Taller sobre prácticas aplicadas a la preservación digital en un repositorio institucional

Las siguientes notas forman parte de un taller realizado el 11 de julio de 2022 en el marco del Evento conmemorativo de los 10 años de la Red Brasileña de Servicios de Preservación Digital - CARINIANA, Brasilia.

OAIS. ISO 14721.

INTRO

Un OAIS (Open Archival Information System) se compone de una organización de personas y sistemas que tienen como objetivo asegurar la preservación y la accesibilidad a largo plazo a la información para una comunidad designada. El modelo de referencia que indica cómo debe estar compuesto un sistema OAIS está definido por la norma ISO 14721 y está considerado como el estándar para crear y mantener en el largo plazo un repositorio digital/un archivo en términos de preservación digital. El modelo define una serie de entidades involucradas, un modelo de información y un modelo funcional. El entorno OAIS requiere de 4 actores involucrados: los productores de la información, los consumidores de la información, los encargados de la gestión (data management) y el propio sistema. El modelo de información está compuesto por objetos de datos, los cuáles son ítems que pueden ser tanto físicos como digitales y están representados por paquetes de información (IP) que cambian desde su ingreso al repositorio, con los procesos internos (para asegurar la preservación) y para entregarlos a un usuario. El modelo funcional está compuesto por **6 entidades funcionales**: Ingesta, Almacenamiento, Gestión de datos, Administración, Planificación de la preservación y Acceso, las cuales pueden verse en la figura 4-1.

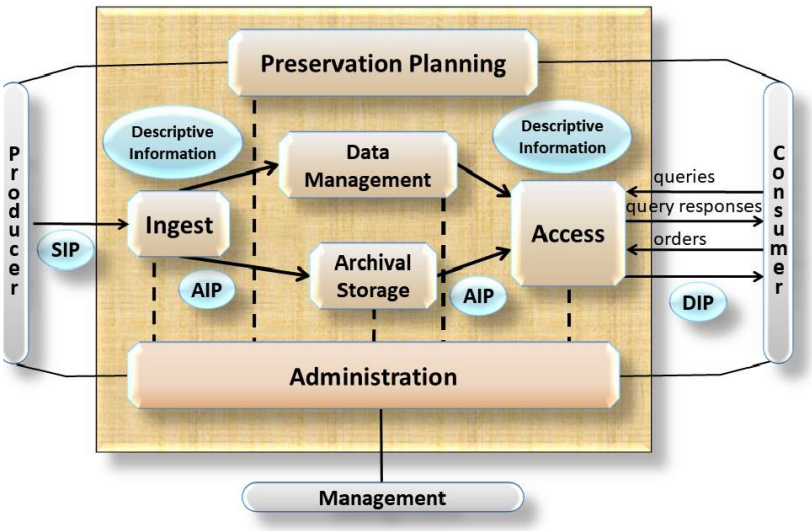


Figure 4-1: OAIS Functional Entities[1]

[1] Imagen extraída de: *Management Council of the Consultative Committee for Space Data Systems (CCSDS). (2019). OAIS final v3 draft with changes wrt OAISv2 20190924-rl.docx.* *https://cwe.ccsds.org/moims/_layouts/15/WopiFrame.aspx?sourcedoc={61C755A7-2C54-4D0D-A8F0-7B6A4228D74C}&file=OAIS%20final%20v3%20draft%20with%20changes%20wrt%20OAISv2%2020190924-rl.docx&action=default .* P 4-1.

ENTIDADES

INGEST

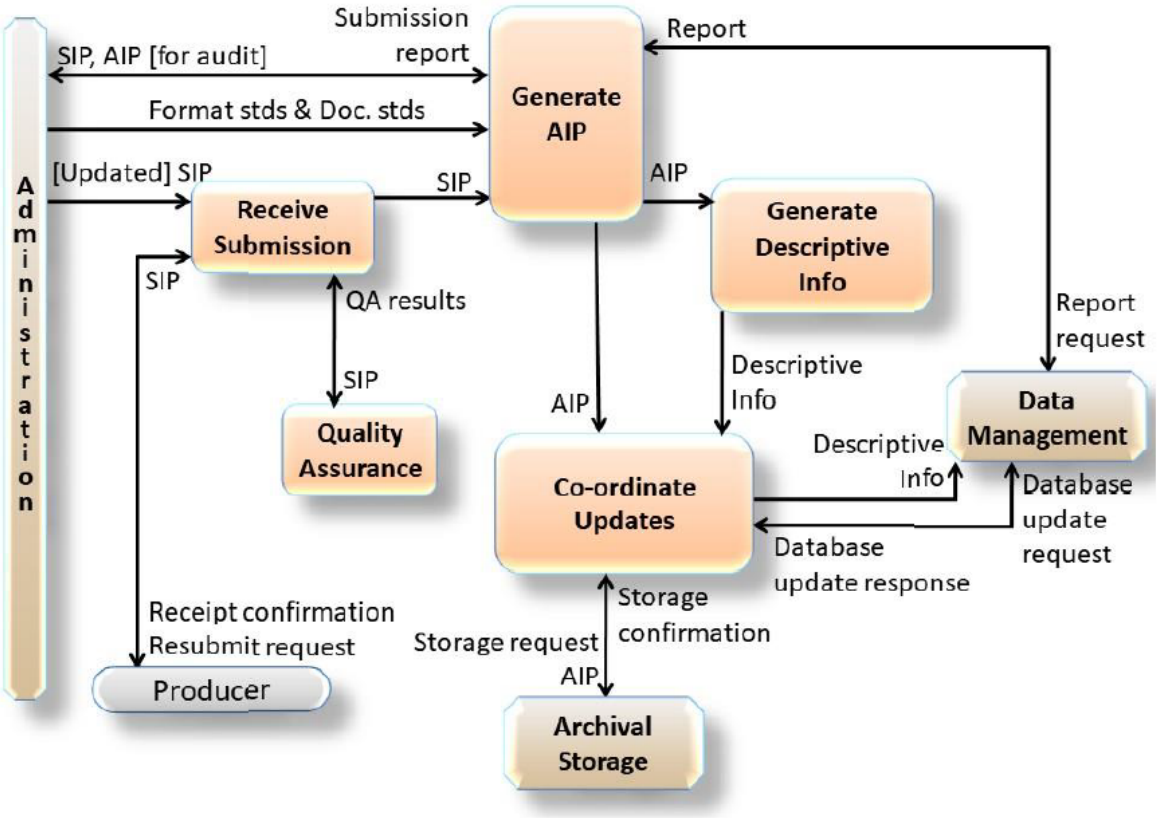


Figure 4-2: Functions of the Ingest Functional Entity [2]

- Provee los servicios y funciones para aceptar un SIP por parte de los Productores o bajo el control de la Administración: puede ser una transferencia de custodia (e incluir la licencia adecuada) e incluso puede requerir del agregado de permisos especiales para el acceso.
- 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.
- La evidencia de autenticidad puede tomar muchas formas. La evidencia está diseñada para respaldar la afirmación de que el objeto es lo que se supone que es. La evidencia inicial es proporcionada por el Productor como parte de la PDI en la presentación/envío y esta evidencia es mantenida, actualizada y/o incrementada por el Archivo a lo largo del tiempo. Con el tiempo, es posible que sea necesario cambiar el objeto de alguna manera para que siga siendo comprensible de forma independiente para la comunidad designada. Es importante que estos cambios se documenten como parte de la Información de procedencia

[2] Op. cit. p. 4-5.

(provenance information) del objeto para que el objeto pueda rastrearse hasta el objeto original enviado al Archivo por el Productor. También es importante que cualquier cambio en el objeto no cambie la información del contenido hasta el punto de que ya no transmita la información prevista del objeto original. Un método para proporcionar evidencia que respalde la afirmación de la autenticidad del objeto modificado es el uso de descripciones de propiedades de la información.

- La norma habla de que el productor o la administración pueden proveer información descriptiva de la información y pone de ejemplo un libro indicando que esta información incluiría cosas tales como título, márgenes, palabras y puntuaciones.

Sin embargo si se observa un item de un repositorio desde la administración es fácil ver que si bien existen algunos elementos de la PDI como el checksum o el identificador persistente, la sintaxis que provee Dspace no es la adecuada en relación a lo que se precisa para asegurar una identificación clara del tipo de cambios realizados sobre un objeto digital o sobre quién realizó esos cambios, para constatarlo se muestra una captura de pantalla de un item del repositorio SEDICI donde es fácil observar lo inadecuado de la descripción de "agente" y "evento" en términos del diccionario de datos de PREMIS[3].

