

ACCESO ABIERTO
COMUNICACIÓN CIENTÍFICA
Y PRESERVACIÓN DIGITAL

Noviembre 13 -16 Barranquilla, COLOMBIA Universidad del Norte

Las actividades y el Planeamiento de la Preservación en un Repositorio Institucional

Marisa De Giusti, Ariel Lira, Néstor Oviedo, Gonzalo Villarreal, José Daniel Texier



La preservación digital

Conjunto de prácticas de naturaleza política, estratégica y acciones concretas destinadas a asegurar el acceso a largo plazo a los objetos digitales.

Primer paso hacia un plan de preservación para **SeDiCI**.

Índice

- Sección 1 Conceptos básicos
- Sección 2 Directrices sobre preservación
- Sección 3 Metadatos de preservación
- Sección 4 OAIS – Dspace
- Sección 5 Lo que viene después...

Índice

Sección 1 Conceptos básicos

Sección 2 Directrices sobre preservación

Sección 3 Metadatos de preservación

Sección 4 OAIS – Dspace

Sección 5 Lo que viene después...

El repositorio

Un Repositorio es una infraestructura web capaz de brindar un conjunto de servicios a una comunidad, destinados a recopilar, gestionar, difundir y preservar contenidos a través de una colección organizada y accesible en abierto que debe estar provista de facilidades que le permiten interoperar con otros repositorios similares.

El desarrollo de los servicios en un repositorio no es parejo: gestión y difusión avanzadas, recopilación compleja, preservación: reciente.

Interoperabilidad

Capacidad de un sistema de información para comunicarse con otros sistemas heterogéneos e intercambiar datos de manera transparente, con mínima pérdida de valor y de funcionalidad y a través de una interconexión automática y sin perder la interfaz propia.

Interoperabilidad

Ejemplo: DRIVER: intenta llegar a la interoperabilidad en dos capas:

- *La sintáctica* (y estructural): directrices para gestores y administradores de repositorios sobre cómo exponer los recursos científicos digitales utilizando los metadatos Dublin Core y el protocolo OAI-PMH y homogeneizando los outputs
- *La semántica* a través de la utilización de vocabularios que permitan un significado común.

Claves: estándares abiertos para la interacción y descripción de los recursos para asegurar su reutilización desde herramientas de búsqueda y acceso heterogéneas.

Índice

Sección 1 Conceptos básicos

Sección 2 Directrices sobre preservación

2.1 Modelo para la preservación

2.2. Aproximaciones a la preservación

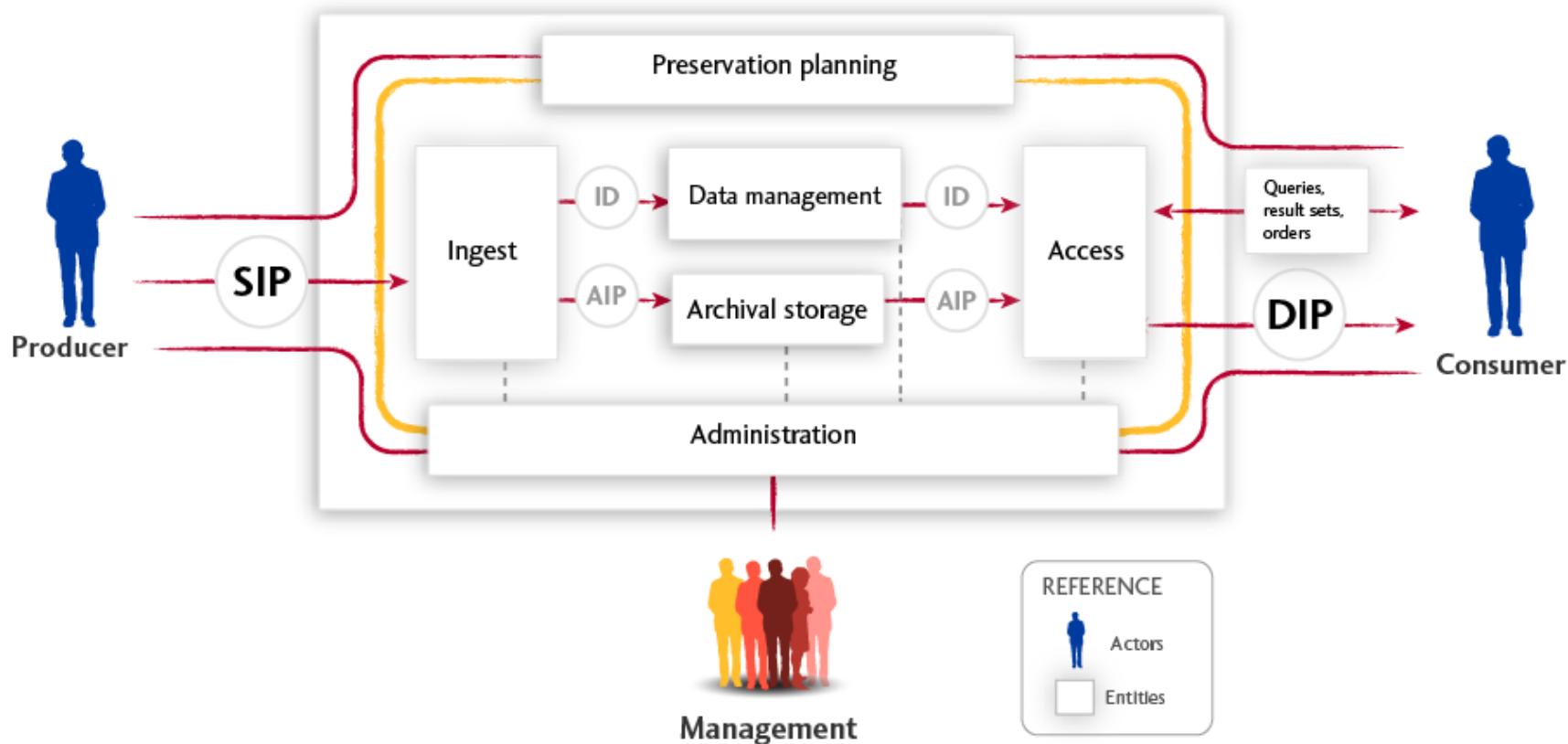
Sección 3 Metadatos de preservación

Sección 4 OAIS – Dspace

Sección 5 Lo que viene después...

Directrices de la Norma ISO 14721. Última versión Junio de 2012.

ISO Reference Model for an Open Archival information System (OAIS)



Tipos de paquetes de información

- **Archival Information Package (AIP):** contiene, como mínimo, suficiente información de un objeto como para garantizar la preservación a largo plazo. Busca mantener la mayor calidad posible de información descriptiva de preservación y de representación de los objetos.
- **Submission Information Package (SIP):** es el paquete que proviene del productor y se va a incorporar al OAIS. Suele contener menos información que el AIP.
- **Dissemination Information Package (DIP):** es el paquete que se entrega a un Consumidor en respuesta a una solicitud.

