PrEBi

SeDiCI SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

Bibliotecas y Repositorios Digitales

Tecnología y Aplicaciones

http://sedici.unlp.edu.ar

Bibliotecas y repositorios digitales

Capítulo 7

Interoperabilidad: ventajas y dificultades. La recolección desde otros repositorios y la exposición por diversos protocolos. El protocolo OAI-PMH. Problemas derivados del volumen y heterogeneidad de los datos recolectados. Directrices de interoperabilidad..

PREBI

Contenido

Introducción

Niveles de interoperabilidad

Formas de interoperar

Formatos de metadatos

OAI-PMH

Recolección de recursos

Directrices de interoperabilidad

Introducción

Introducción

¿Qué es la interoperabilidad?

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

Introducción

¿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.

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.

Introducción

Agregadores de recursos

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

Hispana: más de 4 millones de registros recolectados de entre más de 200 repositorios de España. http://hispana.mcu.es

Europeana: más de 32 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

Introducción

Agregadores de recursos

- Ofrecen servicios de valor agregado.

Ofrece acceso a conjuntos de todos tipos de materiales del patrimonio bibliográfico español, reunido de diversos repositorios institucionales.

Reúne contribuciones culturales digitalizadas en instituciones de la Unión Eurpea. Incluyen: libros, películas, pinturas, periódicos, mapas, manuscritos, etc.

Permite un servicio de búsqueda regional de publicaciones científicas, integrando el material de repositorios en ocho países de Latinoamérica.

PREBI

Introducción

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.

Directrices Driver 2.0: http://www.driver-support.eu

Directrices OpenAIRE: https://guidelines.openaire.eu

Directrices SNRD: http://repositorios.mincyt.gob.ar

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

PREBI

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
- Exposición Remota
- 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

Formas de interoperar

Búsqueda remota: OpenSearch

URLs de Ejemplo

- http://www.juntadeandalucia.es/medioambiente/servtc5/ventana/busquedaRSS.do?q={searchTerms}
  - Donde searchTerms es el término de búsqueda deseado
- http://<host>:<port>/alfresco/service/api/search/keyword?q={searchTerms}&p={startPage?}&c={count?}&l={language?}
- http://sedici.unlp.edu.ar/open-search/discover?rpp=100&format=atom&sort_by=2&order=desc&query=sedici.creator.person:"{name}"
  - Donde {name} es el nombre del autor a buscar

Formas de interoperar

Recolección de registros: OAI-PMH

Open Archives Initiative - Protocol for Metadata Harvesting

Establece un conjunto de reglas a partir de las cuales puede realizarse el intercambio de registros 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 registros: 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 registros 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

Exposición remota: Formatos XML

RSS

Siglas de **Really Simple Syndication**, un formato XML para sindicar o compartir contenido en la web.

Permite distribuir contenidos sin necesidad de un navegador, utilizando un software diseñado para leer estos contenidos

Es parte de la familia de los formatos XML, desarrollado específicamente para todo tipo de sitios que se actualicen con frecuencia

Formas de interoperar

Exposición remota: Formatos XML

Atom

El Protocolo de Publicación Atom (resumido en Inglés AtomPub o APP) es un protocolo simple basado en HTTP para crear o actualizar recursos en Web

Surge como alternativa a RSS

Se basa en tres conceptos:

- Fuente web
- Sitios web consumidores
- Agregadores

Formas de interoperar

Exposición remota: Formatos XML

Ejemplo de uso

El portal de revistas científicas de la UNLP, cuyo desarrollo y mantenimiento esta a cargo del equipo del SEDICI, expone de manera remota los contenidos que corresponden a las revistas del portal y que se encuentran en el repositorio.

Esto se logra mediante la utilización de los formatos xml antes mencionados.

