cliente-servidor

Formas de interoperar

Depósito remoto: SWORD

Múltiples usos potenciales

- Depósito simultáneo en múltiples repositorios
- Depósito automático por parte de equipamiento científico
- Depósito desde aplicaciones externas al repositorio (escritorio, OJS, etc)

Es un estándar que se limita a la transferencia de un objeto desde el cliente al servidor, sin imponer restricciones en cuanto a los objetos que se transportan.

Esto lo hace suficientemente flexible como para ser usado en cualquier tipo de repositorio.

Formatos de metadatos

Formatos de metadatos

Existen muchos estándares de formatos de metadatos

Cada repositorio decide que formato de metadatos usar (incluso puede usar un formato propio)

Los repositorios que deciden interoperar deben estar de acuerdo en cuanto a un formato de metadatos que ambos puedan manejar

Formatos de metadatos

En todas las formas de interoperar presentadas existe un rol de proveedor de recursos y un rol de receptor de recursos.

¿Qué sucede cuando el proveedor de recursos utiliza un formato de metadatos que no es manejado por el receptor?

¿Como se gestiona este problema?

Formatos de metadatos

Algunas de las alternativas aplicables en cualquiera de los dos roles mencionados pueden ser:

- Se decide no interactuar con ese repositorio en particular
- Extender el software para así agregar soporte para un formato de metadatos en particular
- Realizar mapeos entre formatos de metadatos
  - También dependen de la flexibilidad del software

Formatos de metadatos

Mapeos entre formatos de metadatos

En algunos casos, las entidades responsables de un formato de metadatos recomiendan cómo deben realizarse los mapeos a otros formatos. Ejemplo de esto es MODS:

Conversión de DC (sin calificar) a MODS:

http://www.loc.gov/standards/mods/dcsimple-mods.html

Conversión de MODS a DC (sin calificar):

http://www.loc.gov/standards/mods/mods-dcsimple.html

Formatos de metadatos

Mapeos entre formatos de metadatos

Manual: es un trabajo muy costoso, ya que puede tratarse de miles de registros

Automático: la transformación desde un formato complejo/jerárquico a uno simple/plano implica pérdida de información. La transformación inversa puede generar recursos deficientes en cuando a la descripción (campos incompletos, imposibilidad de uso de la especificidad de un formato complejo). No hay un humano tomando decisiones.

OAI-PMH

Open Archives Initiative
Protocol for Metadata Harvesting

OAI-PMH

Introducción

Protocolo para la recolección de metadatos

- Ampliamente adoptado por repositorios digitales en todo el mundo
- Es muy simple de entender y utilizar
- Funciona sobre XML y HTTP
- Se centra en establecer un marco de reglas para la transferencia eficiente de recursos
- No impone (*casi*) ninguna restricción en cuanto al contenido a transmitir

http://www.openarchives.org/OAI/openarchivesprotocol.html

PREBI

OAI-PMH

Introducción

Las peticiones al servidor se hacen por medio de un *verbo* y un conjunto de parámetros, codificados en una URL

http://host/oai?verb=ListRecords&metadataPrefix=oai_dc&from=2011-05-01&until=2011-10-01

http://host/oai?verb=ListRecords&resumptionToken=1320093034051

Un verbo es una *orden* que indica al servidor lo que se requiere, refinando algunos aspectos de ese requerimiento a través del uso de parámetros.

PREBI

OAI-PMH

Introducción

La respuesta a una petición OAI-PMH es un documento XML.

Se compone de dos secciones:

- *Información de la petición:* fecha, hora, verbo y parámetros (común para cualquier verbo)

- *Cuerpo con la respuesta:* datos con una estructura acorde a la información solicitada (específico para cada verbo)

OAI-PMH

Funcionamiento

Los verbos disponibles son:

- Identify
- ListRecords
- ListMetadataFormats
- ListSets
- ListIdentifiers
- GetRecord

OAI-PMH

Funcionamiento

Verbo *Identify*

Retorna información del repositorio e información acerca de la implementación del OAI Data Provider.

