Antología: DSpace

García Rojas Alan
Liévana Poy Erick
Lima Estrada Efraín
Reyes Reyes Julián
BUAP Puebla - México

2021

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
Facultad de Ciencias de la Computación
Administración de Redes

Antología: DSpace

García Rojas Alan
Liévana Poy Erick
Lima Estrada Efraín
Reyes Reyes Julián
Profesor Larios Gomez Mariano

Dedicatoria

Este trabajo se realizo gracias al esfuerzo en conjunto de los alumnos: García Rojas Alan, Liévana Poy Erick, Lima Estrada Efraín, Reyes Reyes Julián y a la gran colaboración del profesor Larios Gomez Mariano. Asi mismo agradecemos a la Benemérita Universidad Autónoma de Puebla, por el apoyo brindado en estos tiempos dificiles.

Índice general

Capítulo 1
Introducción

1.1. Repositorios

Un repositorio es un espacio centralizado donde se almacena, organiza, mantiene y difunde información digital, habitualmente archivos informáticos, que pueden contener trabajos científicos, conjuntos de datos o software. Son sistemas de información que preservan y organizan materiales científicos y académicos como apoyo a la investigación y el aprendizaje, a la vez que garantizan el acceso a la información
Los repositorios abiertos tienen sus inicios en los años 90, en el área de la física y las matemáticas, donde los académicos aprovecharon la red para compartir sus investigaciones con otros colegas. Este proceso era valioso porque aceleraba el ciclo científico de publicación y revisión de resultados
Los datos almacenados en un repositorio pueden distribuirse a través de una red informática, como Internet, o de un medio físico, como un disco compacto. Pueden ser de acceso público o estar protegidos y necesitar de una autenticación previa. Los repositorios más conocidos son los de carácter académico e institucional. Los repositorios suelen contar con sistemas de respaldo y mantenimiento preventivo y correctivo, lo que hace que la información se pueda recuperar en el caso que la máquina quede inutilizable o ciertos formatos queden obsoletos con el paso del tiempo.

1.2. DSpace

Es un software de código abierto que provee herramientas para la administración de colecciones digitales a través de repositorios abiertos. Responde a la necesidad específica como sistema de archivos digitales centrado en el almacenamiento, acceso y preservación a largo plazo de contenido digital.
Comúnmente usada para la administración de colecciones digitales o repositorio bibliográfico. Soporta una gran variedad de datos entre los que destacan:

- Libros
- Tesis
- Fotografías
- Filmes
- Vídeos
- Datos de Investigación

1.2.1. Historia

La primera versión de DSpace fue liberada en noviembre de 2002, siguiendo un esfuerzo conjunto por los desarrolladodres del MIT y HP Labs en Cambridge, Massachusetts. En marzo de 2004 tuvo lugar el primer DSpace User Group Meeting en MIT, un grupo de instituciones interesadas formó la Federación DSpace,

que determinó la gobernanza del futuro desarrollo de software mediante la adopción del modelo de desarrollo comunitario de la Fundación Apache y el establecimiento del Grupo DSpace Committer.
En julio de 2007, a medida que la comunidad de usuarios de DSpace crecía, HP y MIT formaron conjuntamente la Fundación DSpace, una organización sin fines de lucro que brindó liderazgo y apoyo. En mayo de 2009, la colaboración en proyectos relacionados y las crecientes sinergias entre la Fundación DSpace y la organización Fedora Commons llevaron a la unión de las dos organizaciones para perseguir su misión común en una organización sin fines de lucro llamada DuraSpace. DuraSpace y LYRASIS se fusionaron en julio de 2019. Actualmente, el software DSpace y la comunidad de usuarios reciben liderazgo y orientación de LYRASIS.
Dspace es usado por mas de 2500 instituciones al lo largo del mundo, entre las que destacan:

- El Banco Mundial
- Universidad de Cambridge
- Universidad de Hardvard
- Instituto de Tecnología de Massachusetts
- Imperial College London
- La Organización Mundial de la Salud

