

## Actividades de la Clase 2

Ing. Richard Astudillo Sarmiento, MSc.

1. **Defina concisamente las tres capas principales que componen la arquitectura de DSpace. Describa el modelo de contenidos completo de DSpace (colecciones, comunidades, items, etc.).**

En la actualidad, casi la mitad de los repositorios digitales del mundo están soportados por el software DSpace y lo mismo sucede en igual o mayor medida con los nuevos repositorios que se siguen creando día a día (OpenDoar, 2014 ).

La estructura de **DSpace** consta de tres capas básicas, cada una de ellas ofrece servicios a la capa superior por medio de APIs y utiliza los servicios de la capa inferior. El código fuente se organiza en paquetes que representan esta arquitectura en las siguientes capas:

### Capa de Aplicación

La capa de Aplicación incluye todas las herramientas que permiten al exterior (usuarios u otros sistemas) hacer uso del repositorio; por ejemplo, XMLUI, JSPUI, módulo OAI, SWORD Server, entre otros (Schema, 2014).

Esta capa consta de los siguientes criterios:

- Interacción con la base de datos
- Ítems y sus metadatos
- Personas y grupos
- Información de autorización
- Trabajos en curso (workflow)
- Índices de búsqueda y exploración

### Capa de Lógica del Negocio

Esta capa ofrece:

- Administración
- Búsqueda
- Exploración
- Gestión de usuarios y grupos
- Autorización
- Carga de documentos
- Workflow
- Handle manager

- Abstracción en Comunidades, Colecciones e Ítems

### Capa de Almacenamiento

Conjunto de módulos que permiten la interacción con el mundo exterior y ofrece:

- Aplicación web: JSPUI y XMLUI
- OAI-PMH Data Provider
- Estadísticas
- Importar/Exportar
- MediaFilter

### Descripción del Modelo de Contenidos de DSpace

Una instalación DSpace ofrece varios puntos de configuración y extensión mediante diversos archivos, que definen parte del modelo de datos del repositorio, los flujos de trabajo, las características de interfaz y muchísimos puntos más. Sin embargo, muchos de estos archivos requieren no solo comprender el funcionamiento de la aplicación sino también entender su implementación y otros conceptos vinculados a las herramientas complementarias que utiliza DSpace para funcionar (De Giusti, Adorno, & Lira, 2014)

Estructura típica:

1. El repositorio se organiza en una o más **comunidades de nivel base** que se organizan jerárquicamente en **subcomunidades**, siendo estos considerados como espacios de trabajo.
2. Las **colecciones** son los “estantes” dentro de las comunidades, que agrupan contenido relacionado.
3. Los **ítems** son las obras que van en los estantes y que se pretende que el público encuentre. Los ítems están vinculados a sus bitstreams a través de entidades llamadas Bundles. Los bundles agrupan bitstreams bajo determinados criterios:
  - El criterio **original** contiene los bitstreams a publicar;
  - El criterio **Thumbnails**, son archivos extraídos a partir de los bitstreams originales;
  - El criterio **Text**, que contiene el texto completo (full-text) de otros bitstreams y se genera a partir de la extracción automática de texto sobre otros bitstreams y se usa durante la indexación para mejorar los resultados de búsqueda;
  - El criterio **License**, que contiene la licencia que el usuario aceptó al depositar el contenido y,
  - El criterio **CC License**, que contiene la Licencia CC (si es que existe) seleccionada por el usuario durante la carga.

Los ítems del repositorio se componen de objetos digitales que representan la obra en sí que se quiere publicar. Los objetos digitales pueden ser:

- ✓ Audios
- ✓ PDF
- ✓ Documentos de texto (.doc, .odt., etc)
- ✓ Diapositivas de presentaciones
- ✓ Planillas de cálculo
- ✓ Imágenes
- ✓ Videos etc.

4. Los **metadatos** describen al recurso, siendo estos datos estructurados que describen otros datos y también datos sobre datos. Uno de los esquemas estandarizados de metadatos más utilizado es el Dublin Core Element Set (DCES). Un repositorio puede utilizar un perfil de metadatos basado en esquema propio, estándar y/o mixto.

5. Los **bitstreams** son la representación digital del recurso, cuyo almacenamiento puede ser local, cuando éste se realiza en el sistema de archivos local al servidor en el que funciona la aplicación y Storage Resource Broker (SRB), el cual permite tener un sistema de archivos distribuido. En el lenguaje de DSpace, estos objetos digitales reciben el nombre de BITSTREAMS.