Formas de interoperar

Exposición remota: Formatos XML

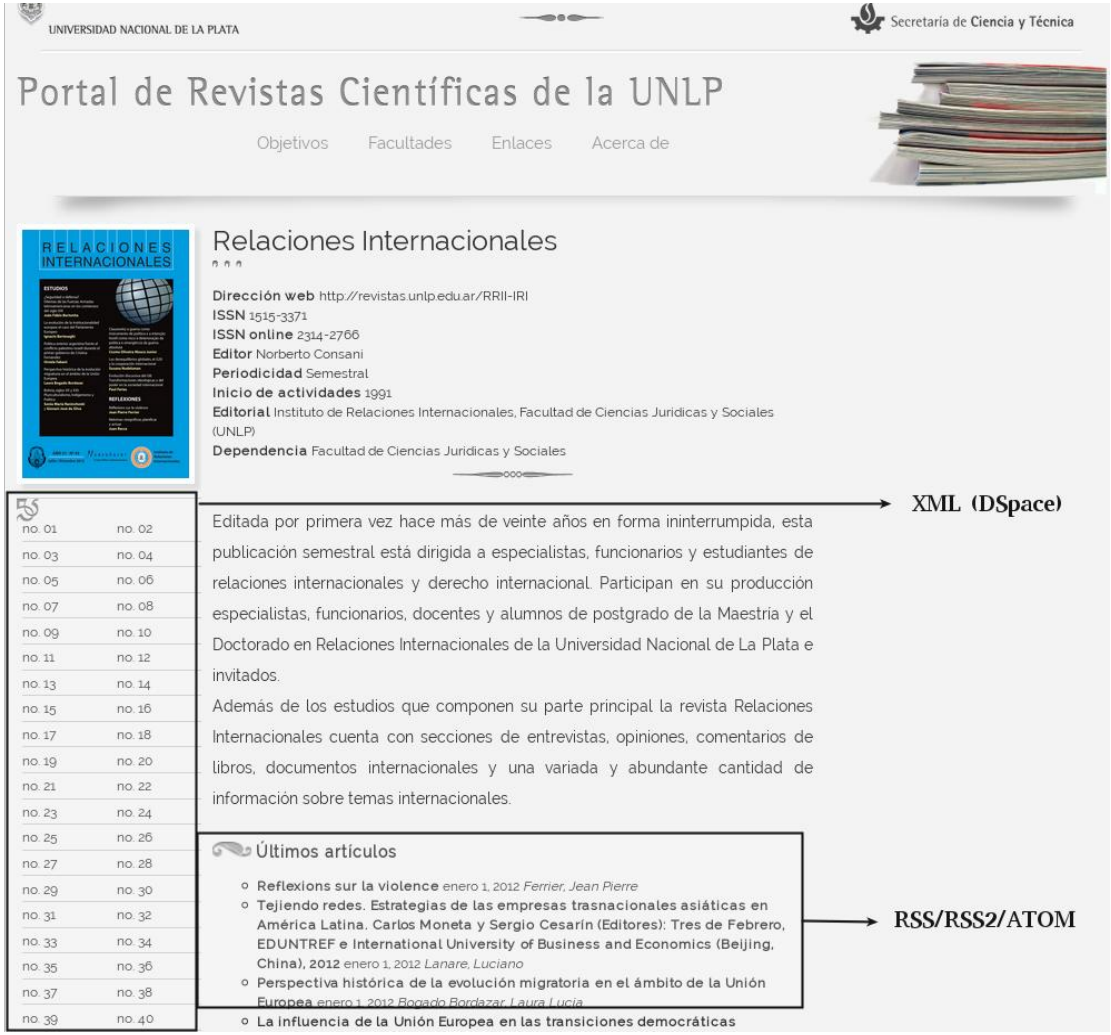
Recuperación y exposición de números y artículos de una revista (Wordpress)

Listado de números en XML
http://sedici.unlp.edu.ar/handle/10915/402?XML

RSS últimos artículos

http://sedici.unlp.edu.ar/open-search/discover?rpp=5&format=atom&sort_by=2&order=desc&scope=10915/402