Capítulo 2

Características

DSpace está construido con aplicaciones web Java, muchos programas y un almacén de metadatos asociado. Las aplicaciones web proporcionan interfaces para administración, depósito, ingesta, búsqueda y acceso. El almacén de activos se mantiene en un sistema de archivos o un sistema de almacenamiento similar. Los metadatos, incluida la información de acceso y configuración, se almacenan en una base de datos relacional y admiten el uso de PostgreSQL y la base de datos Oracle. Las existencias de DSpace están disponibles principalmente a través de una interfaz web. Las versiones más recientes de DSpace también admiten la función de búsqueda y exploración por facetas mediante Apache Solr.
Las principales características de DSpace son:

- Sofware de Código libre y abierto
- Manejo del FrontEnd y BackEnd
- Base de Datos
- Motor de Búsqueda
- Uso de Metadatos

2.1. Arquitectura de la Aplicación

El sistema DSpace está organizado en tres capas, cada una de las cuales consta de varios componentes.

- **Capa de aplicación:** responsable del almacenamiento físico de metadatos y contenido.
- **Capa de lógica empresarial:** se ocupa de la gestión del contenido del archivo, los usuarios del archivo, la autorización y el flujo de trabajo.
- **Capa de almacenamiento:** contiene componentes que se comunican con el mundo fuera de la instalación individual de DSpace, por ejemplo, la interfaz de usuario web y el protocolo Open Archives Initiative para el servicio de recolección de metadatos.

Cada capa solo invoca la capa debajo de ella; la capa de aplicación puede no usar la capa de almacenamiento directamente, por ejemplo. Cada componente de las capas de almacenamiento y lógica empresarial tiene una API pública definida. La unión de las API de esos componentes se conoce como la API de almacenamiento (en el caso de la capa de almacenamiento) y la API de DSpace Java (en el caso de la capa de lógica empresarial), y la API de DSpace REST (en el caso de la capa de aplicación). En la capa de aplicación, vale la pena señalar que la interfaz de usuario web solo accede a DSpace a través de la API REST.
Es importante tener en cuenta que se confía en cada capa. Aunque la lógica para autorizar acciones está en la capa de lógica empresarial, el sistema se basa en aplicaciones individuales en la capa de aplicación para autenticar a las personas electrónicas de forma correcta y segura. Si se permitiera que una aplicación 'hostil'