Los bitstreams son alojados en un Bitstore, que puede estar alojado de forma local o en la nube (según la configuración en `dspace/config/spring/api/bitstore.xml`).

- 2. Describa al menos 2 protocolos de interoperabilidad utilizados en DSpace. Basándose en el estándar del protocolo OAI-PMH (<https://www.openarchives.org/OAI/openarchivesprotocol.html>), especifique los tipos de proveedores existentes y la definición de Repositorio.**

La interoperabilidad determina la capacidad de intercambiar información entre dos o más sistemas, este módulo se localiza en la capa de aplicación de DSpace, en contacto con el mundo exterior. De acuerdo a estudios realizados, OAIPMH es el protocolo más utilizado para implementar interoperabilidad de metadatos en el ámbito de repositorios digitales (COAR, 2014).

### **Interoperabilidad - RSS y OpenSearch**

RSS se utiliza para la sindicación de noticias y contenidos en línea OpenSearch, es un conjunto de tecnologías que permite publicar los resultados de una búsqueda en un formato adecuado para la sindicación y agregación. Además permite que otras aplicaciones y sitios web expongan contenidos del repositorio y se integra fácilmente mediante RSS/Atom.

## **Interoperabilidad – REST**

REST está basado sobre HTTP, es muy simple y es utilizado en aplicaciones web modernas ya que permite obtener datos o ejecutar operaciones sobre los datos además, del intercambio de información en cualquier formato: XML, JSON, etc, ya que evita las abstracciones adicionales de otros protocolos de intercambio de mensajes.

## **Tipos de proveedores existentes**

### **OpenAIRE en EOSC**

La European Open Science Cloud (EOSC) surge del objetivo de la CE de promover el acceso y la reutilización de los datos de investigación que provienen de investigaciones financiadas con fondos públicos. Como tal, la implementación de políticas de Open Science (OS), los flujos de trabajo y las infraestructuras en todos los rincones de la esfera de investigación europea son fundamentales para que EOSC funcione (OpenAire, 2020)

OpenAIRE tiene una gran red en cada estado miembro y más allá. Está formado por expertos de Open Science en IES, centros de datos y consorcios de infraestructura que brindan capacitación y asesoramiento a nivel nacional sobre todos los aspectos de Open Science. Promueven activamente el concepto de SO y apoyan nuevas formas de comunicar ciencia e incorporar nuevos flujos de trabajo en la práctica diaria.

## **Driver**

Digital Repository Infrastructure Vision for European Research (Visión de infraestructura de repositorios digitales para la investigación europea), es un proyecto realizado por un consorcio financiado por la Unión Europea que está creando un marco de trabajo tecnológico y organizativo para implementar una capa paneuropea de datos, que permita el uso avanzado de los recursos de contenido en el ámbito de la investigación y la educación superior. DRIVER desarrolla una infraestructura de servicios y una infraestructura de datos. Ambas están concebidas para orquestar los recursos y los servicios existentes en la red de repositorios.  
<https://www.openarchives.org/OAI/openarchivesprotocol.html#ProtocolFeatures>

## **Características del protocolo**

### **Incrustación HTTP de las solicitudes de OAI-PMH**

Las solicitudes de OAI-PMH se expresan como solicitudes HTTP. Una implementación típica utiliza un servidor web estándar que está configurado para enviar solicitudes OAI-PMH al software que maneja estas solicitudes.

### **Formato de solicitud HTTP**

Las solicitudes de OAI-PMH deben ser presentadas usando los métodos HTTP GET o POST. El POST tiene la ventaja de no imponer limitaciones a la longitud de los argumentos. Los repositorios deben soportar tanto el método GET como el POST. Hay una única base URL para todas las solicitudes. La URL base especifica el host y el puerto de Internet, y opcionalmente una ruta, de un servidor HTTP que actúa como repositorio. Los repositorios exponen su URL base como el valor del elemento baseURL en la respuesta de identificación. Tenga en cuenta que la composición de cualquier ruta está determinada por la configuración del servidor HTTP del repositorio.