http://revistas.unlp.edu.ar/RRII-IRI/gateway/plugin/WebFeedGatewayPlugin/rss2



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

Distintos usos posibles

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

Formas de interoperar

Depósito remoto: SWORD

Ejemplo de uso

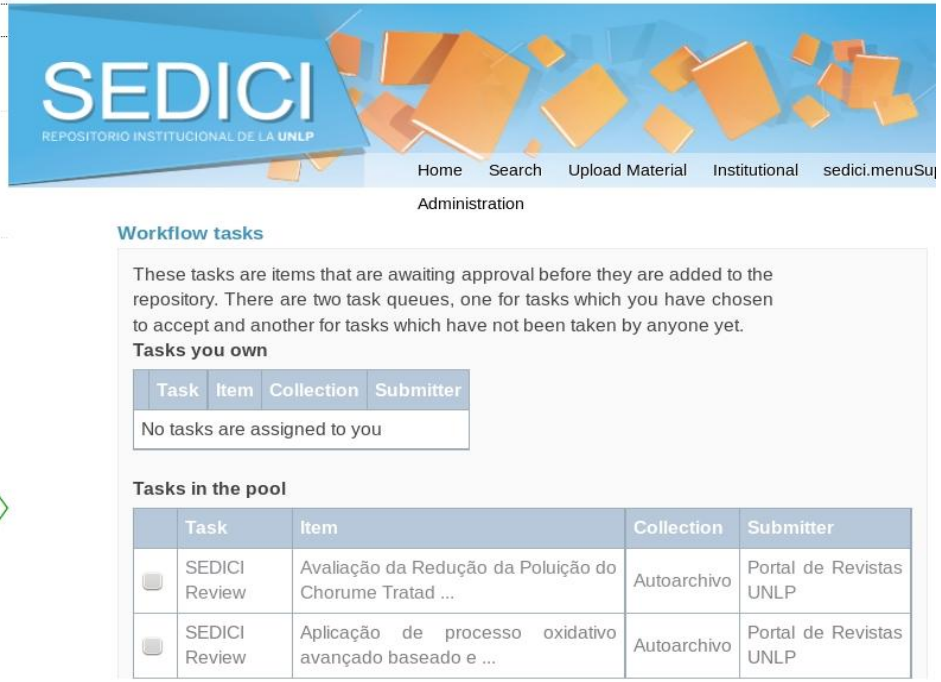
Se utiliza para depositar la producción académica del portal de revistas de la UNLP en el repositorio institucional.

En este caso:

- El Repositorio Institucional tiene implementado un sword-server
- El Portal de Revistas, implementado con OJS, posee un plugin que permite la utilización de un sword-client
- El depósito es unidireccional y diferido.
- Se creó un flujo de trabajo en el cual se deposita el contenido, se verifica el resultado de este depósito, se completan algunos metadatos que no se añaden de forma automática y se publica dicho contenido en el Repositorio.