Paquete de Información: contenedor conceptual

Información de contenido (CI)

Objeto de contenido (CDO)
Uno o más recursos físicos o digitales

Información de
representación (RI)
Sobre los objetos de contenido

Información descriptiva de preservación (PDI)

Information de procedencia

Información de contexto

Información de referencia

Información de integridad

Información de empaquetado (PI)

Establece una relación, y, en ciertos casos une las partes del IP

Información descriptiva (DI)

Metadatos adicionales sobre el objeto de contenido y el paquete para facilitar el acceso.

El primer paso: qué preservar

La “Guía Unesco: *Directrices para la preservación del Patrimonio Digital*” [1], sostiene en su artículo 5.2.1 sobre Patrimonio que **“El patrimonio digital está constituido únicamente por aquellos que se considera que poseen un valor permanente”**. Este es el primer punto de un plan de preservación: definir materiales y tiempos de preservación de cada uno.

[1] UNESCO “Directrices para la preservación del Patrimonio digital | Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura,” 2012.

No sólo lo técnico

Tránsito Ferreras Fernández [2] : “los esfuerzos para preservar los objetos digitales no se acaban con lo técnico, sino que existen otros retos: legales (permiso de los autores), económicos e institucionales (compromiso permanente) para llegar al resultado de perdurabilidad”.

[2] T. Ferreras Fernández, “Preservación digital en repositorios institucionales: GREDOS,” 2010. [Online].

Preservación: motivos y tendencias

- Miles de objetos digitales sometidos a la obsolescencia tecnológica: fragilidad de los objetos digitales y obsolescencia de los medios informáticos.
- La necesidad de metadatos apropiados para mantener el acceso a ese patrimonio.

Tendencias:

- ➡ Digital curation (ciclo de vida)
- Digital preservation (a largo plazo)
- Web archiving...

Algunos proyectos

Repository Preservation and Interoperability (Preserv2):

The **technical registry**
PRONOM

Droid

DROID (Digital Record and Object
Identification)

JISC KeepIT

Roles de los metadatos

- **Identificación y descripción de los recursos.**
- **Búsqueda, recuperación, ubicación.**
- Descripción de derechos patrimoniales y de uso.
- Preservación.
- Powder.
- WAI –accesibilidad de los contenidos.
- Modalidades de acceso...

Índice

Sección 1 Conceptos básicos

Sección 2 Directrices sobre preservación

Sección 3 Metadatos de preservación

3.1 Generalidades

3.2 PREMIS

3.2.1 Modelo de datos PREMIS

3.3 METS

3.4 Otros esquema de metadatos y más
posibilidades en la preservación

Sección 4 OAIS – Dspace

Sección 5 Lo que viene después...

Metadatos y metadatos de preservación

Metadatos tradicionales

- 
- ID
 - Título
 - Autor
 - Idioma
 - Fecha
 - etc

documento

Metadatos de preservación

- Procedencia,
- Autenticidad
- Contexto
- Gestión de derechos
- Actividades de preservación

Metadatos y metadatos de preservación

- **Metadatos tradicionales:** describen, explican, localizan, facilitan la localización y gestionan un recurso, pueden ser descriptivos, administrativos o estructurales (hay modos diferentes de clasificarlos).
- **Metadatos de preservación:** documentan el proceso de preservación de los recursos digitales, también son descriptivos, administrativos y estructurales (si se sigue la misma taxonomía previa)

OD y metadatos de preservación

1. Debe mantenerse en el repositorio de manera **segura**.
2. Se debe conocer su **Creador**
3. Si hay un cambio, debe saberse **quién** lo efectuó
4. Debe poder localizarse y entregarse al usuario.
5. Su soporte debe ser **compatible** con los sistemas actuales
6. Las estrategias de **emulación y migración** requieren datos sobre los objetos originales y sus entornos
7. Deben soportar la **Autenticidad** mediante la documentación de su procedencia
8. El repositorio debe tener los derechos suficientes para sostener el **acceso** al objeto
9. Deben guardarse las relaciones que vinculen al documentos con otros, en caso de existir



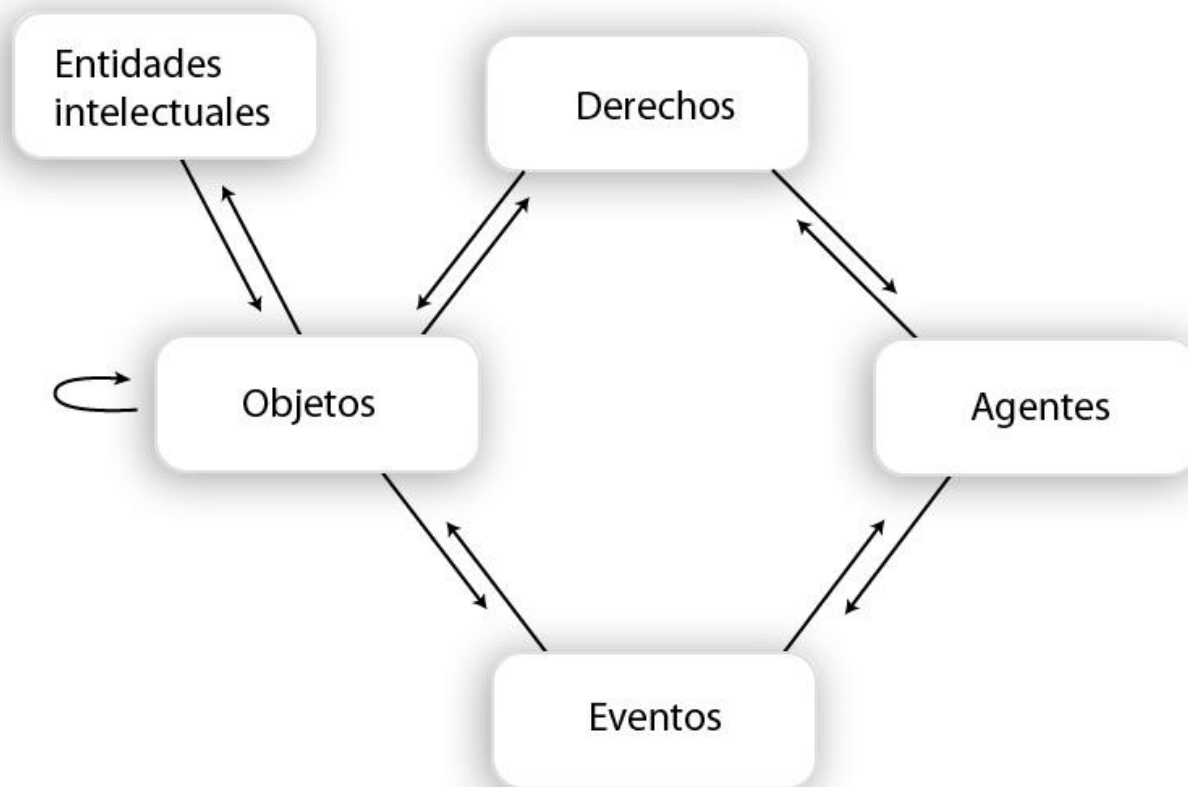
Metadatos de preservación

En resumen los metadatos de preservación están destinados a almacenar los detalles técnicos sobre el formato, la estructura, el acceso y el uso de los contenidos digitales, la historia de todas las acciones realizadas en el recurso, incluyendo los cambios, la información de autenticidad, las características técnicas o la historia de la custodia y las responsabilidades y la información sobre los derechos con que se cuenta para realizar las acciones de preservación.