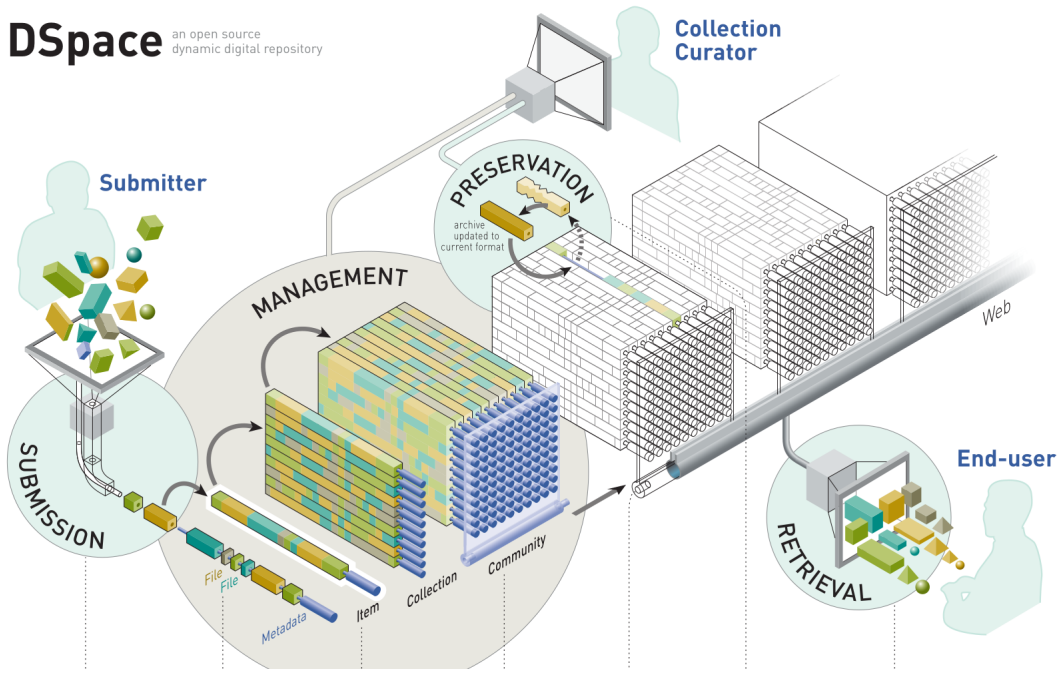

Figura 2.1: Diagrama del funcionamiento general de DSpace

o insegura invocara la API de Java directamente, podría realizar acciones muy fácilmente como cualquier persona electrónica en el sistema.
La razón de esta elección de diseño es que los métodos de autenticación variarán ampliamente entre diferentes aplicaciones, por lo que tiene sentido dejar la lógica y la responsabilidad de eso en estas aplicaciones.
El código fuente está organizado para ser coherente muy estrictamente con esta arquitectura de tres capas.

2.2. Motor de Búsqueda Integrado

DSpace puede procesar contenido basado en texto cargado para búsquedas de texto completo. Esto significa que no solo se podrán buscar los metadatos que proporcione para un archivo determinado, sino que también se indexará todo su contenido. Esto permite a los usuarios buscar palabras clave específicas que solo aparecen en el contenido real y no en la descripción proporcionada.
Tiene integrado Apache Solr, un motor de búsqueda con interfaces de busqueda personalizables que permite:

- Búsqueda y Recuperación
- Filtro de resultados
- El texto completo de los archivos es indexado
- Búsqueda a través de metadatos

2.3. Base de Datos

Un sistema gestor de base de datos es un conjunto de programas que permiten el almacenamiento, modificación y extracción de la información en una base de datos. Los usuarios pueden acceder a la información usando herramientas específicas de consulta y de generación de informes, o bien mediante aplicaciones al efecto.
Estos sistemas también proporcionan métodos para mantener la integridad de los datos, para administrar el acceso de usuarios a los datos y para recuperar la información si el sistema se corrompe. Permiten presentar

la información de la base de datos en variados formatos. La mayoría incluyen un generador de informes. También pueden incluir un módulo gráfico que permita presentar la información con gráficos y tablas.
Generalmente se accede a los datos mediante lenguajes de consulta, lenguajes de alto nivel que simplifican la tarea de construir las aplicaciones. También simplifican las consultas y la presentación de la información. Un SGBD permite controlar el acceso a los datos, asegurar su integridad, gestionar el acceso concurrente a ellos, recuperar los datos tras un fallo del sistema y hacer copias de seguridad. Las bases de datos y los sistemas para su gestión son esenciales para cualquier área de negocio, y deben ser gestionados con esmero.
En el caso de DSpace se hace uso de la base de datos PostgreSQL, la cual como muchos otros proyectos de código abierto, el desarrollo de PostgreSQL no es manejado por una empresa o persona, sino que es dirigido por una comunidad de desarrolladores que trabajan de forma desinteresada, altruista, libre o apoyados por organizaciones comerciales.
PostgreSQL se ha ganado una sólida reputación por su arquitectura probada, confiabilidad, integridad de datos, conjunto de características robustas, extensibilidad y la dedicación de la comunidad de código abierto detrás del software para brindar soluciones innovadoras y de alto rendimiento de manera consistente. PostgreSQL se ejecuta en todos los principales sistemas operativos, ha sido compatible con ACID desde 2001 y cuenta con potentes complementos como el popular extensor de base de datos geoespacial PostGIS.
Entre sus mayores caracteristicas:

- Acepta casi todos los tipos de datos
- Integridad de la información
- Concurrencia
- Alto rendimiento
- Recuperación de información
- Seguridad
- Extensible