### **Formato de respuesta XML**

Todas las respuestas a las solicitudes de la OAI-PMH deben ser documentos de instancia XML bien formados. La codificación del XML debe utilizar la representación UTF-8 de Unicode. Deben utilizarse referencias de caracteres, en lugar de referencias de entidades. Las referencias de caracteres permiten que las respuestas XML sean tratadas como documentos independientes que pueden ser manipulados sin depender de declaraciones de entidades externas al documento.

**Un repositorio** es un servidor accesible en red que puede procesar las solicitudes de OAI-PMH de la manera descrita en este análisis. Un repositorio es administrado por un proveedor de datos para exponer los metadatos a los recolectores. <https://www.openarchives.org/OAI/openarchivesprotocol.html>

### **3. Describa la relación existente entre un metadato y una autoridad vinculada. Enumere las ventajas de utilizar metadatos controlados por autoridad.**

El término control de autoridades no es nuevo, aunque hace pocos años se ha comenzado a utilizar y se le ha dado mucha relevancia dentro de la automatización de los catálogos. Según (Herrero, 2012) el control de autoridades es el proceso de unificar, mediante la utilización de una forma normalizada, los puntos de acceso de los catálogos automatizados y mostrar además las relaciones entre los distintos puntos de acceso. Es decir, supone la normalización de los nombres de personas, entidades, títulos uniformes o materias, que pueden constituir el punto de acceso principal o los secundarios de un catálogo automatizado.

Entre los Vocabularios controlados gestionados en otros sistemas tenemos los Tesauros, Sistemas de clasificación / Taxonomías, Lista de encabezamientos de materias, Bases de datos de investigadores, Jerarquía de instituciones y Grados alcanzados. En DSpace estos vocabularios controlados son denominados como Autoridades.

La ubicación de las autoridades pueden ser externas a DSpace. Se puede facilitar la integración de nuevas autoridades sin modificar el código en DSpace.

## **Control de Autoridades en DSpace - API**

Para implementar el control de autoridades DSpace ofrece un API con las siguientes clases e interfaces:

- Choice: Clase que contiene los atributos authority, label, confidence y value.
- Choices: Clase que contiene un conjunto de Choice.
- Choice Authority: Interfaz para suplir el mecanismo del control de autoridades.

## **Control de Autoridades en DSpace - Niveles de confianza**

El nivel de confianza representa la calidad o confianza de un valor de autoridad.

1. ACCEPTED - Código 600.
2. UNCERTAIN - Código 500.
3. AMBIGUOUS - Código 300.
4. FAIL - Código 200.
5. REJECTED - Código 100.
6. NOVALUE - Código 0.
7. UNSET - Código -1.

## **Control de Autoridades en DSpace - Base de datos**

1. No es un reemplazo del valor de los metadatos.
2. Es configurado mediante un campo en la base de datos.

## **Catálogo de autoridades**

Es un conjunto organizado de registros de autoridad que contienen los datos relativos a los puntos de acceso de un catálogo bibliográfico. La información que contiene un catálogo de autoridades es:

- Forma autorizada (autoridad)
- Uno o más formas no autorizadas (variantes).
- Fuentes de información sobre la forma autorizada (fuentes)
- Fuentes de información sobre las formas no autorizadas (fuentes)
- Notas de información hacia referencias (véase, además).
- Notas hechas por el catalogador durante el proceso de creación de la autoridad (datos biográficos, administrativos, etc.).

## **Ventajas de utilizar metadatos controlados por autoridad.**

DSpace nos da la posibilidad de utilizar autoridades, siendo éstas un conjunto de valores fijos identificado por una clave. El registro de autoridad es la información asociada con uno de los valores de la autoridad y la clave de autoridad es un identificador persistente

que se corresponde con el registro de autoridad. Entre las ventajas de utilizar metadatos controlados por autoridad, podemos mencionar:

1. Permite probar que dos valores son idénticos comparando por la clave de las autoridades.
2. Ayuda a completar metadatos con valores correctos.
3. Permite mejorar la calidad de los metadatos.
4. Mejora la interoperabilidad compartiendo un nombre de autoridades con otra aplicación.
5. Reduce el tiempo de carga

4. **Describa brevemente qué es un identificador persistente, explique las ventajas de su uso. Investigue 3 proveedores de identificadores persistentes, indique el tipo de recursos sobre los que se utilizan (artículos, imágenes, personas, etc).**

La enorme expansión del ámbito digital ha hecho precisa la creación de instrumentos destinados a facilitar el manejo de los objetos digitales (Ibáñez Hernández, 2020), entre los más utilizados tenemos:

**URL, [Localizador de Recursos Uniforme]** no es más que una dirección, unas coordenadas que apuntan una localización, independientemente de los recursos que en ella se encuentran en cada momento.

