



PONTIFICIA
**UNIVERSIDAD
CATÓLICA**
DEL PERÚ



Conferencia Internacional
BIREDIAL - ISTE
22-23-24-25 octubre 2018

Ciencia abierta, Interoperabilidad,
Preservación digital, Visibilidad
académica, Gobierno abierto

VIII Conferencia Internacional sobre
Bibliotecas y Repositorios Digitales
de América Latina (BIREDIAL-ISTEC)

Implementación de la preservación en un repositorio institucional

Dra. Marisa R. De Giusti



Esta obra está bajo una [Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional](#).

Implementación de la preservación en un RI - Hoja de ruta (parte 1)

La preservación digital en el contexto de un RI

- Conceptos generales sobre preservación digital
- Gestión de recursos en un repositorio:
 - Definición de políticas
 - Colecciones y comunidades
 - Modelo de contenidos y catalogación de recursos
 - Tipos de recursos y formatos de archivos

Implementación de la preservación en un RI - Hoja de ruta (parte 2)

Preservación digital

- Definiciones, objetivos, etapas
- Problemas relativos a la PD
- Estrategias
- Selección de formatos
- El formato PDF/A
- El Modelo OAIS (ISO 14721:2012)
- PREMIS
- Preservación en DSpace
- Otros modelos:
 - Archivemática+DSpace
- Otros modelos: RODA

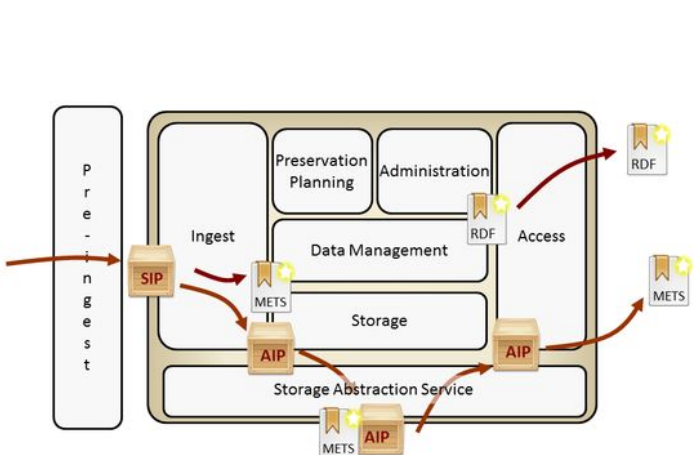
Tratamiento de documentos

- Descargas y mejora de documentos
- Generación de PDF/A
- Optical Character Recognition

Digitalización

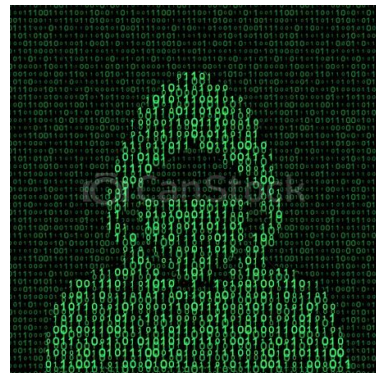
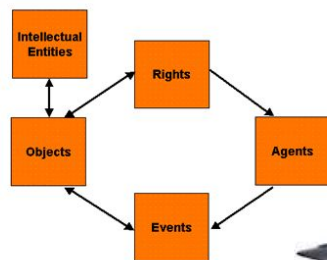
- Introducción a la digitalización de documentos
- Captura de imágenes
- Edición de imágenes
- Formatos de exportación

Preservación digital



PREMIS PRESERVATION METADATA MAINTENANCE ACTIVITY

PREMIS data model



Objetivos del taller

- Crear conciencia en los profesionales, usuarios, funcionarios de bibliotecas y archivos, políticos, investigadores, etc., sobre los riesgos que conlleva mantener en el tiempo los objetos digitales y dar accesibilidad permanente a los mismos.
- Analizar estándares e implementaciones para cumplir con el objetivo de preservación.
- La digitalización sus dificultades y la generación de nuevos materiales para preservar.

Las razones de la preservación digital

Una de las bases de nuestra cultura es preservar la producción del presente de nuestra instituciones y también la histórica.

La gran panacea de estos tiempos es la información digital, que, para mantener el principio enunciado más arriba conlleva mucho trabajo.

Los objetos digitales dependen de la tecnología informática.



Contexto y definiciones



El entorno de la preservación

- Toda la información es pasible de reducirse a un código binario en el que se pueden representar números, letras, sonidos e imágenes.
- Todos los documentos y los datos pueden ser digitalizados y gestionados por medios informáticos gracias al software y al hardware.
- El avance tecnológico mantiene a la informática sometida a cambios constantes.



El entorno de la preservación

- Los aparatos informáticos: computadoras, discos, memorias, software se transforman en objetos de museo.
- Los datos digitales pueden pervivir y continuar siendo usados, sobreviviendo al software y al hardware, por eso es que no basta con conservarlos, hay que preservarlos.



Formatos

Los datos se codifican y almacenan en archivos que pueden tener estructuras distintas llamadas formatos que permiten adaptarse mejor a determinado tipo de datos y a determinados usos.

Los formatos están íntimamente ligados al software y al hardware porque se plasman en ambos.

Los formatos se vuelven obsoletos por los cambios en software, hardware y en las prestaciones que piden los usuarios



Definiciones

Conservación: pone el énfasis en establecer condiciones ambientales, de almacenamiento y de uso. Son claves en los documentos analógicos y no pueden obviarse en los digitales, aunque parecen más obvias.



Definiciones

Seguridad Informática: establece **políticas** para analizar, detectar y solucionar riesgos informáticos que pueden sufrir los datos. Se ocupa de crear copias de seguridad, de controlar las autorizaciones de acceso a los datos, de analizar la confiabilidad de software y hardware, de detectar virus o actividades de pirateo, de prever contingencias por cortes eléctricos, incendios.

Se piensa en el presente y a corto plazo. la normativa por ejemplo puede ser la ISO 27001.

No se piensa en la legibilidad y comprensión a largo plazo.



Definiciones

Preservación digital: debe asegurar el acceso y el uso a futuro de los documentos nacidos digitales y de los digitalizados. Incluye: políticas de conservación, seguridad informática y añade además otras como migraciones, emulaciones que permiten cumplir con los objetivos a largo plazo.



Resultante: Problemas en la preservación de OD

1. La propia naturaleza de los objetos digitales los hace efímeros.
2. La obsolescencia de los medios informáticos: dado que los OD siempre están mediados por la tecnología que cambia constantemente; una inadecuada vigilancia o falta de transformaciones puede dejarlos inaccesibles. La incompatibilidad entre sistemas nuevos y antiguos sumado a que los formatos, medios de soporte, software y hardware quedan obsoletos en poco tiempo.

Respuesta: ¡Resiliencia!

- Los cambios en el software, el hardware y los formatos son inevitables y hay que convivir con ellos.
- El problema aparece cuando el documento, la imagen o cualquier dato digital debe tener una vida útil más prolongada que sus soportes.



Respuesta: prever los cambios e incorporarlos en la gestión de datos. Este es pues el objetivo de la **preservación digital**.



Los problemas no se terminan

Los archivos digitales viven en problemas debido a los cambios en hardware, software y formatos.

Además los archivos digitales se pueden cambiar o incluso borrar de manera intencional o involuntaria o incluso por negligencia (no hay copia de seguridad) o debido a un virus circulando por el sistema.



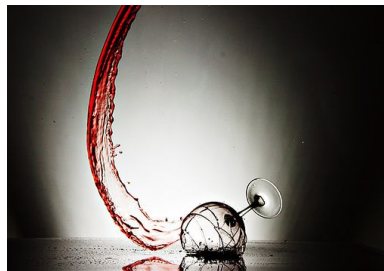
Edito un documento, edito la versión equivocada, ¿cuál es el último?



Fragilidad de los datos

Se hace evidente durante el período de creación y de mayor uso de los datos, pero...sigue a lo largo de toda su vida útil. En cualquier momento puede ocurrir que al sistema entre un virus o que alguien corrompa la información.

La fragilidad se asocia a la integridad y a la autenticidad de los archivos.



El alcance del daño en un documento en papel se puede limitar a un detalle del texto o de la imagen, restituible y además, circunscripto al punto.

Un error en un archivo digital puede dejar inutilizado al archivo. la restauración (en el concepto tradicional) no existe en el ámbito digital.



La preservación de los contenidos

En los documentos en papel se habla de “negligencia benigna”: el olvido de un manuscrito en un arcón, puede que lo preserve. En los digitales, la negligencia benigna no sirve: un disco olvidado 5 años... no sirve. La preservación digital DEBE tomar un rol activo



- No a la negligencia benigna.
- No a la preservación basada en las condiciones ambientales.
- Se debe asegurar la integridad y autenticidad del recurso.
- Obliga a pensar lo que se va a preservar y a qué plazo.
- No necesariamente se conserva la integridad externa del documento sino las propiedades significativas y ¡hay que pensar en ellas!



Preservación digital

- La preservación digital supone, en relación con la conservación de los documentos en papel, un importante **reto tecnológico**, pero también de otros tipos:
- **legal**, permisos de los autores para realizar las transformaciones necesarias
- **económico**, ¿quién financia el personal y las acciones para la preservación?,
- **organizativo** ¿de quién es la responsabilidad de cada acción? ¿cómo se asegura la continuidad de las decisiones?)

(Keefer; Gallart, 2007).

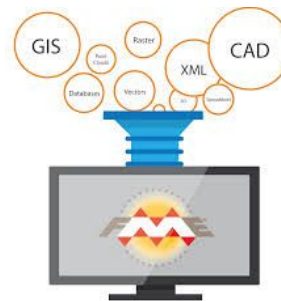


Retos legales

No se encuentran resueltos todos los aspectos jurídicos de la preservación digital.

La legislación de derecho de autor impide cambiar el soporte y el formato de los documentos sin la autorización de quien tiene los derechos. Se lo debe poner en la licencia de difusión que firma el autor, por ejemplo.

Otro tema legal, pero diferente y a tener en cuenta es la protección de los datos ante usos indebidos, por ejemplo datos personales.



Retos económicos

Aunque resulta fácil lograr un financiamiento como proyecto de innovación puntual, la preservación hay que sostenerla en el tiempo.

Con el tiempo las colecciones crecen y el problema aumenta.

Algunos estudios muestran que el 50% de los costos de un repositorio están en la ingesta, pero las acciones de preservación y acceso casi se reparten el resto.

Dentro de lo que hace a la preservación hay que considerar el almacenamiento, discos y demases, pero no son la parte más importante. El gasto en almacenamiento se vincula, claro, a los formatos...



Retos organizativos

El primer punto a tener en cuenta en un proyecto de preservación a largo plazo es qué información digital debe preservarse.

El segundo punto es quién es el responsable por la preservación. Puede ocurrir, incluso que a quien creó el documento, ya no le interese y menos preservarlo.

El párrafo previo implica la vigilancia en todo el ciclo de vida del objeto digital porque el mismo pasa por distintos gestores: autor, administrador...



Retos organizativos en el repositorio institucional

El repositorio es un archivo web donde se depositan, en formato digital, materiales derivados de la producción científica o académica de una institución (universidades, centros de investigación).

El objetivo de crear repositorios es facilitar el acceso de la comunidad académica y científica internacional a los resultados de la investigación realizada por sus miembros y aumentar la visibilidad de la producción científica de la institución. Asimismo, contribuir a la preservación de los documentos digitales allí depositados.



Organización en el repositorio institucional



**Muchas
Tipologías**

Revistas,
libros,
sonidos,
capítulos de
libros, tesis,
actas de
congresos,
imágenes fijas
software,
patentes...

Tipologías y formatos

Artículo: Un artículo es un texto en prosa no ficcional en el que se desarrolla un tema en particular. Está escrito por un autor o varios autores, quienes llevaron adelante una investigación y escriben para publicar los resultados de investigación en una revista científica o académica.

Sonido: Recurso diseñado principalmente para ser oído. Ejemplos: formato de archivo de audio, CD-ROM y grabaciones de sonido. (AUDIO EN CIC)

Comunicación de congreso: Comunicación de congreso que se presenta en el congreso y se expone ante la audiencia. La comunicación del congreso se publica en las actas.

Contribución a revista: Texto (por ejemplo: artículo, columna, editorial u otra contribución) escrito para ser publicada en una revista académica o científica.

Imagen fija: Representación visual estática. Este tipo de imágenes incluye diagramas, dibujos, diseños gráficos, planos, mapas, fotografías e impresiones.

Informe: Se denomina informe al registro de hallazgos de una investigación terminada o en progreso, o de otros hallazgos técnicos. Por lo general, está identificado por un número de informe y a veces también por un número de subsidio asociado a la agencia de financiamiento.

Libro: Una publicación no seriada que está completa en un volumen o en un número definido de volúmenes. En su formato físico, un libro es una serie de hojas de papel, pergamino u otro material encuadernadas en uno de sus extremos con tapas. Los libros se suelen identificar con un ISBN.

Capítulo de libro: Un capítulo o sección de un libro, por lo general con un título o número de capítulo que lo identifica.

Publicación Periódica: Publicación seriada con su propio título identificador, caracterizada por una variedad de contenidos y contribuciones (por ejemplo, artículos, editoriales, reseñas, columnas), que se publica con una periodicidad regular.

Reseña: Una reseña del trabajo de un tercero.

Tesis: una tesis o disertación es un documento presentado para postularse a un título académico o a un cargo profesional, en la que el autor presenta un tema de investigación y los resultados. Pueden ser de grado, maestría o doctorado.

Organización en el repositorio institucional

Se deben tener políticas claras.

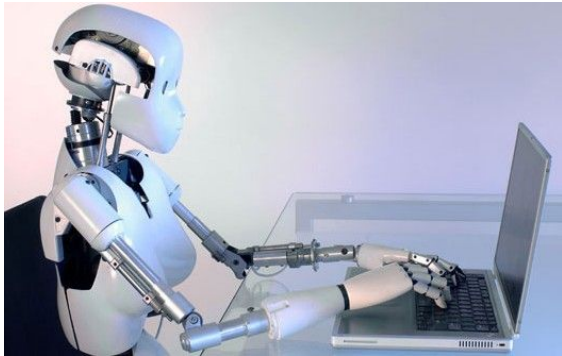
- Política de contenidos
- Política de acceso a datos
- Política de depósito
- Política de preservación
- Política de Metadatos
- Política de Servicios

Usar muchos metadatos



Cosas en las que hay que pensar

Para mantenerse y crecer se debe aumentar en la mayor medida posible la automatización de la ingesta de contenidos.



Se deben tener políticas muy claras sobre preservación de contenidos, pero, a veces generar estas políticas toma mucho tiempo en un universo cambiante, de ahí que a veces se decida preservar todo de manera indefinida.



Problema adicional: escenario cambiante de la preservación digital

Si bien el escenario recuerda al de un archivos históricos e inamovible, se está en las antípodas.

Los sistemas de preservación digital están pensados para crecer de manera indefinida y en paralelo a los cambios en las necesidades de los actores involucrados.

Los documentos son frágiles y las técnicas también: un error, una falla basta.

En algunos casos la institución, incluso, resguarda para terceros...

Para sostener la preservación y que sea confiable se debe revisar y **auditar**.

How I learned to stop worrying and love
ISO 16363

Karel Van Marrewijk | *ISO 16363*
PUBLISHED
October 25, 2019
ISO 16363:2016
© Karel Van Marrewijk

nestor

DRAMBORA interactive

Digital Repository Audit Method
Based on Risk Assessment

Etapas en la preservación

1. Archivar los documentos digitales
gestión documental
2. Preservar el *bitstream*
3. Garantizar el acceso a largo plazo

Los supuestos de la preservación digital

- Los datos se mantendrán en el repositorio sin sufrir daños, sin perderse o sin ser alterados de forma malintencionada/o no.
- Los datos podrán ser localizados y entregados al usuario.
- Los datos podrán ser interpretados y comprendidos por el usuario.
- Las metas 1, 2 y 3 serán realizables a largo plazo sobre los objetos digitales elegidos y con distintos plazos.

Objeto digital

Acciones en su ciclo de vida para mantener el acceso

OD Y METADATOS DE PRESERVACIÓN

Debe mantenerse
en el repositorio
de manera **segura**

Deben guardarse
las relaciones que vinculen
al objeto con otros

El repositorio debe tener
los derechos suficientes
para sostener el **acceso** al objeto

Si hay un cambio debe
saberse **quién** lo efectuó



Autenticidad
Mediante la documentación
de su procedencia

Debe conocerse
su **creador**

Debe poder ser **localizado**
y **entregado** al usuario

Su soporte deber ser **compatible**
con los sistemas actuales

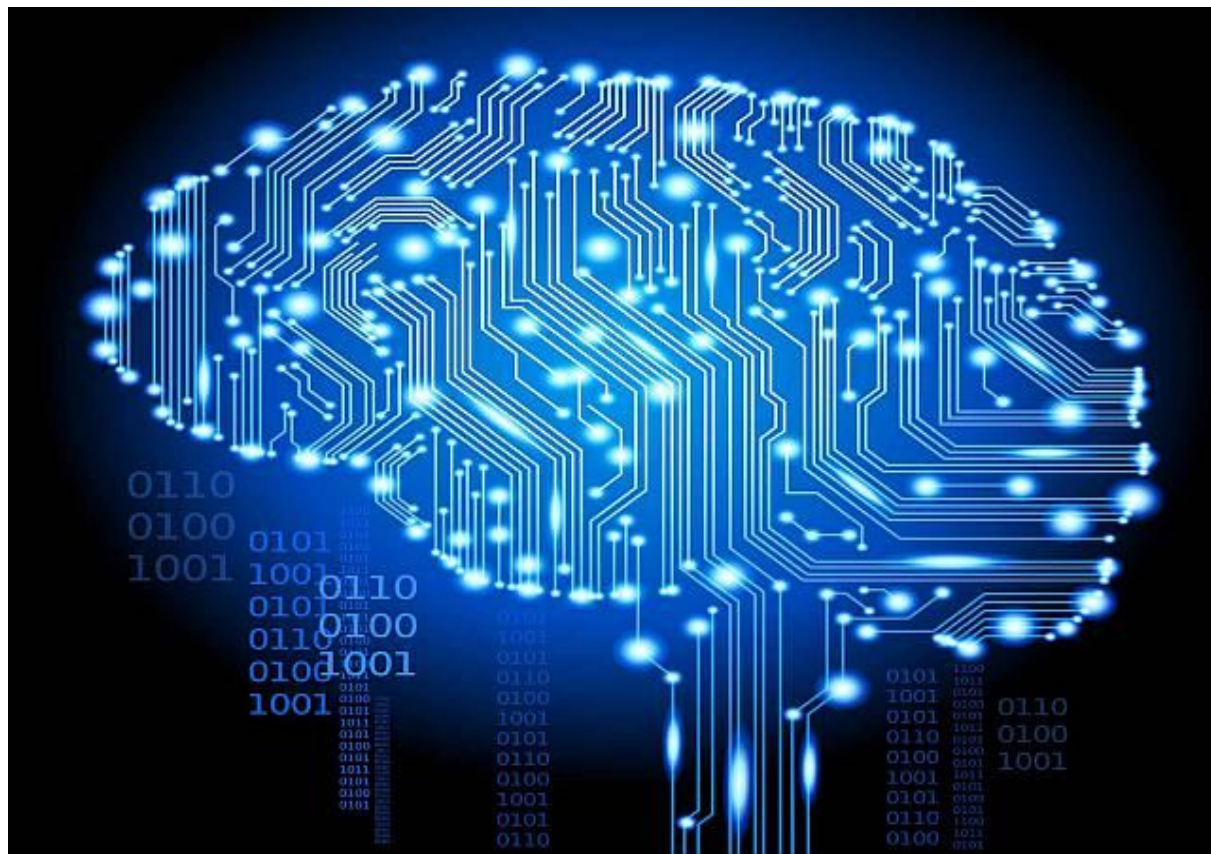
Las estrategias de **emulación** y **migración**
requieren datos sobre los objetos originales
y sus entornos

“UNA METODOLOGÍA DE EVALUACIÓN DE REPOSITORIOS DIGITALES PARA ASEGURAR LA PRESERVACIÓN EN EL TIEMPO Y EL ACCESO A LOS CONTENIDOS”

Autora: Ing. Marisa R. De Giusti

Directora: Dra. Silvia Gordillo

Tecnologías



Preservación de contenido: aspectos técnicos

- Hay una muy importante necesidad de preservar el contenido digital en el tiempo, con el objetivo de conservarlo accesible frente a riesgos como: incendios, inundaciones, robos, problemas de hardware (rotura de discos, etc.) y cambios tecnológicos constantes.
- *Es un proceso continuo.*

Obsolescencia

Es el resultado de la evolución de las tecnologías: a medida que surgen nuevas tecnologías, las viejas van quedando en desuso y se vuelven obsoletas.



Mantener tecnologías obsoletas en funcionamiento puede ser justificado en casos particulares, pero no en la mayoría.

Cornell University Library creó la "Cámara de los horrores"

<http://dpworkshop.org/dpm-eng/oldmedia/chamber.html>

Preservación de contenido. “Obsolescencia digital”

Mantener tecnologías obsoletas requiere conservar

- Hardware
- Software (aplicaciones, librerías, sistema operativo, etc)
- Documentación (manuales, instructivos, etc)
- Personal con la capacitación y las habilidades necesarias para trabajar en ese entorno obsoleto

Suelen ser opciones muy difíciles de mantener y muy costosas.

Preservación de contenido. Estrategias

Las formas de atacar los problemas de preservación, y en particular los problemas de obsolescencia, son:

- Migración
- Adhesión a estándares internacionales
- Emulación
- Encapsulamiento
- Metadatos de preservación
- Políticas de backup

Preservación de contenido. Migración continua

Migrar la información de una tecnología a la siguiente de forma continua, evitando así la obsolescencia.

- Es una de las opciones de mayor uso
- Asegura el acceso en todo momento (los datos son siempre accesibles mediante una tecnología actual)
- Requiere transformación de los datos originales
- Decisiones sobre qué se desea preservar

Preservación de contenido. Encapsulamiento

Se basa en agrupar cada objeto a preservar junto con todos los elementos (incluso software) necesarios para asegurar su acceso en el tiempo.

Como elementos a encapsular podemos tener:

- Especificaciones del formato de archivo.
- Instructivos relacionados a la emulación necesaria.
- Información de configuración de alguna herramienta en particular.
- Software de emulación.
- Especificaciones de hardware.

Preservación de contenido. Emulación

Se trata de imitar las características y capacidades de un software y/o hardware, de modo que los procesos "crean" que están funcionando en la plataforma original.

- No hay necesidad de modificar los datos originales (como en la migración), manteniendo la integridad de la información.
- Una vez que se archivaron los datos, solo hay que asegurarse que el soporte físico utilizado siga siendo accesible.
- Se puede usar un mismo emulador para múltiples objetos del mismo tipo.

Preservación de contenido. Análisis forense digital

Las técnicas y metodologías de preservación digital pretenden que todos los archivos serán correctamente tratados y por lo tanto interpretados en el futuro.

A pesar de todo, habrá archivos que llegarán al futuro sin tratarse.

A menudo se trata de sistemas de almacenamiento que ya no pueden ser leídos: Discos lomega, discos de 5 ¼, que, a su vez guardan archivos grabados en formatos antiguos: WordStar...

Se debe poder leer el soporte y recuperar el contenido!

Preservación de contenido. Análisis forense digital

Las técnicas aplicadas de recuperación a veces se conocen con el nombre de arqueología digital, pero desde un punto de vista riguroso son un subconjunto de las técnicas de análisis forense digital, normalmente usadas por las fuerzas policiales y por privados para investigar delitos en el ámbito informático.



Preservación de contenido. Análisis forense digital

EnCase: es una herramienta forense muy completa que permite hacer proceso de carving (recuperación de archivos borrados) sobre los dispositivos, que permitiría acceder a datos a muy bajo nivel, como fragmentos de texto, incluso cuando ya no sea posible acceder al archivo del cual formaban parte.

Es muy complejo de utilizar y muy caro. El negocio de la empresa es tanto la venta de la licencia (con llave por hardware) como la venta de las capacitaciones, que rondan las decenas de miles de dólares.

Preservación de contenido. Análisis forense digital

Magnet IEF y Magnet AXIOM: son también productos comerciales, pero no permiten realizar un trabajo tan minucioso como EnCase. Encuentra archivos de tipos conocidos (de texto, documentos de office, imágenes, videos, etc.) y tiene una colección muy grande de artefactos de navegación de internet y correos electrónicos. También es bastante caro y requiere llave por hardware, pero es mucho más simple de utilizar y se puede pedir una versión de prueba.



Preservación de contenido. Análisis forense digital



Caine: es una distribución de Linux que contiene un conjunto bastante amplio de herramientas forenses. Se puede probar sin mucho esfuerzo la herramienta Autopsy, que viene también para Windows y tiene una interfaz gráfica simil EnCase, aunque todo open source. Sirve para recuperar archivos borrados, ver artefactos de uso de la PC, y hacer carving.

Preservación de contenido. Análisis forense digital

DEFT: distribución de linux similar a Caine.



Recuva: se usa mucho para recuperar archivos eliminados, una vez terminado el proceso, se puede filtrar por tipo o por extensión de lo que se quiera recuperar. Es gratis y sólo hace eso.



**Normas, estándares,
recomendaciones**



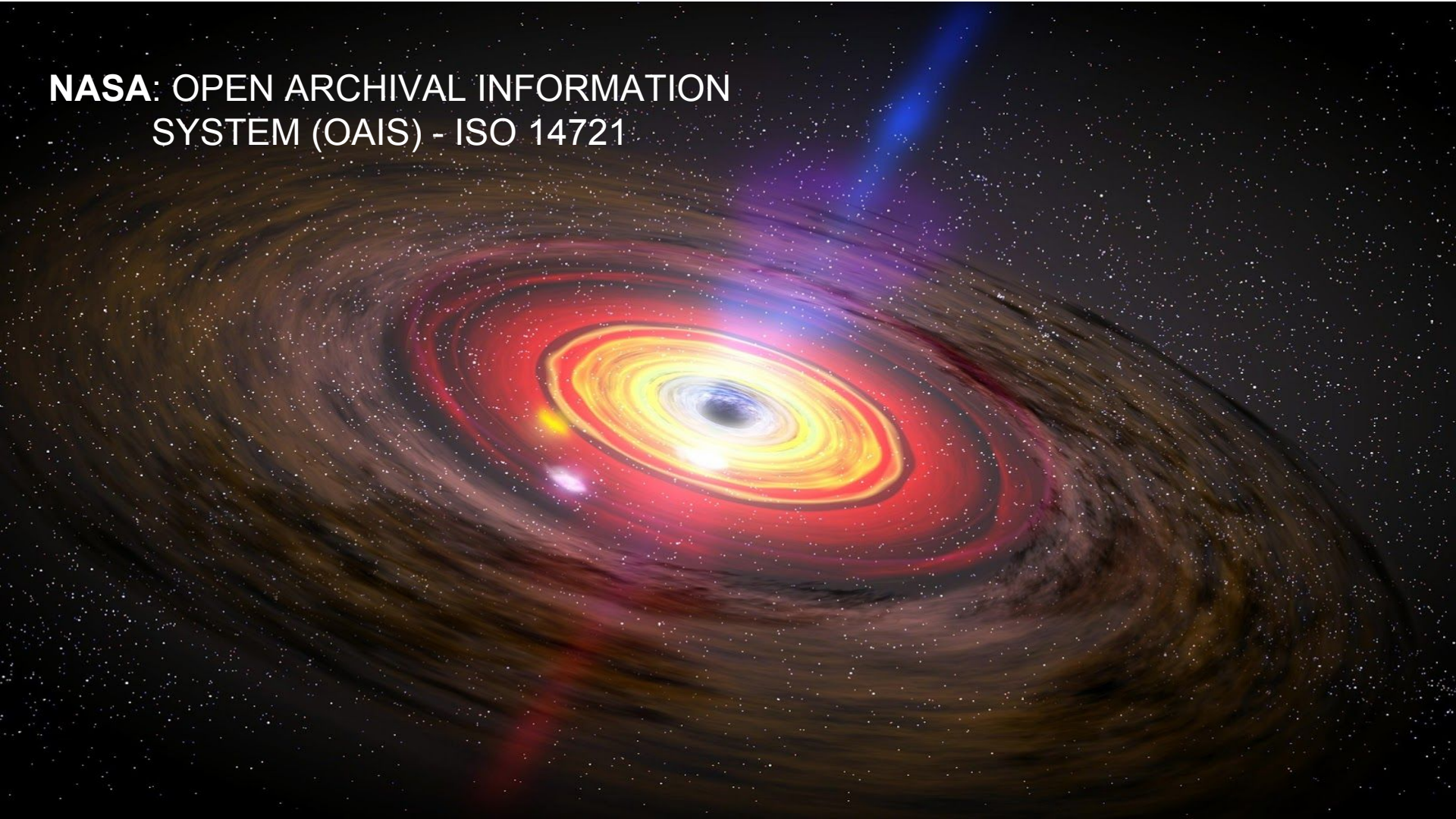
Noción de preservación de UNESCO



“La preservación digital puede definirse como el conjunto de los procesos destinados a garantizar la continuidad de los elementos del patrimonio digital durante todo el tiempo que se consideren necesarios”.

“La mayor amenaza para la continuidad digital es la desaparición de los medios de acceso. No puede decirse que se han conservado los objetos digitales si, al haber dejado de existir los medios de acceso a ellos, resulta imposible utilizarlos. El objetivo de la preservación de los objetos digitales es mantener su accesibilidad, es decir, la capacidad de tener acceso a su mensaje o propósito esencial y auténtico”. (UNESCO, 2003: p. 37)

**NASA: OPEN ARCHIVAL INFORMATION
SYSTEM (OAIS) - ISO 14721**



El Modelo OAIS

Modelo de Referencia
para un Sistema Abierto de
Archivo de Información.

ISO 14721: 2012

ISO Reference Model
of an Open Archival
Information System (OAIS).