2.4. Tipos de Archivos

DSpace puede acomodar cualquier tipo de archivo cargado. Si bien DSpace es más conocido por albergar materiales basados en texto, incluida la comunicación académica y las tesis y disertaciones electrónicas, hay muchas partes interesadas en la comunidad que usan DSpace para multimedia, datos y objetos de aprendizaje. Si bien se aplican algunas restricciones, DSpace puede incluso servir como almacén de archivos HTML.
Los archivos que se han subido a DSpace a menudo se denominan "Bitstreams". La razón de esto es principalmente histórica y se remonta a la implementación técnica. Después de la ingestión, los archivos en DSpace se almacenan en el sistema de archivos como una secuencia de bits sin la extensión del archivo.
De forma predeterminada, DSpace solo reconoce tipos de archivos específicos, como se define en su Registro de formato Bitstream. El registro de formato Bitstream predeterminado reconoce muchos formatos de archivo comunes, pero se puede mejorar en su institución local a través de la interfaz de usuario de administrador.
Optimizado para la indexación de Google
La comunidad de Duraspace fomenta una estrecha relación con Google para garantizar una indexación óptima del contenido de DSpace, principalmente en la Búsqueda de Google y los productos de Google Scholar. A los efectos de la indexación de Google Scholar, DSpace agregó metadatos específicos en las etiquetas de encabezado de página para facilitar la indexación en Scholar. Se puede obtener más información en la página Asignaciones de metadatos de Google Scholar. Los repositorios populares de DSpace a menudo generan más del 60 % de sus visitas desde las páginas de Google.
Entre los tipos de archivos mas importantes se hallan:

- DOC
- PDF
- XLS

- PPT
- JPEG
- MPEG
- TIFF

2.5. Metadatos

DSpace usa Qualified Dublin Core(QDC) un sistema de 15 definiciones semánticas descriptivas que pretenden transmitir un significado semántico a las mismas. Estas definiciones:

- Son opcionales
- Se pueden repetir
- Pueden aparecer en cualquier orden

Este sistema de definiciones fue diseñado específicamente para proporcionar un vocabulario de características "base", capaces de proporcionar la información descriptiva básica sobre cualquier recurso, sin que importe el formato de origen, el área de especialización o el origen cultural
En términos generales, DSpace contiene tres tipos de metadatos sobre el contenido archivado:

- **Metadatos descriptivos:** DSpace puede admitir varios esquemas de metadatos planos para describir un elemento.
- **Metadatos administrativos:** esto incluye metadatos de preservación, datos de políticas de autorización y procedencia.
- **Metadatos estructurales:** esto incluye información sobre cómo presentar un elemento, o flujos de bits dentro de un elemento, a un usuario final y las relaciones entre las partes constituyentes del elemento.

2.6. Seguridad

DSpace tiene integrado su propio sistema de autenticación y autorización. De igual forma permite el uso de otros sistemas de autentificarían como LDAP o Shibboleth. Permite la asignación de permisos para leer/escribir, en todo el sistema, por comunidad o por colección, por ítem o por archivo. Igual permite la asignación de permisos administrativos por comunidad o por colección

2.6.1. Autorización

La autenticación es cuando una sesión de aplicación se identifica positivamente como perteneciente a un Usuario y/o Grupo. En DSpace, se implementa mediante un mecanismo llamado Autenticación apilable: la configuración de DSpace declara una "pila"de métodos de autenticación. Una aplicación (como la interfaz de usuario web) llama al administrador de autenticación, que prueba cada uno de estos métodos a su vez para identificar la persona electrónica a la que pertenece la sesión, así como los grupos adicionales. Los métodos de autenticación de Usuario se prueban sucesivamente hasta que uno tiene éxito. Cada autenticador de la pila tiene la oportunidad de asignar grupos adicionales. Este mecanismo ofrece las siguientes ventajas:

- Separa la autenticación de la interfaz de usuario web, por lo que se utilizan los mismos métodos de autenticación para otras aplicaciones, como los servicios web no interactivos.
- Modularidad mejorada: los métodos de autenticación son independientes entre sí. Los métodos de autenticación personalizados se pueden .ªpilar"sobre el método predeterminado de nombre de usuario/contraseña de DSpace.
- Soporte más limpio para la autenticación ïmplícita"donde el nombre de usuario se encuentra en el entorno de una solicitud web.