dc.description.provenance	Submitted by Nancy Martini (nancymartini@quimica.unlp.edu.ar) on 2020-11-19T13:42:16Z workflow start=Step: SeDiCILevelReview - action:claimaction No. of bitstreams: 1 Tesis doctoral Nancy Martini 2020 .pdf: 13328499 bytes, checksum: fdaf83eca57a2cb81d91189118344beb (MD5)	en
dc.description.provenance	Step: SeDiCILevelReview - action:editaction Approved for entry into archive by Analía Pinto(aprumiante@gmail.com) on 2020-11-19T14:17:11Z (GMT)	en
dc.description.provenance	Made available in DSpace on 2020-11-19T14:18:06Z (GMT). No. of bitstreams: 1 Tesis doctoral Nancy Martini 2020 .pdf-PDFA.pdf: 13856207 bytes, checksum: 877fd2b4dedae1da315caa854a27c7e2 (MD5) Previous issue date: 2020	en
dc.description.provenance	Submitted by Nancy Martini (nancymartini@quimica.unlp.edu.ar) on 2022-07-04T11:24:22Z workflow start=Step: SeDiCILevelReview - action:claimaction No. of bitstreams: 3 Tesis doctoral Nancy Martini 2020 .pdf-PDFA.pdf: 13856207 bytes, checksum: 877fd2b4dedae1da315caa854a27c7e2 (MD5) Tesis doctoral Nancy Martini 2020 .pdf-PDFA.pdf.txt: 560515 bytes, checksum: 3cf2f3164bbcd4ef4723b9da7a5df333 (MD5) Tesis doctoral Nancy Martini 2020 .pdf-PDFA.pdf.jpg: 3801 bytes, checksum: 622108bfd8a2c072d9e9e0c095e1652c (MD5)	en
dc.description.provenance	Step: SeDiCILevelReview - action:editaction Approved for entry into archive by Analía Pinto(aprumiante@gmail.com) on 2022-07-04T13:12:49Z (GMT)	en

Captura de pantalla de los metadatos provenance del item
http://sedici.unlp.edu.ar/handle/10915/109472

[3] Congress, L. of, & Committee, P. E. (s. f.). *PREMIS Data Dictionary for Preservation Metadata, Version 3.0 (Library of Congress)* [Webpage]. Recuperado 4 de agosto de 2022, de https://www.loc.gov/standards/premis/v3/index.html

Archival storage

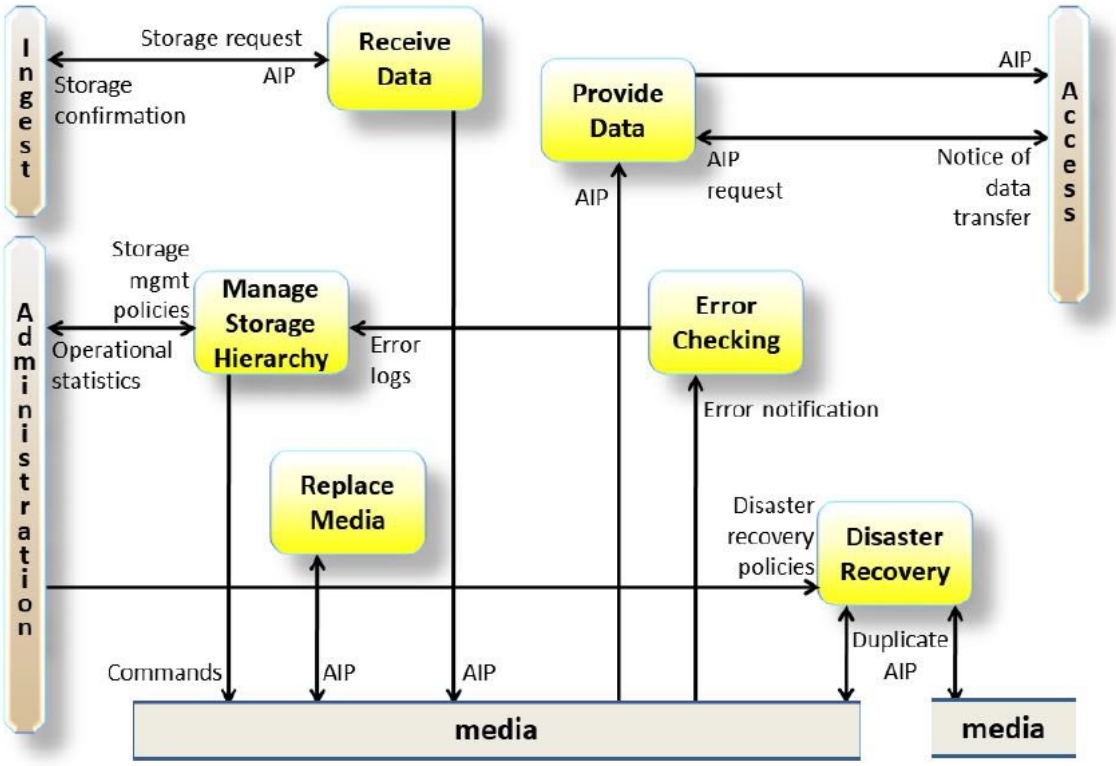


Figure 4-3: Functions of the Archival Storage Functional Entity [4]

- Un aspecto importante es que mueve los contenidos al storage permanente, se habla de que gestiona jerarquías y del tema de backups. Habla aquí de la transformación para que sea accesible en el tiempo, pero en el repositorio esto se hace en ingest. En Dspace no hay jerarquías de guardado.
- Provee copia de las solicitudes a Access.

[4] Op. cit. p. 4-8.

Data management

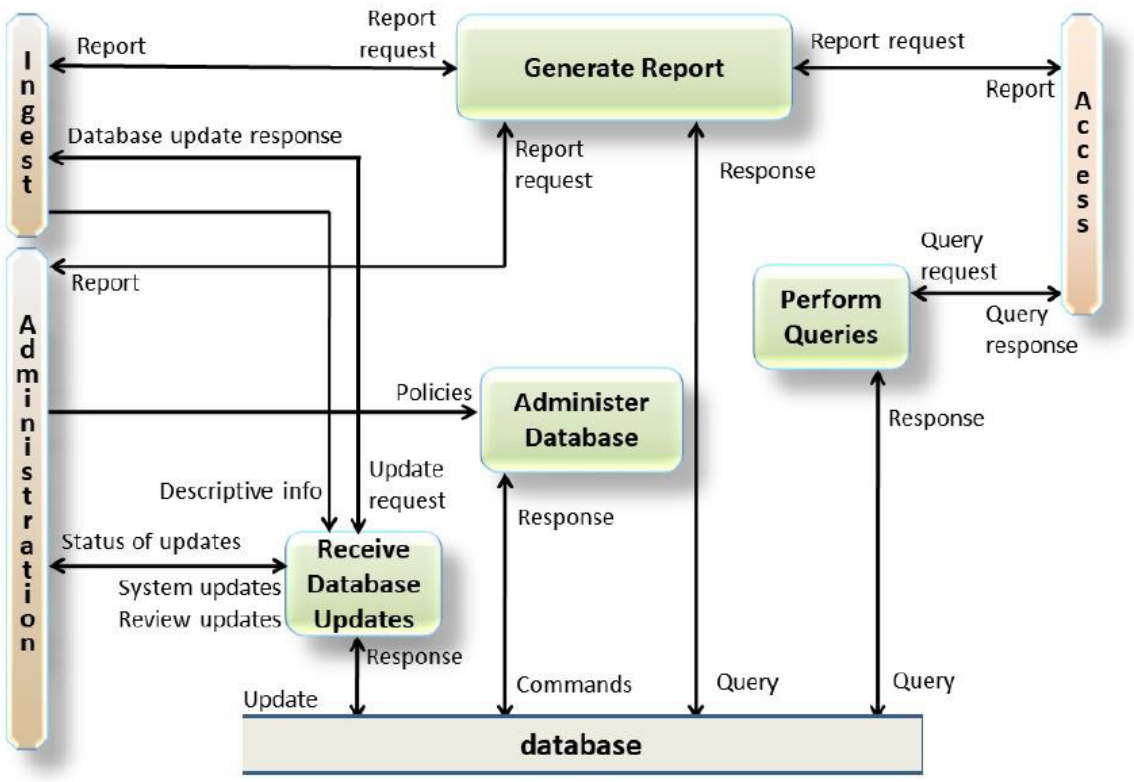


Figure 4-4: Functions of the Data Management Functional Entity [5]

- 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.
- Recibe solicitudes de actualización de los AIP desde Ingest y Administration.

[5] Op. cit. p. 4-10.

Administration

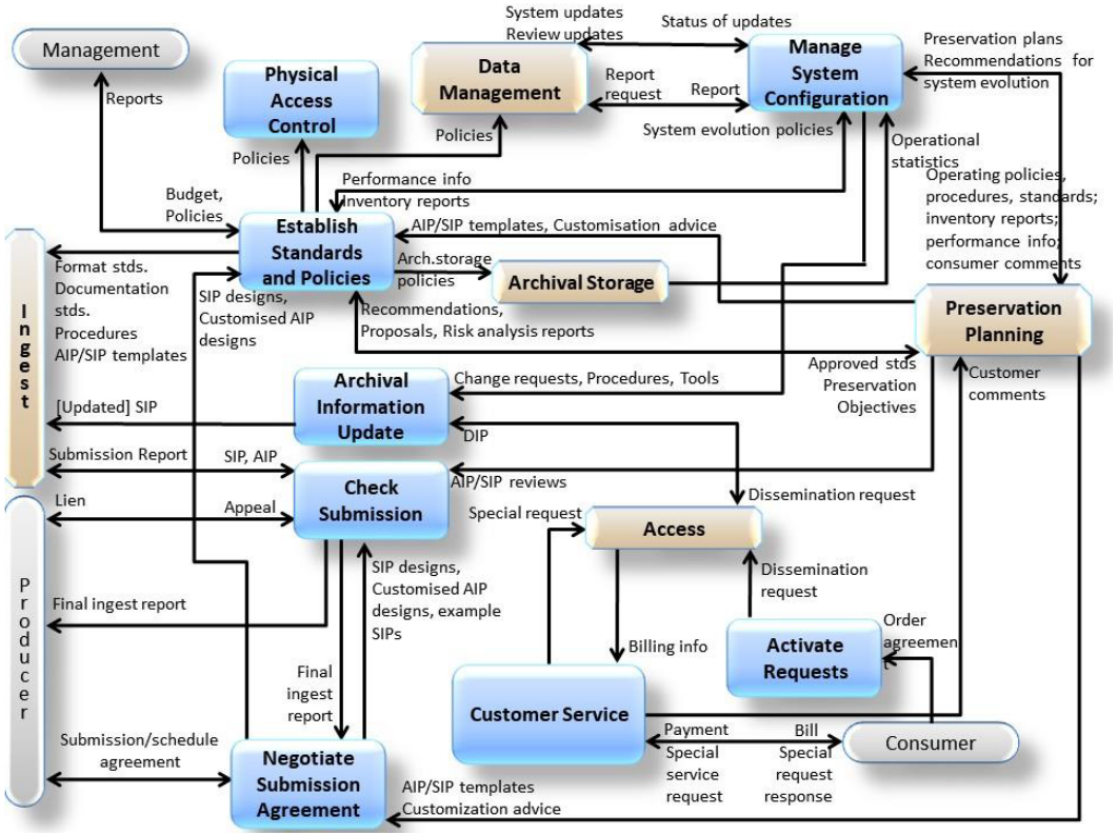


Figure 4-5: Functions of the Administration Functional Entity [6]