PREMIS

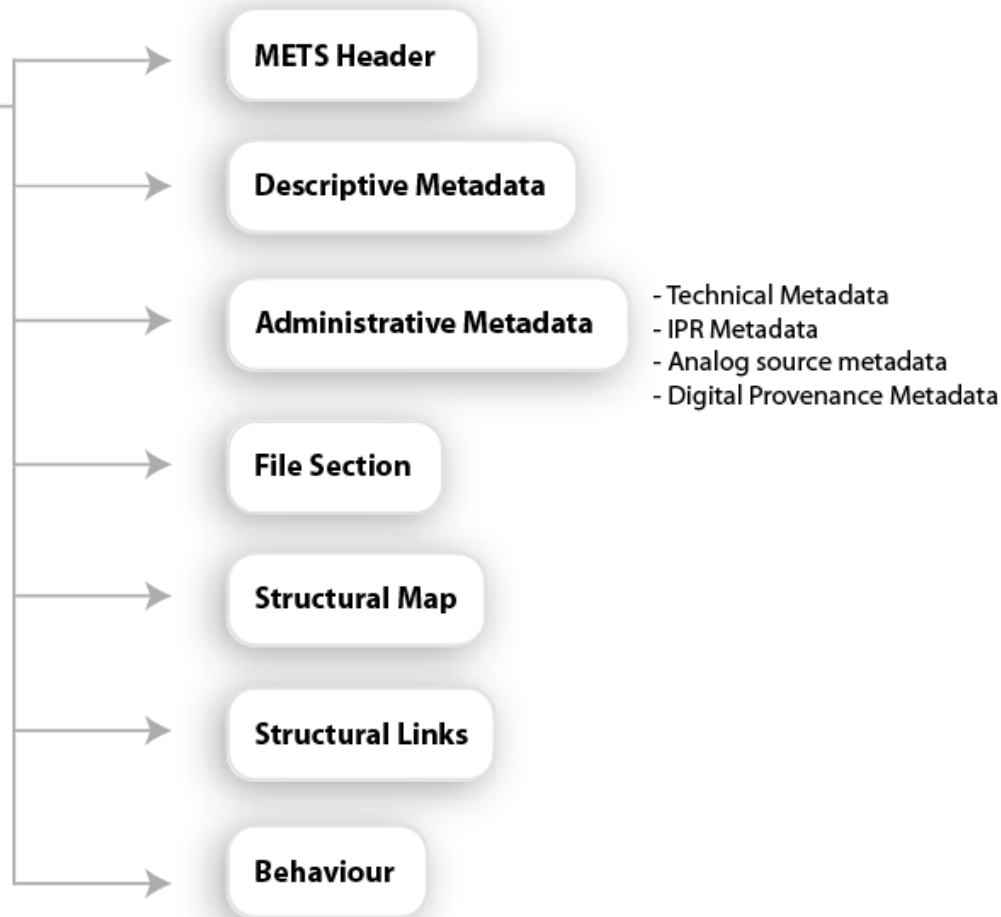
PREMIS es un grupo de trabajo internacional patrocinado por *Online Computer Library Center* (OCLC) y *Research Libraries Group* (RLG) que, como su nombre lo indica, se enfoca en estrategias de implementación de metadatos de preservación en Archivos Digitales. En 2008, este grupo elaboró el Diccionario de Datos PREMIS para Metadatos de Preservación el cual define los metadatos de preservación como “la información que utiliza un repositorio para dar soporte al proceso de preservación digital”. El diccionario define un conjunto de *unidades semánticas, propiedades, e información que la mayoría de los repositorios necesita conocer de sus entidades para asegurar la preservación.*

Modelo de datos PREMIS



Mets

METS



Otros esquemas y sus posibilidades

Si bien el par **METS-PREMIS** es muy adecuado a los fines de la preservación, los esquemas más utilizados de metadatos cuentan con metadatos útiles para ese cometido.

Tratar de detectar en el esquema usado metadatos útiles para proveer datos técnicos sobre el formato, la estructura y el uso de los contenidos digitales, la historia de todas las acciones realizadas en los recursos, los cambios, la autenticidad, las responsabilidades y permisos.

Otros esquemas y sus posibilidades

Si surgen ambigüedades (algunas ampliamente referenciadas), una práctica más elaborada a la hora de incorporar los metadatos, p.e. agregado de subelementos en DC ayuda a la desambiguación. Los esquemas que no están dedicados específicamente a la preservación tienen limitaciones, pero son un comienzo.

DC-PREMIS

DUBLIN CORE	PREMIS DD	DEFINITION
Identifier	Object Entity (1.1 objectIdentifier)	Unique reference to the resource in a given context.
Creator	Agent Entity	Responsible for creating the resource.
Description	Object Entity (1.10 relationship)	Summary of the resource.
Date	Object Entity (1.5.5 creatingApplication)	Time associated to an event in the resource's life cycle.
Format	Object Entity (1.5.4 format)	Physical or digital presentation of the resource.
Relation	Object Entity (1.10 relationship)	A reference to a related resource.
Rights	Right Entity (4.1 RightsStatements)	Information on the legal rights that affect the use of the resource.

MODS-PREMIS

MODS CORE Element	PREMIS Element	DEFINITION
Identifier	Object Entity (1.1 objectIdentifier)	Identifies the resource or work content.
name	Agent Entity	Associated in some way to the resource.
internetMediaType	Object Entity (1.5.4 format)	Physical or digital presentation of the resource.
relatedItem	Object Entity (1.10 relationship)	Summary of resource content.
location	Object Entity (1.7.1 contenLocation)	Identifies the institution or repository where the resource is or a remote location through a URL.
accessCondition	Right Entity (4.1 RightsStatements)	Information about restrictions imposed on access to a resource.

Índice

Sección 1 Conceptos básicos

Sección 2 Directrices sobre preservación

Sección 3 Metadatos de preservación

Sección 4 OAIS – Dspace

4.1 Descripción general de DSPACE

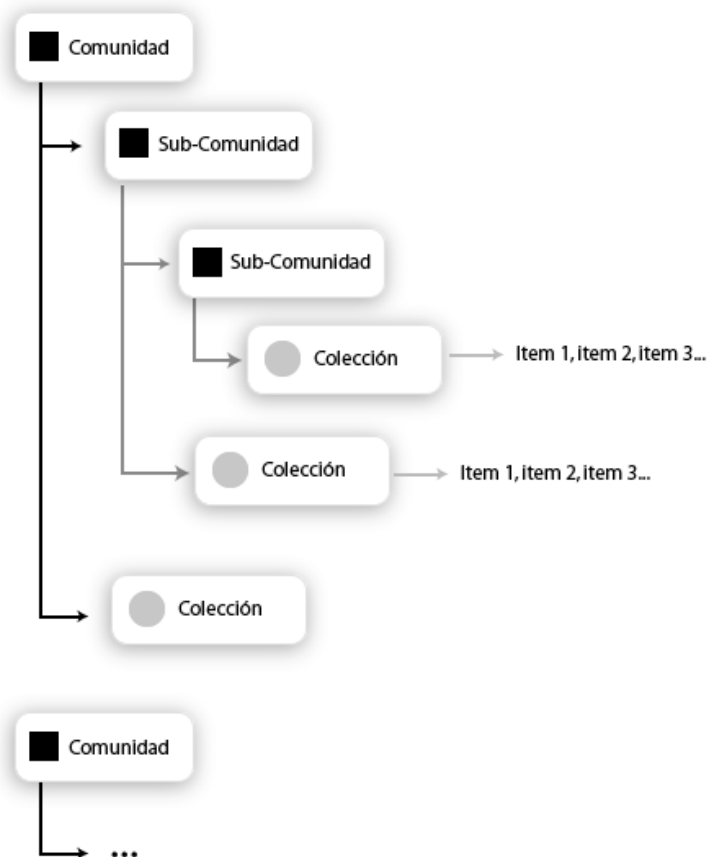
4.2 OAIS en DSpace

4.3 Tabla cruzada

Sección 5 Lo que viene después...