El Modelo OAIS

- Archivo que comprende una organización de personas y sistemas que han asumido el compromiso de preservar a largo plazo y hacer disponible un determinado corpus de información (cualquier tipo de conocimiento a intercambiar) para una comunidad designada.
- Se refiere a la información analógica y a la digital, pero el foco está en esta última.
- Open (abierto): se usa para indicar que esta recomendación ha sido realizada en foros abiertos. No significa que el archivo es de acceso gratuito o irrestricto. Puede ser cualquiera.

El modelo de Referencia OAIS

1. Introducción: propósitos, alcance, campo de aplicación, razones, conformidad, estándares relacionados y definiciones.
2. Conceptos: Medioambiente, información e interacciones externas de alto nivel.
3. Responsabilidades: obligatorias y deslindes.
4. Modelo: funcional, de información, transformaciones.
5. Preservación: de la información y del acceso a la información.
6. Interoperabilidad.

Sección 1

Justificación del Modelo de referencia

- Ninguna discusión sobre la conservación de repositorios y flujos de trabajo estaría completa sin al menos una breve introducción al modelo de referencia OAIS.
- Una introducción a este modelo sirve para mostrar cómo implementa muchos de los procesos de flujos de trabajo y cómo se relaciona con la conservación digital.
- **Se recomienda como la mejor práctica actual.**

Antecedentes

- El Comité Consultivo para los Sistemas de Datos Espaciales (CCSDS, por sus siglas en inglés), un foro para agencias nacionales espaciales interesadas en desarrollar acuerdos de cooperación sobre normas de gestión de datos en la investigación espacial, llevó a cabo el desarrollo inicial de esta norma para permitir el almacenamiento de datos digitales a largo plazo, generados a partir de las misiones espaciales.
- En colaboración con la Organización Internacional para la Normalización ISO, el modelo de referencia fue aprobado como norma ISO en 2002 (ISO-14721).

Funciones del Modelo de referencia

- Las dos funciones principales del modelo son **conservar** la información y **garantizar el acceso** a la misma.
- El modelo funcional OAIS, que se propone lograr estos objetivos amplios, en cierta medida, define la arquitectura aproximada de cualquier tipo de sistema de software diseñado para cumplir con esta norma y con todo tipo de flujos de trabajo asociados con el repositorio.

Propósito y campo de aplicación

- Es aplicable para cualquier archivo, pero especialmente está enfocada en organizaciones con responsabilidad de hacer que la información esté disponible a largo plazo para una **comunidad designada**.
- Es de interés para aquellos que crean información que puede necesitar preservación a largo plazo.
- No especifica un diseño o una implementación. Cada implementación dará lugar a una funcionalidad distinta.
- El foco primario es la información inherentemente digital.
- El modelo se acomoda para información que no es inherentemente digital pero el modelo y la preservación de esa información no está descrito en detalle.

Propósito y campo de aplicación

- Estandariza las relaciones y los componentes de un sistema de archivos. Es un framework que sirve para entender mejor de qué se habla.
- Establece un vocabulario común.
- Ofrece un marco consensuado internacional para la definición de entidades, procesos y funciones de los archivos de datos.
- Facilita comprender y aplicar conceptos necesarios para la preservación de información digital a largo plazo.

Sección 2

Conceptos en OAIS

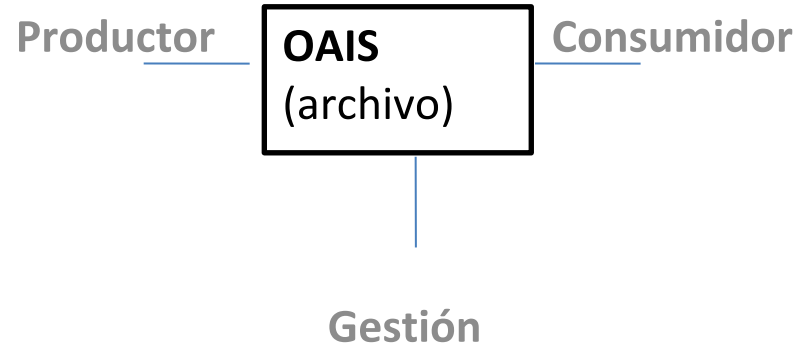
El propósito de esta sección es motivar y describir varios conceptos clave, de alto nivel del OAIS. Un punto de vista más completo y una modelización formal de estos conceptos, se da en la sección 4.

Medioambiente OAIS

- Un productor que provee la información.
- Una política global de gestión (management), NO las operaciones diarias.
- Un consumidor que busca, encuentra y adquiere la información de su interés que ha sido preservada.
- La comunidad designada es el conjunto de los consumidores que son capaces de comprender la información preservada.

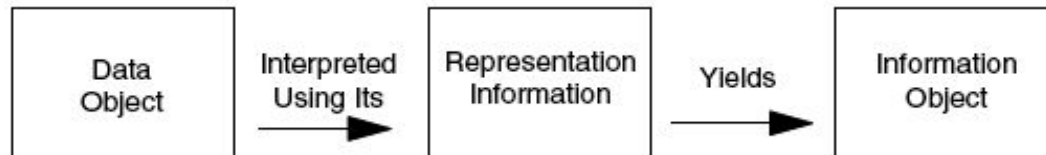
Actores en el modelo

- **Producer-Consumer-Management**



Conceptos en OAIS

Una definición clara de información es central para la capacidad del OAIS para preservar esa información. Una persona o un sistema, tienen una base común de conocimientos (KB) que le permite comprender la información. Se considera información en este campo a cualquier tipo de conocimiento que puede intercambiarse y que se expresa a través de algún tipo de datos: la información en un artículo periodístico, se expresa por caracteres (datos), los cuales bajo el paraguas de un lenguaje (KB), se convierten en información relevante. Si el receptor desconoce la lengua, entonces el artículo tendrá que ser acompañado por información extra, por ejemplo, un diccionario o una gramática.

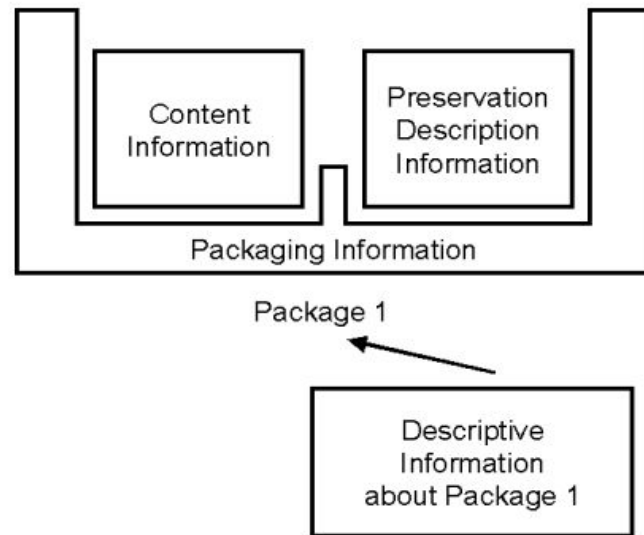


Conceptos en OAIS

- A fin de que este objeto de información se preserve con éxito, es fundamental para un OAIS identificar con claridad y comprender los objetos de datos y la representación de la información asociada.
 - Para la información digital, esto significa que el OAIS debe identificar claramente los bits y la representación de la información que se aplica a los bits.
- El OAIS debe entender la base de conocimientos de su comunidad determinada/designada para comprender la representación de la información mínima que debe mantenerse.

Conceptos en OAIS

La unidad de intercambio entre un OAIS y su medioambiente es el paquete de información, Information Package (IP).



ISO 14721: Fig 2-3: Paquete de información: conceptos y relaciones

Conceptos en OAIS

Un IP contiene 2 tipos de información:

- De contenido.
- De descripción de preservación (PDI).
- La información de contenido y la PDI pueden verse como encapsuladas e identificables por medio de la información de empaquetado.
- El paquete resultante es recuperable en virtud de la información descriptiva: DI.

Conceptos en OAIS

- La información de empaquetado es la información que, ya sea real o lógicamente, une, identifica y relaciona la información del contenido y la PDI.
- La información descriptiva es la información que se utiliza para descubrir qué paquete tiene la información de contenido de interés.

Estructura del Paquete de Información



El paquete de información (IP)

La norma define el IP como un contenedor conceptual con dos tipos de información: de contenido y de preservación. La *información de contenido (CI)* es el objeto mismo que se desea mantener en el tiempo y la *información descriptiva de preservación (PDI)*, debe brindar datos suficientes sobre la **procedencia**, el **contexto**, la **referencia**, la **integridad** y los **derechos de acceso**.

Elementos de la PDI

La **procedencia**, más allá de describir la fuente, incluye los procesos que se han realizado sobre la información: la historia del objeto, cambios, versiones y responsables. El **contexto** muestra las relaciones con otras fuentes de información o contenidos. La **referencia** provee una identificación única del contenido. La **integridad (o fijeza)** provee una protección para que la información no sea alterada de manera intencional /no. Los **derechos de acceso** proveen información sobre los términos de acceso incluyendo preservación, distribución y uso de la información de contenido.

Conceptos en OAIS

- Variantes del paquete de información:
 - Submission Information Package (**SIP**)
 - Archival Information Package (**AIP**)
 - Dissemination Information Package (**DIP**)
- Los paquetes de información variarán dependiendo de su rol:
 - Por ejemplo master file y versiones derivadas (thumbnails, JPEG, PDFs...).

Clases de IP según su función

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.

Clases de IP según su función

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 representados o contenidos.

Clases de IP según su función

Dissemination Information Package (**DIP**): es el paquete que se entrega a un consumidor en respuesta a una solicitud. La información de empaquetado toma muchas formas dado que los usos de OAIS son diversos, puede ser tan completo como los AIP a partir de los cuales se construye o ser sólo una breve descripción del paquete.

OAIS interacciones externas de alto nivel

La figura que sigue es un diagrama de flujo de datos que representa los flujos de información entre productores, consumidores y el OAIS y no incluye flujos que involucren al management.

OAIS interacciones externas

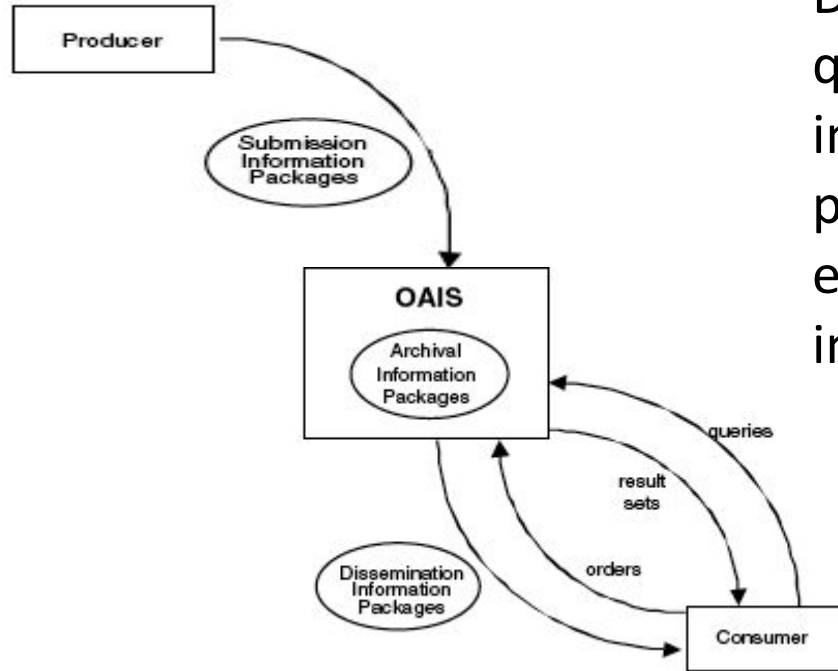


Diagrama de flujo de datos que representa los flujos de información entre productores, consumidores y el OAIS y no incluye flujos que involucren al management.

Visión de alto nivel de las interacciones en un entorno OAIS

- Interacción de la gestión
 - financiación, utilización de recursos, pagos, resolución de conflictos.
- Interacción del productor
 - los acuerdos de ingesta. Acuerdo por los SIPs que va a mandar, tiempo (acuerdo por data submission session)
- Interacción de los consumidores
 - Ayudas, descubrimiento de información, ordenamiento de la información. (Data dissemination session).

Sección 3: Responsabilidades del OAIS

OAIS Responsabilidades obligatorias

- Negociar y aceptar la información adecuada de los productores de información.
- Obtener un control de la información proporcionada al nivel necesario para asegurar la conservación.
- Determinar, por sí o con otras partes, cuáles comunidades deben convertirse en la comunidad designada y, entender la información proporcionada.
- Asegurar que la información que se conserva es comprensible independientemente de la comunidad designada.

OAIS Responsabilidades obligatorias

- Cumplir con las políticas y procedimientos documentados para:
 - garantizar que la información se conserva en contra de todas las contingencias razonables, y
 - permitir que la información sea difundida como copia autenticada de la original, o como **trazabilidad** a la original.
- Lograr que la información preservada esté disponible para la comunidad designada.
 - Secciones 3.1 y 3.2 de la norma.

Mecanismos de deslinde de responsabilidades

1. Las negociaciones para y la información que acepta.
2. El control eficiente de la preservación.
 - Cuestiones de copyright, propiedad intelectual y restricciones legales para el uso.
 - Autoridad para modificar el modo de representación de la información.
 - Acuerdos con organizaciones externas.
3. Determinación de la comunidad designada de consumidores.

Sección 4

OAIS

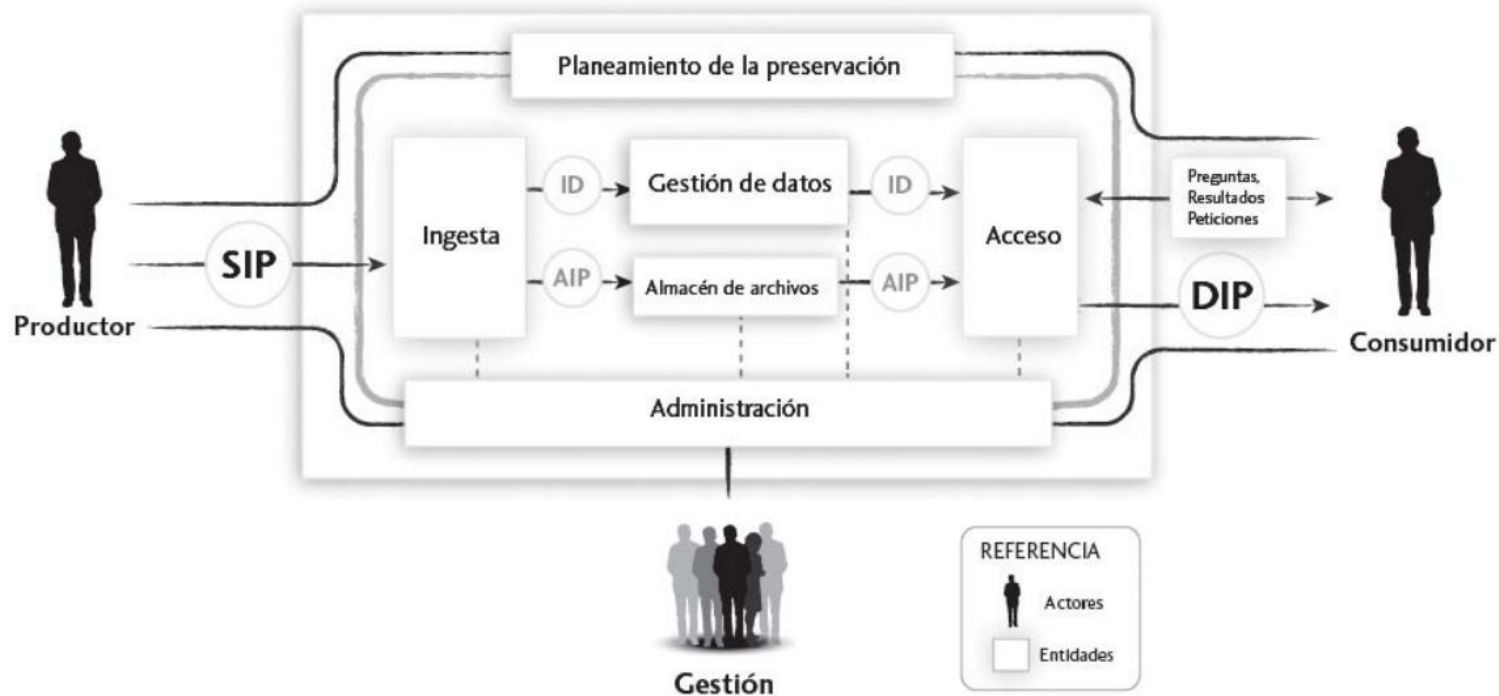
Modelo Funcional

Sección 4.1

OAIS Modelo funcional

Seis entidades funcionales e interfaces relacionadas:

- Ingesta- Ingest
- Almacenamiento de archivos-Archival storage
- Gestión de datos-Data management
- Administración-Administration
- Planeamiento de la preservación-Preservation Planning
- Acceso- Access



Modelo OAIS

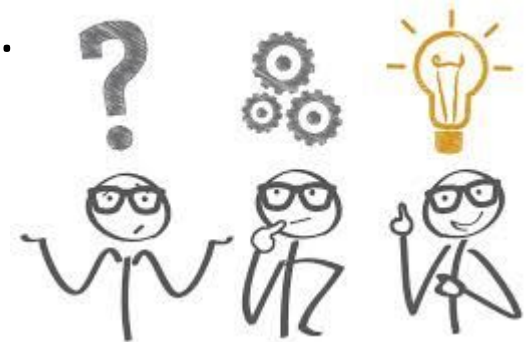
- El productor suministra el SIP a través del *ingest*, que luego se convierte en AIP terminando en la entidad *archival storage*.
- El consumidor busca una información en el sistema, que es entregada como un DIP a través de la entidad *access*, ya que la información está preservada en el sistema previamente.
- Los datos relacionados con los documentos y el repositorio mismo se mantienen organizados a través de la entidad *data management*.

Modelo OAIS

- El módulo de *planificación de la preservación* desarrolla estrategias y normas de conservación, monitorea las últimas novedades y avances en el campo, y monitorea los cambios en la comunidad designada, para que toda la información nueva que se solicite, se pueda adjuntar a los AIP correspondientes.
- *administration* se dedica a la administración y gestión (administradores y responsable del repositorio) y se relaciona con las secciones de ingesta, *gestión de datos*, *almacenamiento de archivos* y *planificación de la preservación*.

Problemas en la preservación: software

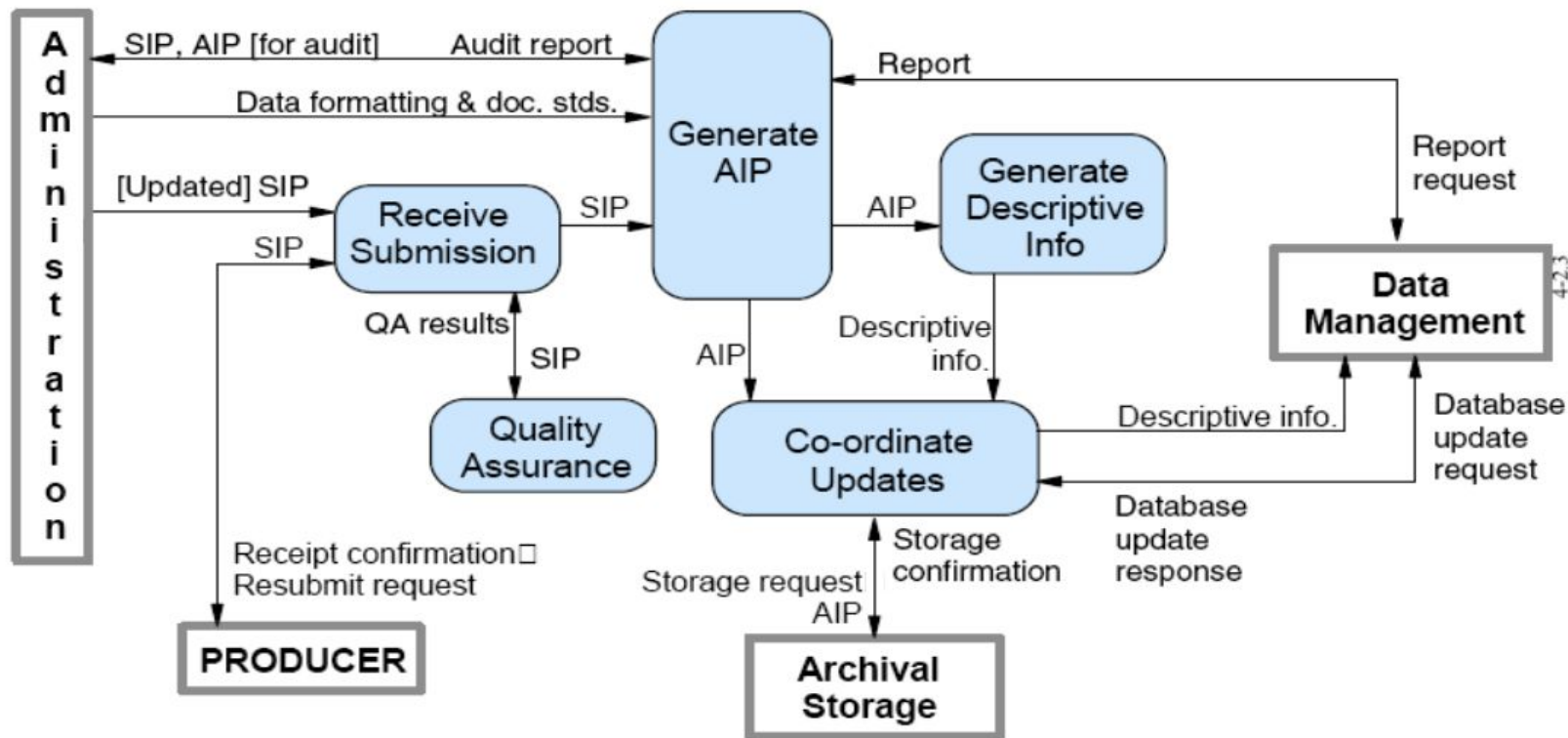
- Muchos problemas en lo relativo a la preservación derivan de una configuración deficiente del software que soporta el repositorio. Es necesario revisar las facilidades del software que soporta el repositorio en comparación con el modelo de preservación OAIS y realizar las personalizaciones necesarias para cumplir con algunos requerimientos del plan de preservación no brindados de forma nativa.



OAIS y DSpace

- DSpace está influenciado por el modelo OAIS, en términos de metodología y funciones. Siempre que es posible, adopta el modelo y vocabulario OAIS.
- El modelo de datos, los metadatos y el módulo de autorizaciones cumplen en gran medida con el modelo de referencia.
- Soporta metadatos de preservación e identificadores persistentes (Handle, DOI) aunque no dispone de un módulo de planeamiento de la preservación.
- Referencias: <https://groups.google.com/forum/#!msg/dspace-community/SP5Z9vwG9kY/4nL9pXMXBAAJ>
 - DSpace as an Open Archival Information System: Current Status and Future Directions, Tansley, Robert; Bass, Mick; Smith, MacKenzie, 2003, <http://hdl.handle.net/1721.1/29464>

Entidad OAIS Ingest



Entidad OAIS Ingest - Descripción

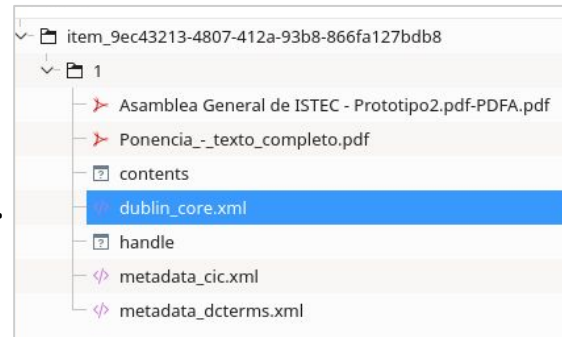
- Provee los servicios y funciones para aceptar el paquete de información (SIP) presentado por los Productores (o a partir de elementos internos bajo control de la administración).
- Prepara los contenidos para almacenamiento y gestión dentro del archivo.
- Realiza el aseguramiento de calidad/validación 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.

Entidad Ingest en DSpace

- La ingesta puede realizarse por importación en masa, cosecha, depósito SWORD o proceso de carga tradicional.
- El proceso de carga tradicional o Submission process, permite configurar para cada colección y para cada tipología documental qué metadatos se requieren de forma opcional u obligatoria, y cuál es la secuencia de etapas que debe cumplir el ítem antes de ingresar al repositorio.
- La cosecha se realiza sobre el protocolo OAI-PMH usando configuraciones específicas al repositorio para adaptar los recursos recolectados al metadata profile interno.
- Los depósitos SWORD se realizan desde clientes autenticados en puntos de depósito autorizados y con un esquema de metadatos preacordado.

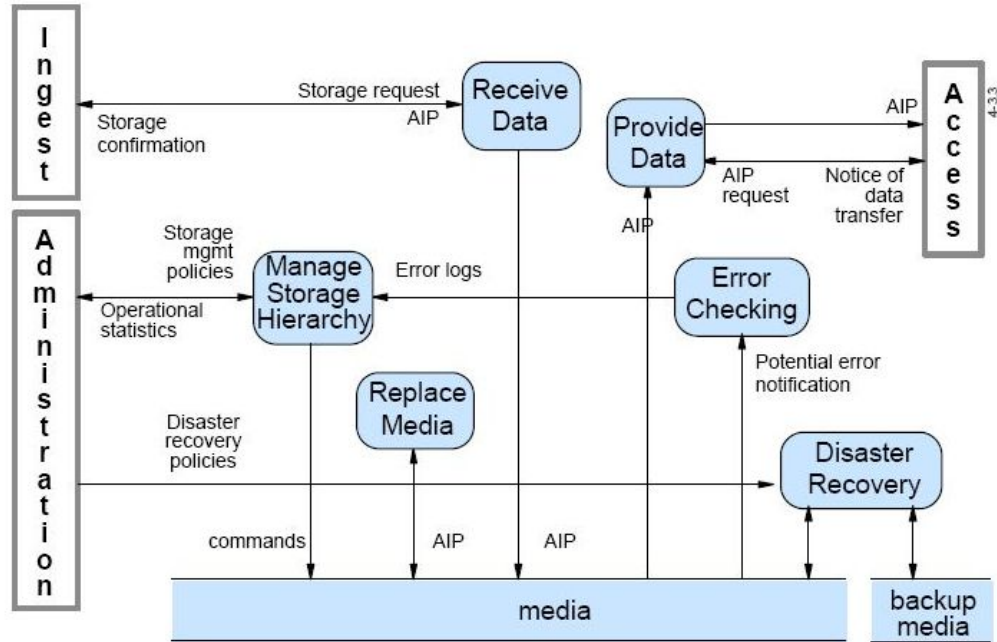
Entidad Ingest en DSpace

- La importación toma SIPs basados en diversos formatos:
 - SimpleArchiveFormat: se basa en un contenedor ZIP con archivo índice, metadatos y binarios,
 - METSSIImporter en formato METS o,
 - en cualquier otro formato, si se desarrolla un packager plugin ad-hoc.



- En todos los casos, al ingestar un recurso con un componente binario (un bitstream) **se hace la extracción de ciertos metadatos técnicos** que luego serán usados por las demás entidades del sistema.

Entidad OAIS Archival Storage



Entidad OAIS Archival Storage - Descripción

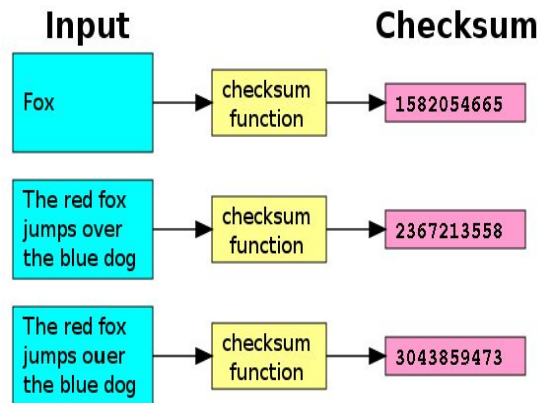
- Provee los servicios y funciones para el almacenamiento, mantenimiento y recuperación de los AIPs.
- Recibe el AIP de la entidad ingest y lo almacena.
- Gestiona las jerarquías de almacenamiento.
- Configura niveles especiales de servicio, seguridad y protección (por ejemplo backups).
- Provee estadísticas 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 copia de los AIPs almacenados a la entidad *access*.

Entidad Archival Storage en DSpace

- Los datos de las entidades, sus permisos y los metadatos se almacenan en una base de datos relacional típica (PostgreSQL u OracleDB).
- Los binarios se almacenan en el sistema de archivos en el servidor (assetstore) o de forma externa, por ejemplo espacio en la nube de Amazon S3.
- El assetstore es una gran jerarquía de directorios y archivos sin encriptación $\Rightarrow$ un copiado directo de estos archivos en cualquier otro entorno permitiría leer su contenido desde otro software (ej. Droid).

Entidad Archival Storage en DSpace

- DSpace permite ejecutar un proceso denominado Checksum Checker que realiza controles de integridad sobre los binarios (bitstreams) de cada ítem. Este mecanismo detecta cambios inesperados en el contenido de cada archivo y, ya sean cambios accidentales o malintencionados, permite actuar y recuperar el binario original de una copia de seguridad.
- Es posible importar y exportar AIPs completos de forma muy simple, generando paquetes totalmente autocontenidos para ítems, colecciones, comunidades e incluso para todo el repositorio. A diferencia de los SIP y DIP, estos AIP contienen todos los datos sobre el recurso en el repositorio.



Entidad Archival Storage en DSpace

En cuanto a la recuperación ante errores o desastres (*disaster recovery*) DSpace no cuenta con un mecanismo automático.

