1. Actividad 2

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.).

Las tres capas principales que componen la arquitectura de **DSpace** son la capa de almacenamiento (*storage layer*), la capa de lógica de negocios (*business logic layer*), y la capa de aplicación (*application layer*).

La función de la capa de almacenamiento es persistir el volumen de información archivada y acceder a este en forma eficiente. Los metadatos se pueden almacenan en diversas bases de datos relacionales, como *PostgreSQL*, *Oracle*, u otras soportadas por *Hibernate* (herramienta de Java para persistir objetos en bases de datos). El volumen de documentos, en cambio, se almacena en el sistema de archivos, ya sea centralizado o distribuído utilizando *storage resource brokers*.

La capa de lógica de negocios provée las funciones básicas de autorización, administración y auditoría, además de gestión de grupos, búsqueda por metadatos o palabras clave (utilizando *Apache Lucene*), y permite el ingreso de documentos al repositorio. En el caso de DSpace se encuentra programada en Java, y admite su extensión por medio de un sistema de *plugins*.

La capa de aplicación agrupa el conjunto de interfaces visuales que permiten al usuario final la realización de sus tareas, como el autoarchivo de documentos, el análisis estadístico, y la importación/exportación de datos. La interoperabilidad programática se encuentra incluída en la capa de aplicación, permitiendo que el contenido del repositorio sea consultado por medio de protocolos como OAI/PMH o que se empuje contenido al mismo a través de protocolos como SWORD.

El modelo de contenidos de DSpace asigna jerárquicamente la localización de los contenidos almacenados. La agrupación de mayor nivel se denomina *community* (comunidad), dentro de la cuál pueden a su vez crearse *subcommunities* (subcomunidades). Cada comunidad o subcomunidad puede representar una facultad o centro de investigación, un departamento o un laboratorio, u otras agrupaciones convenientes; cada comunidad contendrá metadatos descriptivos, tanto sobre sí mismo como sobre las colecciones que contiene. Las comunidades pueden entenderse como espacios de trabajo conjunto.

Así como las comunidades representan, en cierta medida, agrupaciones jerárquicas de personas, las *collections* (colecciones) representan agrupaciones de contenido relacionado. Cada comunidad tendrá colecciones que contengan a su vez ítems o archivos, pero estas colecciones pueden ser compartidas, perteneciendo a múltiples comunidades para expresar un esfuerzo de colaboración. Las colec-

ciones también contienen metadatos descriptivos, tanto sobre sí mismas como sobre los ítems contenidos en ella.

Un *item* (ítem) es la unidad de registro en el repositorio que contiene tanto los metadatos como los archivos relevantes de un cierto contenido, agrupados en *bundles* (manojos). Estos últimos no son más que conjuntos de archivos agrupados y pueden ser de varios tipos, entre ellos *ORIGINAL* (los archivos depositados en el repositorio), *TEXT* (el texto plano de los artículos o documentos, si lo hubiere), *LICENSE* (la licencia de los contenidos, si aplicare). Para cada uno de los archivos almacenados en un bundle se utiliza la denominación *bitstream*.

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

La versión 2.0 del protocolo OIA-PMH, según la descripción del 08/01/2015, utiliza solicitudes HTTP o HTTPS donde la consulta es determinada por los argumentos en la dirección URL, y codifica las respuestas en XML. La definición de esquema XML utilizada puede descargarse de

http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd

Es obligatorio especificar la acción solicitada incluyendo como primer parámetro de la consulta un *verb*, entre seis posibles: *GetRecord* (obtener un registro individual de metadatos), *Identify* (obtener información sobre el repositorio), *ListIdentifiers* (obtener encabezados), *ListMetadataFormats* (obtener listado de formatos de metadatos disponibles), *ListRecords* (obtener registros), y *ListSets* (obtener la estructura de colecciones del repositorio). Como se observa, todos los verbos corresponden a operaciones de consulta y no de modificación, alta, o baja; para tales fines se utilizan protocolos alternativos.