Jerarquía de comunidades y colecciones en DSPACE

Las comunidades y colecciones sirven a los fines administrativos para organizar los ítems dentro del repositorio, aunque en muchos repositorios se los considera como metadatos, dado que representan información estructural sobre los ítems.



DSpace: 3 grupos lógicos de metadatos para el contenido archivado

1) Metadatos descriptivos: soporta múltiples formatos de metadatos no jerárquicos y permite mantener para un mismo ítem, metadatos de diversos formatos, habilitando de esta forma la utilización de uno o más Application Profiles. Para colecciones y comunidades, sólo permite gestionar una serie de campos fijos y básicos aunque probablemente en las próximas versiones de Dspace se permita configurar los metadatos como sucede con los items.

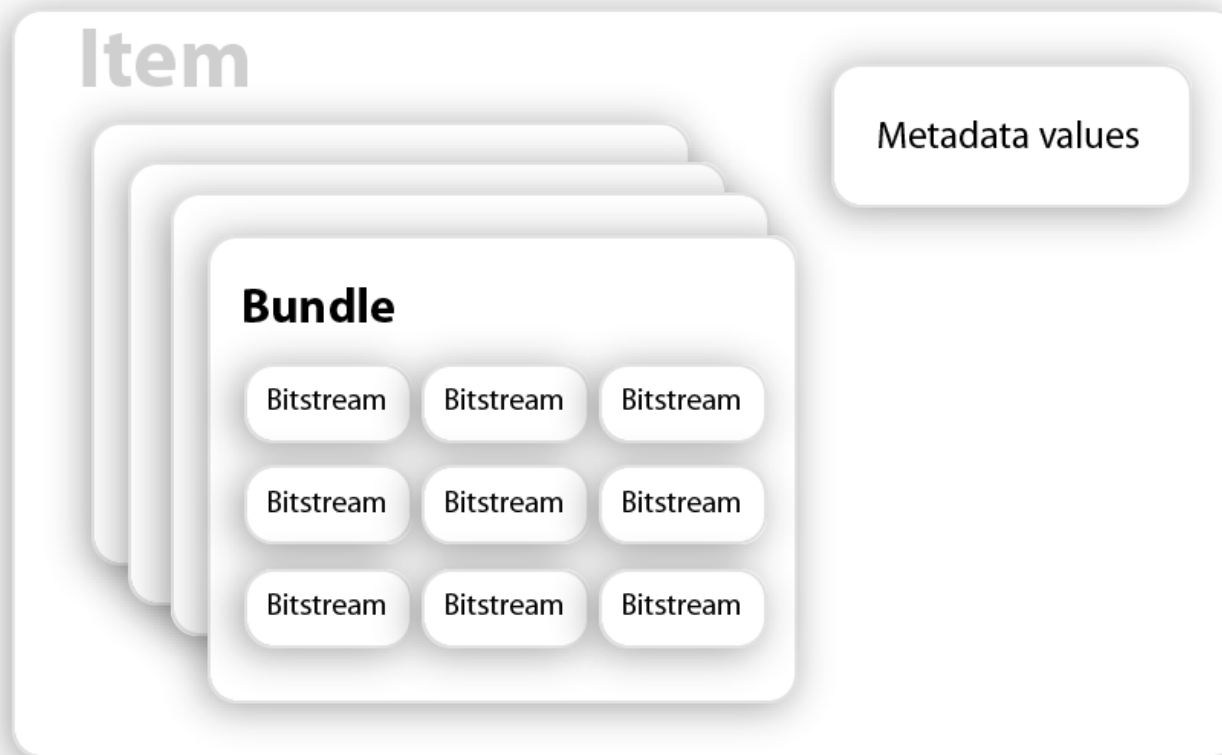
DSpace: 3 grupos lógicos de metadatos para el contenido archivado

2) Metadatos Administrativos: información de procedencia y permisos de acceso y uso sobre cada ítem, colección y comunidad. Se registran campos adicionales sobre varios elementos: tamaño, suma de comprobación y tipo Mime de cada archivo.

DSpace: 3 grupos lógicos de metadatos para el contenido archivado

3) Metadatos estructurales: mantiene sólo unos pocos metadatos estructurales sobre los archivos de cada ítem p.e. la relación de pertenencia entre paquetes de archivos (bundles) y orden. Dependiendo cómo se armen las comunidades y colecciones, pueden considerarse las relaciones de pertenencia a colección como un metadato estructural. Para los ítems, es posible almacenar información estructural en los bitstreams o metadatos, aunque inicialmente no se provee nada.

Modelo de datos en DSPACE



OAIS en DSPACE

DSpace está influenciado por el modelo **OAIS**, en términos de metodología y funciones. Siempre que es posible, adopta el modelo y vocabulario OAIS para articular su propio diseño y metodología. El modelo de datos, los metadatos y el módulo de autorizaciones cumplen con el modelo de referencia. En términos de la preservación digital, el sistema proporciona los metadatos de preservación como en el modelo OAIS y el identificador persistente Handle (<http://www.handle.net/>). Sin embargo, la arquitectura de Dspace no coincide con el modelo funcional de OAIS