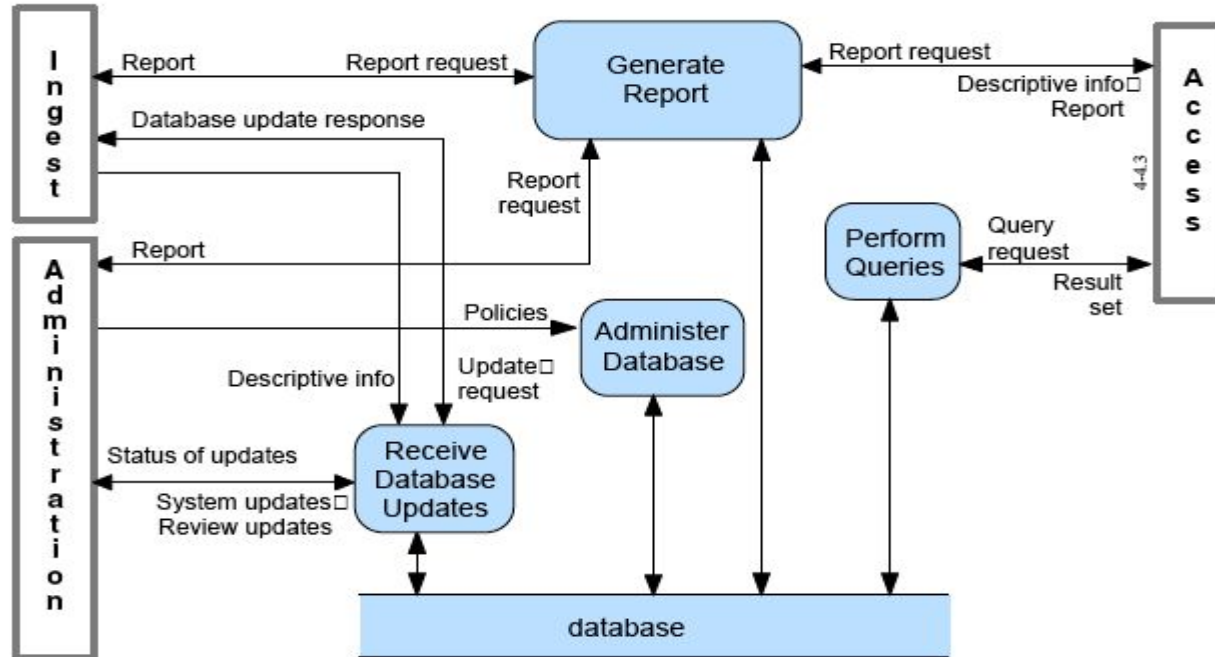
Ante sucesos que generen desastres, el *plan de contingencia* queda a manos de los administradores de los servidores del repositorio. Éstos deben de prever la realización de backups periódicos así como la correctitud de su información.

Mediante estos backups, el repositorio podría recuperarse a un estado anterior, previo al evento del desastre.

Se debe considerar la ejecución frecuente de

- Backup de assetstore
- Backup de base de datos
- Backup de configuraciones
- Backup de registros de uso estadísticos
- Rotación y preservación completa de archivos de logs (para reducir errores de poco espacio en disco)

Entidad OAIS Data Management



Entidad OAIS Data Management

- **Descripción:** provee los servicios y funciones para poblar, mantener y acceder a la información descriptiva que identifica, documenta el contenido del Archivo, y a los datos administrativos usados para gestionarlo.
- **Funciones:**
 - Es responsable de la administración de la base de datos.
 - Recibe solicitudes de la entidad *access* y genera resultados.
 - Recibe pedidos de las entidades *ingest*, *access* y *administration* y genera reportes: p.e.: resúmenes de fondos o estadísticas de uso.
 - También recibe actualizaciones de *ingest* y *administration*.

Entidad Data Management en DSpace

- DSpace dispone de un módulo de estadísticas de uso que permite registrar el uso de sus contenidos, así como de sus servicio de búsqueda y depósito (*workflow*). A partir de estos registros, permite la generación de tablas de reportes.
- Para la búsqueda de contenidos en el repositorio, se dispone del servicio Discovery, que permite búsqueda mediante términos libres, aplicación mediante filtros y refinamiento facetado.
- DSpace permite comprobar el estado de los elementos que se encuentran en un repositorio a través de **tareas de curación** automáticas.

Tareas de Curation - ¿Qué son?

Una **tarea de curación** es un mecanismo que permite aplicar una acción determinada sobre los elementos del repositorio.

DSpace da soporte para definir tareas a partir de una API JAVA y provee algunas tareas predefinidas.

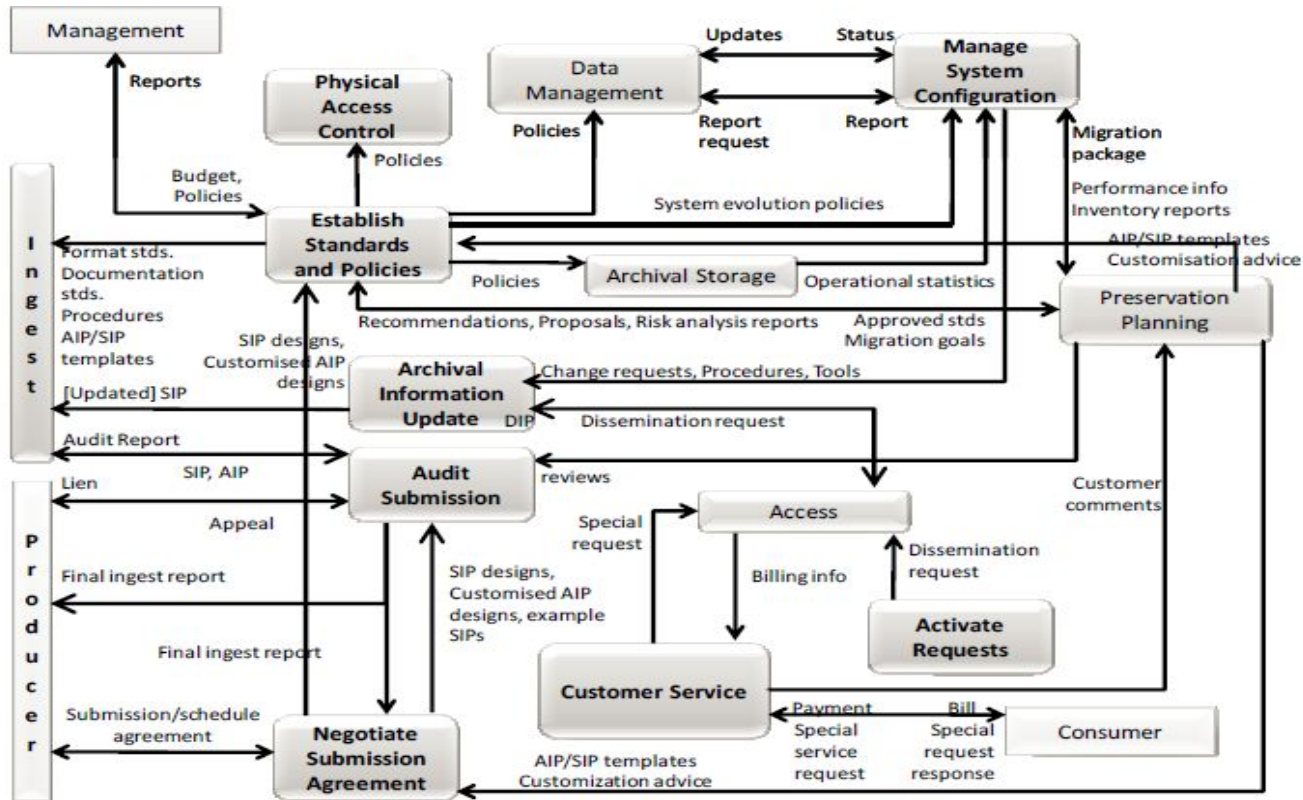
Las tareas de curation pueden ser:

- de sólo lectura o modificación de elementos, para control y transformación respectivamente.
- ejecutadas automática o manualmente
- de ejecución única o periódica.
- aplicadas sobre una colección o sobre todo el repositorio

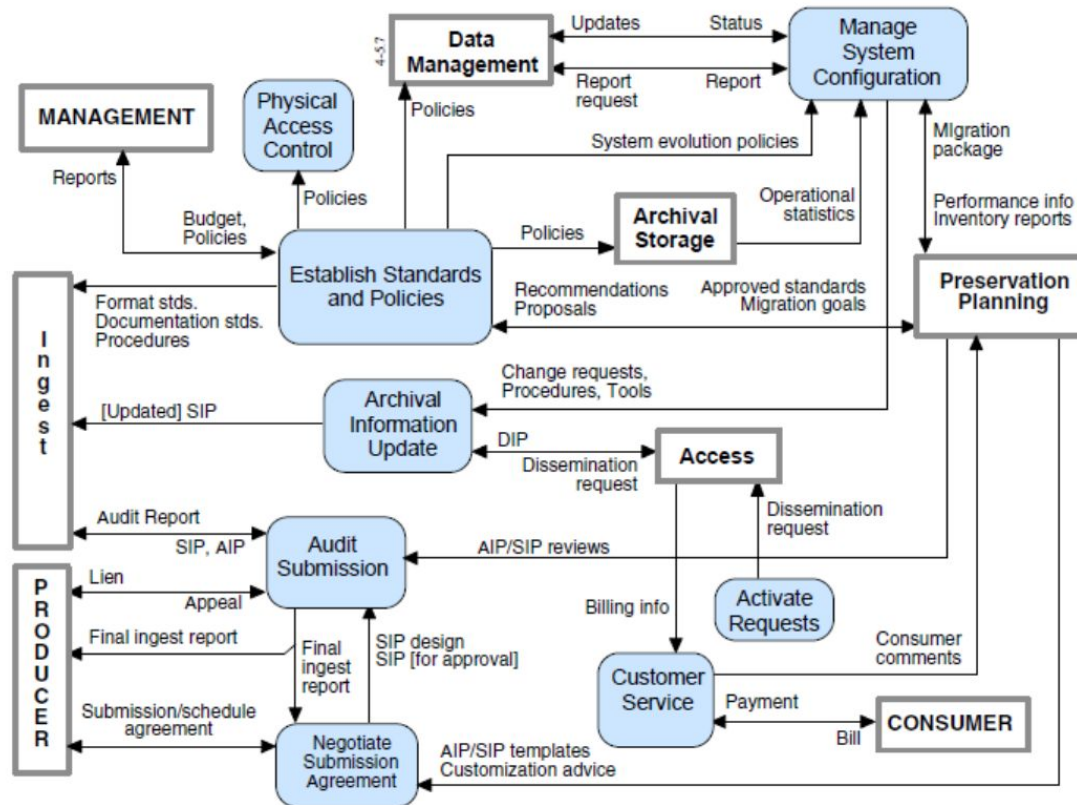
Tareas de Curation - ¿Qué permiten?

- Sobre metadatos:
 - chequear la presencia de metadatos obligatorios y reportar los faltantes. Ej: falta de dc.rights, dc.date, dc.creator, etc
 - validar metadatos existentes para:
 - repararlos (ej corregir caracteres,normalizar valores controlados)
 - dividirlos (en caso de múltiples valores)
 - eliminarlos (ej por duplicación),
- Sobre otros datos:
 - Evaluar el estado y validez de archivos, jerarquías, etc.
 - Recopilar estadísticas de un grupo de elementos o de todo el repo

Entidad OAIS Administration



Entidad OAIS Administration



Entidad Administration

Descripción: Provee los servicios y funciones para la operación global del sistema de Archivo.

Funciones:

- Solicita la información necesaria sobre los archivos y negocia los acuerdos con los Productores.
- Monitorea la funcionalidad del OAIS, 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.
- Provee información sobre performance e inventario a *preservation planning* para establecer políticas y estándares.
- Recibe los paquetes de migración de *preservation planning*.

Entidad OAIS Administration

Funciones (cont):

- Recibe los pedidos de cambio, procedimientos y herramientas para la actualización del archivo.
- Es 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.
- Desarrolla políticas de gestión de archivo por jerarquías, incluyendo políticas de migración.
- Es responsable de la recuperación ante desastres.

Entidad OAIS Administration

Funciones (cont):

- Verifica que los AIP y SIP suministrados sigan las especificaciones. En el caso de SIP y de AIP verifica la comprensión por parte de la comunidad designada.
- Verifica que la Información de representación y la PDI son adecuadas y comprensibles para la comunidad designada.
- Mantiene un registro de solicitudes y revisa periódicamente los contenidos del archivo para determinar si los datos están disponibles.
- Crea/mantiene/borra las cuentas de acceso de los consumidores.

Entidad Administration en DSpace - Authorization

DSpace presenta una sección general de Administración que permite:

- Gestión de cuentas de usuario y grupos
- Esquema de autorización basado en:
 - permisos para usuarios y grupos sobre los elementos del repositorio (comunidades, colecciones, ítems o bitstreams).
 - herencia de permisos sobre colecciones y comunidades a grupos y usuarios. Ej.: se asignan permisos de ADMIN a una comunidad específica para un grupo específico; automáticamente ese grupo tiene permisos de ADMIN sobre las colecciones hijas de esta comunidad
- Permite restringir el acceso a contenidos del repositorio mediante políticas permanentes o temporales de privacidad (embargo).

Entidad Administration en DSpace - Cambios

- Los **cambios** sobre el contenido del repositorio son hechos por una persona de acuerdo a los permisos que la misma posea. Ciertos cambios se pueden hacer sin control de autorización cuando se ejecutan desde el servidor, es decir, cuando es el sistema en sí el que ejecuta la acción.
- Por default no se mantiene registro de los cambios atómicos que se realiza sobre los items, ni de sus metadatos ni de sus bitstreams ni de sus permisos. Sólo se registra en el metadato provenance las transiciones dentro del workflow de edición.
- Es posible habilitar el módulo de versionado de ítems, el cual registra todos los cambios que sufre cada ítem en el repositorio.

Entidad Administration en DSpace: Estado

DSpace dispone de un **Panel de Control** (/admin/panel) en la interfaz de usuario para evaluar el estado general de DSpace:

- Version de JVM utilizada
- Memoria disponible
- Uso de la cache Java
- Parámetros de configuración activos
- Peticiones realizadas al repositorio
- etc.

Panel de control				
Información Java	Configuración de Dspace	Alertas del Sistema	Recolectando	Current Activity
PARAR el registro de la actividad de usuarios anónimos.				
EMPEZAR el registro de la actividad de bots.				
Actividad actual (máximo 250 páginas)				
Marca horaria	Usuario	Dirección IP	Página URL	Navegador
0 s	Facundo Gabriel Adorno	163.10.34.221	/admin/panel	Firefox
2 s	Analia Pinto	163.10.34.197	/handle/10915/50/submit/continue	Chrome
5 s	Analia Pinto	163.10.34.197	/handle/10915/50/submit/continue	Chrome
7 s	Coordinación de egreso FBA UNLP	163.10.39.130	/handle/10915/50/submit/continue	Chrome
15 s	Analia Pinto	163.10.34.197	/handle/10915/61601	Chrome

Entidad Administration en DSpace: Reportes

A partir de la versión 6 de DSpace, se dispone de una herramienta que envía mediante email reportes sobre el estado de “salud general” del repositorio mediante la realización de un conjunto configurable de chequeos:

- cantidad de ítems en workflow, workspaces, retirados, públicos, etc,
- cantidad de comunidades, colecciones, usuarios, grupos, etc.

Esta es una utilidad CLI se llama [HealtCheck](#), que debe activarse mediante tareas programadas del sistema (cronjobs).

```
Resource without policy: 1
Deleted bitstreams: 73
Orphan bitstreams: 0 []

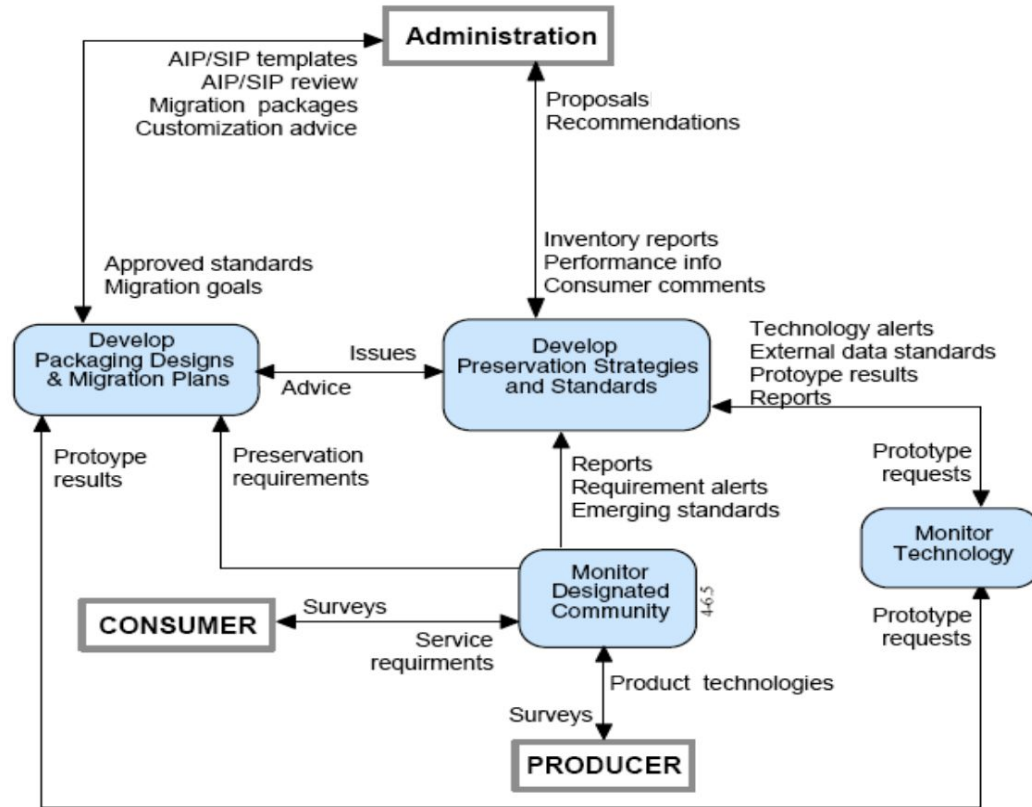
Published items (archived, not withdrawn): 1113
Withdrawn items: 137
Not published items (in workspace or workflow mode): 58
In Stage 1: 31
In Stage 2: 3
In Stage 3: 7
In Stage 4: 2
In Stage 5: 12
Waiting for approval (workflow items): 3
Count bitstream: 695
Count bundle: 286
Count collection: 3
Count community: 3
Count dcvalue: 21301
Count eperson: 208
Count item: 1308
Count handle: 1274
Count epersonsgroup: 15
Count workflowitem: 3
Count workspaceitem: 55
```


Entidad Administration en DSpace: Performance

No realiza control sobre la performance del sistema. Para esto hay que utilizar otras herramientas:

- herramientas propias del servidor (htop, df, du para evaluar el uso del procesador y espacio en disco)
- herramientas propias del gestor de base de datos (*pg_stat_statements* en PostgreSQL para evaluar los tiempos de ejecución por query, el uso del procesador para su resolución, etc.)
- herramientas del servidor (p.e. Tomcat Manager, JavaMelody, etc. para evaluar el uso de recursos de las distintas aplicaciones levantadas para el uso de DSpace-> *servidor oai, servidor solr, interfaz de usuario*, etc...)

Entidad OAIS Preservation Planning



Entidad OAI 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.

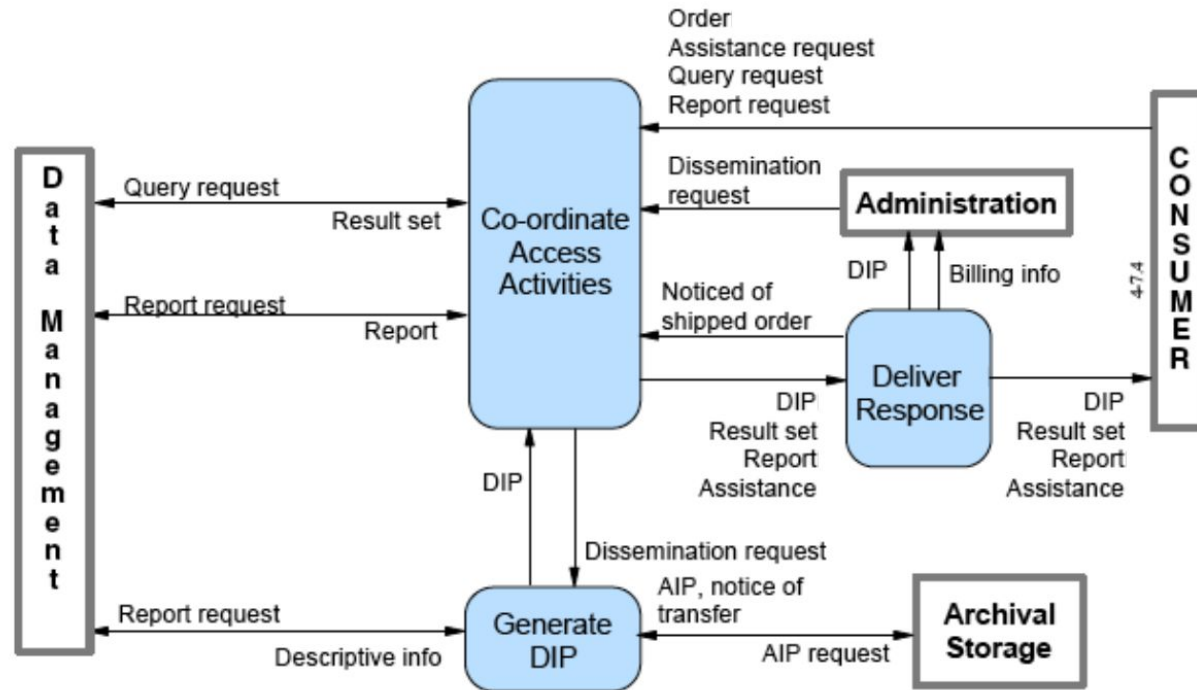
Entidad Preservation Planning en DSpace

DSpace no presenta por defecto ninguna de las funcionalidades relativas a la planificación de la preservación.

Se podría considerar:

- la creación de curation tasks que identifiquen formatos obsoletos
- la modificación *ad-hoc* del software para la implementación de reportes y alertas sobre el estado de obsolescencia general de los bitstreams.
- la vinculación de DSpace con otros sistemas que sí proporcionan nativamente dichas funcionalidades, como lo es el caso del software [Archivematica](#).

Entidad OAIS Access



Entidad OAIS 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*.

Funciones:

- 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.

Entidad Access en DSpace

DSpace provee diversos caminos para exportar los ítems:

- Exportación de metadatos (ej en CSV)
- Exportación empaquetada de metadatos y bitstreams
- Exportación empaquetada de metadatos, bitstreams e info administrativa

Entre los metadatos exportados puede haber algunos destinados a la preservación:

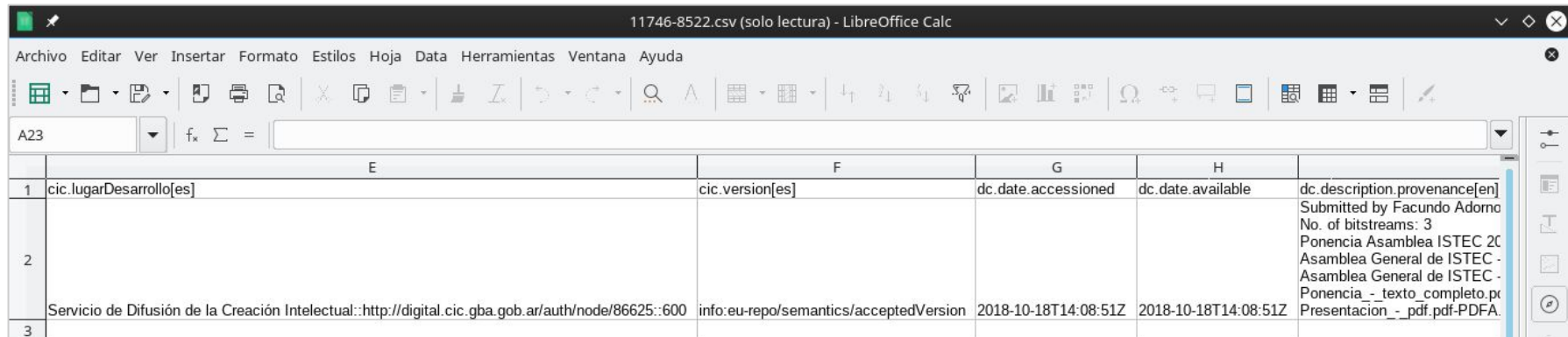
- `dc.description.provenance` presenta información sobre las acciones realizadas sobre el ítem para su depósito en el repositorio.
- `dc.rights` que contiene información de sobre la licencia de uso.
- `handle` identificador persistente del ítem
- `checksum de bitstreams` embebidos en el metadato `provenance`

Entidad Access en DSpace

Ante la exportación única de metadatos se devuelve un archivo CSV que contiene para una columna por metadato exportado.

Cada columna presenta el siguiente formato

`schema.element.[qualifier].[content_language]`



The screenshot shows a LibreOffice Calc window titled "11746-8522.csv (solo lectura) - LibreOffice Calc". The spreadsheet contains the following data:

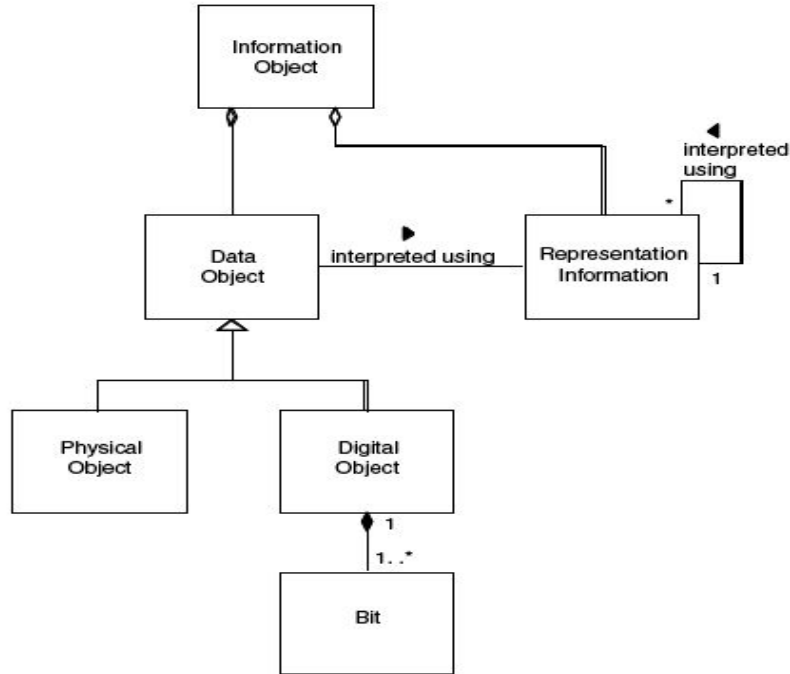
	E	F	G	H	
1	cic.lugarDesarrollo[es]	cic.version[es]	dc.date.accessioned	dc.date.available	dc.description.provenance[en]
2					Submitted by Facundo Adorno No. of bitstreams: 3 Ponencia Asamblea ISTEAC 20 Asamblea General de ISTEAC - Asamblea General de ISTEAC - Ponencia - texto_completo.pr Presentacion_-_pdf.pdf-PDFA.
3	Servicio de Difusión de la Creación Intelectual::http://digital.cic.gba.gob.ar/auth/node/86625::600	info.eu-repo/semantics/acceptedVersion	2018-10-18T14:08:51Z	2018-10-18T14:08:51Z	

OAIS

Modelo de Información

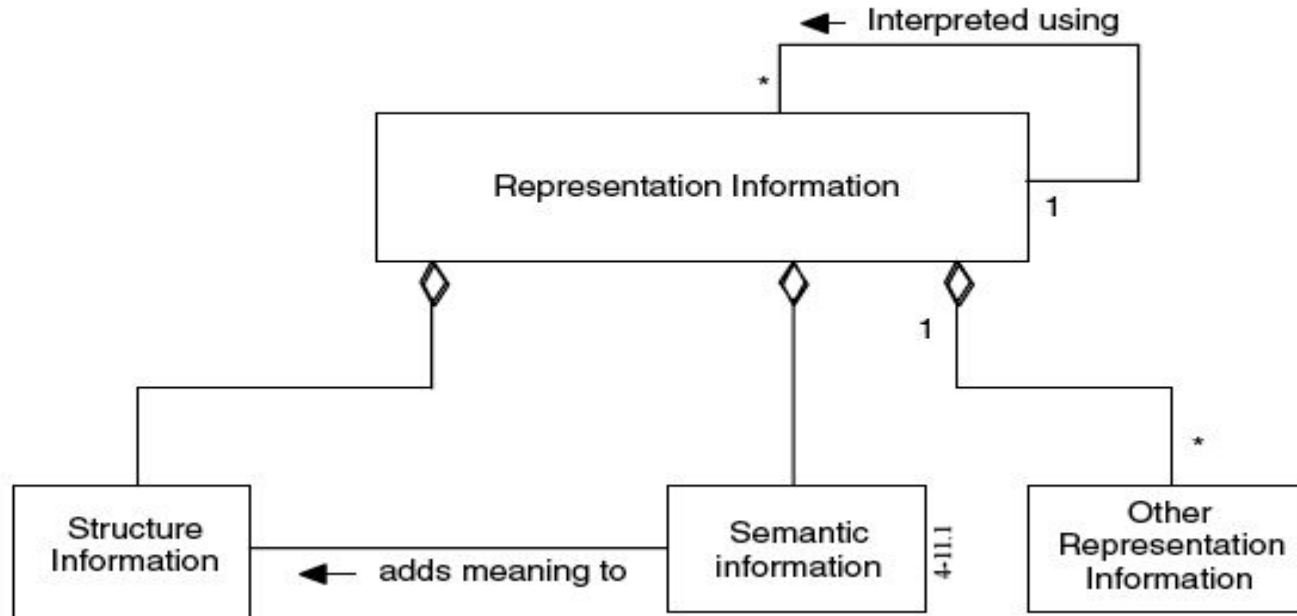
Sección 4.2

OAIS Objeto de información

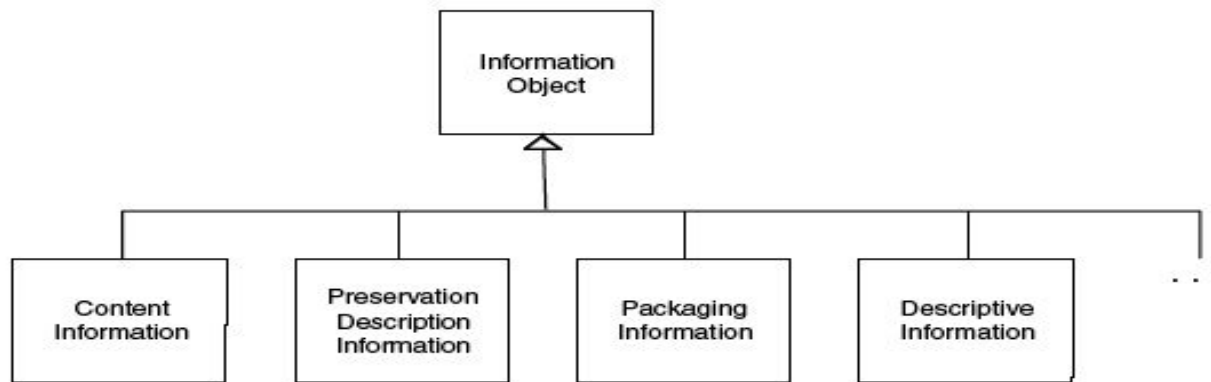


El **Objeto de Información** está compuesto de un Objeto de Datos, que puede ser físico o digital, e Información de Representación que permite la interpretación completa de los datos.