2.6.2. Autentificación

El sistema de autorización de DSpace se basa en asociar acciones con objetos y las listas de e-person que pueden realizarlas. Las asociaciones se denominan Políticas de recursos y las listas de Usuarios se denominan Grupos. Hay dos grupos integrados: .ªdministradores", que pueden hacer cualquier cosa en un sitio, y .ªnónimos", que es una lista que contiene a todos los usuarios. Asignar una política para una acción en un objeto a anónimo significa dar permiso a todos para realizar esa acción. Los permisos deben ser explícitos: la falta de un permiso explícito da como resultado la política predeterminada de 'denegar'. Los permisos tampoco 'conmutan'; por ejemplo, si un usuario tiene permiso de LECTURA en un elemento, es posible que no necesariamente tenga permiso de LECTURA en los paquetes y flujos de bits de ese elemento. Actualmente, las colecciones, comunidades y elementos se pueden descubrir en los sistemas de exploración y búsqueda independientemente de la autorización de LECTURA.

2.7. Recuperación

DSpace permite exportar todo el contenido del repositorio como Paquetes de Información de Archivo(AIP). Estos AIP pueden ser usados para restaurar el sitio completo, comunidades, colecciones o items.
Asi mismo se hace uso de un checador de suma para verificar que el contenido de un repositorio de DSpace no se haya corrompido o manipulado. La funcionalidad se puede invocar de forma ad-hoc desde la línea de comandos o configurarse a través de cron o similar. Existen opciones para admitir grandes repositorios que no se pueden verificar por completo en una ejecución de la herramienta. La herramienta se puede ampliar a nuevos enfoques de prioridad de informes y verificación.
Al recibir un archivo DSpace calcula y guarda una suma de verificación para cada archivo. DSpace usa estas sumas de verificación para validar la integridad de los archivos.

Capítulo 3

Funcionamiento

3.1. Propuesta de Información

De forma predeterminada, el flujo de trabajo de una colección puede tener hasta tres pasos. Cada colección puede tener un grupo de usuarios asociado para realizar cada paso; si no hay ningún grupo asociado con un determinado paso, ese paso se omite. Si una colección no tiene grupos de personas electrónicas asociados con ningún paso, los envíos a esa colección se instalan directamente en el archivo principal. Sin embargo, tenga en cuenta que este es solo el comportamiento predeterminado y que el proceso de flujo de trabajo se puede configurar/personalizar fácilmente, consulte Flujo de trabajo configurable.
En otras palabras, la secuencia predeterminada es la siguiente: La colección recibe un envío. Si la colección tiene un grupo asignado para el paso 1 del flujo de trabajo, se invoca ese paso y se notifica al grupo. De lo contrario, se omite el paso 1 del flujo de trabajo. Del mismo modo, los pasos 2 y 3 del flujo de trabajo se realizan si y solo si la colección tiene un grupo asignado a esos pasos.
Cuando se invoca un paso, el envío se coloca en el "grupo de tareas"del grupo asociado al paso. Un miembro de ese grupo toma la tarea del grupo y luego se elimina del grupo de tareas, para evitar la situación en la que varias personas del grupo pueden estar realizando la misma tarea sin darse cuenta.
El miembro del grupo que ha tomado la tarea del grupo puede realizar una de estas tres acciones:

- Aceptar o Rechazar la propuesta.
- Puede editar los metadatos proporcionados por el usuario con el envío, pero no puede cambiar los archivos enviados. Puede aceptar el envío para su inclusión o rechazar el envío.
- Puede editar los metadatos proporcionados por el usuario con el envío, pero no puede cambiar los archivos enviados. Luego debe comprometerse a archivar; no puede rechazar la presentación.

Si es rechazada se informa al usuario a través de un correo la razón del rechazo, el usuario puede hacer las modificaciones necesarias para reiniciar el proceso
Si es aceptada, entonces se inicia el siguiente paso en la metodología de trabajo. De no haber otro paso la información es guardada en la colección

3.2. Estructura

La forma en que se organizan los datos en DSpace pretende reflejar la estructura de la organización que utiliza el sistema DSpace. Cada sitio de DSpace se divide en comunidades, que se pueden dividir en subcomunidades que reflejan la estructura universitaria típica de colegio, departamento, centro de investigación o laboratorio. Las comunidades contienen colecciones, que son agrupaciones de contenido relacionado. Una colección puede aparecer en más de una comunidad.