Esta entidad tiene una gran cantidad de módulos y funciones que exceden lo que se tiene en la Administración de un repositorio. Tal vez las que son de algún modo asimilables son:

- The negotiate submission agreement porque puede haber acuerdos con productores distribuidos de información (humanos o no) e incluso porque está relacionada con formatos aceptados por el repositorio que determinan acuerdos entre productores y la administración del repositorio. También es posible pensar que ante cantidades muy grandes de información a ingestar en el repositorio se precise efectivamente acordar límites y tiempos.
- The manage system configuration es un módulo que se encargaría de vigilar el funcionamiento correcto y la funcionalidad del archivo entendido como la base de datos/disco o lo que sea donde se guardan los contenidos.
- The archival information update, provee un mecanismo para actualizar contenidos del archivo tanto se trate de SIPs como de DIPs y eso sí es posible, sea por versiones a cambiar, o que reemplazan n archivo previo o porque se necesita un cambio en el DIP (que debe ser consecutivo a un cambio en el AIP) para que la información sea comprendida por los usuarios.
- Sobre Establish standards and policies atiende a varias cuestiones vinculadas al tema de formatos permitidos, obsolescencia y detección de errores.
- The Check Submission process se encarga de vigilar que la información de representación del OD sea adecuada y del tema de autenticidad.

[6] Op. cit. p. 4-12.

Preservation Planning

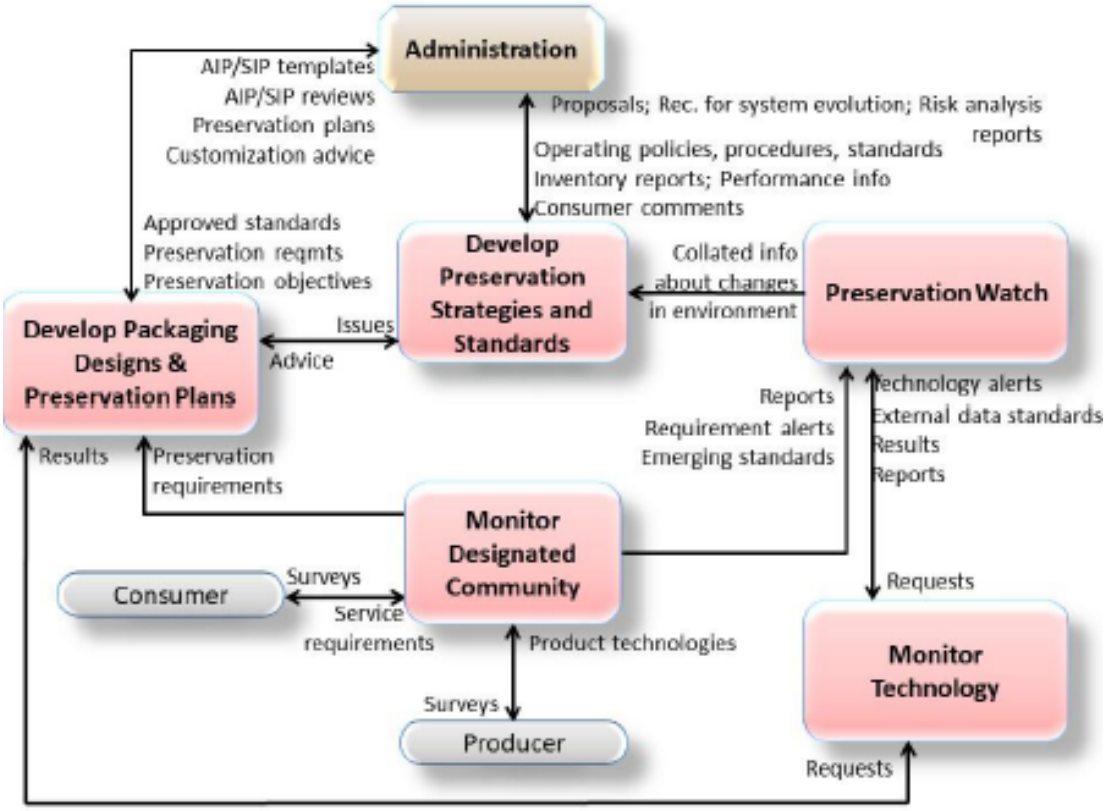


Figure 4-6: Functions of the Preservation Planning Functional Entity [7]

- The Monitor Designated Commnunity realiza un seguimiento en los cambios de los requisitos de servicio de la comunidad designada, así como de aspectos vinculados al acceso y la comprensión de la información por parte de la comunidad designada.
- The Monitor technology, es responsable de rastrear tecnología emergentes.
- The Preservation Watch trae informes, alertas de requisitos, recopila información relacionada con la preservación.
- The develop preservation strategies and standards es responsable de recomendar estrategias y estándares y evaluar riesgos que pueden incluir métricas.

[7] Op. cit. p. 4-15.

Access

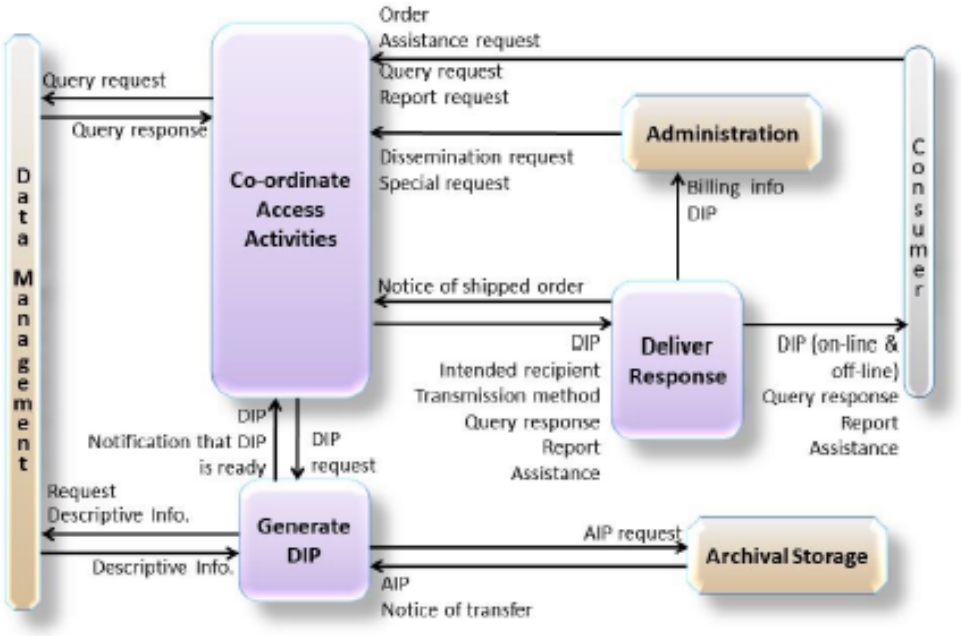


Figure 4-7: Functions of the Access Functional Entity [8]

- The coordinate access activity: provee la interfaz para búsqueda, navegación y para acceder a los holdings del archivo.
- Se distinguen tres categorías de solicitudes del consumidor cuyas respuestas son entregadas por Deliver response: solicitudes de consulta, solicitudes de informes y pedidos.
- The generate DIP recibe un AIP y solicita la información descriptiva necesaria para generar el DIP. Puede realizar otras operaciones, incluso transformaciones y verificaciones de accesos según el usuario.

[8] Op. cit. p. 4-18.