Representation Information Object



Tipos de objetos de información



Los objetos de información se clasifican por su contenido y función como : objetos de información de contenido, de descripción de la preservación, de empaquetado y de información descriptiva.

Información de contenido

- La información de contenido es el conjunto de información que es el objetivo original de la preservación del OAIS.
- La información de contenido es el contenido de datos del objeto, junto con su representación de la información.
- Los objetos de datos contenidos en la información de contenido puede ser un objeto digital o un objeto físico (por ejemplo, una muestra física de microfilm,).
- Cualquier objeto de información puede servir como información de contenido.

Información descriptiva de preservación (PDI)

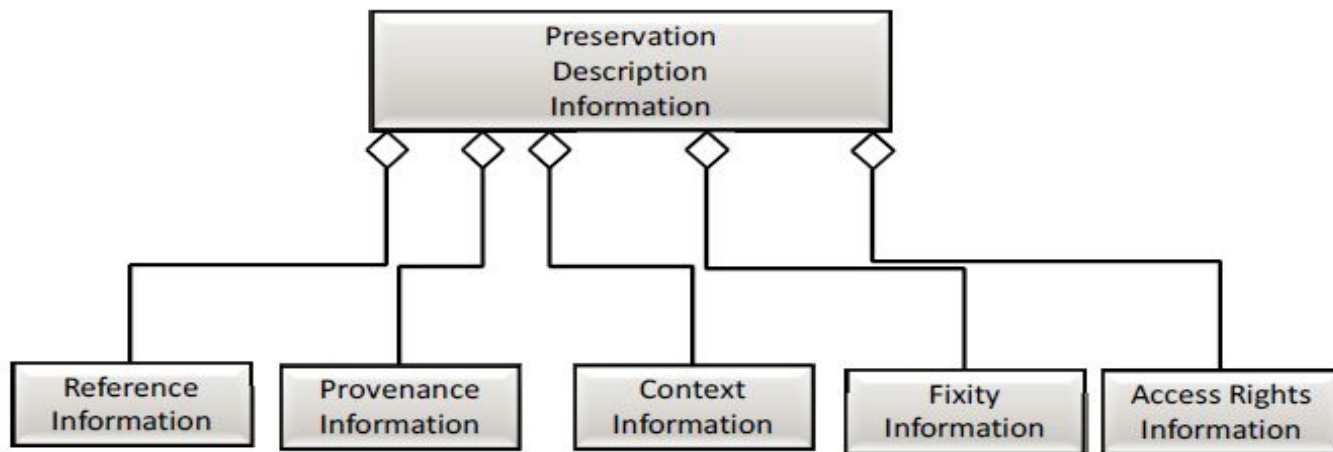


Figure 4-16: Preservation Description Information

Información descriptiva de preservación

Información de referencia: identificación y descripción de uno o más mecanismos para proporcionar los identificadores asignados para la información del contenido. También proporciona los identificadores.

Información de contexto: documenta las relaciones de la información de contenido con su entorno (¿por qué la información de contenido fue creada y cómo se relaciona con otra información de contenido).

Información descriptiva de preservación

Información de procedencia: los documentos de la historia de la información de contenido (origen o fuente, los cambios y la custodia) de procedencia puede ser visto como un tipo especial de información de contexto.

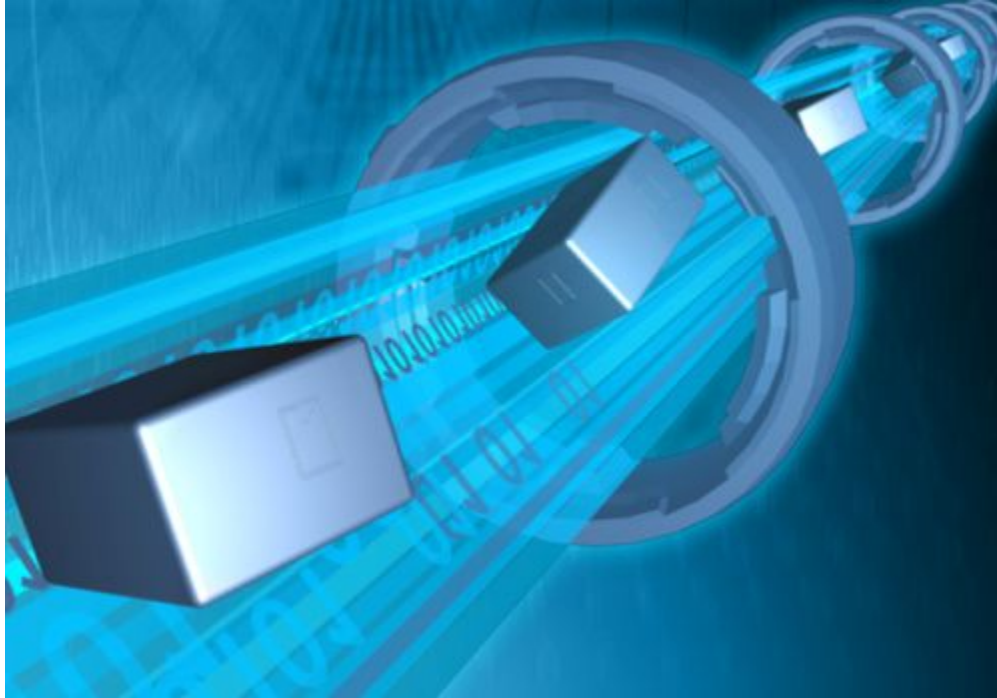
Información de la fijeza: proporciona los controles de integridad de los datos o claves de validación usados para asegurar que la información de contenido no ha sido alterada.

Información de sobre derechos de acceso: proporciona los permisos de uso de la información de contenido.

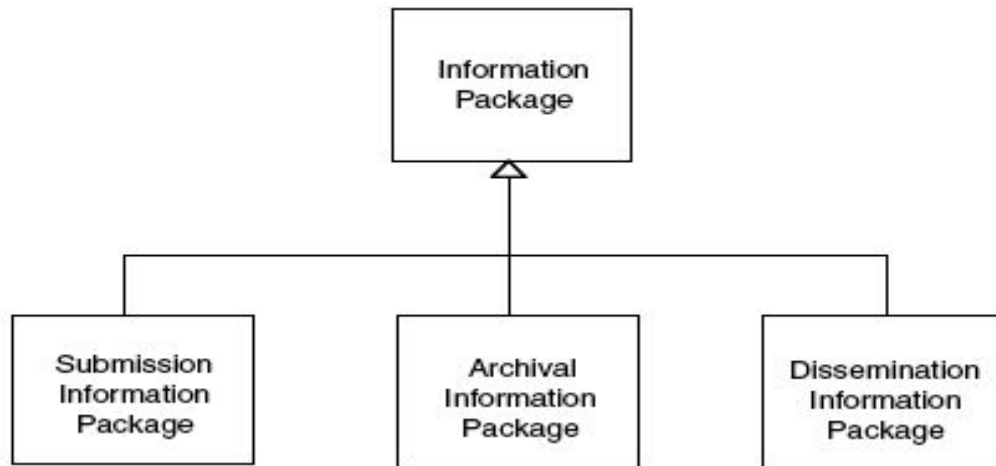
Paquetes de información en OAIS

- Las estructuras de información conceptual necesarias para cumplir las funciones OAIS.
- Un paquete de información es un contenedor.
- Hay varios tipos de paquetes de información que se utilizan en el proceso de archivo. Estos paquetes de información pueden ser utilizados para:
 - estructurar y almacenar las participaciones OAIS (AIP);
 - para transportar la información desde el productor hasta el OAIS (SIP)
 - para el transporte de la información requerida entre el OAIS y Consumidores (DIP).

Paquetes de información en OAIS



Tipos de paquetes de información



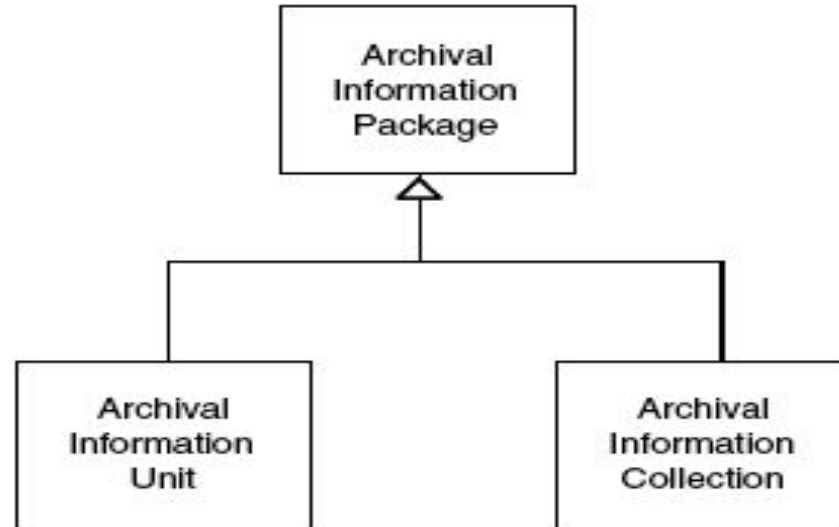
SIP

- La forma y el contenido detallado de un SIP típicamente se negocia entre el productor y el OAIS.
- La mayoría de los SIPs se tiene alguna información de contenido y algunas PDI, pero se puede requerir varios SIPs para proporcionar un conjunto completo de información de contenido y PDI asociados.
- Si hay varios SIPs que utilizan el mismo Repositorio de información, éste sólo se proveerá una vez?
- Dentro de la OAIS, uno o más SIPs se transforman en uno o más AIPs para su conservación.

AIP

- Un Paquete de Información de Archivo es una especialización del Paquete de Información. El PIA se define para proporcionar una forma concisa de referirse a un conjunto de información que tiene, en un principio, todas las cualidades necesarias para una Conservación a Largo Plazo de un determinado Objeto de Información, de forma permanente o indefinida. El PIA es en sí mismo un Objeto de Información que contiene otros Objetos de Información.

Tipos de AIPs

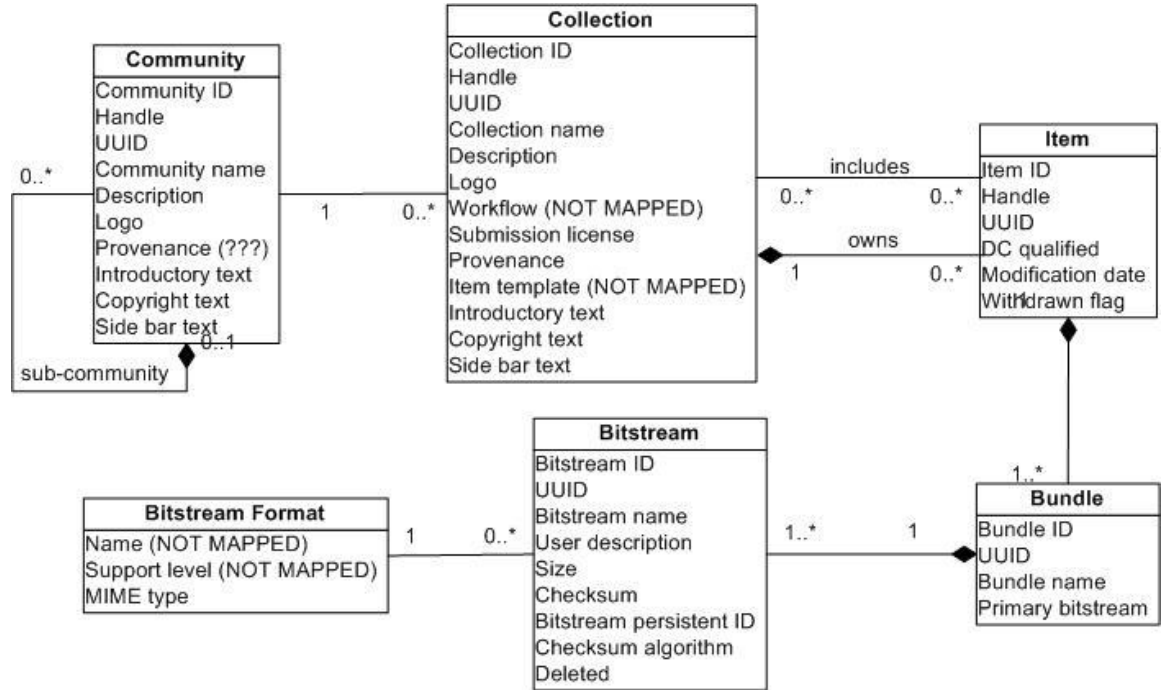


DIP

- En respuesta a una petición, el **OAIS** ofrece la totalidad/parte de la AIP a un consumidor en la forma de un DIP.
- El DIP también puede incluir las colecciones de la AIP, según el acuerdo de difusión entre **OAIS** y Consumidores.
- La información de paquetes siempre estará presente para que el consumidor distinga claramente la información solicitada.
- El propósito de la información descriptiva de un DIP es dar al consumidor información suficiente para reconocer el DIP de entre los posibles paquetes similares.

DSpace Data Model

- El repositorio se organiza en comunidades y subcomunidades
- Las comunidades poseen colecciones
- Cada colección posee ítems
- Los ítems se componen de bundles
- Cada bundle posee los bitstreams (archivos)



Modelo de contenidos - Bundles

Los ítems están vinculados a sus bitstreams a través de entidades llamadas Bundles. Los **bundles** agrupan bitstreams bajo determinados criterios:

- **ORIGINAL**: contiene los bitstreams a publicar.
- **THUMBNAILS**: son archivos con thumbnails extraídos a partir de los bitstreams originales (p.e. thumbnails de PDFs, imágenes, etc.).
- **TEXT**: contiene el texto completo (full-text) de otros bitstreams. Se genera a partir de la extracción automática de texto sobre otros bitstreams y se usa durante la indexación para mejorar los resultados de búsqueda.
- **LICENSE**: contiene la licencia que el usuario aceptó al depositar el contenido.
- **CC_LICENSE**: contiene la [Licencia CC](#) (si es que existe) seleccionada por el usuario durante la carga.

Modelo de contenidos - Bitstreams

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

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

En el lenguaje de DSpace, estos objetos digitales reciben el nombre de **BITSTREAMS**.

The screenshot displays the CIE DIGITAL repository interface. At the top, the logo 'CIE DIGITAL' is visible next to the text 'Repositorio Institucional' and 'Comisión de Investigaciones Científicas'. Below this is a navigation bar with links: 'Inicio', 'Explorar', 'Aportar Material', 'Mas información', and 'Contacto'. On the right side of the navigation bar are links for 'Mi cuenta', 'Estadísticas', and 'ES'. The main content area shows a document titled 'El Falcon Verde como lugar simbólico y la posibilidad de su activación para un uso turístico responsable' by 'Sahores Avalis, Virginia'. A 'Resumen:' section follows, containing text about the symbolic value of the Falcon Verde and its potential for sustainable tourism. To the right of the document title is a box indicating the year 'Año 2013'. Below the document title, there is a 'Descargas' section with a red Adobe PDF icon and the text 'Documento completo' and 'Archivo PDF (5.525Mb)'. A red rectangular box highlights the 'Documento completo' link and the PDF icon.

Modelo de contenidos - BitStore

Los bitstreams son alojados en un **Bitstore**, que puede estar alojado de forma local o en la nube (según la configuración en [DSpace/config/spring/api/bitstore.xml](#)):

- **ASSETSTORE:** si tenemos esta forma de almacenamiento, los bitstreams se alojan en un directorio local en el servidor:

{DSpace-install-dir}/assetstore/

- **S3Store:** si tenemos esta forma de almacenamiento, los bitstreams se alojarán en la nube, más específicamente en el Cloud Storage de [Amazon S3](#). Para que esto funcione debemos configurar diferentes parámetros como, p.e, una *accessKey* al sistema de Amazon.

Modelo de contenidos - Metadatos

dc.format.extent	116 p.	es
dc.language	es	es
dc.title	Un metalenguaje de programación orientado al diseño de interfaces gráficas	es
dc.type	Tesis	es
sedici.creator.person	Santana, Carlos Ariel	es
sedici.creator.person	Coni, Cintia Vanesa	es
sedici.subject.materias	Ciencias Informáticas	es
sedici.subject.keyword	arquitectura dirigida por modelos	es
sedici.subject.keyword	transformaciones de modelos	es
sedici.subject.keyword	interfaz de usuario	es
sedici.description.fulltext	true	es
mods.originInfo.place	Facultad de Informática	es
sedici.subtype	Tesis de grado	es
sedici.rights.license	Creative Commons Attribution 4.0 International (CC BY 4.0)	
sedici.rights.uri	http://creativecommons.org/licenses/by/4.0/	
sedici.contributor.director	Pons, Claudia	es
thesis.degree.name	Licenciado en Sistemas	es
thesis.degree.grantor	Facultad de Informática	es

Los metadatos

- son datos estructurados que describen otros datos
- son datos sobre datos

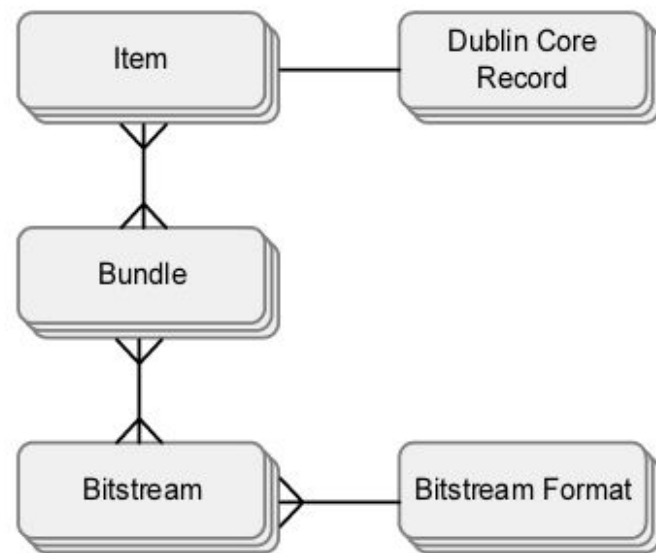
← Ej: representación del recurso

<http://sedici.unlp.edu.ar/handle/10915/47008>

OAIS en DSpace

OAIS describe una especialización de AIP llamada *Archival Information Unit* (AIU) que representa los "átomos" de información que el archivo está almacenando.

En DSpace, la unidad básica de almacenamiento de archivos, que constituye una AIU en el modelo OAIS, es el **Item**. Un Item es puede formarse por muchos *Bitstreams* (p.e. archivos PDFs).



OAIS en DSpace

Los **Bitstreams** son archivos individuales de contenido digital, como una imagen JPEG, un documento HTML, un archivo PDF, un archivo de audio MP3 o un archivo de película MPEG.

Cada Bitstream está asociado con exactamente un Formato de Bitstream del Bitstream FormatRegistry del sistema.

Para cada Bitstream, también se almacenan algunos metadatos técnicos: el **tamaño** y la suma de comprobación MD5 (**checksum**) para el *control de integridad*.

OAIS en DSpace

El AIP surge en DSpace como un objeto lógico que se conforma de datos dispersos en:

- registros de varias tablas dentro de una base de datos relacional,
- registros de uso en un indexador de textos (solr) y
- archivos en el sistema de archivos.

La llamada unidad de información OAIS en DSpace es el *ítem* y se gestiona como AIP en formato METS. Para cumplir con los requerimientos mínimos de preservación sobre los archivos.

➤ Se puede exportar en METS pero internamente no está en METS.

OAIS y DSpace

DSpace mantiene tres grupos lógicos de metadatos para el contenido:

1) Metadatos descriptivos: soporta múltiples formatos de metadatos no jerárquicos y permite mantener metadatos de distintos formatos para un mismo ítem. Para colecciones y comunidades, es posible gestionar metadatos de la misma manera aunque por default la interfaz sólo soporta una serie de campos fijos.

2) Metadatos Administrativos: incluye metadatos de preservación, información de procedencia, colección y comunidad. Se registran campos adicionales sobre varios elementos: tamaño, suma de comprobación y tipo Mime de cada archivo.

OAIS y DSpace

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.

Es posible almacenar información estructural en los bitstreams o metadatos, aunque DSpace no provee nada implementado para tal fin.

Participantes

- El productor es el autor o quien lo presenta, y suministra los artículos para el archivo a través de los procedimientos de entrada (ingest/ingesta) que constituiría el **flujo de trabajo de presentación**.
- El paquete de información presentada resultante (SIP, Submission Information Package) se convierte en el paquete de información archivada (AIP, Archival Information Package) a través del proceso del **flujo de trabajo de post-presentación** y por lo tanto pasa al almacenamiento de archivos.

Participantes

- Sección especializada para la administración adjunta a la gestión: **administradores**.
- Se relaciona con la sección de gestión de datos y la de planificación de la conservación.
- Esto permite una gestión estructural y también ayuda a mantener los AIPs a lo largo del tiempo.

Participantes

Para satisfacer los diversos requisitos detallados que exige este modelo de referencia, un sistema de repositorio debe captar todos los metadatos relevantes para convertir el SIP en un AIP con garantía de calidad y rastros de auditoría colocados al momento de la presentación, además de la información asociada como por ejemplo las normas del formato de archivo y otro tipo de metadatos técnicos.

Participantes

El AIP debe ser colocado en el archivo de almacenamiento, y se deben mantener referencias actualizadas en el sistema de gestión de datos. El almacenamiento del archivo debe permitir el uso de técnicas de almacenamiento tradicionales y verificadas, por ejemplo copias de seguridad y la verificación del contenido a lo largo del tiempo y la migración a otros medios de almacenamiento.

Participantes

- La **administración** del sistema requiere la creación de políticas y autorizaciones para permitir el acceso, y la gestión de la configuración del sistema.
- Relacionada con el proceso de ingesta, la auditoría de presentación se define dentro de su alcance y en última instancia pasa a formar parte del AIP, y también la negociación del acuerdo de presentación, que está muy asociado al tema de las licencias.
- OAIS recomienda que los administradores manejen los pedidos de diseminación y se encarguen de resolver los problemas de atención al cliente en caso de que surgieran o fueran relevantes al manejo del repositorio.

Participantes

El **acceso** a los materiales se garantiza al consumidor, quien se define según el modelo como un miembro de la comunidad designada, este es un concepto que detalla quién debe comprender el material: si la búsqueda archivada está en el campo de la física, la comunidad designada se especificará como “físicos” y los metadatos y los documentos relacionados respecto del significado del contenido se omiten por la razón de que la comunidad designada podrá comprender el material sin recurrir a estos.

Participantes

- La comunidad se asigna con el DIP, que puede contar con la mediación de los administradores o puede ser manejado exclusivamente por el sistema.
- El DIP se obtiene realizando una búsqueda en el módulo de gestión de datos, que a su vez ofrece referencias a los AIPs que deben convertirse y entregarse.
- El modelo recomienda mantener un registro de todas las solicitudes de contenido que se agregarán al rastro de auditoría del AIP.

Participantes

El módulo de **planificación de la conservación** abarca todas estas secciones, y su trabajo es desarrollar estrategias y normas de conservación, monitorear las últimas novedades y avances en el campo, y monitorear los cambios en la comunidad designada, para que toda la información nueva que se solicite se pueda adjuntar a los AIP correspondientes.

Participantes

Los resultados de este módulo servirán como pautas para que los administradores diseñen sus políticas, y en última instancia, guiarán las actividades de conservación de los materiales. Debe tenerse en cuenta que la migración y demás políticas de cambio de formatos, exigen la generación de nuevos AIP, y de ninguna manera deben modificarse los ya existentes.

Sección 5: Perspectivas sobre preservación

- 5.1 Información para la preservación.
 - Motivadores para la migración.
 - Contexto.
 - Tipos de migración: refresco, replicación, reempaquetado, transformación.
 - Versiones de los AIP.
- 5.2 Preservación del acceso.

Saliendo de la 14721



Aproximaciones a la preservación

Existen numerosas estrategias para asegurar la preservación de la información:

- Metadatos de Preservación
- METS
- El estándar PREMIS
- UNESCO: “Directrices para la preservación del patrimonio cultural”.
- Servicio PRONOM
- Herramienta DROID

METS

[illegible]

Metadatos



Metadatos

Son la información descriptiva que se aplica sobre un recurso para facilitar su organización, son datos que describen otros datos, información estructurada para describir, explicar, localizar o facilitar la obtención, uso o administración de un recurso de información.

Según la norma ISO 15489-1 del 2001 define los metadatos, en el contexto de la gestión de documentos, como: "datos que describen el contexto, contenido y estructura de los documentos, así como su gestión a lo largo del tiempo (...) Como tales, los metadatos son información estructurada o semiestructurada que posibilita la creación, registro, clasificación, acceso, conservación y disposición de los documentos a lo largo del tiempo y dentro de un mismo dominio o dominios diferentes."

Papel de los metadatos

- Recuperación de la información
- Administración de documentos
- Gestión de derechos, autoría y propiedad intelectual
- Estado de archivo
- Control y descripción de procesos

Papel de los metadatos

- Seguridad y autenticación
- Valoración de contenidos
- Preservación y conservación
- Visibilidad de la información
- Actualización de la información

Tipos de Metadatos

Tipo	Uso	Ejemplos
Administrativo	Usados en la identificación, gestión y administración de recursos de información	Adquisición de información Derechos y reproducción Requerimientos legales para el acceso Localización de información Criterios de selección para la digitalización Control de la versión
Descriptivo	Utilizados para representar recursos de información	Registros catalográficos Proporcionar ayuda en la búsqueda Índices especializados Hiperenlazar relaciones entre recursos Anotaciones de los usuarios
Preservación	Para salvaguardar los recursos de información	Informar sobre las condiciones de uso de los recursos físicos Informar sobre las acciones llevadas a cabo para preservar versiones físicas y digitales de recursos
Técnico	Relativos a cómo funcionan los sistemas o el comportamiento de los metadatos	Documentación de hardware y software Digitalización de la información (formato, ratio de compresión...) Autentificación y datos de seguridad (encriptación, passwords, etc.) Control de tiempo de respuesta de sistemas
Uso	Relativos al nivel y tipo de uso que se hace con los recursos informativos	Información sobre versiones Reutilización del contenido del recurso

Metadatos – Dublin Core (DCES)

Dublin Core DC es un esquema de metadatos elaborado por la DCMI (Dublin Core Metadata Initiative). Actualmente es el esquema de datos más utilizado. Consta de 15 elementos básicos.

Contenido	Propiedad	Instanciación
	Intelectual	
Title	Creator	Date
Subject	Publisher	Type
Descripción	Contributor	Format
Source	Rights	Identifier
Lenguaje		
Relation		
Coverage		

Metadatos – Dublin Core cualificado (DCTERMS)

El [Dublin Core cualificado](#) es una extensión de esquema Dublin Core básico donde algunos de sus elementos están especializados:

- **Título:** dc.title ⇒ dcterms.alternative, dcterms.title
- **Relación:** dc.relation ⇒ dcterms.isVersionOf, dcterms.isPartOf, ...
- **Fecha:** dc.date ⇒ dcterms.created, dcterms.available

Otros tipos de esquema son:

- Networked Digital Library of Theses and Dissertations (ETD)
- Metadata Objects Description Schema (MODS)
- Categories for the Description of Works of Art (CDWA)
- Cataloguing Cultural Objects (CCO)

Networked Digital Library of Theses and Dissertations (ETD)

- Universidad de Utah State
 - Versiones electrónicas de tesis o disertaciones
 - Metadatos utilizados en SEDICI
-
- thesis.degree.grantor (Institución garante)
 - thesis.degree.name (Grado alcanzado)

Metadata Objects Description Schema (MODS)

- Desarrollado por la Oficina de Desarrollo de Redes y Normas MARC de la Biblioteca del Congreso (LC)
- MODS es un conjunto de metadatos que tiene el mismo objetivo que DC, pero que busca superar los problemas y limitaciones detectados en la aplicación de DC. Para ello, toma como punto de partida el formato MARC, del que selecciona diferentes En SEDICI se utilizan:
 - **mods.location** (URL de acceso al catalogo de la biblioteca, o información de acceso al recurso físico)
 - **mods.origenInfo.place** (Institución donde se creo el documento)
 - **mods.recordInfo.recordContentSource** (Registra el nombre de la entidad y/o sitio web desde donde se obtuvo la información del registro)

Categories for the Description of Works of Art (CDWA)

Directrices para la documentación de los bienes patrimoniales desarrolladas por un grupo de trabajo dirigido por el Getty y la College Art Association of America (CAA),

Cataloguing Cultural Objects (CCO)

Manual para documentar arte, arquitectura y objetos culturales desarrollado y diseminado por el Visual Resources Association (VRA).