Cada verbo, excepto Identify, incluye un conjunto de parámetros adicionales para refinar la consulta, en algunos casos opcionales. Por ejemplo, GetRecord requiere los parámetros *identifier* y *metadataPrefix*, que especifican respectivamente el identificador único del elemento cuyo registro es solicitado y una cadena que determina el formato de metadatos. Los valores soportados por el repositorio pueden obtenerse con el verbo ListMetadataFormats, posiblemente restringiendo los resultados utilizando el parámetro opcional *identifier*, si se requiere uno solo de ellos.

Construiremos un ejemplo concreto, suponiendo que nuestro repositorio digital se encuentra alojado en la dirección *nuestro.repo.digital*, y que el punto de entrada se denonomina *dataprovider*. Así, obtener el detalle del formatos de metadatos *oai_dc* se realizaría invocando la dirección

http://nuestro.repo.digital/dataprovider/verb=ListMetadataFormats&identifier=oai_dc

que devolvería algo similar a

Listing 1: Resultado de ListMetadataFormats

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
         http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-02-08T14:27:19Z</responseDate>
  <request verb="ListMetadataFormats"
    identifier="oai:perseus.tufts.edu:Perseus:text:1999.02.0119">
    http://www.perseus.tufts.edu/cgi-bin/pdataprov</request>
  <ListMetadataFormats>
   <metadataFormat>
     <metadataPrefix>oai_dc</metadataPrefix>
     <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd
       </schema>
     <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/
       </metadataNamespace>
   </metadataFormat>
 </ListMetadataFormats>
</OAI-PMH>
```

Los dos tipos de proveedores, o participantes, definidos en el marco de OAI-PMH son los *proveedores de datos* y los *proveedores de servicios*. Los primeros administran sistemas que soportan el protocolo OAI-PMH como una medio para exponer metadatos, mientras que los segundos utilizan los metadatos recolectados a través de OAI-PMH como sustrato para crear servicios de valor agregado.

El protocolo SWORD es una especialización del protocolo *Atom Publishing Protocol* que permite a un repositorio digital aceptar publicaciones de documentos de sus clientes utilizando formatos diversos. SWORD es un protocolo liviano que ha sido implementado en DSpace como un módulo de extensión. Existen tres versiones de SWORD, que pueden consultarse en

http://swordapp.org/

La presente descripción utilizará la versión v3 del protocolo. SWORD especifica la operación a realizar utilizando tanto el verbo HTTP (GET, POST, PUT, o DELETE) como una URL de servicio. Así, por ejemplo, la URL de servicio *Metadata-URL* puede servir tanto para obtener los metadatos de un objeto por medio del verbo GET como para reemplazarlos por medio del verbo PUT. Idénticamente, la URL de servicio *Object-URL* soporta lectura, agregado, reemplazo, y eliminación de objetos respectivamente por medio de los verbos GET, POST, PUT, y DELETE, en ese orden. Las URL de servicio soportadas son *Service-URL* (obtener o establecer el documento de servicio), *Object-URL*, *Metadata-URL* (anteriormente descriptos), *FileSet-URL* (operaciones de colecciones), *File-URL* (operaciones de archivo), *Staging-URL* (creación de URL temporaria para subir archivos segmentados), y *Temporary-URL* (operaciones con archivos segmentados).

SWORD hace uso extensivo de los mecanismos disponibles en HTTP para no crear elementos *ad hoc*. Los *headers* de HTTP son empleados en lugar de parámetros en la URL; por ejemplo, *Authorization* para enviar las credenciales de autorización requeridas. Asimismo, los códigos de error y las redirecciones tienen sentido propio dentro del protocolo. Por ejemplo, los códigos 403 (forbidden, es decir "prohibido") y 404 (not found, es decir "no encontrado"), equivalen a ausencia de autorización para realizar la operación solicitada y a inexistencia del recurso solicitado.

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