Paquete de Información

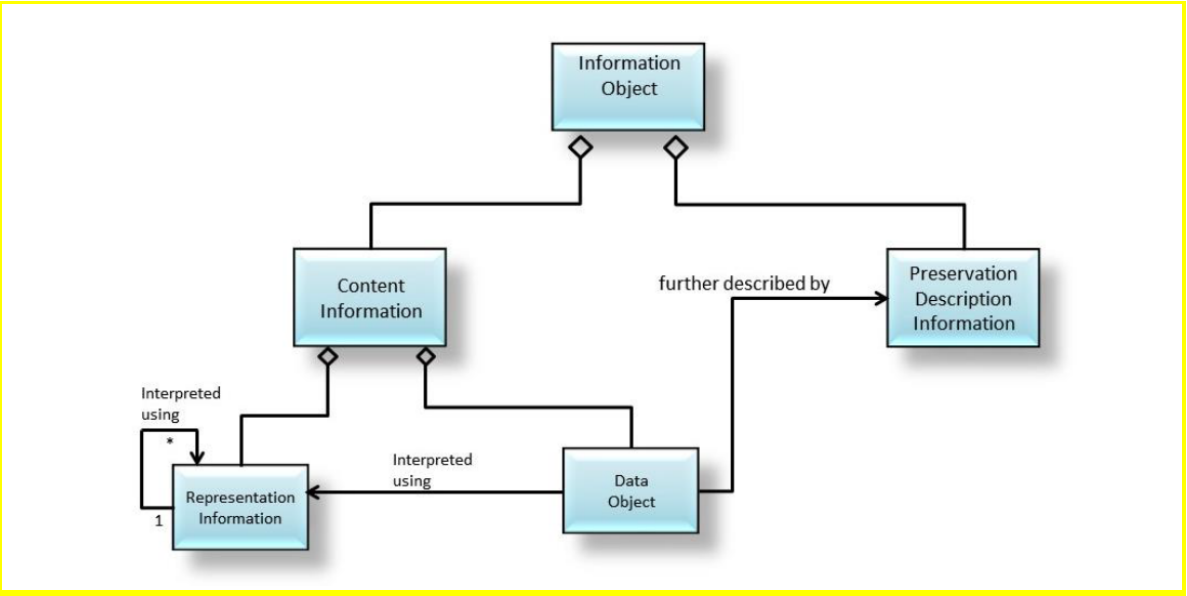


Figure 4-14 Example of an Information Object made up of Content Information and PDI [9]

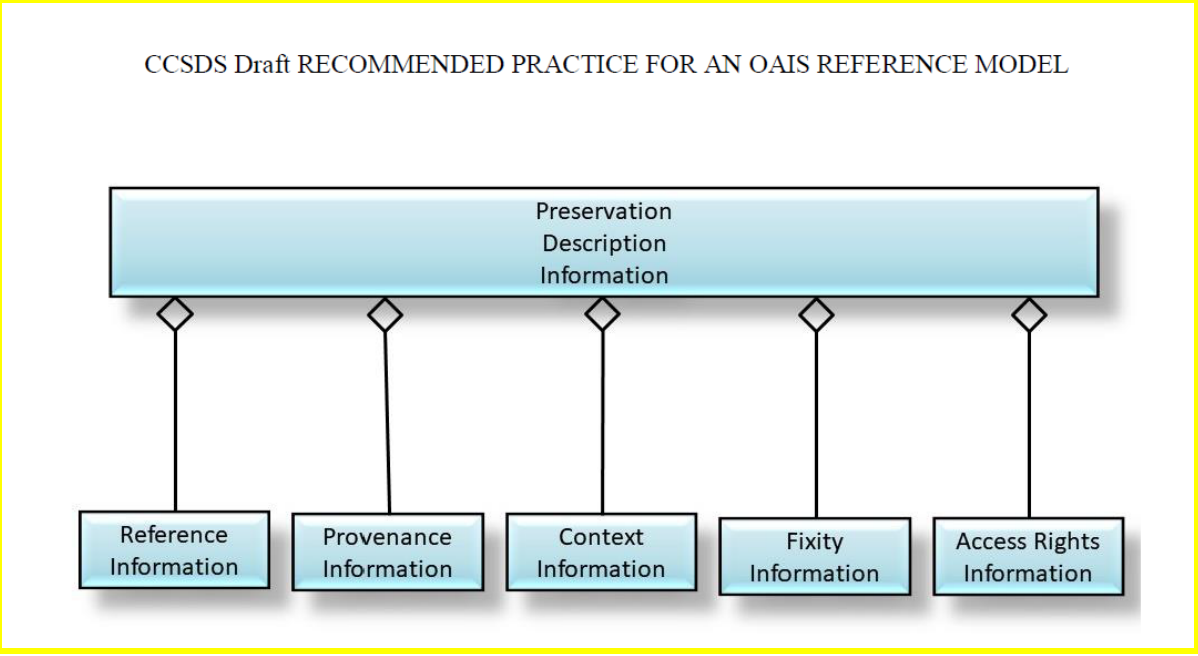


Figure 4-17: Preservation Description Information [10]

Provenance: documenta la historia del objeto digital. Del bitstream.
Context: documenta las relaciones del bitstream.

[9] Op. cit. p. 4-37.
[10] Op. cit. p. 4-41.

Cambios en la Norma ISO 14721[11]

- El cambio sustancial de la norma parece ser que la PDI refiere ahora al content data object (directamente al bitstream) en lugar de a la content information. Lo que significaría que el archivo tendría que tener asociados los 5 elementos de la PDI, es decir un identificador persistente propio, licencias, …
- La información de representación debería tener asociado también un fixity. Según los autores mencionados previamente, cuando la información de representación es un objetivo de la preservación, se la debe ver como parte de la información de contenido.
- Se hace una definición puntualizada de los objetivos de preservación, que fuera originalmente vinculada al término usabilidad pero sin mayores precisiones: Objetivos de preservación: “Un objetivo alcanzable específico que se puede llevar a cabo utilizando el objeto de información” y que se utiliza para la definición de otros términos como es el caso de la información de representación y de la frase: “comprensible de manera independiente”.
- Información de Representación: La información que mapea un Objeto de Datos en conceptos más significativos para que el Objeto de datos pueda entenderse de formas ejemplificadas por los objetivos de conservación.
- Independientemente comprensible: una característica de la información que es lo suficientemente completa como para permitir que la comunidad designada la entienda, como lo ejemplifican los objetivos de conservación asociados, sin tener que recurrir a recursos especiales que no están ampliamente disponibles, incluidos los individuos designados.
- Los Objetivos de Preservación están destinados a permitir que el depósito/repositorio posibilite probar y demostrar si la información es realmente comprensible independientemente por los miembros de la Comunidad Designada ahora y en el futuro.
- La nueva norma provee algunos ejemplos de preservación digital:
    - La capacidad de reproducir documentos, imágenes, vídeos o sonidos de forma suficientemente similar al original. Esto podría verificarse verificando que, por ejemplo, el documento sea legible o la imagen sea visible. También se podría comparar un análisis de los colores. Se podría realizar un análisis espectral de los sonidos y compararlo con el original.
    - La capacidad de procesar un conjunto de datos y generar los productos de datos esperados. Esto podría verificarse comparándolo con algo generado anteriormente, por ejemplo en Ingest.
    - La capacidad de comprender un conjunto de datos y usarlo en herramientas de análisis para generar resultados, por ejemplo, la densidad de electrones en la atmósfera superior o la estructura de una molécula, dadas ciertas medidas. Estos podrían compararse con los resultados generados anteriormente.
    - La capacidad de volver a realizar una actuación artística. Esto podría compararse con una grabación de una actuación anterior.

[11] Zierau, E., Giaretta, D., Garrett, J., Conrad, M., Longstreth, T., Hughes, J. S., ... & Felix, E. (2019, September). OAIS Version 3 Draft Updates. In *The 16th International Conference on Preservation of Digital Objects*.

- ☐ Las principales actualizaciones del Modelo de Información llevan adelante los cambios que se han descrito precedentemente en el apartado III (del artículo) donde la PDI se conecta al objeto de datos en lugar de la información de contenido, estos se explicitan en el diagrama V-1 expuesto más abajo en el mismo trabajo.
- ☐ Actualizaciones para la interoperabilidad de archivos: un cambio importante en la discusión de varios tipos posibles de interacciones de archivo es la forma en que se puede describir la distribución de la funcionalidad OAIS. Algunas categorías posibles de asociaciones de archivos de conjuntos de tres categorías tiene grados sucesivamente más altos de interacción organizacional o bien diferenciados de acuerdo a las funciones internas/externas:
    - ☐ Independientes: Archivos motivados por preocupaciones locales sin gestión ni interacción técnica entre ellos.
    - ☐ Cooperando: Archivos con Productores comunes potenciales, estándares de presentación comunes y estándares de difusión comunes, pero sin ayudas de búsqueda comunes.
    - ☐ Federados: Archivos con una Comunidad Local (es decir, la Comunidad Designada original atendida por el Archivo) y una Comunidad Global (es decir, una Comunidad Designada ampliada) que tiene intereses en los fondos de varios Archivos OAIS y ha influido en esos Archivos para proporcionar acceder a sus existencias a través de uno o más instrumentos de búsqueda comunes.
    - ❖ All In-house: Archivos/repos que realizan todas las funciones internamente.
    - ❖ Recursos compartidos: Archivos que han firmado acuerdos con otras organizaciones para compartir recursos, tal vez para reducir costos. Esto requiere varios estándares internos del Archivo (como estándares de interfaz de almacenamiento de ingesta y almacenamiento de acceso), pero no altera la visión de la comunidad de usuarios del Archivo.
    - ❖ Distribuido: Archivos que han distribuido la funcionalidad OAIS ya sea geográficamente u organizacionalmente. Son posibles diferentes niveles, formas y organización de la distribución. En todos los casos, se requiere que el Archivo supervise y administre el uso del Archivo de las funciones distribuidas, pero no altera la visión de la comunidad de usuarios del Archivo.