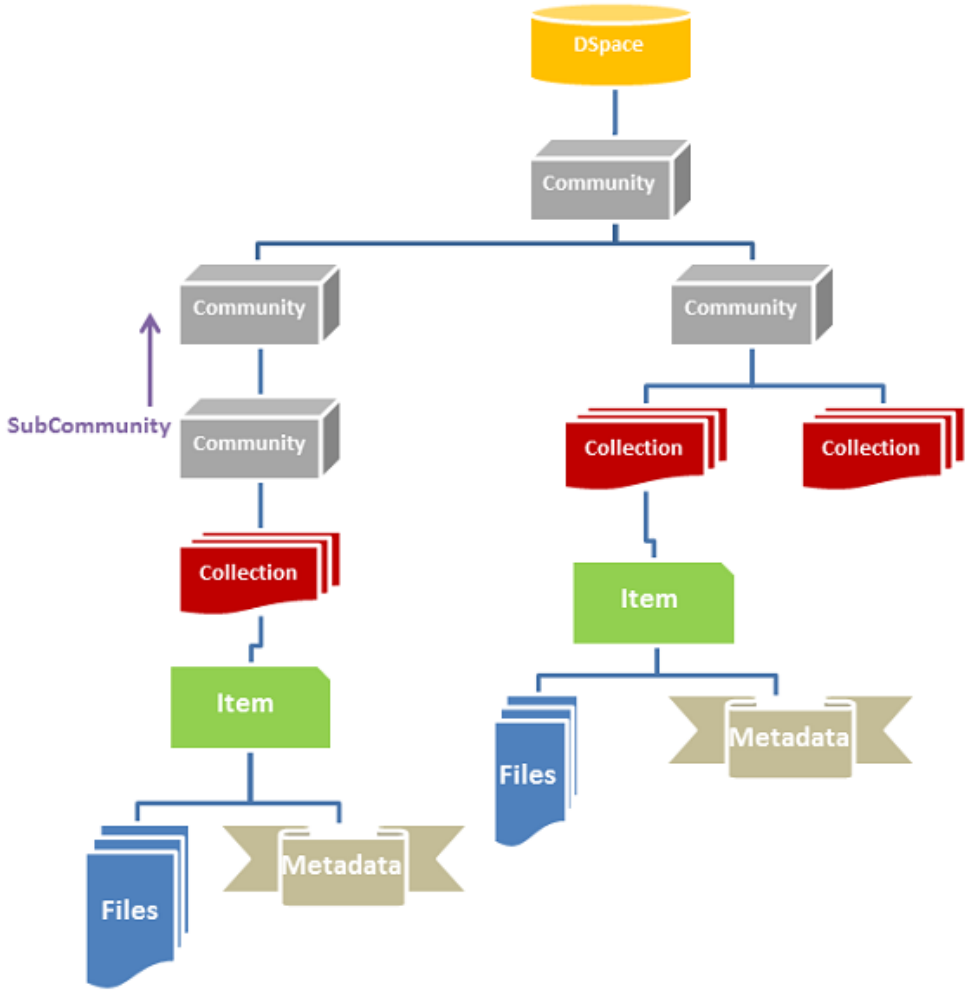

Figura 3.1: Estructura de un repositorio DSpace

Cada colección se compone de elementos, que son los elementos archivísticos básicos del archivo. Cada artículo pertenece a una colección. Además, un artículo puede aparecer en colecciones adicionales; sin embargo, cada artículo tiene una sola colección.
Los elementos se subdividen a su vez en paquetes de flujos de bits con nombre. Los flujos de bits son, como su nombre indica, flujos de bits, generalmente archivos de computadora ordinarios. Los flujos de bits que de alguna manera están estrechamente relacionados, por ejemplo, los archivos HTML y las imágenes que componen un solo documento HTML, se organizan en paquetes.
En la práctica, la mayoría de los elementos tienden a tener estos paquetes con nombre:

- **ORIGINAL:** el paquete con los flujos de bits depositados originales
- **THUMBNAILS:** miniaturas de cualquier flujo de bits de imágenes
- **TEXT:** texto completo extraído de flujos de bits en ORIGINAL, para indexación
- **LICENSE:** contiene la licencia de depósito que el remitente otorgó a la organización anfitriona; en otras palabras, especifica los derechos que tiene la organización de acogida
- **CC_LICENCE:** contiene la licencia de distribución, si la hubiera asociada con el artículo. Esta licencia especifica lo que los usuarios finales que descargan el contenido pueden hacer con el contenido.