Metadatos - Esquema de metadatos SEDICI

SEDICI utiliza un perfil de metadatos propio: dichos metadatos están traducidos al inglés y al portugués, y el sistema los transforma al formato [Dublin Core](#) para que sean accedidos por otros repositorios vía [OAI-PMH](#).

Ej.: [tesis](#)

dc.format.extent	110 p.	es
dc.language	es	es
dc.title	Familias y estrategias de crianza	es
dc.type	Tesis	es
sedici.title.subtitle	Prácticas de tránsito de niños y niñas por diversos grupos de crianza en sectores populares de Viedma	es
sedici.creator.person	Belmartino, María Belén 🇺🇦	es
sedici.subject.materias	Trabajo Social 🇺🇦	es
sedici.subject.decs	Familia 🇺🇦	es
sedici.subject.decs	Crianza del Niño 🇺🇦	es
sedici.subject.decs	Población Suburbana 🇺🇦	es
sedici.description.fulltext	true	es
mods.origininfo.place	Facultad de Trabajo Social 🇺🇦	es
sedici.subtype	Tesis de maestría	es
sedici.rights.license	Creative Commons Atribución-NoComercial-CompartirIgual 2.5 Argentina (CC BY-NC-SA 2.5)	
sedici.rights.uri	http://creativecommons.org/licenses/by-nc-sa/2.5/ar/	
sedici.contributor.director	Colangelo, María Adelaida 🇺🇦	es
sedici.contributor.juror	Colangelo, María Adelaida	es
sedici.contributor.juror	Fuentes, Pilar	es
sedici.contributor.juror	Velurta, Marcela	es
thesis.degree.name	Magister en Trabajo Social 🇺🇦	es
thesis.degree.grantor	Facultad de Trabajo Social 🇺🇦	es
sedici.date.exposure	2012-06-04	

Metadatos y metadatos de preservación

Los objetos digitales cambian, y dichos cambios deben registrarse y validarse para asegurar la autenticidad del objeto, por lo que también es preciso incorporar metadatos de procedencia y autenticidad. Dado que cualquier actividad de preservación está limitada por los derechos de propiedad intelectual, se hace necesario incluir metadatos para la gestión de los mismos.

Metadatos

Los metadatos se clasifican en distintas categorías de acuerdo con las funciones que cumplen: los **descriptivos** ayudan a describir y recuperar los recursos; los **administrativos** gestionan un recurso: mantenimiento, almacenamiento y entrega, incluyendo datos técnicos sobre la creación, control de acceso y calidad, gestión de derechos, utilización y condiciones de preservación, migración, etcétera; y los **metadatos estructurales** refieren la estructura interna del recurso y los elementos que lo integran, indican cómo reunir objetos digitales complejos para que se puedan utilizar, por ejemplo: página, sección, capítulo, numeración, índices, tablas de contenidos, entre otros.

Los **metadatos de preservación** soportan los datos necesarios para cumplir con una serie de requerimientos de preservación con el objetivo de asegurar la utilización a largo plazo de un recurso digital. A continuación se incluyen algunos de estos requerimientos sobre cada objeto digital:

- Debe mantenerse en el repositorio de manera segura sin perderse ni ser modificado sin autorización.
- Se debe conocer su creador.
- Si cambia se debe conocer quién realizó el cambio.
- Debe poder localizarse y entregarse al usuario.
- Debe almacenarse en soportes que puedan leer los sistemas actuales de manera que el usuario pueda comprenderlos.

- Del mismo modo las estrategias de emulación y migración requieren metadatos sobre los formatos de los objetos originales y los entornos de hardware y software que los soportan.
- Soportar la autenticidad mediante la documentación de la *procedencia digital* a través de su cadena de custodia y el historial de cambios autorizados.
- El repositorio debe disponer de los derechos suficientes como para llevar adelante las transformaciones necesarias para mantener el acceso al objeto.
- Si el objeto está relacionado con otros del repositorio o de otros depósitos externos, estas relaciones deben guardarse.

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.

Metadatos de preservación: METS



Mets (<http://www.loc.gov/standards/mets/>) es un esquema XML para codificar metadatos relacionados con la preservación de los objetos digitales, es mantenido por la library of Congress.

Explica cómo codificar metadatos descriptivos (título autor editor), de propiedad intelectual (procedencia y derechos de uso), técnicos (profundidad de color, resolución, frecuencia de nuestro, valor de verificación...) y estructurales (qué archivos representan el OD y cómo se relacionan entre ellos).

Metadatos de preservación: METS

Es extensible y permite que se puedan codificar los distintos tipos de metadatos según estándares XML externos, como pueden ser DC o PREMIS.

Su punto fuerte son las secciones de archivo y estructural en ellas los archivos digitales compuestos, formados por diversos archivos (por ejemplo libros con páginas escaneadas por separado) tienen marcada su jerarquía (cuáles actúan como índices) y ordenación (como usarlos). METS también mantiene la relación entre distintas manifestaciones de una misma obra (un libro en PDF, TIFF, JPEG, EPUB y TXT) y la función (miniatura de previsualización, versión de consulta por pantalla, versión de archivo, versión de descarga y texto OCR para búsquedas).

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”*.

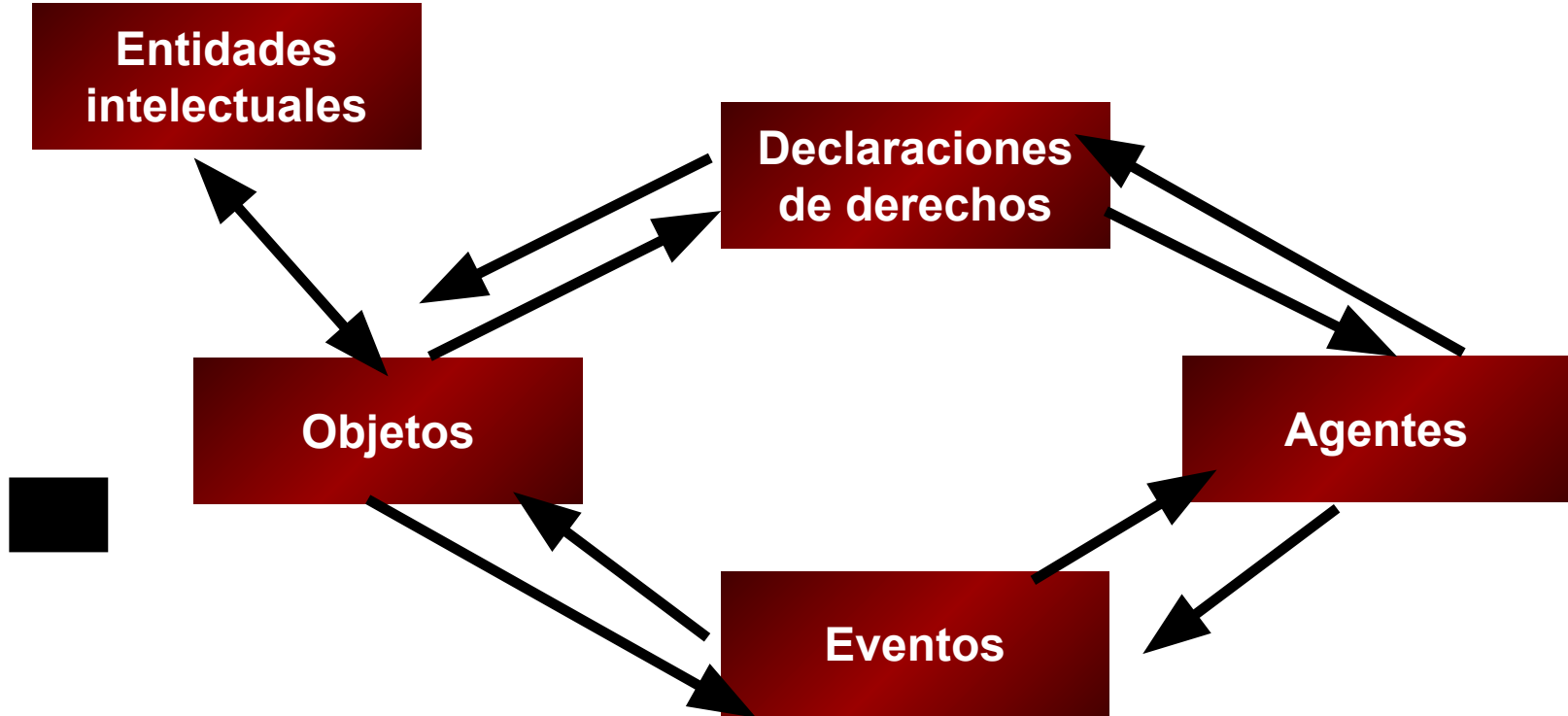
Diccionario de datos PREMIS

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.

PREMIS plantea la necesidad de representar las unidades semánticas de forma abstracta, aunque no regula su implementación ni representación.

Modelo de datos

P R E M I S



Modelo de Datos PREMIS

Las entidades que este modelo define se denominan:

- Entidades intelectuales
- Objetos
- Derechos
- Agentes
- Eventos

Entidad intelectual

Una ***entidad Intelectual*** es un conjunto de contenidos que se considera como una unidad intelectual individual al propósito de gestión y descripción. El diccionario de datos no determina los metadatos descriptivos a vincular a una entidad intelectual, sino que deja abierta la elección a cualquier formato deseado. Una entidad intelectual puede ser: un sitio web, puede incluir una página web, una página web puede incluir una fotografía.

Una entidad intelectual puede tener una o más representaciones...

PREMIS

Objeto en PREMIS difiere de la definición de objeto digital normalmente utilizada en la comunidad de las bibliotecas digitales, que entiende el término “**digital object**” como una combinación de *identificador + datos + metadatos*. No es en absoluto un conflicto. La entidad objeto en el modelo de PREMIS es una abstracción definida sólo para agrupar atributos (unidades semánticas) y clarificar relaciones.

Objetos

Los **Objetos** son unidades discretas de información en forma digital, que se clasifican en tres tipos: **archivo (file)**, **representación (representation)** y **cadenas de bits (bitstream)**. El objeto *archivo* es tal cual entendemos normalmente, es decir un archivo PDF de un capítulo de un libro, un archivo JPEG, etc. El objeto *representación* es el conjunto de todos los archivos que se necesitan para representar la entidad **Intelectual** (un libro, una foto, un mapa, un sitio web), incluyendo los metadatos estructurales. Los objetos *cadenas de bits* son subconjuntos de archivo con propiedades útiles a la preservación, en el ejemplo del archivo **JPEG** cada imagen puede tener sus propios identificadores y metadatos. La información que se puede registrar en los objetos incluye: un identificador, la integridad, el tamaño, información sobre la creación, sobre el entorno, el soporte y la relación con otros objetos y otros tipos de entidades.

Eventos

La entidad **Eventos** agrega información sobre acciones que un agente, o varios, lleva adelante sobre los objetos de los repositorios, por ejemplo: el identificador del acontecimiento (no repetible), el tipo (creación, migración, etc), la fecha de ocurrencia del evento, la descripción y el resultado codificado del acontecimiento así como los agentes.

Agentes

Los **Agentes** pueden ser personas, organizaciones o aplicaciones de software con actividades o responsabilidades en los eventos. El Diccionario de datos aconseja como información: un identificador único, el nombre del agente y su tipo (por ej. persona).

Derechos

La entidad ***Derechos*** agrega información sobre los permisos y derechos sobre los objetos que le han sido otorgados al repositorio por parte su poseedor. Se debe incluir: identificador único, un agente que concede, datos sobre la licencia y las acciones permitidas.

Ejemplo de metadatos PREMIS

[usuarios autorizados](#)
[Alertas](#)
[Editar perfil](#)
[Cuestiones y comentarios](#)
[Ayuda](#)

Palabras clave: Sistemas de colas
Llegadas en bloque
Atención en bloque
Tiempo discreto
Distribución estacionaria
Queuing systems
Batch arrivals
Bulk service
Discrete-time queue
Stationary distribution

Materia: Estadística matemática
Mathematical statistics
Empresas-Gestión
Industrial management

Derechos: Este documento está sujeto a una licencia de uso Creative Commons, por la cual está permitido hacer copia, distribuir y comunicar públicamente la obra siempre que se cite al autor original y no se haga de él uso comercial ni obra derivada

Licencia: <http://creativecommons.org/licenses/by-nc-nd/4.0/>

URI : <http://hdl.handle.net/10259/4507>

Aparece en las colecciones: [Tesis Matemáticas y Computación](#)

Ficheros en este ítem:

Fichero	Descripción	Tamaño	Formato
Lorente_Marín.pdf		9,6 MB	Adobe PDF

[Visualizar/Abrir](#)

[Mostrar el registro Dublin Core completo del ítem](#)

[Muestra el registro PREMIS del ítem](#)

Este ítem está sujeto a una licencia Creative Commons [Licencia Creative Commons](#)



Los ítems del Repositorio Institucional de la Universidad de Burgos están protegidos por copyright, con todos los derechos reservados, a menos que se indique lo contrario.

<http://riubu.ubu.es/handle/10259/4507>

Ejemplo de metadatos PREMIS

<http://riubu.ubu.es/handle/10259/4507>

```
</dim:dim>
</mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
▼ <mets:amdSec ID="amd_2">
  ▼ <mets:techMD ID="techMD_file_30476_PREMIS">
    ▼ <mets:mdWrap MDTYPE="PREMIS">
      ▼ <mets:xmlData>
        ▼ <premis xmlns:premis="http://www.loc.gov/standards/premis" xmlns="http://www.loc.gov/standards/premis">
          ▼ <object>
            ▼ <objectIdentifier>
              <objectIdentifierType>URL</objectIdentifierType>
              ▼ <objectIdentifierValue>
                http://riubu.ubu.es/bitstream/10259%2F4507/6/Lorente_Mar%C3%ADn.pdf.txt
              </objectIdentifierValue>
            </objectIdentifier>
            <objectCategory>File</objectCategory>
          ▼ <objectCharacteristics>
            ▼ <fixity>
              <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
              <messageDigest>19ca483d4bc1ca3877f5d5a5013d0d70</messageDigest>
            </fixity>
            <size>343799</size>
          ▼ <format>
            ▼ <formatDesignation>
              <formatName>text/plain</formatName>
            </formatDesignation>
          </format>
          </objectCharacteristics>
          <originalName>Lorente_Marín.pdf.txt</originalName>
        </object>
      </premis>
    </mets:xmlData>
```

Diccionario de dato: Objeto

- El identificador único del objeto (tipo y valor),
- Fijeza de la información, como la suma de verificación (mensaje cifrado) y el algoritmo utilizado para obtenerla,
- El tamaño del objeto,
- El formato del objeto, que puede especificarse directamente o mediante un enlace a un registro de formatos,
- El nombre original del objeto,
- Información sobre su creación,
- Información sobre los inhibidores,
- Información sobre sus propiedades significativas,
- Información sobre su entorno (véase más abajo),
- Dónde y en qué soporte se almacena,
- Información sobre la firma digital,
- Relación con otros objetos y otros tipos de entidades.

Diccionario de dato: Eventos

La información sobre los acontecimientos que se puede registrar incluye:

- El identificador único del acontecimiento (tipo y valor),
- El tipo de acontecimiento (creación, ingesta, migración, etc.),
- La fecha y hora en la que ocurrió el acontecimiento,
- La descripción detallada del acontecimiento,
- El resultado codificado del acontecimiento,
- Una descripción más detallada del resultado,
- Los agentes implicados en el acontecimiento y sus funciones,
- Los objetos implicados en el acontecimiento y sus funciones.

Diccionario de dato: Agente

El diccionario de datos incluye:

- Un identificador único para el agente (tipo y valor),
- El nombre del agente,
- La designación del tipo de agente (persona, organización, software).

Diccionario de dato: Derechos

La información que puede registrarse en una mención de derechos incluye:

- El identificador único de la mención de derechos (tipo y valor),
- Si la base para la reclamación de los derechos es el copyright, una licencia o una ley,
- Información más detallada sobre el estado del copyright, los términos de la licencia o la ley, si es aplicable,
- La(s) acción (es) que permite la mención de derechos,
- Cualquier restricción sobre la(s) acción(es),
- Los derechos otorgados o el período de tiempo durante el que se aplica la mención,
- El (los) objeto(s) a los que se aplica la mención,
- Los agentes implicados en la mención de derechos y sus funciones.

Objeto

Editar ítem

[Estado del ítem](#)[Archivos del ítem](#)[Metadatos del ítem](#)[Ver ítem](#)[Curar](#)

Bienvenido a la página de gestión de ítems. Desde aquí podrá retirar, reintegrar o borrar el ítem. También podrá modificarlo o añadir nuevos metadatos/archivos en las otras pestañas.

ID interno del ítem:

72104

Handle:

10915/68404

Última modificación:

2018-10-19 12:44:55.655

Página del ítem:

<http://sedici.unlp.edu.ar/handle/10915/68404>

Autorizaciones del ítem:

[Editar autorizaciones](#)

Retirar el ítem del repositorio:

[Retirar...](#)

Mover el ítem a otra colección:

[Mover...](#)

Poner Privado al ítem:

[Hacerlo Privado...](#)

Borrar completamente el ítem:

[Borrar permanentemente](#)

[Volver](#)

Objeto

```
<object>  
  <objectIdentifier>  
    <objectIdentifierType>URL</objectIdentifierType>  
    <objectIdentifierValue>  
      http://riubu.ubu.es/bitstream/10259%2F4507/6/Lorente_Mar%C3%ADn.pdf.txt  
    </objectIdentifierValue>  
  </objectIdentifier>  
</object>
```

Editar ítem

Estado del ítem

Archivos del ítem

Metadatos del ítem

Ver ítem

Curar

Bienvenido a la página de gestión de ítems. Desde aquí podrá retirar, reintegrar o borrar el ítem. También podrá modificarlo o añadir nuevos metadatos/archivos en las otras pestañas.

ID interno del ítem:

72104

Handle:

10915/68404

Última modificación:

2018-10-19 12:44:55.655

Página del ítem:

http://sedici.unlp.edu.ar/handle/10915/68404

Autorizaciones del ítem:

Editar autorizaciones

Retirar el ítem del repositorio:

Retirar...

Mover el ítem a otra colección:

Mover...

Poner Privado al ítem:

Hacerlo Privado...

Borrar completamente el ítem:

Borrar permanentemente

Volver

Objeto en DSpace

<http://sedici.unlp.edu.ar/handle/10915/68404>

☐	dc.description.abstract	El objetivo de la presente tesis es el desarrollo de tests de screening de la lectura destinados al monitoreo de los aprendizajes y a la detección de desarrollar dificultades eventuales. Los tests de screening son un componente
☐	dc.description.provenance	Submitted by Nancy China (madekumi@gmail.com) on 2018-08-13T21:26:36Z workflow start=Step: SeDiCILEvelReview - action:claimaction No. of bitstreams: 1 Tesis doctoral Nancy China.pdf: 1944320 bytes, checksum: 8813a543659d4243efe58e0dc083a3c1
☐	dc.description.provenance	Step: SeDiCILEvelReview - action:editaction Approved for entry into archive by Analía Pinto(aprumiante@gmail.com) on 2018-08-14T13:25:56Z (GMT)

```
<objectCategory>file</objectCategory>
<objectCharacteristics>
  <fixity>
    <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
    <messageDigest>19ca483d4bc1ca3877f5d5a5013d0d70</messageDigest>
  </fixity>
  <size>343799</size>
  <format>
    <formatDesignation>
      <formatName>text/plain</formatName>
    </formatDesignation>
  </format>
</objectCharacteristics>
<originalName>Lorente_Marín.pdf.txt</originalName>
</object>
```

Evento

Agente

Objeto

Derechos

☐	sedici. description. fulltext		es
☐	sedici. rights. license	Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)	
☐	sedici. rights. uri	http://creativecommons.org/licenses/by-nc-nd/4.0/	
validez			

Maipo Dublin Core - 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.

Mapao 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.

Buenas prácticas y recomendaciones





Sobre Formatos de Archivo

Preservación de contenido

- Los riesgos de pérdida de datos por eventos desafortunados siempre son posibles.
- Para disminuir esos riesgos es necesario contar con un sistema de backups (datos, configuración, documentación, etc)
- También es necesario elegir los formatos de acuerdo a los criterios que se mencionaron: uso de una gran comunidad, apertura, licenciamiento libre...

Formatos digitales

Un formato digital es un sistema particular de codificación de la información para su almacenamiento o tratamiento en un soporte informático o digital.

Los formatos más conocidos son el *.doc* para texto, *.tiff* y *.jpeg* para imágenes, *.mp3* y *.wav* para audio, *.avi*, *.mpg* para video, etc.



Los objetivos perseguidos por un formato específico pueden ser diversos:

- Almacenar un solo tipo de contenido plano sin ninguna codificación adicional. Ejemplo: .txt
- Incorporar especificaciones para codificar la información (principalmente para su compresión, transmisión o cifrado) Ejemplo: .pdf
- Combinar y sincronizar varios tipos de contenido en un solo archivo. Ejemplo: los archivos .mpeg o .AVI que incluyen pistas de audio, vídeo, subtítulos, metadatos, etc.

Cuestiones a tener en cuenta

- **¿Su uso está generalizado?** ¿Existen varios programas para leer este formato? ¿Está utilizado por otras instituciones como formato de preservación?
- **¿Está abierto?** El uso del formato no debe ser regido por patentes.
- **¿Está documentado?** ¿Fue publicada su documentación (un formato puede estar documentado sin estar necesariamente abierto)? ¿Fue normalizado por instituciones como W3C o ISO? Esta documentación permitirá construir nuevos programas para leer el formato si sus vendedores ya no se encargan de su mantenimiento.
- **¿Existen programas para validarlo y caracterizarlo?**

Dificultades para la elección de formatos

- Cuando uno elige el formato más adecuado al contenido, debe valorar los riesgos de **obsolescencia** que corre el formato.
- También se debe tomar en cuenta si las políticas del repositorio o la legislación de la institución impondrán que sean aceptados todos los datos, sea cual sea su formato.
- La finalidad del archivo digital y su ubicación en la cadena de valor: formato inicial de **creación**, formato intermedio para **edición**, formato final para preservación o **difusión**. A veces el formato mostrado al público no es el mismo en el que se preserva.

Niveles de Riesgo de un Formato (BnF)

Nivel	Descripción
Almacenado	Formato no identificado para el cual sólo se asegura la preservación de la cadena de bits.
Identificado	Formato identificado por una herramienta de identificación para el cual no fue definido ninguna estrategia de preservación.

Nivel	Descripción
Conocido	Formato para el cual la BnF tiene por lo menos una herramienta de referencia, realiza una vigilancia tecnológica y define una estrategia de preservación, sea emulación o migración a un formato controlado.
Controlado	Formato documentado por el cual la BnF tiene herramientas de referencia, realiza una vigilancia tecnológica e impone requisitos a los productores.

Factores de sostenibilidad a tener en cuenta en la selección de los formatos

Divulgación/Transparencia: hace mención al grado en el que las especificaciones técnicas de los formatos son accesibles (es decir son de código abierto).

Apertura: se refiere a la dependencia de un formato determinado a una patente. En el caso de las entidades dedicadas a la preservación y/o difusión del patrimonio cultural, en general, se tiende a utilizar formatos abiertos, es decir formatos de archivo que no se hallan sujetos a patentes o derechos de autor. Ejemplo: Empresas como Apple, Amazon, Microsoft conforme con criterios comerciales y no necesariamente de preservación, pueden imponer formatos propietarios, es decir, protegidos por patentes o derechos de autor.

Dependencia/Interoperabilidad: se trata del grado de dependencia de un formato determinado respecto a un hardware, a un software o a un sistema operativo específico.

Estabilidad/Compatibilidad: grado en el que un formato mantiene su funcionalidad e integridad con versiones anteriores o posteriores.

Aceptación: hace referencia al grado de utilización de los formatos por parte de los creadores, distribuidores y usuarios de los recursos.

Estandarización: adecuación formal a los procesos o especificaciones establecidos por un organismo de normalización (ISO, NISO, W3C...) con el objetivo de garantizar la calidad de los archivos y su interoperabilidad.

Mecanismos de protección técnica: los mecanismos de protección técnica como por ejemplo el cifrado, utilizados habitualmente para proteger la propiedad intelectual, no deben dificultar la recuperación de datos, migración de los contenidos o su adaptación a nuevas necesidades derivadas de la evolución tecnológica.

Selección de formatos: generalidades

Para asegurar la preservación de la información, el formato elegido debe ser legible por una aplicación durante el mayor tiempo posible. Esto implica evitar formatos propietarios cerrados, como documentos en Word (en cualquiera de las versiones). Lo recomendable es utilizar formatos propietarios pero abiertos, como el formato [TIFF](#), o mejor aún, formatos no propietarios como el ASCII.

Selección de formatos: generalidades

La utilización de un formato de codificación simple y universal como [XML](#) permite perpetuar los documentos electrónicos. XML es el formato ideal ya que además de ser un formato no propietario, y por tanto ofrecer garantía de preservación de la información (ASCII), permite estructurar la información y el intercambio de información a todos los medios.

Selección de formatos: generalidades

Para asegurar la integridad de los documentos que contienen objetos electrónicos (imágenes, sonidos, modelos, fórmulas, hiperenlaces..) se debe emplear la misma filosofía que con la información textual. Los formatos imagen considerados mejores para la conservación son el [TIFF \(Tagged Image File Format\)](#) que su compresión no experimenta ninguna pérdida de calidad, el [PNG \(Portable Network Graphics\)](#), cuya compresión experimenta apenas pérdidas en la resolución y además es muy ligero y el [JPEG](#).

Selección de formatos: generalidades

En cuanto a los Formatos mixtos (contenedores) los mejores son el [Postscript](#), que puede ser enviado a cualquier periférico que soporte este lenguaje, sin tener en cuenta su resolución, produciendo un resultado adaptado a cada tipo de periférico y el [PDF \(Portable Document Format\)](#), basado en el Postscript, propietario pero abierto de la casa Adobe y que facilita un programa gratuito para poder leer este tipo de documentos. Para la preservación, se recomienda especialmente el [PDF/A](#)



Formatos de texto

El formato más difundido para la publicación de textos es el formato de documento portátil o **PDF**, pero no es raro encontrar autoarchivos de materiales hechos en formatos **.doc** o **.docx** u otro tipo de formatos de texto editable como **.odt**. En estos casos siempre es recomendable la transformación del material al formato PDF. El formato PDF fue creado por Adobe, y es ahora un estándar abierto y oficial reconocido por la Organización Internacional para la Estandarización (ISO). A los fines de la preservación digital el formato recomendado es el PDF/A. El **PDF/A** es el estándar más común para los documentos de texto con formato, pero muchas entidades que ofrecen contenidos en formatos de texto electrónico en formato **EPUB**. Ambos formatos están basados en XML.

Formatos recomendados para texto

El **PDF/A** se presenta como el estándar aceptado para la creación de documentos digitales accesibles online y susceptibles de ser impresos, tanto aquellos basados en texto como los que incluyen imágenes, gráficos, etc. que requieren de un diseño preciso.

El **EPUB** es el estándar de facto recomendado para el texto electrónico. Aunque puede soportar imágenes está más orientado a la publicación de texto, por ello no es el formato más adecuado para documentos que requieren un diseño preciso o están basados en imágenes.

Ventajas del PDF/A

PDF/A es, de hecho, un subconjunto de PDF obtenido excluyendo aquellas características superfluas para el archivado a largo plazo de forma similar a como se ha definido el subconjunto PDF/X para la impresión y artes gráficas. Además, el estándar impone una serie de requisitos a los programas para la visualización de archivos PDF/A.

Un programa de visualización que se ajuste a los requisitos debe seguir ciertas reglas incluyendo la conformidad con las directrices en cuanto a la gestión de color, el uso de fuentes integradas a la hora de la visualización, o la posibilidad de realizar anotaciones por parte del usuario.

Sobre PDF/A

PDF/A es un estándar para codificar documentos en un formato “impreso”, que es portable entre sistemas y ampliamente usado para distribución y archivado de documentos. Sin embargo, la pertinencia de un archivo PDF para preservación depende de las opciones elegidas cuando el PDF fue creado: en particular, si se embebieron las fuentes necesarias para renderizar el documento, si se usa encriptación y si se preserva información adicional del documento original, más allá de lo que se precisa para imprimirlo.

Sobre PDF/A

El estándar PDF/A no define una estrategia de archivado o los objetivos de un sistema de archivado. Sí identifica un “perfil” para documentos electrónicos que asegura que los documentos pueden ser reproducidos exactamente de la misma manera durante años. Un elemento clave para esta reproductibilidad es que los documentos PDF/A deben ser 100% auto-contenidos: esto significa que toda la información necesaria para mostrar el documento de la misma manera cada vez, debe embeberse dentro del archivo. Esto incluye (pero no se limita a) todo el contenido (texto, imágenes rasterizadas, gráficos vectorizados), fuentes, información de color, etc. Un documento PDF/A no puede jamás depender de información de fuentes externas.

Otros elementos de la compatibilidad con PDF/A

- El contenido de audio y video está prohibido.
- Java script y enlaces a archivos ejecutables están prohibidos.
- Todas las fuentes deben estar embebidas, y también deben ser legalmente embebibles para renderización ilimitada y universal. Esto significa para un usuario poder abrir el documento y que los caracteres se muestren de manera correcta (de aquí a X años) aunque no tenga esa tipografía en su computadora.
- Los espacios de colores deben ser especificados de una manera independiente del dispositivo.
- Se prohíbe la encriptación.
- El uso de metadatos basados en estándares se mantiene.

Otros elementos de la compatibilidad con PDF/A

- Las referencias a contenidos externos están prohibidas.
- La compresión de imágenes LZW y JPEG2000 están prohibidas en PDF/A1, pero JPEG 2000 se permite en PDF/A2.
- Capas y objetos transparentes están prohibidos en PDF/A1 pero no en PDF/A2.
- Firmas digitales provisionales se permiten en PDF/A2.
- Los archivos embebidos están prohibidos en PDF/A1, pero PDF/A2 permite embeber archivos PDF/A.
- PDF/A3 permite embeber cualquier formato como XML, CSV, CAD, archivos de Word, planilla de cálculo, otros PDF/A, etc. como objetos archivados completos.