No recibe parámetros.

http://sedici.unlp.edu.ar/oai/request?verb=Identify

http://bdigital.uncu.edu.ar/OAI/index.php?verb=Identify

OAI-PMH

Funcionamiento

Elementos importantes que se desprenden del *Identify*

- Fecha/hora de creación del recurso mas viejo
- Granularidad de las peticiones
- Gestión de registros eliminados
- Compresión de los datos a transferir
- OAI Friends
- Descripción del repositorio

OAI-PMH
Funcionamiento

Verbo *ListRecords*

- Retorna un listado de recursos que cumplen con los parámetros especificados en la petición:
  - metadataPrefix *(obligatorio)*
  - resumptionToken *(opcional)*
  - set *(opcional)*
  - from *(opcional)*
  - until *(opcional)*

http://sedici.unlp.edu.ar/oai/request?verb=ListRecords&metadataPrefix=oai_dc&from=2011-01-01

OAI-PMH

Funcionamiento

Cosechas incrementales

por fecha (from y until)

Información clasificada

por conjuntos (set)

Paginación de resultados

resumptionToken

OAI-PMH

Funcionamiento

Registro de respuesta

```
<header>
   <identifier>ARG-UNLP-TPG-0000000006</identifier>
   <datestamp>2010-07-14</datestamp>
</header>
<metadata>
   <oai_dc:dc xmlns:...>
      <dc:title>Simulación numérica de difusión ...</dc:title>
      <dc:creator>Zyserman, Fabio Iván</dc:creator>
      <dc:subject>Física</dc:subject>
      <dc:contributor>Plastino, Angel L.</dc:contributor>
      <dc:date>2000</dc:date>
      <dc:type>Tesis de Posgrado</dc:type>
  </oai_dc:dc>
</metadata>
<about>
   <rights/>
   <provenance/>
</about>
```

OAI-PMH

Funcionamiento

Verbo *ListMetadataFormats*

Lista todos los formatos de metadatos soportados por el repositorio.

OAI-PMH obliga a exportar, por lo menos, Dublin Core sin calificar.

Se indica el *prefix* que identifica el *namespace* del formato de metadatos.

Parámetro opcional *identifier*

http://sedici.unlp.edu.ar/oai/request?verb=ListMetadataFormats

OAI-PMH

Funcionamiento

Verbo *ListSets*

- Lista los distintos Sets soportados por el repositorio
- Son una forma de organizar la información dentro del repositorio
- Poseen un nombre y una clave que los identifica
- Parámetro opcional *resumptionToken*

sedici.unlp.edu.ar/oai/request?verb=ListSets

bdigital.uncu.edu.ar/OAI/index.php?verb=ListSets

OAI-PMH

Funcionamiento

Verbo *ListIdentifiers*

- Lista los encabezados de todos los registros que se corresponden con los parámetros especificados.
- Recibe los mismos parámetros que ListRecords
- Se suele usar para determinar la cantidad y estado de los registros (borrado o no) que coinciden con ciertos parámetros, sin necesidad de descargar sus metadatos

http://sedici.unlp.edu.ar/oai/request?verb=ListIdentifiers&metadataPrefix=oai_dc&from=2011-11-01

OAI-PMH

Funcionamiento

Verbo *GetRecord*

Retorna el registro completo (encabezado y metadatos) de un recurso específico.

Recibe los parámetros

identifier

metadataPrefix

Recolección de recursos

Utilizando OAI-PMH

Recolección de recursos

Cuando se recolectan recursos desde múltiples repositorios, se presentan varios problemas.

- Políticas de catalogación independientes
- Diferencia de formatos de metadatos (y por lo tanto de especificidad de la información)
- Múltiples términos para el mismo concepto (ej.: idiomas)
- Uso de múltiples vocabularios controlados (tesauros, sistemas de clasificación, etc)
- **La gran mayoría expone sus recursos sólo en Dublin Core sin calificar**

Recolección de recursos

Problemas a solucionar

Formatos de metadatos

Mapeos a un formato común

- ¿cuál?

Diferencias en la codificación de caracteres

Presencia de caracteres inválidos:

- ¿se descarta el caracter inválido?
- ¿se descarta el documento completo?
- ¿se utiliza un caracter de reemplazo?

Recolección de recursos

Problemas a solucionar

Autores

- Distinción entre apellido y nombres (considerar el uso de iniciales)
- Muchas veces se incluye a la institución como autor
- Unificación de autores

Instituciones

- Identificación de instituciones (generalmente aparecen junto con personas)
- Unificación de instituciones

Recolección de recursos

Problemas a solucionar

Idiomas

Identificación del idioma: eng, en, en_US

Muchas veces no se indica el idioma (se necesita aplicar una detección automática)

Unificación de idiomas

Tipología documental

Múltiples formas de referenciar el mismo tipo de recurso

Artículo, ART, Article

Unificación de tipologías documentales

Recolección de recursos

Problemas a solucionar

Tipología documental

articulo
artículo
articulos
artículos
articl
paper,Artículo
article
Article
Peer-reviewed Article
PeerReviewed
ARTICULO
Artículo revisado por pares
journal article