Cada flujo de bits está asociado con un formato de flujo de bits. Debido a que los servicios de preservación pueden ser un aspecto importante del servicio DSpace, es importante capturar los formatos específicos de los archivos que envían los usuarios. En DSpace, un formato de flujo de bits es una forma única y coherente de referirse a un formato de archivo en particular. Una parte integral de un formato de flujo de bits es una noción implícita o explícita de cómo se puede interpretar el material en ese formato. Por ejemplo, la interpretación de los trenes de bits codificados en el estándar JPEG para la compresión de imágenes fijas se define explícitamente en el estándar ISO/IEC 10918-1. La interpretación de los flujos de bits en formato Microsoft Word 2000 se define implícitamente, mediante referencia a la aplicación Microsoft Word 2000. Los formatos de flujo de bits pueden ser más específicos que los tipos MIME o los sufijos de archivo. Por ejemplo, application/ms-word y .doc abarcan varias versiones de la aplicación Microsoft Word, cada una de las cuales produce flujos de bits con características presumiblemente diferentes.
Además, cada formato de flujo de bits tiene un nivel de soporte, lo que indica qué tan bien es probable que la institución anfitriona pueda preservar el contenido en el formato en el futuro. Hay tres niveles de soporte posibles que la institución anfitriona puede asignar a los formatos de flujo de bits. La institución de acogida debe determinar el significado exacto de cada nivel de apoyo, después de una cuidadosa consideración de los costos y requisitos.
Cada elemento tiene un registro de metadatos Dublin Core calificado. Otros metadatos pueden almacenarse en un elemento como un flujo de bits serializado, pero almacenamos Dublin Core para cada elemento para su interoperabilidad y facilidad de descubrimiento. Los usuarios finales pueden ingresar el Dublin Core cuando envían contenido, o puede derivarse de otros metadatos como parte de un proceso de ingesta.

Objeto	**Ejemplo**
Comunidad	Laboratorio de Ciencias de la Computación; Centro de Investigaciones Oceanográficas
Colección	Informes técnicos de LCS; Conjuntos de datos estadísticos ORC
Item	Un informe técnico; un conjunto de datos con una descripción adjunta; una grabación de video de una conferencia
Bundle	Un grupo de flujos de bits de imágenes y HTML que componen un documento HTML
Bitstream	Un solo archivo HTML; un solo archivo de imagen; un archivo de código fuente
Bitstream Format	Microsoft Word versión 6.0; Formato de imagen codificado JPEG

Los elementos se pueden eliminar de DSpace de una de estas dos formas: se pueden retirar", lo que significa que permanecen en el archivo pero están completamente ocultos a la vista. En este caso, si un usuario final intenta acceder al artículo retirado, se le presenta una "lápida"que indica que el artículo ha sido eliminado. Por el motivo que sea, un elemento también puede ser 'borrado' si es necesario, en cuyo caso todos los rastros del mismo se eliminan del archivo.

3.2.1. Comunidades

Una comunidad esta formada por un conjunto de colecciones. A su vez una comunidad corresponde a partes de la organización como departamentos, laboratorios, centros de investigación o escuelas. El diseño modular de DSpace le permite la creación de repositorios grandes e ínter-disciplinarios, que pueden ser usados por múltiples organizaciones. DSpace esta diseñado para preservar la información funcional y actualizada. Adaptándose a los formatos, medios y paradigmas que van evolucionando respecto al tiempo.

Capítulo 4

Referencias

[1] Página DSpace: link

[2] Especificaciones DSpace: link

[3] Diagrama DSpace: link

[4] Smith M., Barton M., Bass M., Branschofsky M., McClellan G., Stuve D., Tansley R. & Harford J.. (2003). DSpace An Open Source Dynamic Digital Repository. Septiembre 2021, de D-Lib Magazine Sitio web: link