Niveles de cumplimiento

PDF/A posee dos niveles de cumplimiento:

PDF/a aplica corrección semántica y estructura. Cada carácter debe tener su equivalente Unicode. La estructura se expresa por medio de etiquetas.

PDF/b aplica integridad visual.

Generación de PDF/A

Un formato de preservación para documentos de texto es el estándar PDF/A, descrito en las normas ISO 19005-(1-2-3).

Este formato está basado en el estándar PDF 1.4, al que le incorpora algunos requerimientos adicionales, por ejemplo:

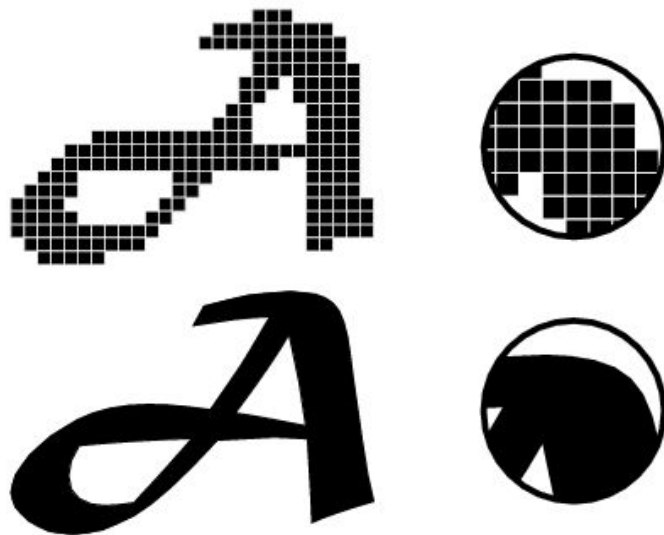
- Especificaciones sobre los metadatos y la estructura del archivo.
- La paleta de colores (incluyendo escala de grises y blanco/negro) no deben ser representados en un espacio de color de dispositivo (DeviceRGB, DeviceCMYK, DeviceGray).
- Las fuentes usadas en texto visibles deben estar embebidas (incluidas dentro del archivo).

Uno de los propósitos de los requerimientos del estándar PDF/A es de proveer soporte para personas con capacidades diferentes, por ejemplo, incorporando la información requerida y necesaria para aplicaciones que hagan el pasaje de texto a voz.

Formatos de Imagen

Existen dos tipos básicos de formatos de imagen:

- Gráficos o imágenes pixeladas o de **mapa de bits** (raster images): las imágenes están representadas por píxeles. Es el tipo de formato que se utiliza comúnmente para captar imágenes con cámaras digitales y escáneres.
- imágenes **vectoriales**: están definidas matemáticamente y utilizan formas geométricas para representar las imágenes. En este caso la información que almacenan no hace referencia a los píxeles. Se utilizan principalmente en actividades como el diseño gráfico, ingeniería, etc. El tamaño de los ficheros es muy pequeño.



Las dos características básicas que determinan la calidad de la imagen, principalmente en las imágenes pixeladas, son la **resolución** y la **profundidad de color** o profundidad del bit.

La **resolución** es la cantidad de puntos que componen una imagen, es decir, la capacidad de distinguir los detalles espaciales más finos. Habitualmente se expresa en píxeles por pulgada (ppp o ppi en inglés). Lo importante es que a mayor número de píxeles, mayor calidad de la imagen y mayor peso del archivo.

La **profundidad de color** o **profundidad del bit**, se refiere al número de bits utilizados para describir el color de cada píxel. Cuanto mayor sea la profundidad de bits mayor será la cantidad de tonos representados

Número de colores visibles por bit

Profundidad de Color	Nº de colores visibles
1 bit (monocromo)	2 tonos
4 bit	16 tonos
8 bit (indexed color)	256 tonos
24 bit (true color)	16.777.216 tonos

Formatos de imagen recomendados (IFLA)

Profundidad de Color	Nº de colores visibles
Preservación documentos impresos y manuscritos y otros materiales especiales (fotos, planos, grabados, mapas...)	Formato TIFF sin compresión / Resolución mínima 400 ppp / Color: Escala grises 256 gamas / 16,7 millones colores
Preservación prensa histórica	Formato TIFF sin compresión / Resolución mínima 400 ppp / Color: Escala grises (256 gamas)
Preservación prensa moderna	Formato TIFF sin compresión / Resolución mínima 300 ppp / Color: Escala grises (256 gamas)
Difusión	archivos con compresión JPEG

** En la digitalización de imágenes, el TIFF sigue siendo el formato más utilizado en la preservación de documentos, aunque el JPEG 2000 está comenzando a ser usado siempre que la compresión sea sin pérdida.*

Formatos

de



Se distinguen dos categorías elementales de formatos:

- Por un lado, aquellos que recogen en forma de datos el sonido creado directamente a través de dispositivos electrónicos como ordenadores, sintetizadores, etc. para que después sean interpretados por el software utilizado para su reproducción.
- Por otro, los formatos más comunes y que interesan en este caso, los utilizados en las grabaciones musicales, de voz o de cualquier otro tipo, como por ejemplo el WAV, que almacena muestras de la onda de sonido para su posterior reproducción (formatos waveform).

Formatos de audio

En contenidos de audio son dos las características principales a considerar:

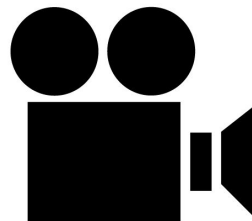
- **Frecuencia de muestreo:** número de muestras que se toman de la onda de sonido en un período de tiempo, se expresa en Khz.
- **Profundidad de bits:** número de bits utilizados para representar la onda sonora.

A mayor frecuencia de muestreo y profundidad de bits, mayor calidad del sonido. Por ejemplo, para obtener una calidad de CD de audio, la grabación se realiza a 44,1 Khz y 16 bits. En cambio, cuando sólo se va a difundir, se eliminan espacios vacíos y algunas frecuencias del archivo.

Formatos de audio recomendados

Recomendaciones IASA (International Association of Sound and Audiovisual Archives) para preservación
Formato WAVE y BWF con fines archivísticos
Calidad recomendable: 96 kHz / 24 bit
Calidad mínima: 48 kHz / 24 bit
Las mismas especificaciones para las grabaciones de voz

Formatos de video



Los formatos de vídeo son en general, **formatos contenedores** o formatos de compresión de información audiovisual ya que contienen varios tipos de información: pistas de vídeo, pistas audio, en ocasiones texto o incluso imágenes incrustadas para los subtítulos y los metadatos necesarios para su sincronización.

En el caso de los contenidos audiovisuales resulta más complejo establecer unas directrices de uso común para la preservación por los diferentes tipos de información (audio, imagen...) y los productos y canales de difusión específicos para los que están diseñados: grabación, edición, cine, DVD, internet.

Resolución de la imagen: definida por el número de píxeles («puntos») que componen la imagen.

Velocidad de la imagen (frame rate): número de imágenes estáticas o fotogramas por segundo, se expresa en fps (frames per second). Existen distintos estándares para cine (24 fps) y los distintos sistemas de televisión, etc.

Tasa de bits (bit rate): velocidad de transferencia de los datos contenidos en el archivo audiovisual. Se expresa en bits por segundo (bps) y determina la nitidez de la imagen: a mayor velocidad más información puede contener la imagen.

Profundidad de color: al igual que en las imágenes estáticas, se refiere al número de bits utilizados para describir el color de cada píxel.

Relación de aspecto o tamaño del cuadro: proporción entre el ancho y la altura de la imagen.

Formatos de video recomendados

Preservación	Difusión
JPEG2000 (ISO/IEC 15444-4) (.mj2)	JPEG 2000 MPEG4 AVI MOV MP4 MKV OGG

En síntesis:

- Para la preservación se utilizan en mayor medida los formatos no propietarios, reconocidos como estándares. Cuando se trata de difundir existe una mayor flexibilidad.
- Los formatos presentados como válidos para la preservación, también lo son para la difusión de los contenidos aunque se suele aplicar algún tipo de compresión o se reduce de calidad en pos de su funcionalidad.
- La principal divergencia encontrada en los formatos utilizados en la industria y en el ámbito de la preservación, reside en la utilización de medios de protección técnica.

Formatos sugeridos

Tipo de contenido	Preservación	Difusión
IMAGEN	TIFF (sin compresión) JPEG2000 (sin pérdida)	JPEG JPEG2000 (con pérdida) TIFF (con compresión)
TEXTO	EPUB (libro electrónico) PDF/A1 (ISO 19005-1)	EPUB PDF/A PDF HTML XHTML Open office Texto plano (ISO 8859-1)
AUDIO	BWF WAVE	MP3
VIDEO	JPEG2000 (ISO/IEC 15444-4) (.mj2)	JPEG 2000 MPEG4 AVI MOV

HERRAMIENTAS INFORMATICAS

DE PERFILAMIENTO DE FORMATOS DE ARCHIVO



Perfilamiento

Para que un repositorio sea responsable de la gestión y preservación de sus archivos es imprescindible que conozca en qué **formatos** están codificados y cuáles son los peligros de conservación, especialmente si son de uso común y si se dispone de un software para su utilización, por ello es que en la fase de ingesta el OAIS propone la comprobación de formato, a la vez que ese metadato técnico sea pasado a la entidad de gestión de datos.

Preservación en el repositorio

Basado en el servicio de PRONOM provisto por The National Archives (TNA) y la herramienta DROID (Digital record object identification service) que usa los perfiles de formato de más de 200 repositorios del registro PRONOM. DROID permite clasificar y evaluar los riesgos de los distintos formatos que usa un repositorio y de este modo elaborar un **plan activo** de preservación que identifique el formato o sugiera el cambio.

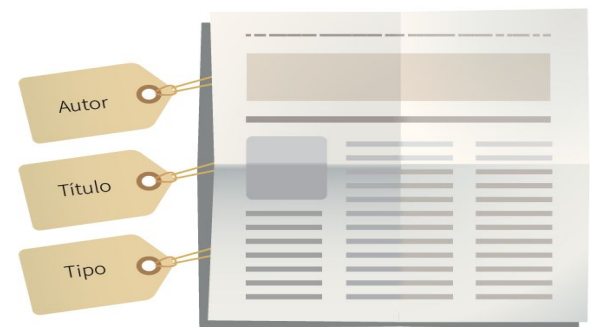
Formatos. ¿Cómo conocer lo que tiene un RI?

Perfilamiento automatizado de los objetos del repositorio: esto involucra al objeto de contenido (CDO) con sus propiedades significativas y a la información de representación de ese objeto (RI). Realizar el perfil con DROID que contrasta con el registro PRONOM y brinda un reporte.

El punto 1 es una de las 3 partes que se consideran importantes a la luz de cumplir con la ISO 14721 y realizar una evaluación del repositorio en los aspectos de preservación y accesibilidad

¿Qué acciones se proponen?

Nombre	Descripción	Formato	Ver	Orden
Bloque: TEXT				
 Tesina de Licen ... mazan Maria Belen.pdf.txt	Extracted text	Text	[Ver]	1 (Anterior:1)
 presentación.xps).pdf.txt	Extracted text	Text	[Ver]	2 (Anterior:2)
Bloque: ORIGINAL				
 Tesina de Licenciatura - Almazan Maria Belen.pdf (principal)	Documento completo	Adobe PDF	[Ver]	1 (Anterior:1)
 ...	Presentación	Adobe	[Ver]	2



Información descriptiva
(DI)

De Giusti, Marisa R. (2014). Tesis doctoral: “UNA METODOLOGÍA DE EVALUACIÓN DE REPOSITARIOS DIGITALES PARA ASEGURAR LA PRESERVACIÓN EN EL TIEMPO Y EL ACCESO A LOS CONTENIDOS”. Disponible en:

<http://hdl.handle.net/10915/43157>

Registro PRONOM

El registro de los archivos nacionales británicos se ha consolidado como el más completo.

 The National Archives

Search our website 

About usEducationRecordsInformation managementArchives sector

You are here: [Home](#) > [Information management](#) > [Our projects and work](#) > [Digital preservation](#) > PRONOM

**The technical registry
PRONOM**

Welcome  Welcome : [About](#)  [Add an entry](#)
 [Search](#) ? [Help](#)  [Information resources](#)

 **Welcome to PRONOM**

[PRONOM changes and DROID signature file release notes.](#)

[DROID signature files.](#)

Work has been undertaken to create a linked data version of PRONOM. We hope to revisit this model for PRONOM in the future. As part of a broad reorganisation of The National Archives web services the current iteration of Linked Data PRONOM has been removed.

The online registry of technical information. PRONOM is a resource for anyone requiring impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value. Find out about the future of PRONOM on our [Information resources](#) page.

 **Search PRONOM** >

Get support on our [Discussion forum](#) >

Tools and Services
Free PRONOM tools and services to support digital preservation, including DROID, the automatic file format

Contribute to PRONOM
Contribute new information to PRONOM via our online [submission form](#) >
Find out more about [PRONOM's creators](#)

 **New to PRONOM?**

- [What is PRONOM?](#) >
- [How do I search PRONOM?](#) >
- [Who is PRONOM for?](#) >
- [How do I find out more?](#) >

DROID



File profiling tool (DROID)

DROID stands for Digital Record Object Identification. It's a free software tool developed by The National Archives that will help you to automatically profile a wide range of file formats. For example, it will tell you what versions you have, their age and size, and when they were last changed. It can also provide you with data to help you find duplicates.

Profiling your file formats helps you to manage your information more effectively. It helps you to identify risks (and therefore plan mitigating actions). It can also help you to save money, for example by supporting data reduction.

You can [download our latest version of DROID for free](#). For previous versions go to <http://droid.sourceforge.net/>. For more information, see our [PRONOM resource](#).

If you are experiencing any problems using DROID, contact us at pronom@nationalarchives.gov.uk.

Also in Managing digital continuity >

- > [What is digital continuity?](#)
- > [Step-by-step guidance](#)
- > [Risk assessment](#)

Herramientas para revisar los formatos, registros...



<http://jhove.openpreservation.org/>



<http://openpreservation.org/technology/products/fido/>



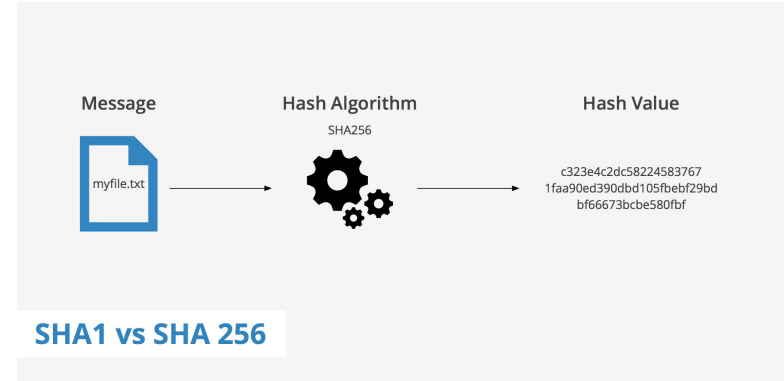
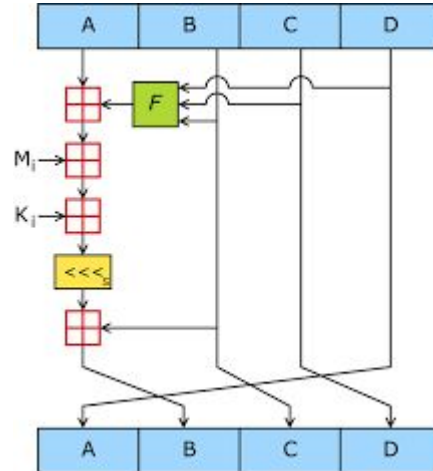
<http://www.nationalarchives.gov.uk/information-management/managing-e-information/preserving-digital-records/droid/>

Herramientas de comprobación de la integridad

Sirven para comprobar que los bytes que forman un archivo no han sido alterados.

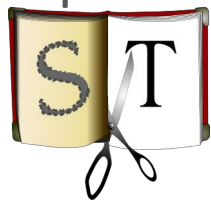
El método más usado es la suma de comprobación, que calcula la equivalencia del conjunto de bits de un archivo a un valor discreto, por ejemplo MD5.

En los sistemas de preservación digital lo correcto es que se integre la suma y la comprobación.



Documentos
Electrónicos
Tratamiento

Una vez que se recibe material digital en el repositorio es necesario hacer un correcto relevamiento de los metadatos a partir de la obra, así como analizar los típicos problemas que pueden presentarse según la tipología del material y el formato del archivo. A continuación se presentarán brevemente algunos casos y posibles soluciones que permiten llevar a cabo de un modo más simple la tarea de descripción documental.



Video Downloader GetThemAll

ofrecido por FVD Inc



Adobe Acrobat DC

Tratamiento y mejoras de archivos PDF

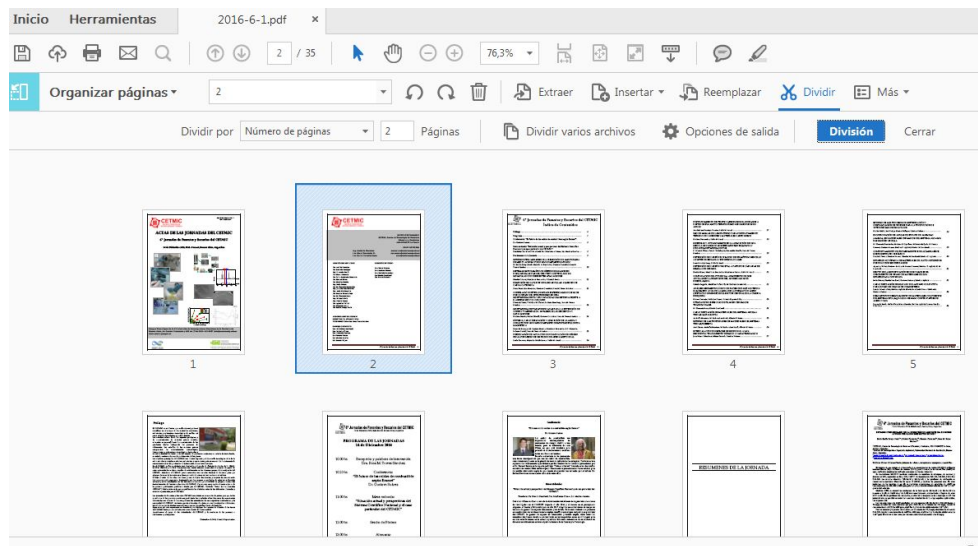
Suele ocurrir que el repositorio recibe archivos PDF de revistas o eventos que deben ser divididos para que cada artículo se cargue por separado. Una de las opciones disponibles es el Acrobat DC.



Adobe Acrobat DC

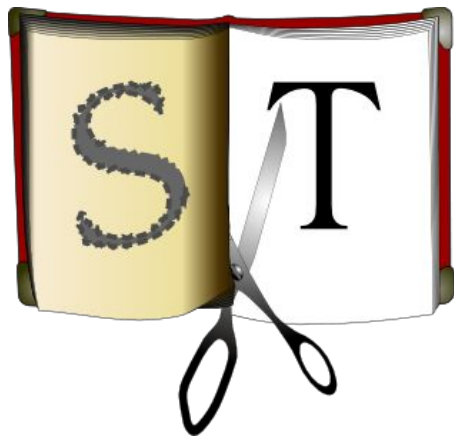
Acrobat DC - División de Archivos

Este programa permite dividir por cantidad de páginas o por marcadores. En el caso de que el archivo no cuente con marcadores es posible agregarlos manualmente de manera fácil para poder dividir el archivo correctamente.

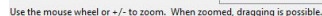


Scan Tailor

Otra posibilidad es que lleguen pdf con hojas escaneadas en modo apaisado o que los editores de una revista o un libro envíen la versión para imprimir. En este caso se necesita contar con un programa, por ejemplo Scan Tailor que permite dividir al medio las hojas y organizar el pdf de manera correcta.

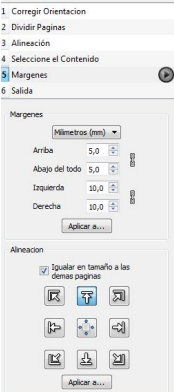
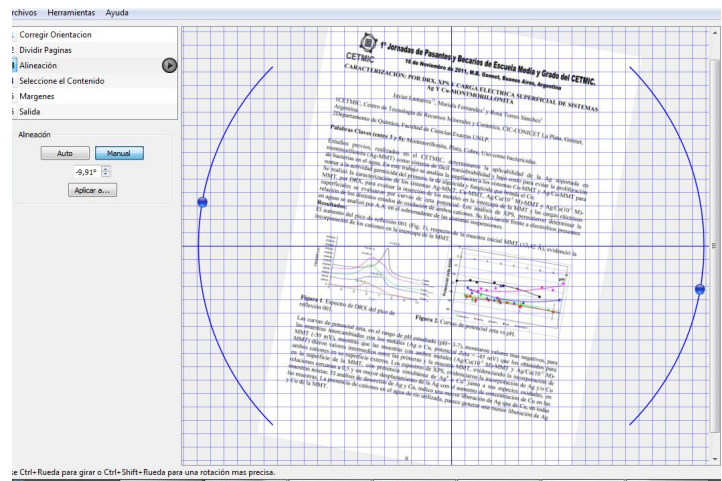


Se puede elegir manualmente el área correspondiente a cada hoja o el programa puede hacerlo de manera automática.



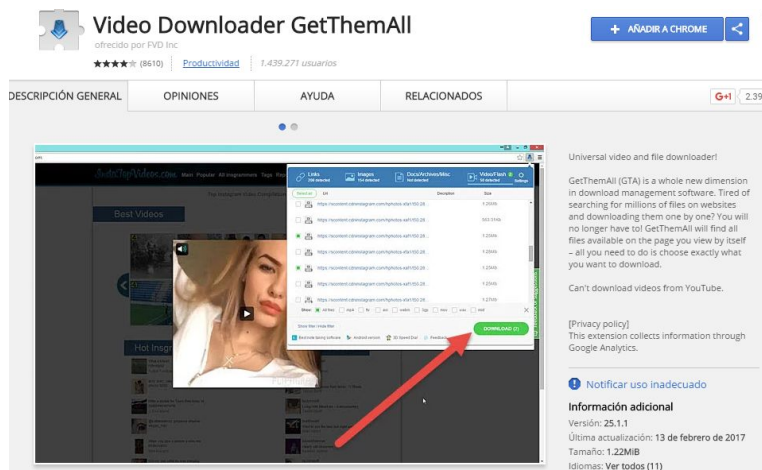
Scan Tailor

También es posible corregir la orientación de las páginas, seleccionar los márgenes y la sección del contenido para realizar OCR y por último ajustar el color y contraste de las páginas.



Descarga masiva de archivos PDF

Muchas veces ocurre que se deben descargar archivos PDF de manera masiva, por ejemplo, un volumen de una revista que se encuentra en una página web. Existen numerosas herramientas, en SEDICI se utiliza una extensión de Chrome llamada GetThemAll



Video Downloader GetThemAll
desarrollado por FWD Inc.
★★★★☆ (8610) | [Productividad](#) | 1.439.271 usuarios

[+ AÑADIR A CHROME](#)

DESCRIPCIÓN GENERAL OPINIONES AYUDA RELACIONADOS

Universal video and file downloader!

GetThemAll (GTA) is a whole new dimension in download management software. Tired of searching for millions of files on websites and downloading them one by one? You will no longer have to! GetThemAll will find all files available on the page you view by itself - all you need to do is choose exactly what you want to download.

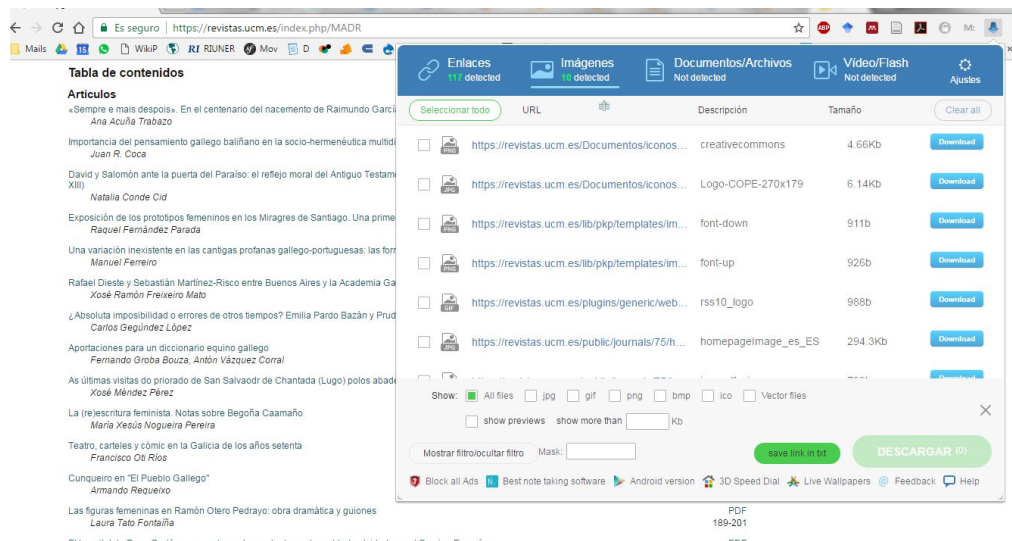
Can't download videos from YouTube.

[Privacy policy]
This extension collects information through Google Analytics.

Información adicional
Versión: 25.1.1
Última actualización: 13 de febrero de 2017
Tamaño: 1.22MiB
Idiomas: [Ver todos \(11\)](#)

Descarga masiva de archivos PDF

La extensión GetThemAll detecta automáticamente los archivos pdf de una página web y permite la descarga en breves pasos. Ejemplo: [Journal of Computer Science and Technology](https://revistas.ucm.es/index.php/MADR).



Normalización de Texto

Es muy común que el título de un artículo, su resumen y otros metadatos sean copiados directamente desde el archivo PDF. Esta acción suele presentar problemas de diferente tipo que alteran el texto a pegar, por ejemplo: codificación inválida, exceso de saltos de línea y tabulaciones, textos en mayúsculas, caracteres especiales que no se veían en el PDF, entre otros.

⇒ Normalizador de texto. <http://sedici.unlp.edu.ar/norm.php>

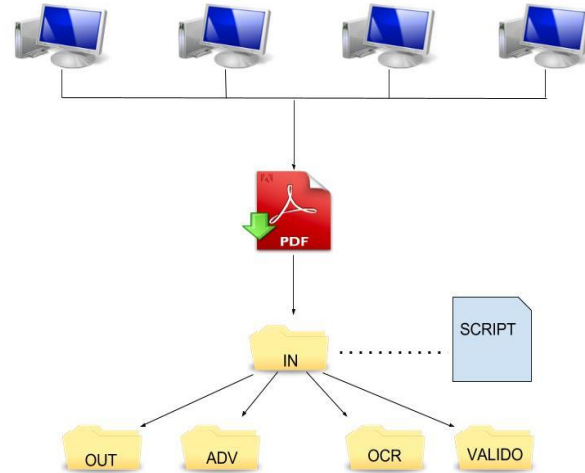
Es una aplicación PHP que a partir de una mínima configuración revisa y corrige el texto ingresado. Por ejemplo:

- multiplicidad indeseada de espacios y saltos de línea
- detección de saltos de línea indebidos
- normalización de caracteres (UNICODE NORM)

3-height

Para trabajar con Documentos PDF se generó un script que utiliza un programa estilo **servidor/cliente**, encargado de:

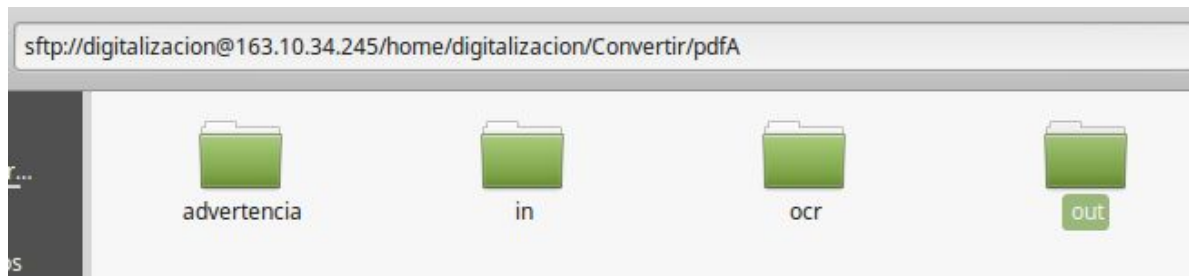
- Deteccion de archivos
- Análisis
- Conversión
- Verificación



3-height

Simplemente se tiene una carpeta compartida con el nombre PDFA que consta de 4 directorios donde los administradores podrán transformar los archivos PDF en PDFA. Los directorios son:

- Una carpeta “in” para ingresar los archivos a procesar
- Una Carpeta “out” donde se depositarán los archivos resultantes.
- Y dos carpetas destinadas a diferentes tipos de errores llamadas “advertencia” y “ocr”



Generación de PDF/A - 3-Height

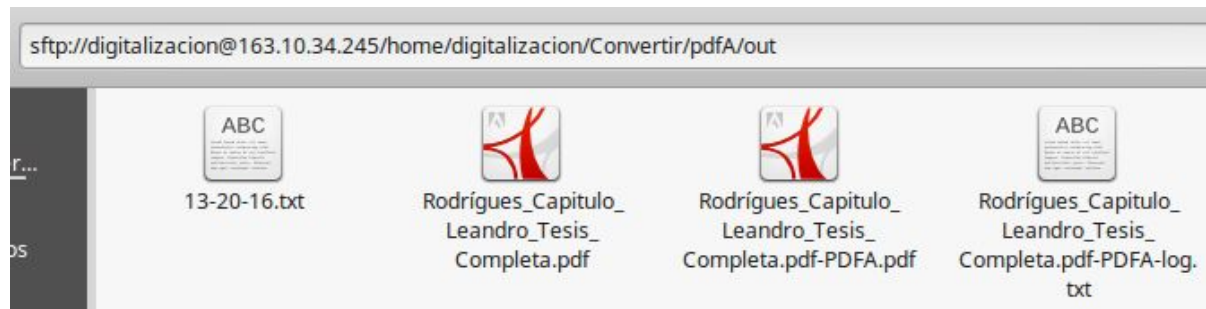
El 3height analiza el pdf y elige en qué versión va a convertirlo. Si la conversión sale bien, en la carpeta out se tienen los siguientes archivos:

El archivo con la fecha 13-20-16.txt presenta el log de la ejecución del script..