Articles
Research paper
ARTÍCULO
Articulo
Artículos
COMUNICACION
Editorial
Comunicación
EDITORIAL
info:eu-repo/semantics/article
DOSSIER
Articulo de Investigación Cientifica

Recolección de recursos

Problemas a solucionar

Acceso al PDF o a los metadatos

Muchos casos en los que la URL apunta a una *jump-page* desde donde se accede al PDF

Otros casos, la *jump-page* no presenta ningún link al PDF

Validación de la URL de acceso al recurso

Muchas veces el enlace de acceso al recurso no funciona (o deja de funcionar un tiempo después)

¿Cómo detectar esos casos y cómo actuar? ¿se descarta el recurso?

Directrices de interoperabilidad

Directrices de interoperabilidad

Son un conjunto de recomendaciones que buscan maximizar la interoperabilidad entre los repositorios.

DRIVER 2.0 es la mas difundida en Europa y la base de muchas otras directrices en el mundo (ej.: LUCIS-MODS, OpenAIRE)

DRIVER 2.0 establece recomendaciones tanto a nivel **sintáctico** y como a nivel **semántico**.

Directrices de interoperabilidad

DRIVER 2.0

Extracto del documento de DRIVER 2.0

Para la comunicación en general es importante que la persona B sea capaz de comprender lo que la persona A está diciendo. Para este entendimiento mutuo, se necesita una base común, un léxico básico con una comprensión del significado de las cosas. A partir de este punto, ya se puede comenzar el razonamiento. Para respaldar la comunicación científica con el uso de repositorios, éstos deberían hablar el mismo idioma y por tanto es fundamental crear una base común.

PREBI

Directrices de interoperabilidad

DRIVER 2.0: características generales

Diseñado sólo para:

- Protocolo OAI-PMH
- Recursos textuales
- Documentos a texto completo
- Documentos en Acceso Abierto
- *Dublin Core sin calificar* como formato de metadatos

Directrices de interoperabilidad

DRIVER 2.0: características generales

Sobre el uso de OAI-PMH

- Se reserva el prefijo *oai_dc* para identificar el formato de metadatos *DC Sin Calificar*

- Los datestamp (tanto en las solicitudes como en las respuestas) debe respetar el formato ISO8601, expresadas en UTC: AAAA-MM-DDThh:mm:ssZ

- La política de registros eliminados debe ser por lo menos *transient* (aunque se recomienda *persistent*).

Directrices de interoperabilidad

DRIVER 2.0: características generales

Sobre el uso de OAI-PMH

Se recomienda que el resumptionToken se mantenga activo por lo menos por 24 horas.

El tamaño del lote debe ubicarse entre 100 y 500 registros.

Si se utiliza un set específico para DRIVER, se recomienda usar *driver* como setSpec.

Es obligatorio indicar un mail de contacto (campo *adminEmail* de la respuesta del verbo *Identify*)

Directrices de interoperabilidad

DRIVER 2.0: características generales

Sobre el uso de Dublin Core

Es obligatorio usar codificación Unicode.

El contenido de los metadatos no puede incluir lenguaje de marcado (HTML ni XML).

Se recomienda que el contenido de los metadatos se encuentre en inglés.

El metadato *dc:creator* debe respetar el estilo bibliográfico APA: *apellido, iniciales (nombre)*

Directrices de interoperabilidad

DRIVER 2.0: características generales

Sobre el uso de Dublin Core

Se recomienda que el metadato *dc:description* contenga un resumen del documento (el abstract).

El metadato dc:date debe respetar el formato de fecha ISO8601. Se recomienda que contenga la fecha de publicación del documento.

Directrices de interoperabilidad

DRIVER 2.0: características generales

Sobre el uso de Dublin Core

El metadato *dc:type* debe pertenecer a un vocabulario definido en un esquema URI (info:eu-repo/semantic)

info:eu-repo/semantics/article

info:eu-repo/semantics/book

info:eu-repo/semantics/bachelorThesis

info:eu-repo/semantics/masterThesis

info:eu-repo/semantics/doctoralThesis

info:eu-repo/semantics/preprint

http://www.info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:eu-repo/

Directrices de interoperabilidad

DRIVER 2.0: características generales

Sobre el uso de Dublin Core

Se recomienda que el metadato *dc:format* sea un MIME-Type incluido en IANA. Ej.: application/pdf

El metadato *dc:identifier* debe respetar un esquema URI, y vincular a:

Identificador persistente (DOI, Handle, etc)

Documento a texto completo (ej.: PDF)

Página de transición (jump-page)