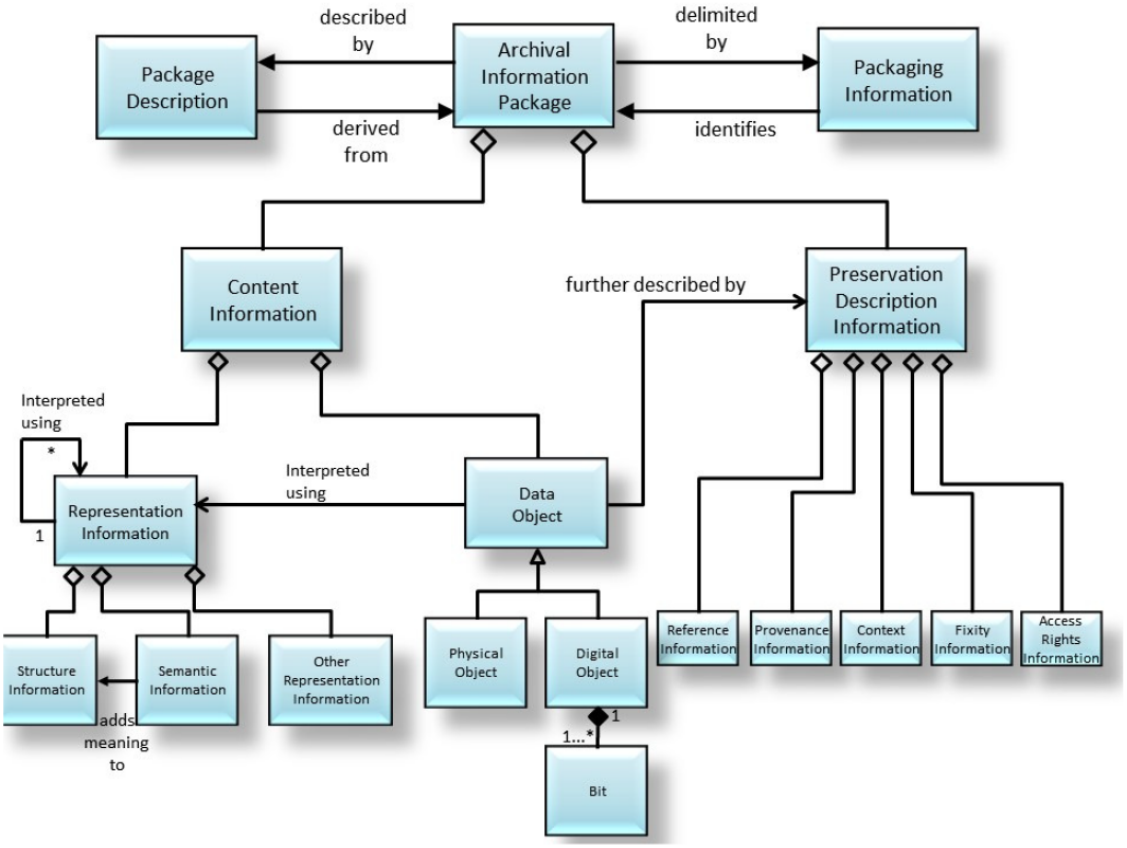


Figure 4-19: Archival Information Package (Detailed View) and its associated Package Description and Packaging Information [12]

En términos de metadatos existen tres tipos de metadatos:

- Descriptivos, con información sobre el ítem, se utiliza un esquema totalmente configurable, por defecto se basa en Dublin Core.
- Administrativos, en donde se incluyen los metadatos de preservación como el provenance, el cual contiene entre otras cosas, información del submitter, el revisor, del número de bitstreams con su nombre y checksum, y la fecha en la que se dió de alta el ítem en el sistema. La sintaxis que utiliza el provenance no se adecua a ningún estándar de metadatos de preservación y la información que almacena no es completa en términos de trazabilidad de los cambios sobre un ítem.
- Estructurales, con información sobre cómo debe presentarse al usuario los objetos digitales, cómo se dividen los distintos bundles en los que se encuentran los bitstreams, cuál es el bitstream primario y un sequence id que identifica unívocamente un bitstream dentro de un ítem y el cual DSpace utiliza como identificador del mismo.

DSpace permite conectar los metadatos con vocabularios controlados o sistemas de autoridades para normalizar y limitar el formato y contenido que pueden tener ciertos metadatos descriptivos.

[12] Op. cit. p. 4-43

Existen herramientas que permiten ver, crear y modificar metadatos de un archivo, en el siguiente enlace es posible ver una presentación y un video que explican el uso de algunas herramientas: http://sedici.unlp.edu.ar/handle/10915/139859

Ciclo de vida del Objeto Digital

***Fuente**: https://www.ticportal.es/glosario-tic/ciclo-vida-documento*

Fuente: Tesis doctoral "Una metodología de evaluación de repositorios digitales para asegurar la preservación en el tiempo y el acceso a los contenidos" [13]

[13] http://sedici.unlp.edu.ar/handle/10915/43157

Estrategia de formatos

Se debe tener una política de contenidos y/o de formatos que determine cuál es la estrategia del repositorio en cuanto al uso de formatos. Algunos de los temas a considerar en estas políticas son:

A. **Qué formatos se piden al usuario que deposita:** se puede tomar una postura permisiva o requerir solo formatos válidos. Esto depende de la comunidad y de la capacidad del repositorio de atender esta labor. Típicamente esto implica extraer datos desde formatos propietarios y/o no estándares.

B. **Qué formatos precisamos para preservar el recurso a largo plazo**: se debe elegir uno o más formatos para cumplir con el fin de garantizar la preservación a largo plazo. En ocasiones esto puede implicar usar formatos poco amigables para el uso público (ej wav, tiff, pdfA).

C. **Qué formatos son necesarios para publicar y permitir el uso por parte de la comunidad general**: dado que los repositorios brindan sus servicios a usuarios que acceden a través de internet y usando dispositivos diversos, se debe contar con formatos aptos para el uso y descarga de todo tipo.

D. **Qué otros formatos son necesarios para la comunidad especializada (si existe)**: en algunos casos, puede tenerse una comunidad específica de usuarios que precisan los recursos en otros formatos, quizás más crudos/puros o quizás en formatos propios de su área. En caso típico son los mapas, para los que suele usarse un visor de mapa para el público pero que requiere la descarga de archivos shapefile y kml para la comunidad específica.

E. **Qué herramientas se usarán para permitir el uso y descarga de los recursos:** En general se suele permitir la descarga directa del recurso desde el repositorio pero en ocasiones resulta conveniente utilizar herramientas complementarias que mejoren la experiencia del usuario, como ser reproductores de audio y video, visores de pdf, etc. De acuerdo al formato y tamaño de los archivos, se debe considerar también la conveniencia de uso de servicios externos como YouTube y vimeo (para streaming de videos), github (para repositorios de datos), entre muchos otros, que permiten servir grandes cantidades de datos a bajo coste. Estas decisiones influyen fuertemente en los formatos elegidos así como en la cantidad de archivos.
    a. SEDICI guarda los archivos de video en sus servidores y utiliza youtube para dar el servicio de streaming (https://www.youtube.com/c/sediciUNLP)

F. **¿Se deben guardar las versiones previas de las conversiones de archivos a otros formatos**? Durante el ciclo de vida de un ítem, se tendrán numerosas conversiones de formato. Algunas se aplican en la recepción, para normalizar los ítems recibidos, hacer OCR, convertir al formato de preservación del momento, pero luego, pueden aplicarse numerosas conversiones por migración de formatos obsoletos o simplemente porque se decide usar nuevos formatos para la publicación. Luego de cada acción de conversión se debe decidir si se guardará o no la versión previa y en caso afirmativo, debe estar claro dónde quedará y quizás, por cuánto tiempo. Si bien suele decirse que lo mejor sería guardar todo dentro del repositorio, puede no ser aplicable en la práctica, más que nada por limitaciones de almacenamiento.

A modo de ejemplo se incluye la siguiente tabla de formatos para recursos de audio, texto y un tipo particular de imágenes:

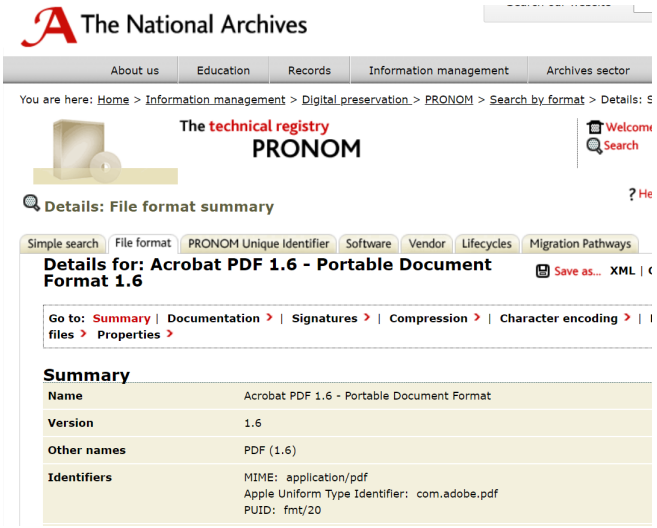
Tipo de recurso	**Formatos aceptados**	**Formato de preservación**	**Formato de publicación**
Audio	*wav* *flac* *mp3* *ogg*	*flac* *wav*	*mp3* (audiencia general) *ogg* (audiencia general)
Texto digital	*pdf* *pdf/A* *DOC* *DOCX* *ODT* *HTML* *txt*	*pdf/A* *txt*	*pdf/A* *epub* *txt*
Texto digitalizado en imágenes	*TIFF* *JPEG* *JP2* *PDF*	*pdf/A en alta definición* *TIFF*	*pdf/A* en tamaño aceptable
Imágenes astronómicas	*FITS* *TIFF* *JPEG*	*FITS* *TIFF*	*JPEG* (audiencia general) *FITS* (audiencia específica)

Tabla de formatos (elaboración propia)

En https://www.loc.gov/preservation/resources/rfs/format-pref-summary.html se puede encontrar un ejemplo mucho más completo de tipos de recursos y sus formatos preferidos.