El archivo pdf original.

El archivo convertido con la terminación: -PDFA.pdf

El último archivo txt da más detalles de la conversión del archivo original

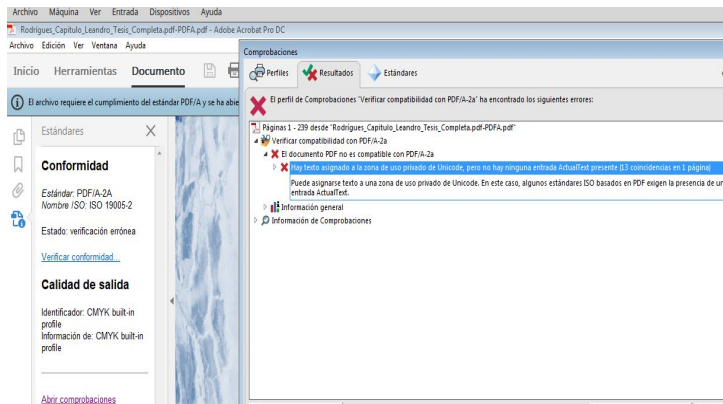


Validación de PDF/A - Acrobat DC

Una vez obtenido el pdf/A de 3-Height es necesario validarlo también en Acrobat DC. Si la verificación es errónea dependiendo el caso de error se puede arreglar desde el mismo Acrobat. Por ejemplo: cuando un archivo no pasa la verificación porque el texto no es unicode en todo el pdf. Generalmente este problema se soluciona transformando el archivo en la versión de pdfA llamada pdf/A-2u.

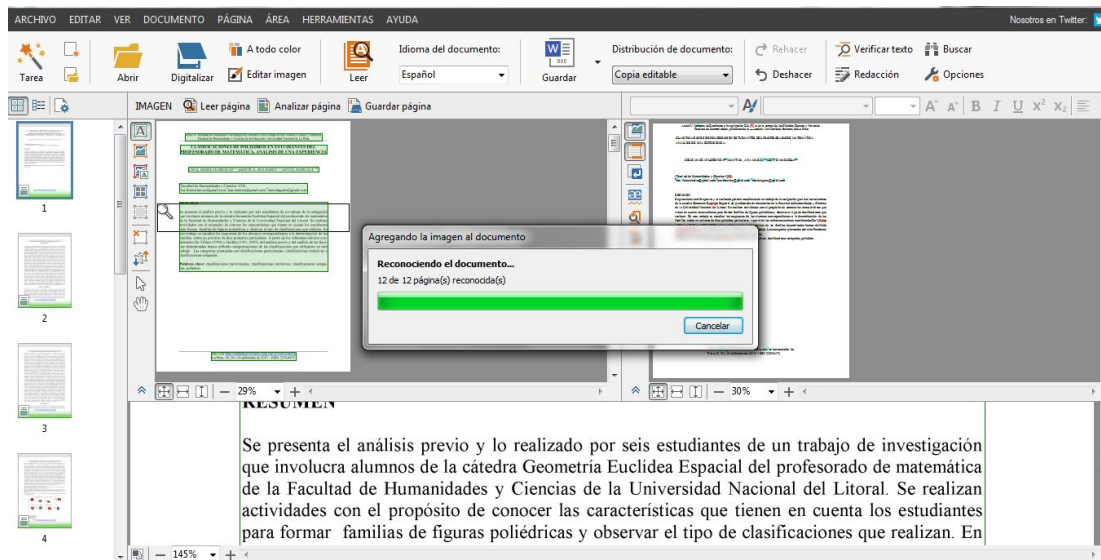


Adobe Acrobat DC



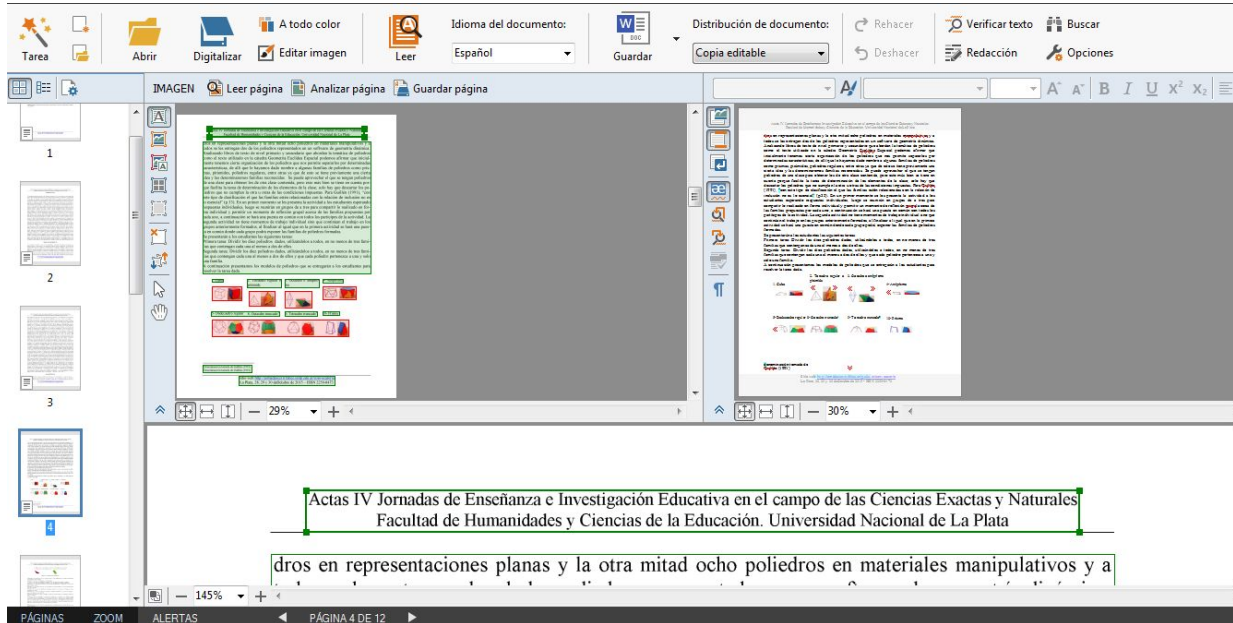
Generación de OCR con ABBYY Fine Reader

ABBYY Fine Reader es un software de OCR que permite trabajar y editar pdf de manera rápida y confiable.



ABBYY Fine Reader

El software permite seleccionar imágenes e indicar qué parte de la página debe ser reconocido y qué no. Es de los softwares con motor de reconocimiento óptico de caracteres más precisos del mercado.



Herramientas y Digitalización



Digitalización

Digitalización y preservación

Digitalización de documentos

Escáner DAL

Proceso de imágenes

Digitalización para dispositivos móviles

Formatos

PDF/A

Digitalizaciones en SEDICI

- 1) Documentos de papel físicos (libros, revistas, tesis, informes, etc.)
- 2) Documentos .doc, .docx, .odt, etc.
- 3) Documentos PDF

1) Digitalización de documentos

- Escáner HP scanjet 7500 / Fujitsu fi-7160

Softwares de escaneos: HP Smart Document Scan Software 3 -
PaperStream

- Escáner DAL

Captura de imágenes

Rotación de la imagen

Renombrado

Combinación

Proceso de imágenes mediante ScanTailor/Photoshop

Impresión en PDF, OCR en Abbyy FineReader 12

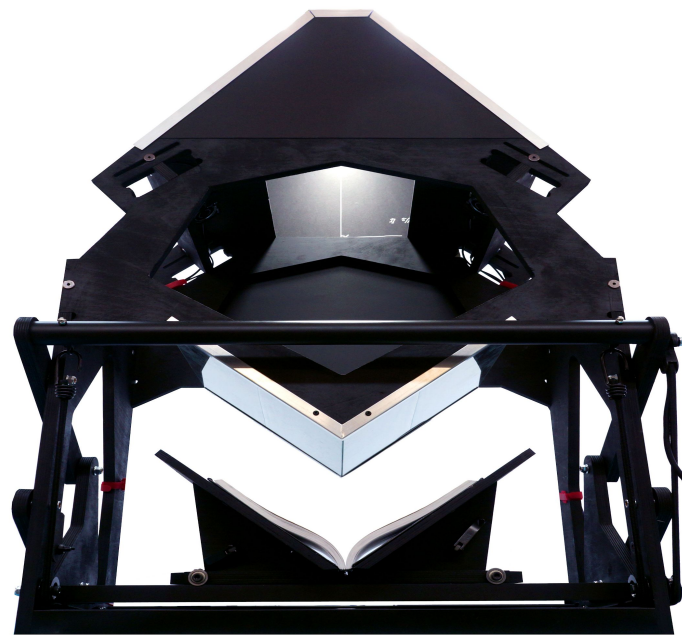
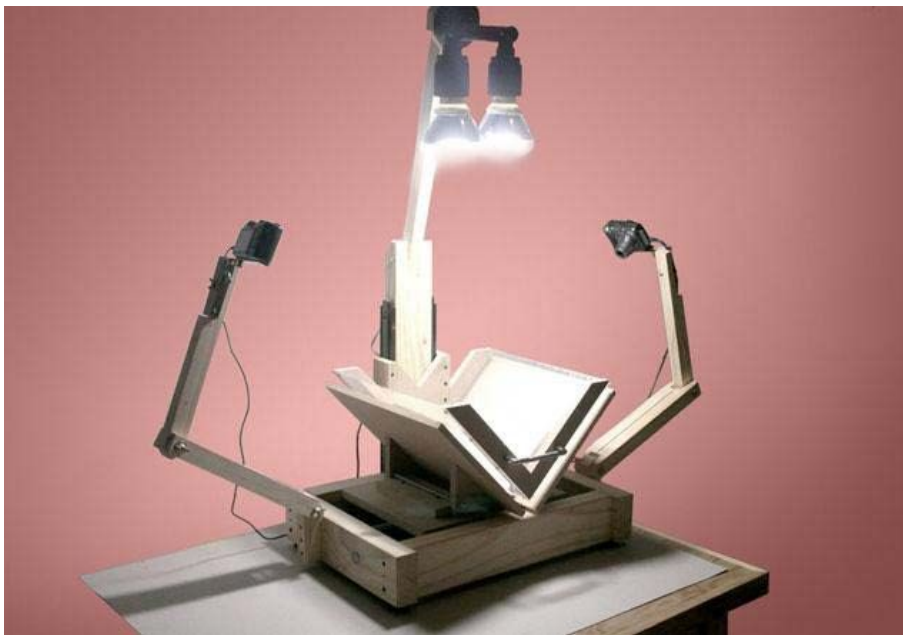
DAL

Derecho a Leer (DaL) (Facultad de Filosofía y Letras - UBA). Es un espacio de difusión de las políticas a favor de la lectura. Inspirados en la experiencia de *Do It Yourself* desarrollaron un escáner para libros de forma artesanal y muy económica, utilizando madera y cámaras digitales estándar. El objetivo es tomar una foto con la mayor precisión posible de cada hoja del libro.

Ejemplo de digitalización realizada en SEDICI:
Academia Nacional de Agronomía y Veterinaria (ANAV).
Tomo I 1932- 1934



Escáneres DAL



Captura de imágenes

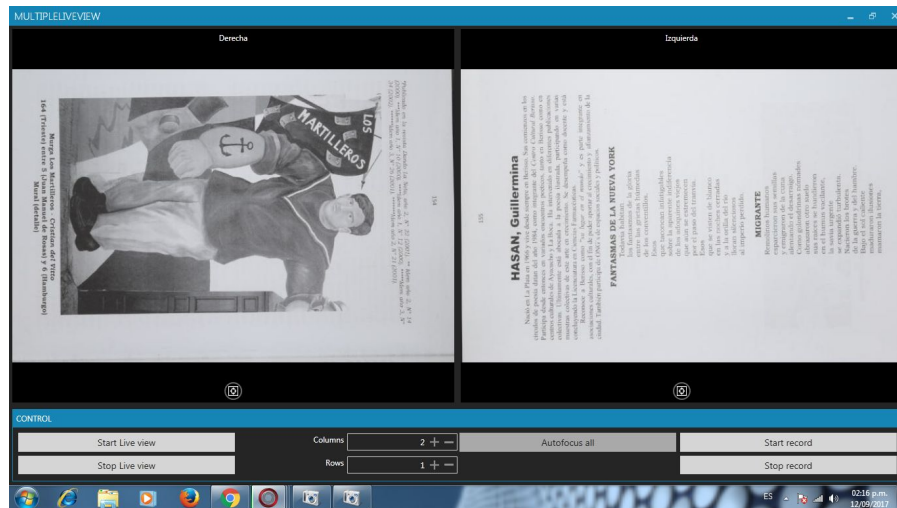
Maquinas Nikon D5300

Software *digiCamControl*

Mecanismo de disparador manual

Configuración a través de software de las cámaras

Previsualización en tiempo real



Pre-procesamiento de imágenes con CHDK

Luego de la descarga de las fotos al PC, se corre un script que rota, renombra y combina las imágenes sacadas de las cámaras.

```
prebi@Digitalización ~ $ ./editar  
MiNuevoProyecto
```

Procesamiento:

Se crea un proyecto en Scan Tailor donde a cada imagen se le hace un proceso de:

- alineación
- selección de contenido
- centralización e igualar tamaño
- blanquear fondo y borrar manchas.

ScanTailor crea imágenes en formato .TIFF Luego de generados los TIFF se crea un nuevo proyecto de AbbyyFineReader, donde se selecciona el texto e imágenes para crear un PDF/A-1a.

El PDF es verificado y optimizado para la vista rápida en web con Acrobat DC.



HP 7500/ Fujitsu fi-7160

Se necesitan las hojas del documento sueltas, sin bordes salientes

Se pueden escanear a una cara o dos

Escaneos en blanco y negro: 600ppp

Escaneos en color/ escala de grises:

400ppp

TIFF con compresión sin pérdida (LZW).

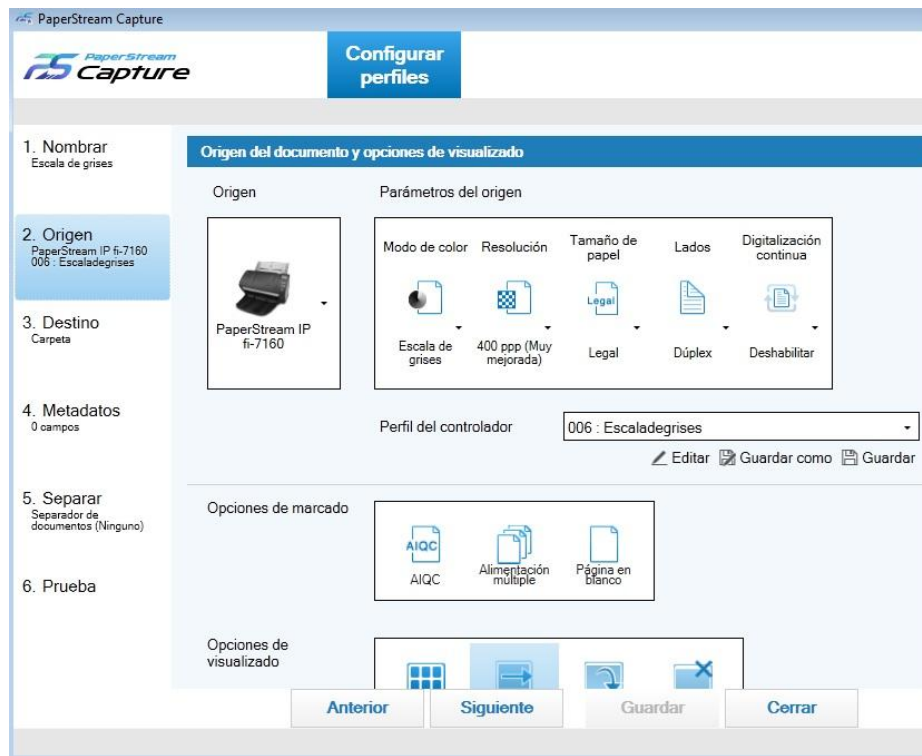
Máximo de hojas por paquete: 100

Software de escaneo: HP Smart

Document Scan Software 3 /

PaperStream

Formato de salida del escaneo: .TIFF



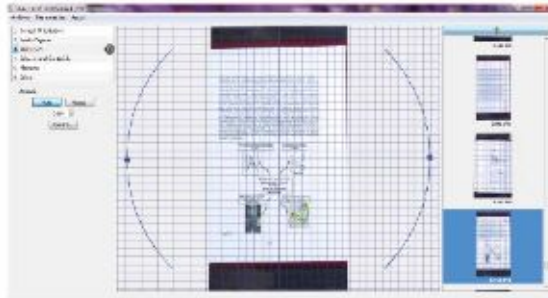
Clasificación del material según relevancia

Dependiendo de la utilidad y el interés del material, los procesos de edición de imagen tienen mayor o menor automatización y revisión posterior. El material de alta relevancia (copias únicas por ejemplo) requieren un proceso de revisión de la edición de imagen y del OCR página por página. Materiales de relevancia media requieren una revisión detallada de la tapa y los índices y una revisión general del resto. En cambio, los materiales de digitalización rápida requieren una revisión general y un proceso casi totalmente automatizado.

Programas utilizados

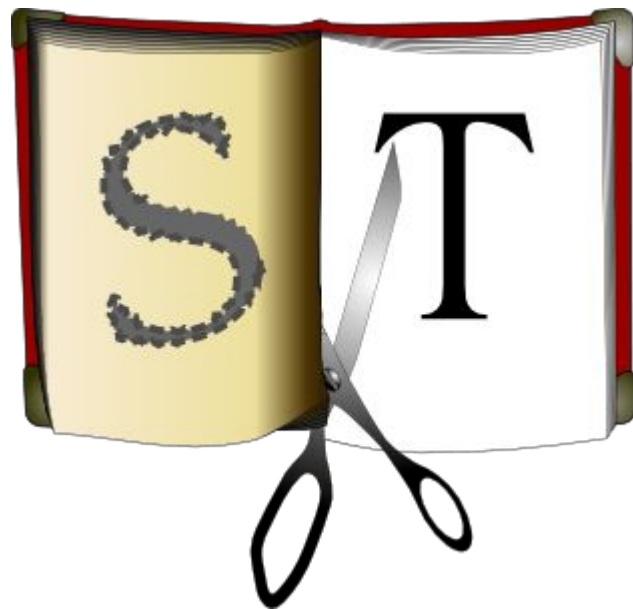


Scantailor



Edición de imagen

Luego de obtener las imágenes escaneadas, se editan en Photoshop o ScanTailor. De esta manera se puede corregir la orientación, dividir y alinear las páginas, seleccionar el contenido, limpiar los márgenes, eliminar las manchas y modificar el contraste.



- 1 Corregir orientación
- 2 Dividir páginas
- 3 Alineación
- 4 Seleccione contenido
- 5 Márgenes

Rotar



Restablecer

Extensión

Aplicar a...



- 1 Corregir orientación
- 2 Dividir páginas
- 3 Alineación
- 4 Seleccione contenido
- 5 Márgenes

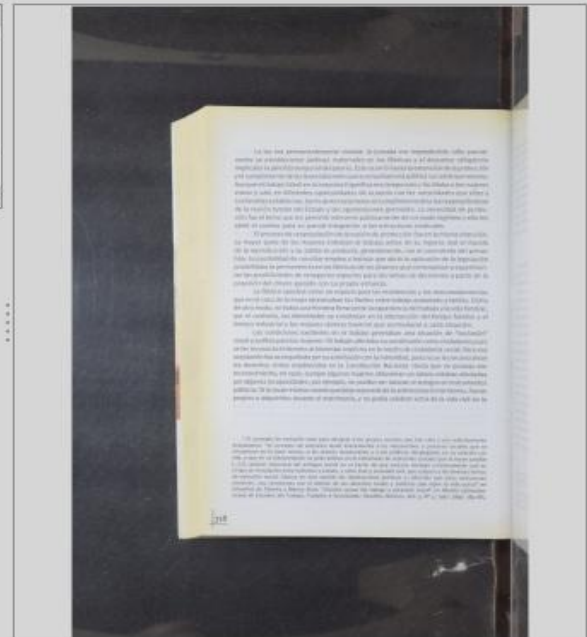
Rotar



Restablecer

Extensión

Aplicar a...



- 1 Corregir orientación
- 2 Dividir páginas
- 3 Alineación
- 4 Seleccione contenido
- 5 Márgenes
- 6 Salida

Diseño de página



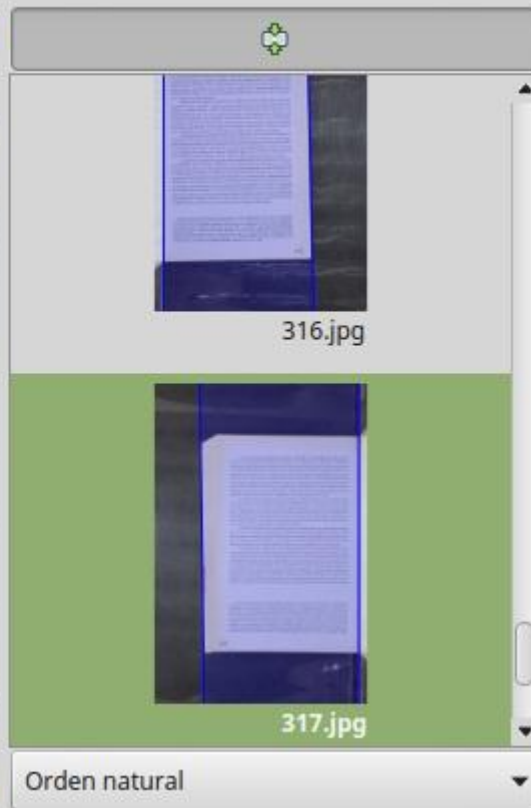
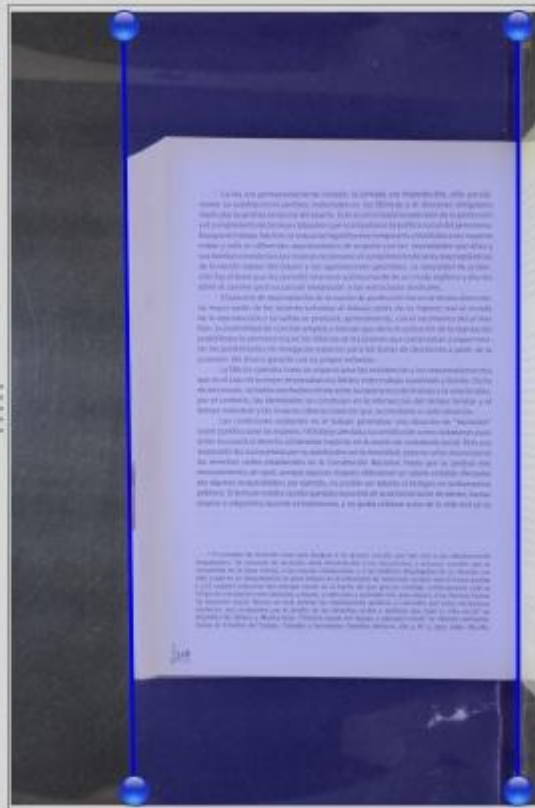
Configurar manualmente

Cambiar...

Línea divisoria

Auto

Manual



- 1 Corregir orientación
- 2 Dividir páginas
- 3 Alineación
- 4 Seleccione contenido
- 5 Márgenes
- 6 Salida

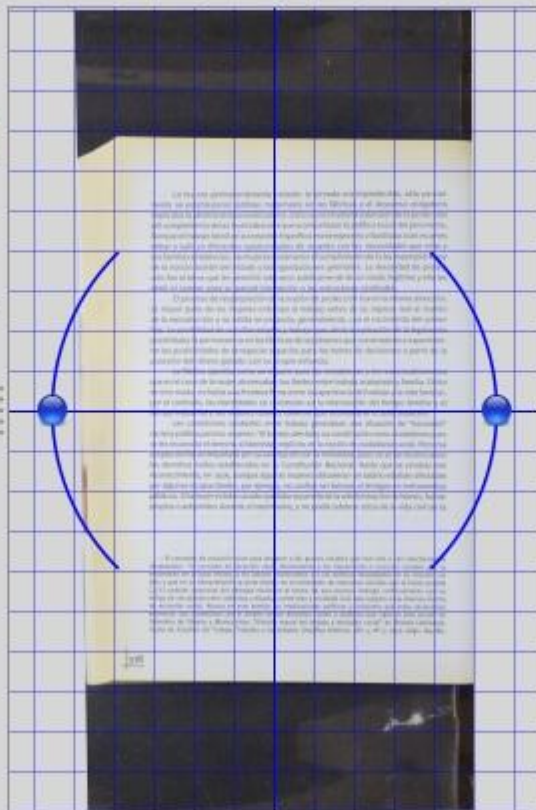
Alineación

Auto

Manual

0,19°

Aplicar a...



316.jpg

317.jpg

- 1 Corregir orientación
- 2 Dividir páginas
- 3 Alineación
- 4 **Seleccione contenido**
- 5 Márgenes
- 6 Salida

Caja de contenido

Auto

Manual

Extensión

Aplicar a...



- 3 Alineación
- 4 Seleccione contenido
- 5 Márgenes

Márgenes

Milímetros (mm)

Arriba 10,0

Abajo 5,0

Izquierda 10,0

Derecha 10,0

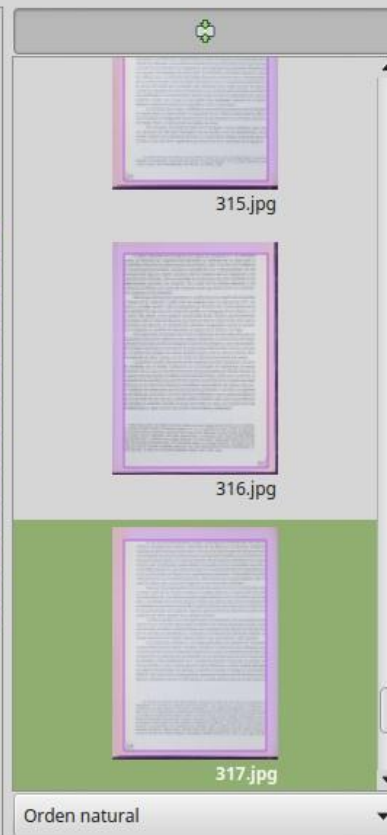
Aplicar a...

Alineación

☒ Igualar en tamaño con las demás páginas



Aplicar a...



- 4 Seleccione contenido
- 5 Márgenes
- 6 Salida

Resolución de salida (DPI)

400

Cambiar...

Modo

Combinado

0



Más fino

Más grueso

Aplicar a...

Antideformación

Apagado

Cambiar...

Eliminar manchas



Aplicar a...

La ley era permanentemente violada: la jornada era impredecible, sólo parcialmente se establecieron jardines maternales en las fábricas y el descanso obligatorio implicaba la pérdida temporal del salario. Esto ocurrió hasta la extensión de la protección y el cumplimiento de las leyes laborales que acompañaron la política social del peronismo. Aunque el trabajo fabril en la industria frigorífica era temporario y facilitaba a las mujeres entrar y salir en diferentes oportunidades de acuerdo con las necesidades que ellas y sus familias establecían, las mujeres reclamaron el cumplimiento de la ley reapropiándose de la noción tutelar del Estado y las organizaciones gremiales. La necesidad de protección fue el tema que les permitió intervenir públicamente de un modo legítimo y ello les abrió el camino para su parcial integración a las estructuras sindicales.

El proceso de reapropiación de la noción de protección fue en la misma dirección. La mayor parte de las mujeres entraban al trabajo antes de su ingreso real al mundo de la reproducción y su salida se producía, generalmente, con el nacimiento del primer hijo. La posibilidad de conciliar empleo y trabajo que abría la aplicación de la legislación posibilitaba la permanencia en las fábricas de las jóvenes que comenzaban a experimentar las posibilidades de renegociar espacios para las tomas de decisiones a partir de la posesión del dinero ganado con su propio esfuerzo.

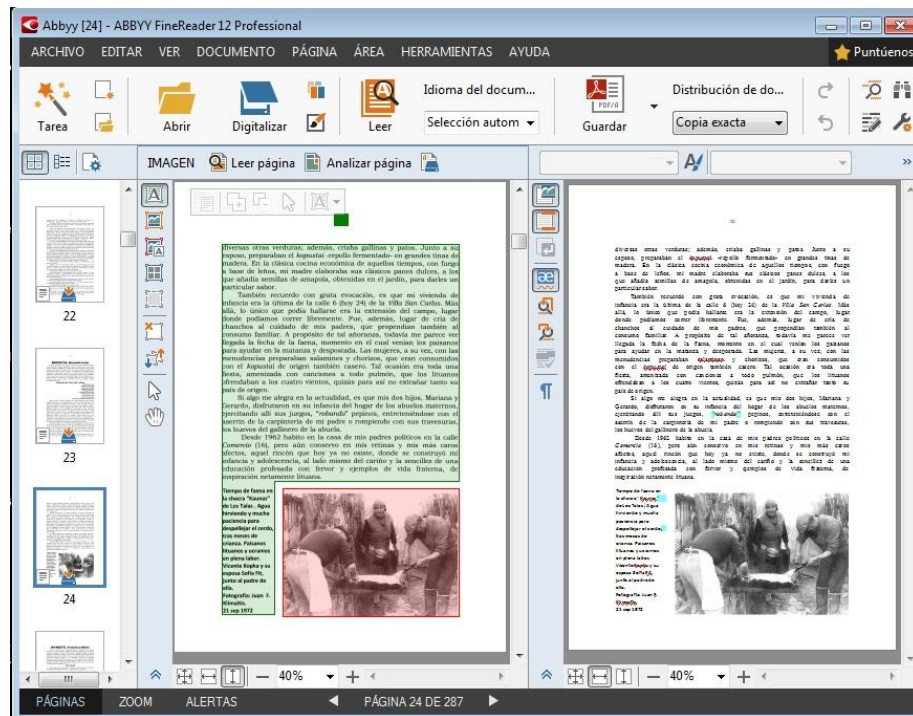
La fábrica operaba como un espacio para las resistencias y los reacomodamientos que en el caso de la mujer atravesaban los límites entre trabajo asalariado y familia. Dicho de otro modo, no había una frontera firme entre la experiencia del trabajo y la vida familiar, por el contrario, las identidades se construían en la intersección del tiempo familiar y el tiempo industrial y las mujeres obreras tuvieron que acomodarse a cada situación.