Ingest

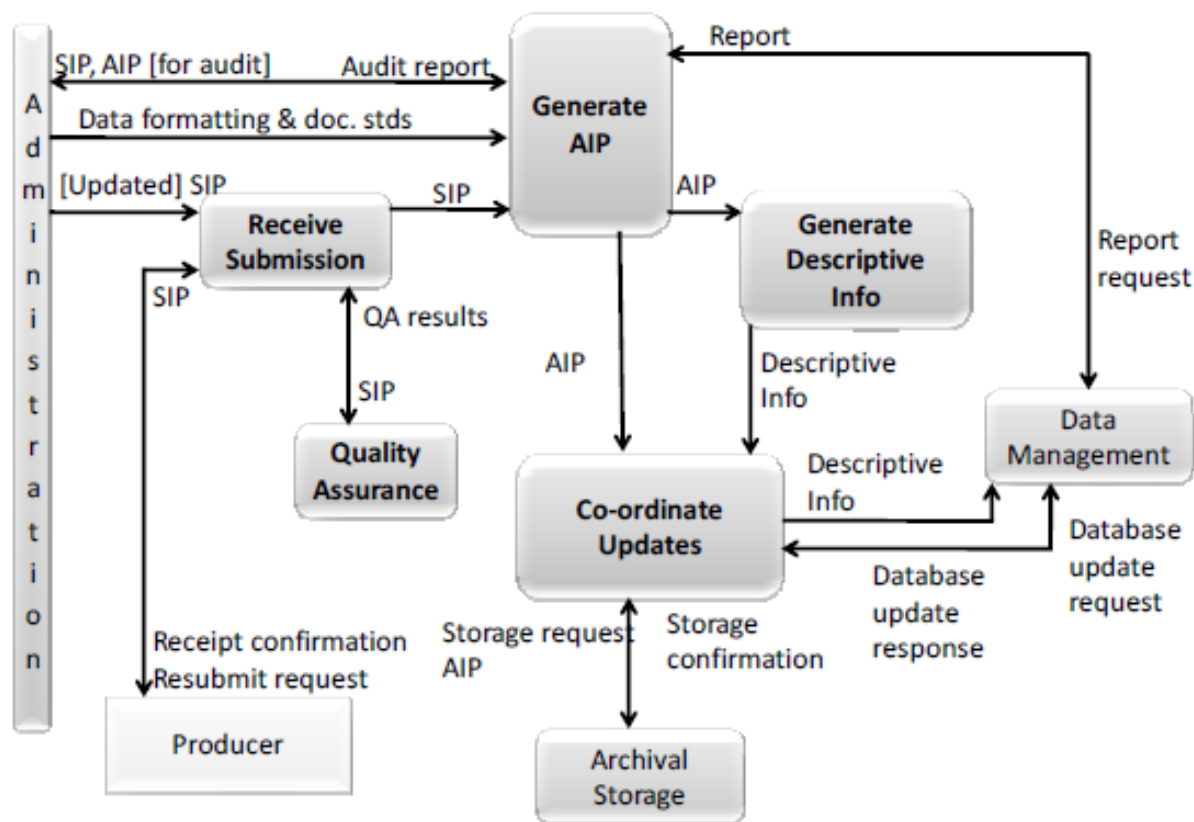


Figure 4-2: Functions of the Ingest Functional Entity

Ingest

Descripción: Provee los servicios y funciones para aceptar un SIP por parte de los Productores o bajo el control de la Administración. Prepara los contenidos para almacenamiento y gestión dentro del archivo. Realiza el aseguramiento de calidad/validación de los SIPs. Genera el AIP que cumple con los estándares de formato de datos y documentos. Extrae la información descriptiva y la envía al *data management*. Coordina las actualizaciones en el *archival storage* y en el *data management* de la base de datos.

Ingest

En DSpace: Se tiene un proceso claramente definido como *Ingest*, que esta compuesto por varios subprocessos. El mismo puede iniciarse por aportes externos al sistema o a través de la interfaz de usuario (vía web) del sistema. Se inicia un elemento *InProgressSubmission* que, dependiendo de la configuración, pasa por varios estadíos durante los cuales se le agregan metadatos descriptivos y administrativos, y se realizan *checksums* a los contenidos. El resultado de esta secuencia es pasado al *item installer*, el cual debe convertir ese objeto en un *Archived Item* de DSpace. El *item installer* se encarga de asignar una fecha de adhesión, fecha de disponibilidad, de ser necesario fecha de publicación, procedencia, un identificador persistente, colección a la que pertenece, y crea los índices para su búsqueda y navegación.

Archival Storage

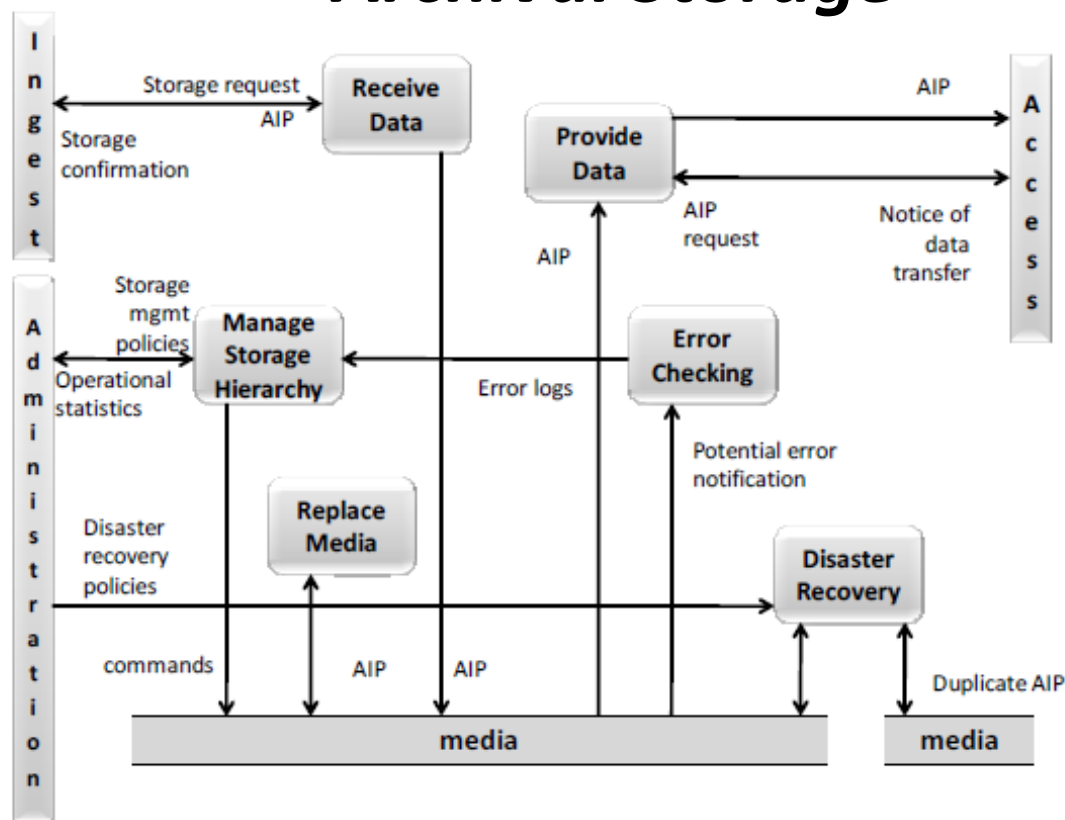


Figure 4-3: Functions of the Archival Storage Functional Entity

Archival Storage

Descripción: Provee los servicios y funciones para el almacenamiento, mantenimiento y recuperación de los AIPs.

Recibe el AIP de la entidad ingest y lo almacena. Gestiona las jerarquías de almacenamiento. Configura niveles especiales de servicio, seguridad y protección (por ejemplo backups). Provee estadísticas de inventario, capacidad disponible, etc. Transforma los datos que constituyen la información de empaquetado para reproducir el AIP en el tiempo. Realiza una verificación de errores. Provee un mecanismo estándar para el seguimiento y verificación de la validez de los datos. Provee un mecanismo de duplicación de los contenidos en una lugar físico separado. Provee copia de los AIPs almacenados a la entidad *access*.