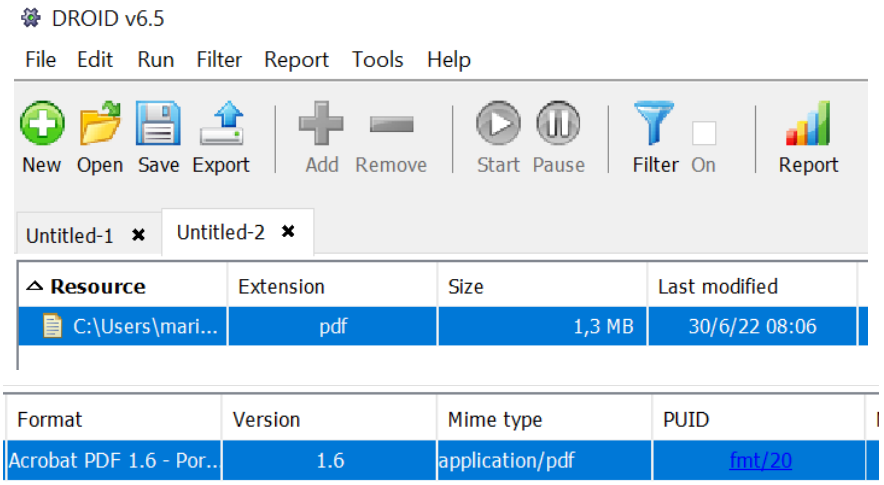
Algunas herramientas útiles para analizar y seleccionar formatos son:

- **Pronom**: catálogo de formatos, del cual se muestra una captura de pantalla del formato PDF 1.6.



Captura de pantalla del formato PDF 1.6 del registro PRONOM [14]

- **DROID**: permite perfilar el formato de uno o más archivos y es una aplicación sencilla de utilizar, a continuación se muestra una vista de los comandos que ofrece.



Vista de comandos y opciones de DROID

Licencias

- Tipos de licencias
  - En una instalación de Dspace por default se utilizan 2 licencias, una de uso y otra de depósito. La idea detrás de esto es que una asigna los derechos de uso y de manipulación del ítem y los objetos digitales relacionados y la otra que contiene las condiciones de depósito en el repositorio.
  - Para la licencia de uso se ofrecen las licencias Creative Commons y se guardan en un metadato ad-hoc denominado dc.rights y opcionalmente en un Bundle llamado “LICENSE-CC”.

[14] https://www.nationalarchives.gov.uk/PRONOM/

    - Para la licencia de depósito se utiliza un texto propio de la institución que una vez aceptado se guarda en un Bundle llamado “License”.
- Asignación de licencias
    - Al igual que las licencias de uso la aceptación de la licencia de depósito ocurre durante el proceso de envío manual, como una parte del formulario de envío.
    - Cuando se ingesta de manera automática no hay una licencia de depósito asociada y se debe elegir una licencia CC válida para cada ítem.
    - Es posible también definir licencias a nivel de comunidades y colecciones. Esto permite reservar ciertos derechos para un conjunto de items que, por ejemplo, pertenezcan a una organización, revista o facultad en la que a todos los ítems, tanto los existentes como los que en un futuro pertenezcan a esas colecciones o comunidades, se les asignen las mismas licencias.

- Caso SEDICI
    - se usan 2 metadatos propios y no se guarda el bundle LICENSE-CC:

sedici.rights.license	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
sedici.rights.uri	http://creativecommons.org/licenses/by-nc-sa/4.0/

    - Además de las opciones de CC normales se permite la selección de varias licencias abiertas, las cuales se obtienen a partir de un vocabulario controlado.

Autoevaluación con NDSA

En el siguiente enlace es posible acceder a una presentación y un video que describen cómo se realizó la autoevaluación con NDSA sobre el repositorio institucional CIC Digital: http://sedici.unlp.edu.ar/handle/10915/139885

Acciones relacionadas con la preservación en el repositorio

Tareas automáticas

Dentro del repositorio se realizan distintas tareas que tienen que ver con la preservación y que están configuradas para ejecutarse en respuesta a determinados eventos, por ejemplo ante la ingesta de un ítem, o de manera recurrente tras un período de tiempo dado con el uso de cronjobs que permiten programar la ejecución de tareas de manera regular.
Entre ellas podemos nombrar:

- Checksum en bitstreams.
    - Cuando se sube un archivo al repositorio DSpace calcula automáticamente el checksum para el objeto digital y almacena el mismo en la base de datos. Además, en los casos de ingesta vía el proceso de autoarchivo o carga

administrativa se registra este checksum junto a la cantidad de archivos en el metadato dc.description.provenance de cada ítem.

- Cronjobs de checksum
    - Además del chequeo de integridad en la ingesta, se tiene una tarea de chequeo de checksum de archivos que se ejecuta a diario y valida que el binario de cada archivo no cambie.
        - https://wiki.lyrasis.org/display/DSDOC5x/Scheduled+Tasks+via+Cron
- Asignación de identificador Handle
    - DSpace permite la configuración de un servidor Handle, el cuál se encuentra incluído con la misma instalación de DSpace. Una vez configurado el server y obtenido el prefijo desde el registro oficial de Handle, para el cual se debe abonar una pequeño monto anual, DSpace asigna automáticamente a cada ítem durante la ingesta de los mismos un identificador persistente Handle con el prefijo adquirido.
- Curation tasks para metadatos, formatos, etc.
    - Las Curation Tasks permiten la ejecución de una misma tarea sobre un grupo de ítems dentro del repositorio, este grupo puede ser desde un ítem o un listado de ítems, una colección, comunidad o el repositorio entero. Se podría, por ejemplo, crear una tarea de curation para el chequeo de calidad de metadatos, y generar un reporte de todos aquellos metadatos que no sigan algún formato específico. Otro ejemplo es una tarea de curation desarrollada en SEDICI, que chequea si los metadatos controlados por autoridad contienen la clave de autoridad adecuada y en caso que no la tengan se las agrega de manera automática. Para crear una de estas tareas se debe crear una nueva clase Java en el código fuente de DSpace que implemente la tarea a realizar, esta clase debe extender el comportamiento de otras clases específicas predefinidas propias de DSpace. Cada tarea se debe habilitar y configurar por separado desde los archivos de configuración, una vez habilitadas se pueden ejecutar desde la CLI de DSpace mediante el comando *dspace curate -t [nombre_tarea]* o también desde la interfaz de usuario como usuario Administrador si la tarea está configurada para que esto sea posible. Si se usan en conjunto con las cronjobs las tareas de curation permiten ejecutar tareas que realicen acciones de preservación o de calidad sobre los datos sobre todo el repositorio o sobre comunidades/colecciones de interés de manera periódica. En SEDICI se desarrolló una tarea de curation para el chequeo de PDFs en formato PDF/A pero no está habilitada por el momento.
    - En el siguiente enlace: http://sedici.unlp.edu.ar/handle/10915/139884 es posible acceder a un video que explica la generación de una tarea de curation.

Tareas manuales

Perfilamiento/Evaluación de archivos y formatos en DSpace

- **¿Qué exportar?**
    - si se quiere hacer un perfilamiento sólo del estado de los archivos digitales, sin tener en cuenta sus metadatos o al ítem al que están relacionados, se puede extraer los archivos directamente desde el assetstore, explorando su

sistema de directorios y obteniendo uno por uno los objetos digitales para luego realizar el procesamiento y el perfilamiento deseado.
    - En cambio si se precisa obtener otro tipo de información sobre los bitstreams, como por ejemplo su formato, el ítem al que pertenece y alguna información que contenga en sus metadatos como su nombre, será necesario complementar el assetstore con alguna consulta a la base de datos, exportaciones csv de los ítems que se desea relevar o una combinación de ambas para obtener la información contextual del bitstream.
    - También se puede hacer una exportación en paquetes SAF o AIP, este último más completo. Las exportaciones se hacen por consola desde el mismo servidor y suelen generar uno o más archivos zip con metadatos y archivos de cada ítem.

- **Ubicación del assetstore:** DSpace permite 2 opciones
    1. En un directorio dentro del servidor del repositorio.
       Se almacenan los objetos digitales en un directorio específico dentro del filesystem local al servidor. Este directorio es configurable y contiene 3 niveles de subcarpetas, para evitar que una sola carpeta tenga demasiados objetos y equilibrar el acceso.
    2. En la nube utilizando el servicio S3 (Simple Storage Service) de Amazon.
       Amazon ofrece almacenamiento ilimitado a través de un “bucket” en la nube, asigna a cada objeto almacenado una clave distinta. S3 es un servicio comercial, permite copias automáticas del contenido y la durabilidad de los objetos es mayor al 99,9%, también ofrece chequeo de integridad. Dado que S3 opera dentro de la red de Amazon y utilizando otros servicios de la compañía, promete una menor latencia de acceso que el almacenamiento local de los bitstreams.