**URN, [Nombre de Recurso Uniforme]** identifica un recurso, pero no indica exactamente dónde se encuentra ese objeto, por lo que no implica su disponibilidad.

**URI, [Identificador de Recurso Uniforme]** puede ser tanto un URL como un URN o una combinación de ambos que identifica un recurso de forma unívoca en un lugar determinado, además de permitir combinación de la información jerárquica que conforma la ruta con otra de estructura no jerárquica para lanzar una solicitud o consulta

### **Ventajas del uso de los Identificadores Persistentes**

Entre las muchas ventajas que ofrece el uso de los IP, se pueden mencionar:

- Las URLs pueden cambiar en el dominio o en la ruta
- Cambios en el software
- Cambios en políticas institucionales, los cuales deben notificarse al manejador de redirecciones, para que éste actualice sus reglas de re direccionamiento.
- Los recursos pueden tener más de un IP: al momento de la publicación o al momento del depósito en el repositorio institucional.
- Los identificadores persistentes pueden aplicarse a los documentos de texto: artículos, tesis, libros; a los datos de investigación: bases de datos, imágenes, audios y a las personas.

- Se usan en la Catalogación y publicación de recursos, en las Citas bibliográficas y en las Menciones en línea (blogs, redes sociales, etc.).

### **Consideraciones adicionales del uso de los IP:**

- Para lograr la persistencia, DSpace usa el sistema del CNRI, sólo a nivel del sitio DSpace, asignándole un identificador. Un usuario vería sus peticiones dirigidas al CNRI, y éste enrutando las peticiones a ese servidor. Por ello, si cambia DSpace de servidor, tendremos que reinstalar el servicio de handle, reenviando determinados ficheros al CNRI, para que actualicen sus tablas de enrutamiento.
- Es responsabilidad de cada instalación preservar la asociación entre identificadores externos y los objetos preservados. Si se hacen muchas exportaciones o importaciones, algunas de las cuáles reasignan identificadores, podría acabar dicha asociación, y los ítems previos serían ilocalizables con los URN antiguos.
- En la actualidad los identificadores persistentes no se asocian ni con bundles ni con bitstreams. Un bitstream tiene un identificador del tipo /filename, siendo x el sequence-id del bitstream, pero no son persistentes, en el sentido de que, si movemos el contenido a otro servidor, la secuencia se alterará. Y la no persistencia, a efectos de preservación, tiene ventajas e inconvenientes.
- Si se va a usar el sistema CNRI-handle, conviene configurarlo correctamente (no hace falta que el servidor handle funcione, simplemente las URLs han de ser las correctas) antes de hacer el sitio público, pues google indexará los handles antiguos.

### **Referencias Bibliográficas**

COAR. (14 de Agosto de 2014). Obtenido de El caso de Interoperabilidad para Repositorios de Acceso Abierto: <https://www.coar-repositories.org>

De Giusti, M., Adorno, F., & Lira, A. (2014). Repositorio DSpace con múltiples contextos OAI PMH. *Universidad Federal de Río Grande do Sul*.

Herrero, P. (2012). El control de autoridades. *Anales de Documentación*. Disponible en: <http://redalyc.uaemex.mx/pdf/635/63500208.pdf>.

Ibáñez Hernández, R. (8 de Enero de 2020). *Tecnologías de Información*. Obtenido de PID, Identificadores Persistentes de obras en formato digital: <https://www.biblogtecarios.es/rafaelibanez/pid-identificadores-persistentes/>

OpenAire. (2020). Obtenido de OpenAire en EOSC: donde contribuimos: <https://www.openaire.eu/openaire-and-eosc>

OpenDoar. (8 de Agosto de 2014 ). *Usage of Open Access Repository Software: Worldwide*. Obtenido de <http://www.opendoar.org/onechart.php>

Schema, I. (10 de AGOSTO de 2014). . Obtenido de XOA XML Schema: <https://github.com/lyncode/xoai/blob/3.x/schemas/xoai.xsd>