Archival Storage

En DSpace: provee dos tipos de almacén de datos para los bitstreams, uno dentro del sistema de archivos del servidor donde se aloja la aplicación y otro fuera, basado en SRB (Storage Resource Broker) que soporta el almacenamiento en forma distribuida y compartida. El resto de los datos del AIP son almacenados dentro de la Base de Datos relacional. Dspace ofrece la posibilidad de definir y ejecutar tareas de *curation* para realizar procesos desatendidos sobre los objetos del repositorio. Esto permite, por ejemplo, realizar validaciones periódicas, generar reportes, verificar los datos y los archivos e incluso realizar transformaciones.

Archival Storage

Para la verificación de los objetos almacenados en el sistema de archivos, provee una herramienta denominada checksum checker que compara la información de fijeza almacenada para el archivo con el el archivo propiamente. Permite efectuar tareas de duplicación sobre los ítems, colecciones y comunidades, generando un paquete AIP totalmente autocontenido. A continuación, los paquetes generados pueden ser restaurados en el almacén y usarse como reemplazo de los existentes.

Data management

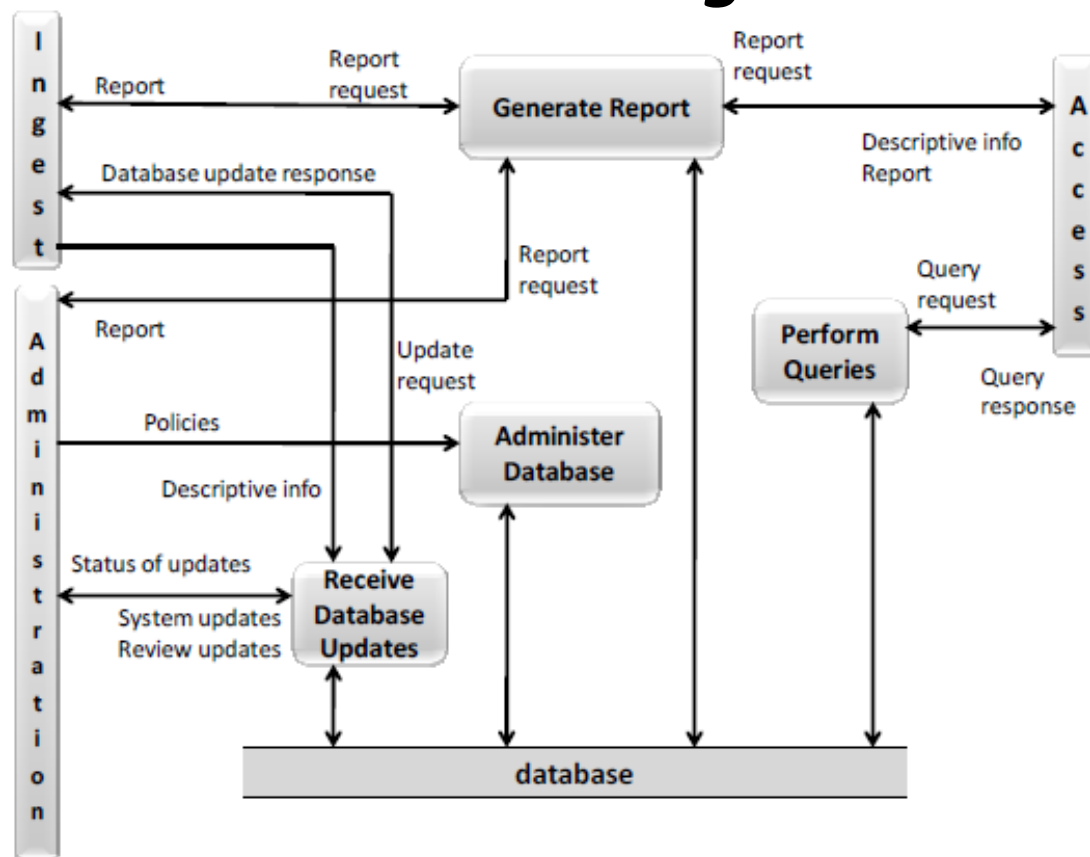


Figure 4-4: Functions of the Data Management Functional Entity

Data management

Descripción: Provee los servicios y funciones para poblar, mantener y acceder a la información descriptiva que identifica y documenta el contenido del Archivo, y a los datos administrativos usados para gestionarlo. Es responsable de la administración de la base de datos. Recibe solicitudes de la entidad *access* y genera un conjunto de resultados.

Recibe pedidos de las entidades *ingest*, *access* y *administration* y genera reportes.

También recibe actualizaciones de *ingest* y *administration*.

Data management

En DSpace: las funciones necesarias para almacenar y recuperar tanto la información descriptiva de los AIP como la información de soporte para las tareas administrativas (usuarios, permisos, etc), son provistas por los componentes asociados a la gestión de la base de datos.

Asimismo, cuando se utiliza el módulo *Discovery* como proveedor del servicio de búsquedas para la interfaz web, se agrega un indexador de texto (denominado Apache Solr) como parte de esta entidad funcional, aportando un modo alternativo y eficiente de acceso a los metadatos descriptivos.