  Asimismo, se permite la configuración de más de un assetstore en simultáneo, cada uno puede estar almacenado de alguna de las dos formas mencionadas anteriormente. Cada assetstore tiene un número asociado comenzando con el 0 (llamado simplemente assetstore), luego cuando este se llena se puede configurar otro assetstore el cual se llamará assetstore1 y así sucesivamente. Para configurar a qué assetstore de todos los declarados van a ir a parar los bitstreams, se debe setear la propiedad *assetstore.incoming*, en el archivo dspace.cfg, con el número de assetstore al que se quiere que los nuevos bitstreams ingresen. Por ejemplo, si se quiere que los bitstreams que se creen de ahora en adelante vayan al assetstore1, entonces la propiedad se debe setear de la siguiente manera: *assetstore.incoming = 1*. Por defecto los bitstreams irán al assetstore0, a menos que se cambie esta propiedad y se reinicie el servidor.

- **Ruta de bitstreams:** La dirección de un bitstream en el assetstore está formada a partir del *internal id* del bitstream en la base de datos, este internal id es distinto del id utilizado como primary key, sirve para indicar dónde se almacena el bitstream. Los primeros dos dígitos del internal id indican el primer nivel de carpetas en donde se encuentra el bitstream, el tercer y cuarto dígito el nombre de la carpeta del segundo nivel y el quinto y sexto dígito el nombre de la carpeta del tercer y último nivel, dentro de este nivel se encuentran los bitstreams cuyo nombre será igual a su internal id.

Por ejemplo, un bitstream con internal id *12345678901234567890123456789012345678* se guardará en el assetstore con la siguiente dirección:

[dspace]/assetstore/12/34/56/12345678901234567890123456789012345678

El uso de este id interno de 38 dígitos generado de manera aleatoria como forma de almacenar a los bitstreams permite una mejor distribución de los objetos en el assetstore que si se utilizara la clave primaria del bitstream en su lugar, mejorando así la eficiencia en el acceso.
https://wiki.lyrasis.org/display/DSDOC5x/Storage+Layer

- **Eliminación de bitstreams**: un punto en el que hay que tener cuidado al acceder directamente al assetstore para obtener los bitstreams es que no hay forma de conocer si los ítems relacionados con esos bitstreams o los mismos bitstreams fueron eliminados del repositorio. Esto ocurre porque DSpace no elimina físicamente los archivos digitales, solamente realiza un borrado lógico de los mismos, y para que se borren permanentemente se debe ejecutar un comando desde la CLI (Interface de línea de comandos) de DSpace. Para saber si un bitstream en cuestión fue eliminado o no, se debe realizar una consulta en la base de datos sobre la tabla en la que se almacenan los metadatos de los bitstreams, en donde existe una columna que indica si el objeto digital fue o no eliminado (lógicamente) del repositorio.

Tareas en lote

Otras tareas dentro del repositorio se realizan sobre un conjunto de ítems o archivos de manera manual, con la ayuda de alguna herramienta, es decir se realiza un procesamiento en tanda o lote. Este procesamiento se puede realizar para chequeos, normalizaciones o transformaciones sobre los datos. Alguna de estas tareas son:

- Validación y conversión de PDFs a PDFAs, con el uso del software 3-Heights.
- Extracción y corrección del fulltext de PDFs conformados por imágenes a través del proceso de OCR utilizando Abby Fine Reader. Eb http://sedici.unlp.edu.ar/handle/10915/139212 se puede se puede acceder a una presentación y un video sobre el proceso.
- Correcciones de metadatos en masa con consultas SQL directamente sobre la base de datos, tanto para agregar un metadato nuevo como para modificar el contenido o generar uno nuevo a partir de la unión de otros.
- Edición de metadatos de una comunidad o colección por parte de los Administradores haciendo uso de la importación y exportación de metadatos en formato csv que provee DSpace.
- Uso de herramientas como ExifTool ó ExifGui para agregar, borrar y modificar metadatos. En http://sedici.unlp.edu.ar/handle/10915/139859 se puede acceder a una presentación y un video sobre el uso de estas herramientas.

Perfilamiento ejemplo de dspace@sedici

A modo de ejemplo se realizó un sampleo de bitstreams de 3 años (2008, 2014 y 2021). Se eligieron 3 años con un intervalo de tiempo considerable entre ellos para poder ver el estado

de SEDICI en sus distintos períodos. Como se puede observar en la siguiente tabla, la cantidad de ítems cargados en cada uno de esos 3 años difieren considerablemente:

Año	Cantidad de archivos que ingresaron a SEDICI	Tamaño de los archivos
2008	1104	3.46 GB
2014	10298	13.89 GB
2021	19639	44.43 GB

Se seleccionaron al azar 1000 bitstreams de cada uno de esos años y luego se analizaron distintas características con el uso de la herramienta DROID con el objetivo de realizar un perfilamiento de los mismos. Se realizó un reporte con DROID por cada uno de estos años con los distintos formatos utilizados y las versiones de esos formatos, el mime-type, fecha de última modificación y otras características. A continuación se realiza un breve análisis de estos reportes:

- Cantidad de archivos, tamaño total y tamaño promedio de los archivos:

Año	Cantidad	Tamaño total	Tamaño promedio
2008	1000	2.44 GB	2.5 MB
2014	1000	3.28 GB	3.36 MB
2021	1000	0.794 GB	0.813 MB

- Cantidad de archivos por formato:

Formato	Muestra		
	2008	**2014**	**2021**
Acrobat PDF 1.1	3	0	0
Acrobat PDF 1.2	16	2	0
Acrobat PDF 1.3	72	44	0
Acrobat PDF 1.4	105	169	0
Acrobat PDF 1.5	136	169	0
Acrobat PDF 1.6	434	383	0
Acrobat PDF 1.7	0	1	0
Acrobat PDF/A 1b	79	88	623
Acrobat PDF/A 1a	134	2	269
Acrobat PDF/A 2a	2	0	7

Acrobat PDF/A 2u	0	6	92
Acrobat PDF/A 2b	2	0	2
Acrobat PDF/A 3a	0	0	1
JPEG 1.01	12	0	0
JPEG 1.02	4	0	0
ePub	1	0	0
MPEG 1/2 Audio Layer 3	0	136	0
Portable Network Graphics 1.0	0	0	1
Portable Network Graphics 1.2	0	0	4
Quicktime (video)	0	0	1

- Cantidad de archivos por MIME type:

MIME type	**2008**	**2014**	**2021**
application/epub+zip	1	0	0
application/pdf	983	864	994
image/jpeg	16	0	0
image/png	0	0	5
audio/mpeg	0	136	0
video/quicktime	0	0	1

Consideraciones sobre el problema del almacenamiento

Hardware

- Para el almacenamiento se debe usar agregaciones de discos que soporten al menos la falla en un dispositivo, es decir, arreglos RAID con redundancia, es decir, que soporte una falla en al menos un disco físico.
- Para cada arreglo RAID es recomendable contar con discos disponibles (spare) que puedan ser agregados al arreglo de forma automática (hot-spare) o al menos manual (cold-spare).

Backups

En el caso de SEDICI y CIC Digital se guarda una primera copia en un directorio dentro del servidor del repositorio. Se realizan copias de seguridad por duplicado, una de ellas se almacena en otro servidor dedicado para backups propio, dentro de la misma infraestructura,

y una tercera copia se guarda en el servicio de almacenamiento de Amazon S3 Glacier Deep Archive. Este servicio ofrece almacenamiento para datos que no son accedidos con regularidad, se almacenan en al menos tres zonas de disponibilidad geográficamente dispersas, con más del 99,99 % de durabilidad, y se pueden restaurar en 12 horas o menos.

Glosario de herramientas que hemos usado y dado pequeños seminarios o talleres

- DROID

DROID es una herramienta desarrollada para realizar una identificación automatizada por lotes de formatos de archivos. Permite a cualquier repositorio digital ser capaz de identificar el formato preciso de todos los objetos digitales almacenados, generar reportes que permitan identificar los formatos y sus dependencias. DROID utiliza firmas internas para identificar e informar del formato y la versión específicos de los archivos digitales. En http://sedici.unlp.edu.ar/handle/10915/103632 se puede acceder a una presentación y un video de cómo instalar y usar DROID.

- Exactly

Es una aplicación de código abierto utilizada para transferir de forma remota y segura cualquier dato digital. Para la integridad de los datos, utiliza el formato de empaquetado de archivos BagIt. En este enlace http://sedici.unlp.edu.ar/handle/10915/110561 se puede acceder a una presentación y un video sobre cómo armar paquetes SIP que brinda infromación sobre esta herramienta.

- Exiftool y Exiftoolgui

Es un software de código abierto que nos permite leer y manipular metadatos (datos EXIF, XMP, IPTC) de imágenes, audios y vídeos. Útil para el borrado de datos sensibles (datos personales, ubicación de vivienda, datos clínicos, etcétera.)
Exiftool se ejecuta por línea de comandos y Exiftoolgui agrega una interfaz gráfica. Sobre estas herramientas ya se agregaron las referencias elaborados (presentación y video) en el apartado titulado “Tareas en lote”.

- Abby

Abby está centrado en el OCR tiene una interfaz que permite editar el ocr y corregirlo a mano, además puede editar las imágenes para mejorarlas tanto de manera manual como automática.
La opción “hot folder” de Abby es la que se usa para tomar un montón de archivos de una carpeta volcados por la administración y de ahí comprimir, pasar a PDFA y volcar en otra carpeta.