Los metadatos están representados por una clave, un valor, y otros campos asociados como por ejemplo el idioma correspondiente al valor del metadato en cuestión. Una autoridad determina el conjunto de valores permitidos para una cierta clave. De esta forma, la utilización de autoridades evita ambigüedades, ya que dos valores idénticos del metadato necesariamente compartirán idéntico valor; el caso es similar al descrito en el punto siguiente con la utilización de ORCID para desambiguar autores, pero no está restringido a personas. La única ventaja de la utilización de autoridades vinculadas a metadatos no es la desambiguación; también lo es la mejora en la calidad descriptiva de los metadatos, pues simplifica el ingreso de valores correctos. Desde el punto de vista de la interoperabilidad, la utilización de autoridades asegura que se puedan compartir valores de manera comprensible entre repositorios o aplicaciones; mientras que una biblioteca digital que codifique el tema de un libro como "HISTORIA" no podría compartir sus metadatos con otra que utilice, por ejemplo, "HISTORIOGRAFÍA" para el mismo fin, ambas sí podrían hacerlo si coincidieran en el uso de la misma autoridad para tal metadato.

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).

Un *identificador persistente* no es más que una referencia duradera al objeto referenciado, ya sea este un documento, un conjunto de datos, una persona, o cualquier otra entidad que necesite ser referenciada sin ambigüedades. En el contexto de los repositorios digitales, un identificador persistente generalmente hará referencia a un documento digital o a alguno de sus autores; sin embargo, su utilización no es exclusiva de los medios digitales y muchos esquemas, como el ISBN, son independientes de tal representación.

La necesidad de identificadores persistentes en el contexto de Internet es producto del choque entre la imposibilidad de asegurar la permanencia a largo plazo de direcciones URL estables y el requisito de apuntar a documentos

u otros recursos sin ambigüedades. Pueden cambiar los dominios de almacenamiento o la ruta interna dentro de un dominio dado puede ser modificada, pero un identificador persistente asegura que el documento o archivo en cuestión será localizado a través de su identificador persistente. El responsable de almacenar el documento deberá notificar al manejador de redirecciones toda vez que la dirección del mismo varíe, y este último asegurará que el identificador persistente apunte a la localización actual.

Como ejemplo de un identificador persistente de existencia previo a Internet se ha mencionado el ISBN, que es un identificador numérico único para libros de circulación comercial, administrado por la *International ISBN Agency*. El ISBN consta de tres grupos numéricos, un dígito de control, y tres dígitos de prefijo para asegurar la compatibilidad con el formato EAN de códigos de barras. Los grupos numéricos están conformados por un elemento de registración, que agrupa a los países por idioma, un registrante -que suele ser la compañia editorial-, y una publicación, que identifica el libro en sus distintas ediciones.

Idénticamente, el *digital object identifier* (DOI) es un identificador persistente digital de uso extendido para una obra publicada, que es administrado por la *International DOI Foundation*, y ha sido estandarizado por ISO. Si bien se utiliza fundamentalmente para publicaciones académicas, no existe una limitación específica y puede utilizarse para otros tipos de archivos. DOI, en contraste con ISBN que se limita a identificar, agrega un mecanismo de resolubilidad que permite encontrar el objeto referenciado incluyendo entre los metadatos información como su dirección URL. Su nomenclatura consta de un prefijo y un sufijo, donde el primero identifica a la entidad registrante del identificador y el segundo es elegido arbitrariamente por el registrante para identificar el documento.

ORCID (*Open Researcher and Contributor ID*) es un identificador persistente para autores científicos y académicos, no para documentos. Su objetivo es desambiguar los nombres personales, que no son únicos, pueden variar por diversos motivos (casamiento, divorcio, motivos legales, cambio voluntario de nombre, etc.) a lo largo del tiempo, y en muchas ocasiones son representados de distintas maneras en distintos alfabetos; permitiendo así trazar la contribución de un autor particular. ORCID es administrado por ORCID, Inc., y su nomenclatura consiste en un subconjunto de ISNI (*International Standard Name Identifier*). Consiste en cuatro grupos de cuatro dígitos decimales separados por guiones, en el rango desde 0000-0001-5000-0007 hasta 0000-0003-5000-0001.