Formas de interoperar

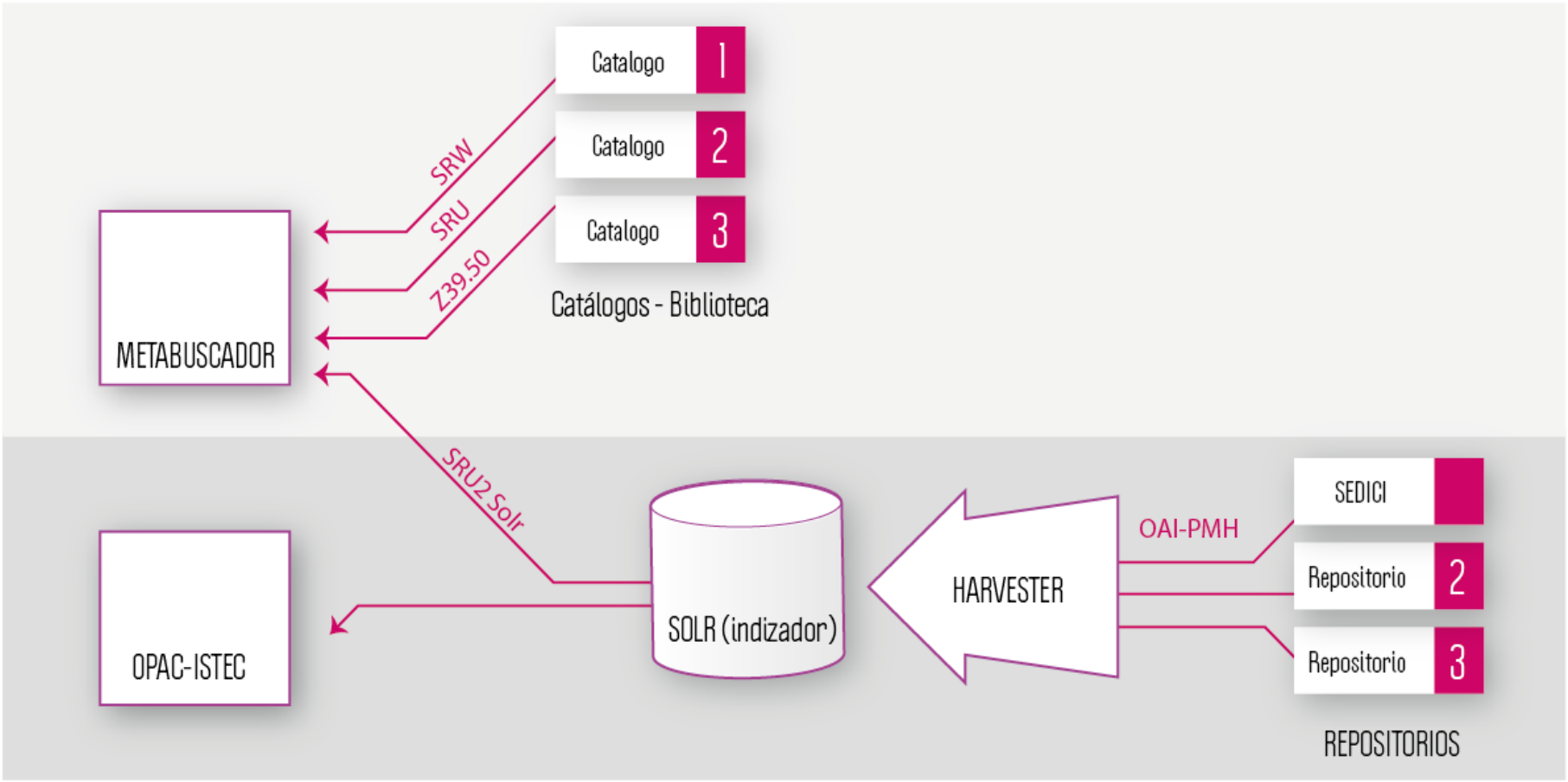
Depósito remoto: SWORD



PREBI

Formas de interoperar

Muestreo



PREBI

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.

Name of the field	Name of the field in DSpace	Name of the field in OJS
DOI	sedici.identifier.doi	doc.doi
Title	dc.title	doc.title
Alternative title	dc.title.alternative	doc.title
Author	sedici.creator.person	doc.author.fullName
Date of publication	dc.date.issued	doc.DatePublished
Language	dc.language	doc.language
Summary	dc.description.abstract	doc.abstract
Type of document	dc.type	doc.type
Issue identification	sedici.relation.journalVolumeAndIssue	journal.year, journal.number, journal.volume
Title of the journal	sedici.relation.journalTitle	journal.title

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet
  <xsl:template
     match="/epdcx:descriptionSet/epdcx:description/epdcx:statement">
<dim:field mdschema="sedici" element="relation" qualifier="journalTitle">
<xsl:value-of select="substring-after(epdcx:valueString,'*')"/>
</dim:field>
  <xsl:template
     match="/epdcx:descriptionSet/epdcx:description/epdcx:statement[@epdcx:p
     ropertyURI='http://purl.org/dc/elements/1.1/title']"><dim:field
     mdschema="dc" element="title" qualifier="alternative">
<xsl:value-of select="epdcx:valueString"/>
</dim:field>
  </xsl:template>
<xsl:template
     match="/epdcx:descriptionSet/epdcx:description/epdcx:statement[@epdcx:p
     ropertyURI='http://purl.org/dc/elements/1.1/title'][1]">
 </xsl:template></xsl:stylesheet>
```

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 registros de metadatos
- No impone (*casi*) ninguna restricción en cuanto al contenido a transmitir

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

OAI-PMH

Introducción

Entidades vinculadas a los metadatos - Definiciones

El protocolo distingue tres tipos de entidades vinculadas a los metadatos:

- **Recurso:** es el objeto o la “cosa”, física o digital, que puede ser descripta mediante metadatos. (P.e: un libro en papel).
- **Ítem:** es un componente en un repositorio que representa a un recurso, constituido por los metadatos que describen al mismo. (P.e: el libro digitalizado y su conjunto de metadatos).
- **Registro:** son metadatos en un formato específico. Es una repuesta en XML a la solicitud de recolección de un ítem específico en un repositorio.

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.

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 (Opcional)
- Descripción del repositorio (Opcional)

OAI-PMH
Funcionamiento

Verbo *ListRecords*

- Retorna un listado de registros 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

PREBI

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

PREBI

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

http://sedici.unlp.edu.ar/oai/request?verb=GetRecord&identifier=oai:sedici.unlp.edu.ar:10915/1063&metadataPrefix=oai_dc

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

Otros problemas:

Acceso al PDF o a los metadatos

Validación de la URL de acceso al recurso

Muchas correcciones a estos problemas se han logrado automatizar mediante un sistema correcto de **tareas de curación**

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**.

PREBI

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)