Administration

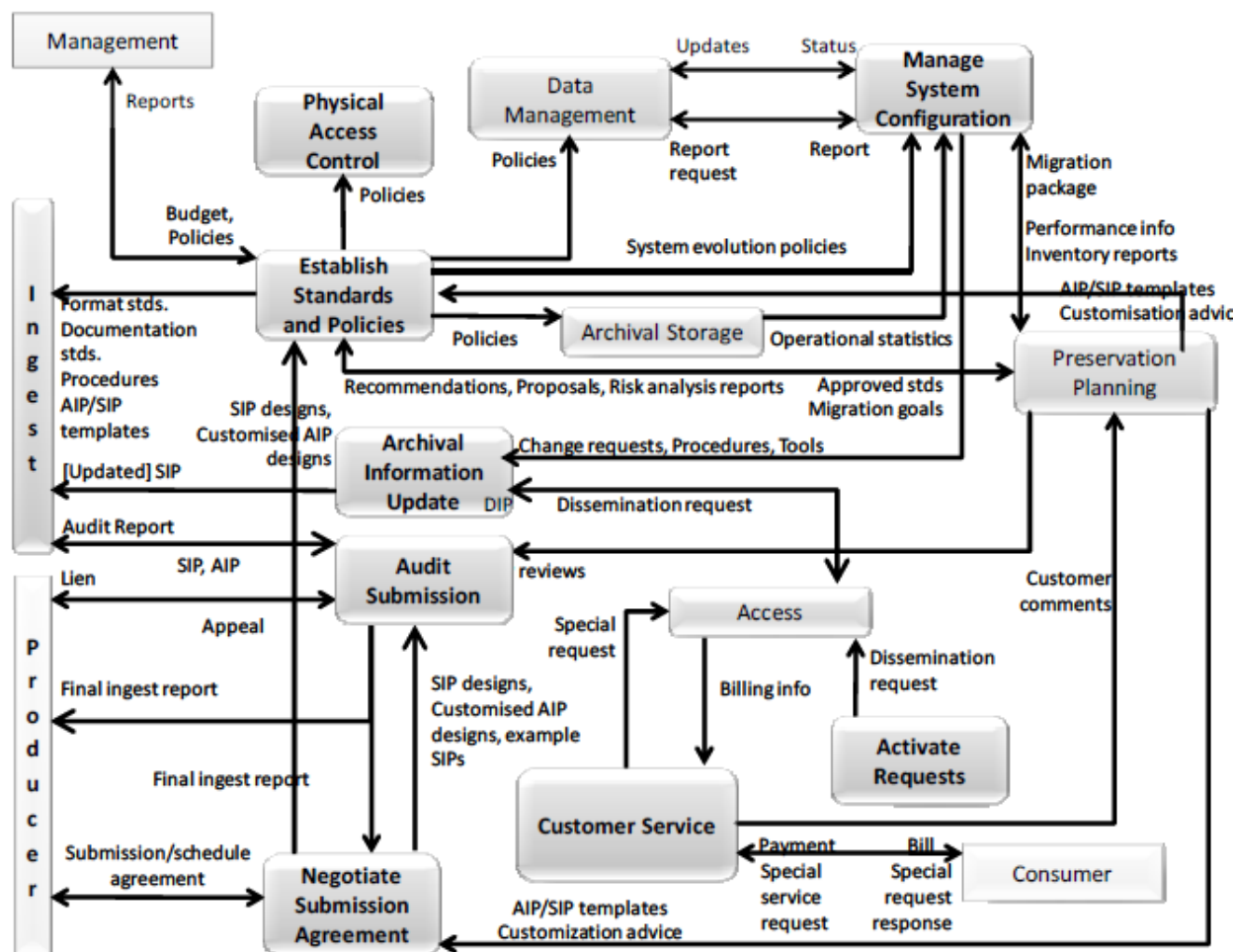


Figure 4-5: Functions of the Administration Functional Entity

Administration

Descripción: Provee los servicios y funciones para la operación global del sistema de archivos. Solicita la información necesaria sobre los archivos y negocia los acuerdos con los Productores. Monitorea la funcionalidad del sistema de archivos, controla los cambios de la configuración y mantiene su integridad y trazabilidad. Audita las operaciones del sistema, performance y uso. Envía reportes al *data management* y recibe reportes de esa entidad. Sumariza todos los reportes y provee información sobre performance del OAIS e inventario y envía esta info a *preservation planning* para establecer políticas y estándares. Recibe los paquetes de migración para *preservation planning*.

Administration

Recibe los pedidos de cambio, procedimientos y herramientas para la actualización del archivo.

Responsable de enviar un pedido de disseminación a *access*, actualizando los contenidos de los DIP y resuministrando los SIP a *ingest*.

Provee mecanismos para restringir/permitir acceso a los elementos del archivo.

Es responsable de enviar información para establecer estándares y políticas.

Administration

Desarrolla políticas de gestión de archivo por jerarquías, incluyendo políticas de migración. Es responsable de la recuperación ante desastres.

Verifica que los AIP y SIP suministrados sigan las especificaciones.

Verifica el PDI según los usos de la comunidad designada.

Revisar periódicamente los contenidos del archivo para determinar si los datos están disponibles.

Crea/mantiene/borra las cuentas de acceso de los consumidores.

Administration

En DSpace:

La incorporación de los SIP está sujeta al cumplimiento de sus partes con las restricciones definidas en la configuración; es posible marcar metadatos requeridos y para algunos tipos de datos, el dominio válido de un metadato. Las autorizaciones de cada usuario dentro del sistema pueden configurarse con un gran nivel de granularidad, permitiendo definir permisos de lectura, adición, revisión, modificación y borrado a nivel de bitstream, item, colección y comunidad.

Administration

Por ello, los usuarios encargados de comunidades pueden mantener de control sobre su contenido en el OAI-S, aliviando la carga administrativa central. Según la tarea de administración a realizar, será preciso utilizar la aplicación Web o ejecutar un comando directamente desde la consola del servidor. Prácticamente todas las tareas administrativas se realizan vía Web, incluyendo la gestión de items con sus archivos y metadatos, la gestión de usuarios y grupos, comunidades y colecciones, y control de ítems dentro de los workflows de carga. Solo algunas tareas deben ejecutarse desde consola, concretamente las creación de administradores, importación de estructura de colecciones y comunidades, eliminación de embargos.

Administration

El módulo de estadísticas provee información de uso y acceso sobre las comunidades, colecciones, items y bitstreams por parte de los visitantes, junto a algunos eventos notables como ser el acceso de usuarios registrados.

Preservation planning

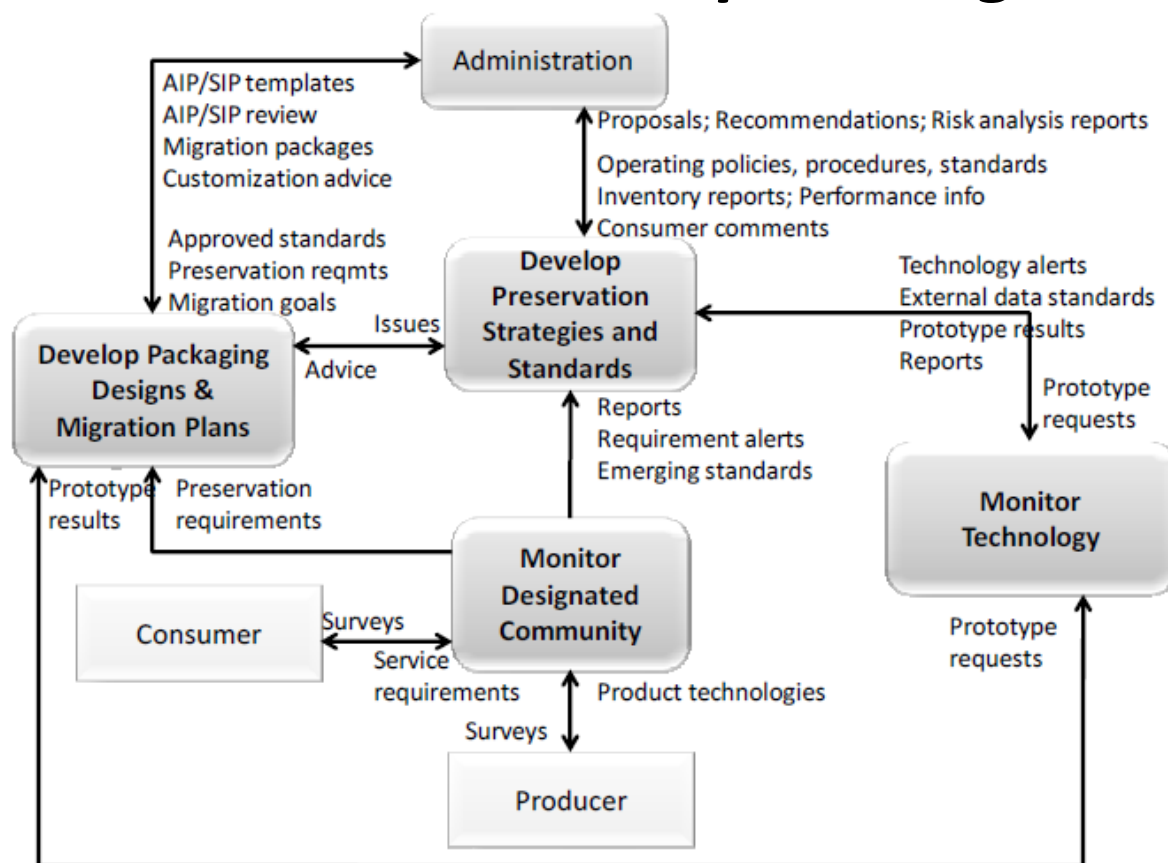


Figure 4-6: Functions of the Preservation Planning Functional Entity

Preservation planning

Descripción: Interactúa con los consumidores y productores de archivos. Proporciona reportes, alertas de requisitos y estándares independientes. Identifica tecnologías que pueden causar obsolescencia. Desarrolla y recomienda estrategias y estándares, que envía a *administration*.