- 3 Heights

En SEDICI se utiliza para convertir formatos y validar de manera automática los PDFAs. El programa cuenta con más funciones y en la actualidad ya está disponible la versión 4 Heights.
Sobre Abby y 3 Heights se incluyó enlace a presentación y video en en el apartado titulado “Tareas en lote”.

Problemas, soluciones y mejoras

Pensar PD de objetos nacidos digitales y de objetos digitalizados

Cambios requeridos de sintaxis

- A pesar de proveer metadatos como el provenance que registran ciertos cambios en el contenido, SEDICI no posee metadatos que se ajusten a algún estándar de preservación existente, o que registren cambios en el ítem durante todo su ciclo de vida. Actualmente, el provenance no registra todas las acciones que se efectúan sobre un ítem y su sintaxis no es la adecuada.
- Una solución podría ser ampliar o modificar los metadatos existentes para que registren dicha información. Por ejemplo, agregar al provenance: agente y evento e incluso licencias.
- Algo similar a la propuesta PREMIS DD

  Unidad semántica Agente:
  - 3.1 agentIdentifier (identificador del agente) (O, R)
  - 3.1.1 agentIdentifierType (tipo de identificador del agente)
  - 3.1.2 agentIdentifierValue (valor del identificador del agente)
  - 3.2 agentName (nombre del agente) (M, R)
  - 3.3 agentType (tipo de agente) (M, NR)

  Unidad semántica Evento: como es muy larga la sintaxis ir a:
http://www.bne.es/es/Micrositios/Publicaciones/PREMIS/002_Diccionario/003_EntidadAcontecimiento/
- Si bien PREMIS propone un esquema jerárquico de metadatos, tal vez pueda plantearse adaptar el diccionario de datos a un esquema plano para integrarlo a SEDICI.

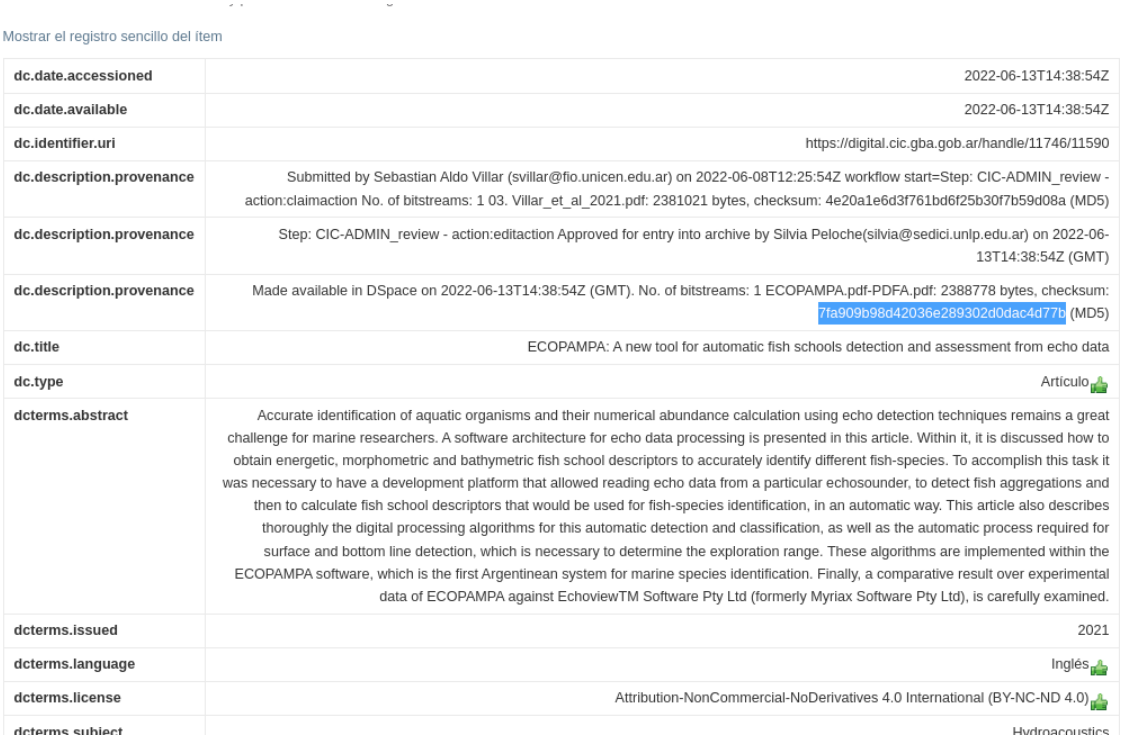
Documentación y creación de un plan de riesgos

- Una falencia de SEDICI es la falta de documentación de medios de almacenamiento, medidas de seguridad, equipos utilizados, etc.
- La creación de un plan integral de riesgos sería un gran avance en ese sentido.
- Allí se deberían listar eventuales problemas, eventos y catástrofes que atenten contra la preservación del contenido.
- Se indica qué acciones se deben tomar ante cada potencial riesgo.
- Por ej, ante un ataque DoS[15] (de denegación de servicio) se tiene definido paso a paso qué hacer para detectar los bots que están dañando al sistema y eventualmente bloquearlos. (http://trac.prebi.unlp.edu.ar/projects/sedici-dspace/wiki/Qu%C3%A9_hacer_frente_a_un_DOS)
- Otra propuesta interesante es la de tener una planificación documentada en lo que se refiere a control de obsolescencia de los medios, equipos y servidores en los que se almacena el contenido y que mantienen al repositorio en funcionamiento. Allí se listarían los distintos equipos con los que se cuenta, el tiempo de vida estimado de cada uno y de qué manera se realizaría su eventual reemplazo en caso de ser necesario.

[15] Un ataque de denegación de servicio (DoS) es un intento malicioso de sobrecargar de tráfico una propiedad web para interrumpir su funcionamiento normal.

Problemas de trazabilidad

1. Ejemplo ítem subido y reemplazado, diferente checksum, pérdida de original.



Mostrar el registro sencillo del ítem

dc.date.accessioned	2022-06-13T14:38:54Z
dc.date.available	2022-06-13T14:38:54Z
dc.identifier.uri	https://digital.cic.gba.gob.ar/handle/11746/11590
dc.description.provenance	Submitted by Sebastian Aldo Villar (svillar@fio.unicen.edu.ar) on 2022-06-08T12:25:54Z workflow start=Step: CIC-ADMIN_review - action:claimaction No. of bitstreams: 1 03. Villar_et_al_2021.pdf: 2381021 bytes, checksum: 4e20a1e6d3f761bd6f25b30f7b59d08a (MD5)
dc.description.provenance	Step: CIC-ADMIN_review - action:editaction Approved for entry into archive by Silvia Peloche(silvia@sedici.unlp.edu.ar) on 2022-06-13T14:38:54Z (GMT)
dc.description.provenance	Made available in DSpace on 2022-06-13T14:38:54Z (GMT). No. of bitstreams: 1 ECOPAMPA.pdf-PDFA.pdf: 2388778 bytes, checksum: 7fa909b98d42036e289302d0dac4d77b (MD5)
dc.title	ECOPAMPA: A new tool for automatic fish schools detection and assessment from echo data
dc.type	Artículo
dcterms.abstract	Accurate identification of aquatic organisms and their numerical abundance calculation using echo detection techniques remains a great challenge for marine researchers. A software architecture for echo data processing is presented in this article. Within it, it is discussed how to obtain energetic, morphometric and bathymetric fish school descriptors to accurately identify different fish-species. To accomplish this task it was necessary to have a development platform that allowed reading echo data from a particular echosounder, to detect fish aggregations and then to calculate fish school descriptors that would be used for fish-species identification, in an automatic way. This article also describes thoroughly the digital processing algorithms for this automatic detection and classification, as well as the automatic process required for surface and bottom line detection, which is necessary to determine the exploration range. These algorithms are implemented within the ECOPAMPA software, which is the first Argentinean system for marine species identification. Finally, a comparative result over experimental data of ECOPAMPA against EchoviewTM Software Pty Ltd (formerly Myriax Software Pty Ltd), is carefully examined.
dcterms.issued	2021
dcterms.language	Inglés
dcterms.license	Attribution-NonCommercial-NoDerivatives 4.0 International (BY-NC-ND 4.0)
dcterms.subject	Hydroacoustics

2. provenance incompleto, sin estructura

dc.description. provenance	Made available in DSpace on 2022-06-13T14:38:54Z (GMT). No. of bitstreams: 1 ECOPAMPA.pdf-PDFA.pdf: 2388778 bytes, checksum: 7fa909b98d42036e289302d0dac4d77b (MD5)

3. otros casos no contemplados: conversiones, transformaciones por curation, versionado ,

Guardar los bitstreams originales junto con el ítem y creación del bundle de preservación

Ahora en SEDICI los bitstreams originales, es decir los archivos que envían los usuarios que no pasaron por el proceso de normalización y transformación de formatos de SEDICI, se almacenan separados del ítem (y por consecuencia del bitstream resultante de su transformación) con el que están relacionados, se guardan en una unidad dedicada para ese fin en Google Drive. Esto hace muy complejo el obtener los archivos originales de un ítem, y además no permite mantener el versionado de los archivos en el mismo ítem.

Una posible solución a este problema es la creación de un bundle especializado en el guardado de información de preservación de los archivos. De esta manera, se podrían depositar allí los archivos originales y mantener un seguimiento de cambios en los archivos

al almacenar las versiones generadas. También permitiría guardar por ejemplo, los checksums de cada una de las versiones y las distintas licencias utilizadas (para las distintas versiones porque la licencia del ítem ya se encuentra en el bundle “License”).