Desarrolla nuevos IP y planes de migración y prototipos, para implementar políticas y directivas de administración de IPs.

Preservation planning

En DSpace: no posee una correspondencia en absoluto con esta entidad funcional. Se debe principalmente a que esta entidad posee un componente evolutivo fundamental. Sin embargo, Dspace provee facilidades para afrontar una gran variedad de cambios que se pueden llevar adelante , por ejemplo: aplicar transformaciones de formatos en forma semi-automática, definir procesos de diagnóstico de los AIP, aplicar cambios en lote sobre metadatos, entre otros.

Access

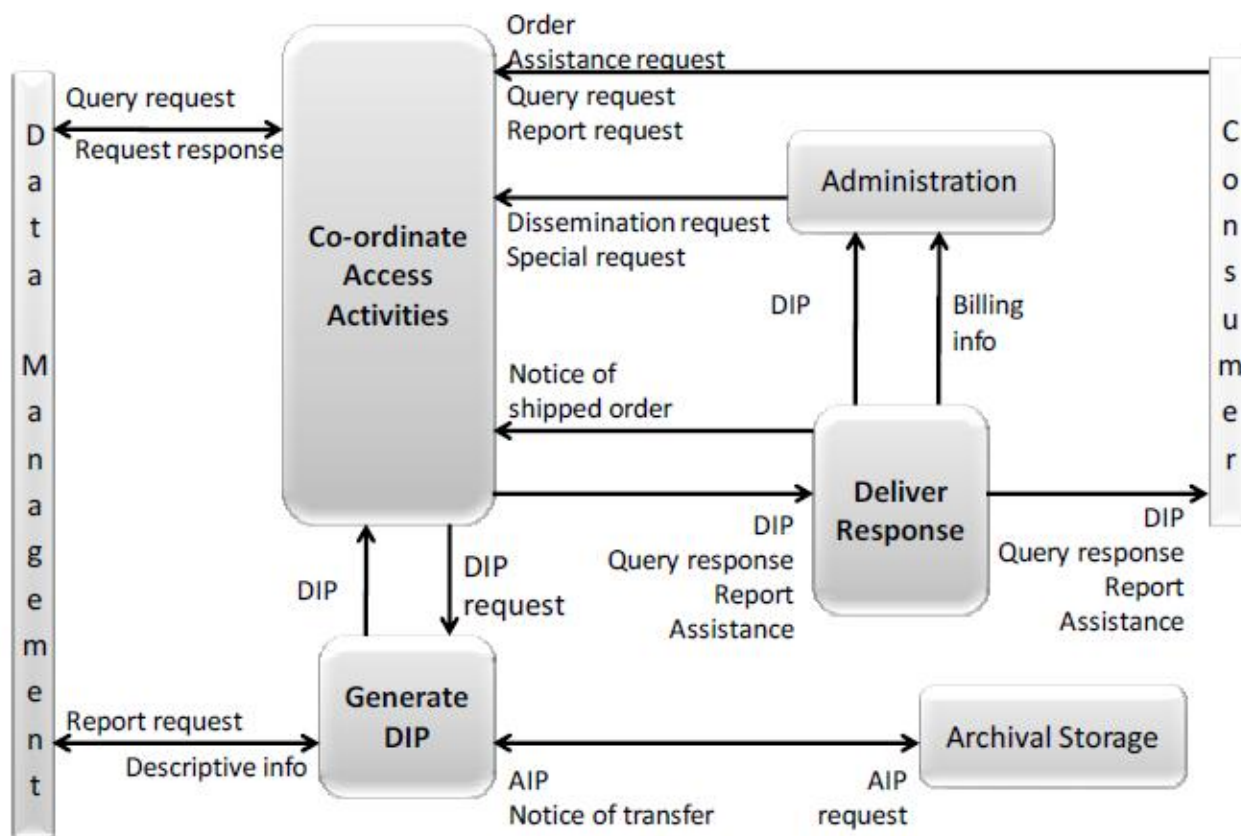


Figure 4-7: Functions of the Access Functional Entity

Access

Descripción: Proporciona una interfaz única de usuario para el acceso a la información de los archivos.

Tiene 3 categorías, los *query requests*, los *result sets* y los *report requests*.

Acepta los requerimientos de los paquetes de diseminación recuperados de los AIP de la entidad *archival storage* y transmite un *report request* al *Data Management* generando un DIP.

Entrega las respuestas en línea y fuera de línea de los consumidores

Access

En DSpace: Se proveen de aplicaciones web para búsqueda y navegación: JSPUI o XMLUI. En ambos casos se disponen de mecanismos de exploración basados en la base de datos interna que permiten navegar el árbol de comunidades, colecciones e ítems del repositorio. En XMLUI se provee una herramienta de búsqueda más completa y eficiente basada en Apache Solr, denominada Discovery [28]. También permite habilitar algunos servicios adicionales como ser OpenSearch, suscripciones a colecciones. Se provee un módulo que cumple con el rol de Data Provider del protocolo OAI-PMH que soporta mapeos configurables hacia cualquier formato de metadatos , haciendo uso de los CrosswalkPlugins.

Access

Para la generación de DIP, se proveen varias alternativas de exportación:

- archivo separado por comas CSV.
- exportación de AIP de comunidades, colecciones e items.
- exportación de items en un formato simple denominado Simple Archive Format
- Mets Dip a partir del packager exporter, que incluye algunos metadatos según el diccionario de datos PREMIS, aunque no todos dado que solo se implementa una porción del modelo de datos PREMIS.

Access

Adicionalmente, es posible extender los mecanismo de exportación e incorporar nuevos ExporterPlugins para generar DIP en otros formatos. Visualización de información estadística generada por el módulo correspondiente.

Índice

Sección 1 Conceptos básicos

Sección 2 Directrices sobre preservación

Sección 3 Metadatos de preservación

Sección 4 OAIS – Dspace

Sección 5 Lo que viene después...

Conclusiones y trabajos futuros

Un trabajo a futuro que resulta de interés para la comunidad de repositorios es relevar el estado, en cuanto a la preservación, de una buena parte de los repositorios más importantes que figuran en los registros internacionales. Finalmente el gran trabajo pendiente del repositorio **SeDiCI-UNLP** es el avance en el plan de preservación del repositorio institucional, el cual está apenas comenzando a hacerse realidad.

¡Muchas gracias!

Prof. Ing. Marisa R. De Giusti

marisa.degiusti@sedici.unlp.edu.ar