Las condiciones existentes en el trabajo generaban una situación de "exclusión" social y política para las mujeres. * El trabajo afectaba su constitución como ciudadanas pues se les reconocía el derecho al bienestar implícito en la noción de ciudadanía social. Pero esa aceptación iba acompañada por su asimilación con la minoridad, pues no se les reconocieron los derechos civiles establecidos en la Constitución Nacional. Hasta que se produjo ese reconocimiento, en 1926, aunque algunas mujeres obtuvieran un salario estaban afectadas por algunas incapacidades; por ejemplo, no podían ser tutoras ni testigos en instrumentos públicos. Si la mujer estaba casada quedaba separada de la administración de bienes, fueran propios o adquiridos durante el matrimonio, y no podía celebrar actos de la vida civil sin la

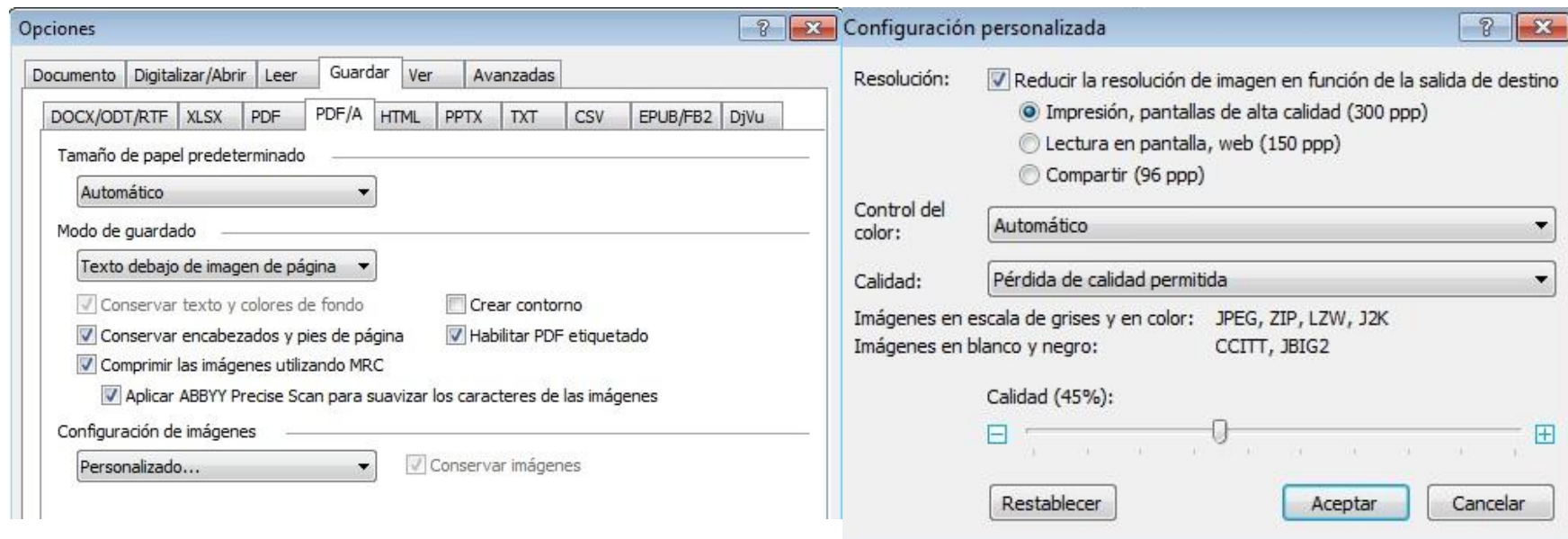
* El concepto de exclusión sirve para designar a los grupos sociales que han sido o son selectivamente desplazados: "el concepto de exclusión alude directamente a los mecanismos o procesos sociales que se encuentran en la base misma, a los actores involucrados y a las políticas desplegadas en su relación con ella; y que en su interpretación se pone énfasis en el entramado de relaciones sociales que la hacen posible [...] El carácter relacional del enfoque reside en el hecho de que procura inteligir continuamente cuál es el tipo de vinculación entre individuos y estado, y entre éste y sociedad civil, que subyace a las diversas formas de exclusión social. Abarca en este sentido las implicaciones políticas y culturales que estas exclusiones encierran, sus conexiones con el ámbito de los derechos civiles y políticos que rigen la vida social" en Orlandina de Oliveira y Marina Ariza: "División sexual del trabajo y exclusión social" en *Revista Latinoamericana de Estudios del Trabajo, Trabajo e Sociedades: Desafíos teóricos*, año 3, Nº 5, 1997, págs. 184-185.

Reconocimiento óptico de caracteres

Luego de editar las imágenes se realiza el OCR con el Abbyy FineReader. En esta etapa del proceso se selecciona el contenido según sea texto, imagen o cuadro. Luego se revisa la corrección del OCR y se generan los archivos PDF/A.



Compresión de PDF



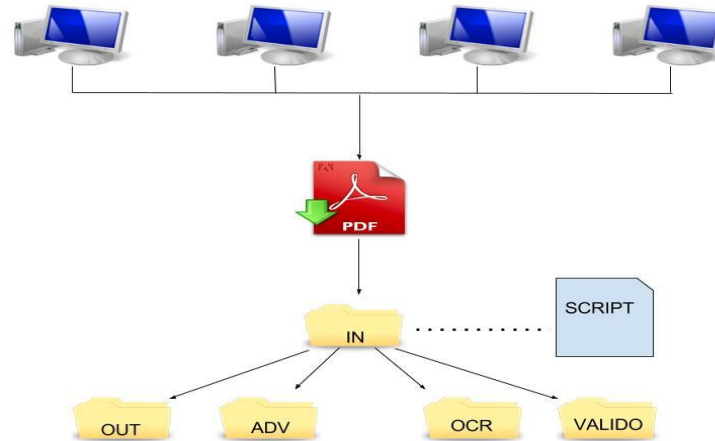
Conversión a PDF/A

2) Conversión de formatos de texto (.doc, odt, etc).

Se guarda el documento como “PDF/A” desde el mismo word/libreOffice, etc.

3) Documentos PDF. Se utiliza un programa estilo servidor/cliente, encargado de:

- Deteccion de archivos
- Análisis
- Conversión
- Verificación

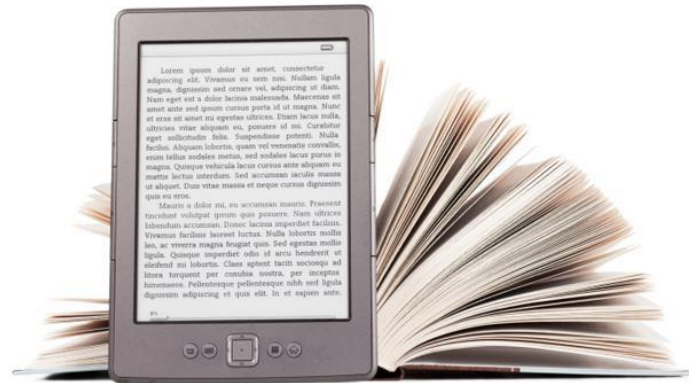


Digitalización para dispositivos móviles

MOBI

Es un formato para archivos de libro electrónico (e-book) creado por Mobipocket S. A. En el formato de libro digital MOBI se marca el contenido, pero no se delimita su formato, que se adapta a los diferentes tamaños de las pantallas de los múltiples lectores de libros electrónicos del mercado.

El formato MOBI está basado en las especificaciones Open eBook.

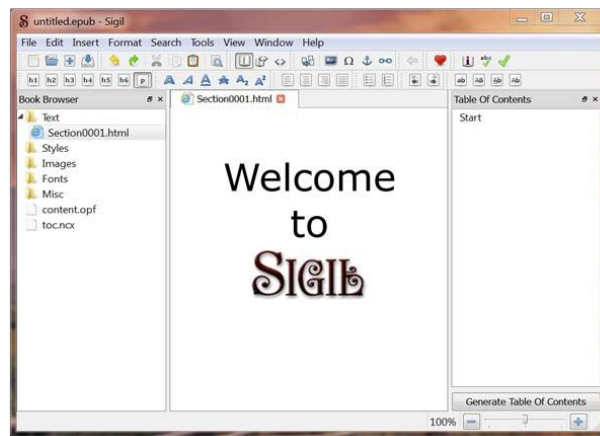


Formatos EPUB

EPUB: El formato se diseñó para adaptarse a distintos tamaños de letra y pantalla. También se puede cambiar el tipo de letra. En el formato de texto para libro digital ePub, se marca el contenido, pero no se delimita su formato, que se adapta a las diferentes pantallas y tamaños de pantalla de los múltiples dispositivos. El formato fue creado por International Digital Publishing Forum (IDPF), como un formato específico para visualizar libros.

EPUB3: El formato, además de las funcionalidades habituales del EPUB ofrece salida de audio en la funcionalidad de texto-a-audio, características de voz, normas de pronunciación, especificación de fallbacks (conversión de un vídeo en texto, para discapacitados).

Programas utilizados en el proceso



Audiolibros

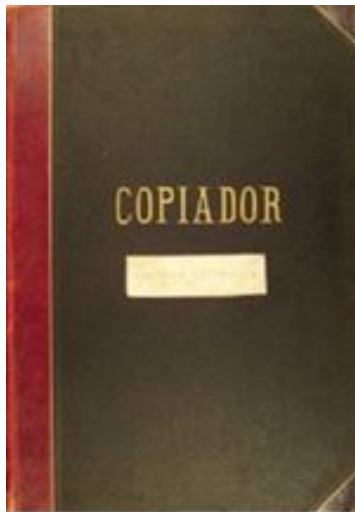
Se trata de grabaciones habladas (o leídas automáticamente por programas informáticos) de libros.

Software usado para convertir texto a voz o archivos de audio: Natural reader
Formato de salida: .MP3 o .WAV



NaturalReader is an easy-to-use software that converts written text into natural sounding spoken words giving educators, professionals and students a powerful aid and learning tool. NaturalReader can assist ESL students (English as a Second Language) and people with learning difficulties such as dyslexia.

DIGITALIZACION DEL LIBRO COPIADOR - FACULTAD DE CS. FÍSICAS, MATEMÁTICAS Y ASTRONÓMICAS (1918-1925)



SEDICI y el **Museo de Física de la Facultad de Ciencias Exactas** de la **UNLP** destinaron personal para la digitalización de un documento archivístico: el libro *Copiador – Facultad de Ciencias Físicas, Matemáticas y Astronómicas (1918-1925)*. Se siguieron los estándares internacionales para la digitalización (IFLA, NARA, FADGI, etc.), pero **muchas de las dificultades que presentó el material no estaban contempladas en la bibliografía.**

Algunas recomendaciones en digitalización

De cada obra digitalizada es recomendable tener al menos 3 versiones:

- a) imagen maestra resultante del escaneo,
- b) documento maestro con imagen derivada de la anterior y el OCR generado sobre ella y
- c) otra(s) derivada(s) a los fines de uso y distribución en la web.

Algunas recomendaciones en digitalización

Como guía en procesos de digitalización pueden usarse:

"Technical Guidelines for Digitizing Cultural Heritage Materials" generado en 2010 por la Federal Agencies Digitization Guidelines Initiative (FADGI).

"Directrices para proyectos de digitalización de colecciones y fondos de dominio público", IFLA (2002).

"Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files - Raster Images" , NARA (2004).

“Recomendaciones para la digitalización de los documentos en archivos”. Junta de Castilla y León (2011).

Enlaces de interés

<http://www.digitizationguidelines.gov/>

[Directrices IFLA para digitalización](#)

[Recomendaciones para la digitalización de archivos. Junta de Castilla y León](#)

**Aplicaciones libres para crear
repositorios de preservación**



DSpace + ARCHIVEMÁTICA



Actividades que realiza la estructura propuesta

- El repositorio en DSpace está encargado del ingreso y la entrega de los contenidos digitales.
- La estructura de Archivematica realiza las actividades de preservación digital a través de la implementación de un conjunto de microservicios, que actúan sobre una estructura conceptual asimilable al paquete de información (IP) en sus distintas versiones.
- La estructura física resultante del paquete de información en sus diferentes versiones (SIP, AIP, DIP) incluye archivos, checksum, logs, documentación de la transferencia y metadatos en una estructura XML.

Archivematica

Su estructura se basa en los microservicios y las foss tools



Archivematica

Usa los esquemas PREMIS, METS y DC.

El Escritorio de Archivematica permite seguir las acciones que suceden en los distintos procesos y microservicios.

El flujo de trabajo comienza tras una transferencia de archivos (pueden venir de un DSpace).

Usa muchas herramientas: Bagit, Fido, Siegrif, Jhove, Fits...

The screenshot shows the Archivematica web interface. At the top, there's a navigation bar with tabs: Transfer, Ingest, Archival storage, Preservation planning, Access, Administration, and a user profile 'sadie' with '1. Tabs'. Below the navigation bar, there's a search area with a text input, a dropdown menu set to 'Any', a 'Keyword' dropdown, and a 'Search transfer backlog' button. Below this is a file management section with a text input containing 'originals', a dropdown menu set to 'arrange', and buttons for 'Hide', 'View File', 'Delete', 'Create SIP', and 'Add Directory'. The main content area is a table titled 'Submission Information Package' with columns for 'UUID' and 'Ingest start time'. The table lists several submission packages, including 'r_hills_j' and 'Cape_fonds'. A dropdown menu is open for the 'r_hills_j' package, showing a list of actions: 'Normalizing for preservation and access', 'Normalizing for preservation', 'Reject SIP', 'Normalize service files for access', 'Do not normalize', 'Normalize manually', and 'Normalize for access'. Red arrows point to various parts of the interface: '2. User login' points to the user profile; '7. Report/Remove icons' points to the report and remove icons in the table; '6. Decision' points to the decision dropdown menu; '5. Jobs' points to the 'Awaiting decision' status; '4. Micro-services' points to the 'Micro-service' column; and '3. Packages' points to the submission packages.

archivematica. Transfer Ingest Archival storage Preservation planning Access Administration sadie 1. Tabs Connected

2. User login

7. Report/Remove icons

6. Decision

5. Jobs

4. Micro-services

3. Packages

Submission Information Package	UUID	Ingest start time
r_hills_j	ffa0e23e-759c-4b4d-a292-efab6499933f	2014-09-30 14:49
Micro-service: Normalize		
Job: Normalize [?]	Awaiting decision	
Job: Resume after normalization file identification tool selected.	Completed successfully	
Job: Identify file format	Completed successfully	
Job: Select pre-normalize file format identification command	Completed successfully	
Job: Move to select file ID tool	Completed successfully	
Job: Set resume link after tool selected.	Completed successfully	
Job: Find options to normalize as	Completed successfully	
Job: Move to workflowDecisions-createDip directory	Completed successfully	
Job: Grant normalization options for no pre-existing DIP	Completed successfully	
Job: Set remove preservation and access normalized files to renormalize link.	Completed successfully	
Job: Check for Access directory	Completed successfully	
Job: Check for Service directory	Completed successfully	
Job: Identify manually normalized files	Completed successfully	
Micro-service: Clean up names		
Micro-service: Remove cache files		
Micro-service: Include default SIP processingMCP.xml		
Micro-service: Rename SIP directory with SIP UUID		
Micro-service: Verify transfer compliance		
Micro-service: Verify SIP compliance		
Micro-service: Approve SIP creation		
Job: Create removal from backlog PREMIS events	Completed successfully	
Job: Approve SIP Creation [?]	Completed successfully	
Cape_fonds	5739137c-1887-4cdd-b78f-8899aa1830f4	2014-09-29 13:17
Micro-service: Upload DIP		

Funciones de Archival Storage

- Archivemática utiliza una estructura de árbol de directorios para almacenar los AIP localmente.
- La estructura en árbol está basada en los identificadores persistentes (16 bits) del AIP; también permite múltiples sitios de almacenamiento locales o remotos e incluso localizaciones LOCKSS.
- El detalle del procedimiento de integración se encuentra en la wiki de Archivemática.

archivemática

Transfer ² Backlog Appraisal Ingest Archival storage Preservation planning Access Administration demo@example.com

Archival storage / Search

Any

Keyword

Show files? ☐ Show AICs? ☐

Search archival storage

Add New

Browse archival storage

Total size: 28.50 MB | Files indexed: 10

AIP	Size	UUID	Date stored	Status	Actions
Images	28.50 MB	754d849a-e5f2-49e0-a87e-3ed6b0695f5a	2016-12-20 14:03	Stored	View

Found 1 entries. Showing 1 to 1.

archivemática

Transfer ² Backlog Appraisal Ingest Archival storage Preservation planning Access Administration demog

Archival storage / Search

Any

Keyword

Show files? ☒ Show AICs? ☐

Search archival storage

Add New

Found 2 entries. Showing 1 to 2.

File(s)

AIP(s)



WFFC01.GIF
de2c0b4c-e691-4b69-bfb3-823a0e97d278



Landing_zone.jpg
592cd421-8961-439d-8205-fae27def3f33



Images 754d849a-e5f2-49e0-a87e-3ed6b0695f5a
([view raw](#))



Images 754d849a-e5f2-49e0-a87e-3ed6b0695f5a
([view raw](#))

Archives Space: descripción y funciones



ArchivesSpace es un software abierto de gestión de archivos que permite dar seguimiento a las sesiones de acceso y la gestión de colecciones. Permite:

- ❑ La incorporación de nuevos registros
- ❑ La publicación de materiales
- ❑ La gestión de autoridades
- ❑ La gestión de lugares
- ❑ La gestión de derechos
- ❑ Brinda servicio de referencia
- ❑ Ofrece generación de informes y reportes
- ❑ Brinda la posibilidad de exportaciones
- ❑ Genera metadatos EAD, MARCXML, MODS, Dublin Core, y METS

RODA



Welcome to RODA!

An open-source digital repository designed for preservation

RODA is a complete digital repository solution that delivers functionality for all the main functional units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to the various types of digital content produced by large corporations or public bodies. RODA is based on open-source technologies and is supported by existing standards such as the Open Archival Information System (OAIS), Metadata Encoding and Transmission Standard (METS), Encoded Archival Description (EAD), Dublin Core (DC) and PREMIS (Preservation Metadata).



Conforms to open standards

RODA follows open standards using EAD for description metadata, PREMIS for preservation metadata, METS for structural metadata, and several standards for technical metadata (e.g. NISO Z39.87 for digital still images).



Vendor independent

RODA is 100% built on top of open-source technologies. The entire infrastructure required to support RODA is vendor independent. This means that you may use the hardware and Linux distributions that best fit your institutional needs.



Scalable

The service-oriented nature of RODA's architecture allows the system to be highly scalable, enabling the distribution of the processing load between several servers. Furthermore, the use of advanced indexing components enable RODA's discovery services to be spread through various servers on a cluster for even greater performance.



Embedded preservation actions

Preservation actions and management within RODA is handled by a job execution module. The job execution module allows the repository manager to run preservation tasks over a given set of data. Preservation actions include format conversions, checksum verifications, reporting (e.g. to automatically send SIP acceptance/rejection emails), virus checks, etc.



Authenticity

RODA uses preservation metadata (PREMIS) to create a trust



Support for multiple formats

RODA is capable of ingesting all sorts of content types.

Roda: Datos generales y de contacto

- Web del proyecto: <http://ww2.roda-community.org/>
- Licencia: GPLv3
- Lenguaje de programación: JAVA
- Sistema de gestión de bases de datos: MySQL y PostgreSQL.
- Plataforma: Multiplataforma.
- Comunidad: Principalmente administración pública portuguesa

<http://roda-community.org/governance>

Roda: Datos generales

- Ámbito de aplicación/audiencia: Administración pública, instituciones académicas, culturales, centros de investigación, archivos y bibliotecas.
- Certificaciones y actuación de acuerdo a otros proyectos: Trustworthy Repositories Audit & Certification y Plato (en proceso).

Roda: Gestión del repositorio

- Gestión de usuarios y permisos: Alta. configuración de permisos y roles.
Permite definir grupos y subgrupos de usuarios y la definición de los niveles de seguridad y acceso puede aplicarse incluso a nivel ítem.
- Interoperabilidad: SOAP o mediante una API.
- Estándares de metadatos: PREMIS, EAD, MIX, METS.
- Escalabilidad: Muy alta. Permite distribuir procesos entre servidores.
- Búsqueda y recuperación: Apache Lucene.

Roda: Gestión de la preservación digital

- Soporte modelo OAIS: Sí, funciona de acuerdo al modelo OAIS y da soporte a cada una de las fases (SIP, AIP, DIP).
- Formatos de archivo: Gran soporte para múltiples formatos y la posibilidad de integrar más a partir de plugins.
- Validación: JHOVE, Droid.
- Trazabilidad: PREMIS.

Roda: Gestión de la preservación digital

- Autenticidad: Sumas de verificación (Checksum).
- Creación y gestión de formatos derivados: GhostScript, JOD Converter and OpenOffice, Mencoder, SoundConverter, gStreamer.
- Gestión de la migración de formatos: Notifica la obsolescencia de formatos. Dispone de un asistente para la migración. Herramienta de evaluación de los objetos resultantes.
- Virus: Comprobación de virus integrada.

Gestión del Acceso Público

- Capa pública: La propia de RODA.
- Facilidad de implementación: Media.



Primary links

- [Home](#)
- [Documentation](#)
- [Download](#)
- [News Archive](#)
- [Contact](#)

07/19/2012 - 13:22

Updated Chapter 5: The
DAITSS Archiving Process

An updated version of the
DAITSS documentation Chapter
5: The DAITSS Archiving
Process is now available from
the Documentation section of
the DAITSS website. The SIP
specifications section is now
consistent with the behavior of
DAITSS v.2.17.19.

[more](#)

Welcome to the DAITSS website!

DAITSS is a digital preservation software application developed by the Florida Center for Library Automation (FCLA) with some support from the IMLS. DAITSS is used by the [Florida Digital Archive \(FDA\)](#), a long-term preservation repository service provided by the [Florida Virtual Campus](#) for the use of the libraries of the eleven publicly-funded universities in Florida.

DAITSS provides automated support for the functions of Submission, Ingest, Archival Storage, Access, Withdrawal, and Repository Management. It is architected as a set of RESTful Web Services and micro-services but enforces strict controls to ensure the integrity and authenticity of archived content. It implements active preservation strategies based on format-specific processing including, where necessary, normalization and forward migration. It is particularly well suited for materials in text, document, image, audio and video formats.

DAITSS was written for a multi-user environment and supports consortial as well as institutional preservation repositories.

DAITSS is available for use through a GPLv3 license. This website provides access to a fully configured VM version of DAITSS that can be downloaded to run under any VM manager, along with sample SIPs (submission packages) and documentation.

For more general information about DAITSS, see:

- [DAITSS, an OAIS-based Preservation Repository](#)
- [DAITSS Grows Up: Migrating to a Second-Generation Preservation System](#)
- [The Florida Digital Archive and DAITSS](#)
- [Chapter 1 of the DAITSS OSS User Manual](#)

For DAITSS documentation, see the [Documentation](#) section of this website, or [Contact Us](#).

Download a DAITSS demo virtual machine *Note that this version of the VM will not work with PREMIS v. 3.0.*

TRY the DAITSS stand-alone [Description Service](#) which performs format identification, validation and characterization on a given file.

Datos generales y de contacto

- Web del proyecto: <http://daitss.fcla.edu/>
- Licencia: GPLv3
- Lenguaje de programación: Ruby
- Sistema de gestión de bases de datos: PostgreSQL
- Plataforma: Multiplataforma
- Comunidad: Proyecto del Florida Center for Library Automation.
- Ámbito de aplicación/audiencia: academia, archivos y bibliotecas.
- Certificaciones y actuación de acuerdo a otros proyectos: TRAC

Gestión del repositorio

- Gestión de usuarios y permisos: configuración de permisos y roles.
- Interoperabilidad: Actúa como “caja negra”. Sólo es posible acceder a los datos del repositorio previa solicitud del DIP a través de la interfaz.
- Estándares de metadatos: PREMIS, DAITSS XML, MODS, MIX, AES, textMD, docMD, METS.
- Escalabilidad: Muy Alta.
- Búsqueda y recuperación: no menciona.

Gestión de la preservación digital

- Soporte modelo OAIS: Sí. Para SIP, AIP y DIP.
- Formatos de archivo: 600 tipos de formatos diferentes.
- Validación: JHOVE.
- Trazabilidad: PREMIS.
- Autenticidad: Sumas de verificación (Checksum) MD5.
- Creación y gestión de formatos derivados: Sí, de los archivos propietarios se crean versiones en formatos abiertos.

Gestión de la preservación digital

- Gestión de la migración de formatos: Sí, almacena una copia del fichero en la versión actual (soporte para los principales formatos del mercado como los de Microsoft, pdf, etc.).
- Virus: Comprobación de virus integrada.
- Zona de cuarentena: Sí.

Gestión del acceso público

- Capa pública: No dispone de capa pública se debería integrar con otro software y no se conoce su nivel de compatibilidad con otras herramientas del mercado.
- Facilidad de implementación: Baja.

Resumiendo PD en RI

Regulación de todos los procedimientos.

Regulación de los derechos de preservación digital sobre los documentos.

Regulación de los formatos admisibles.

Control de formatos en la ingestión.

Formatos de visualización y de preservación

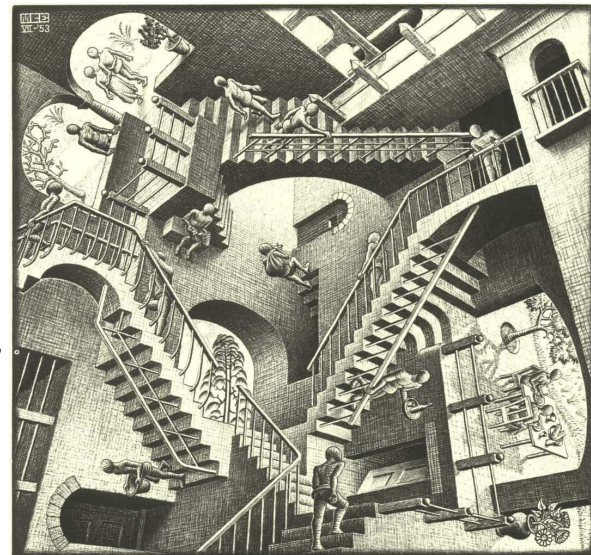
Almacenaje de metadatos técnicos.

Copias sistemáticas externas.

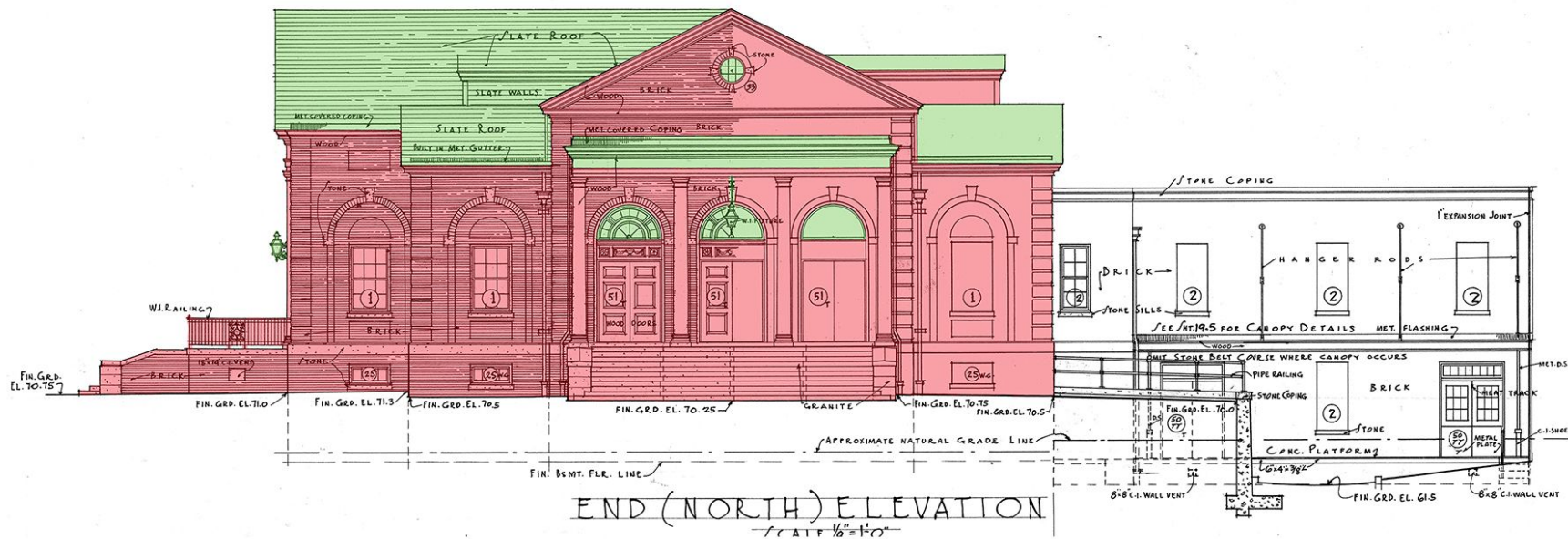
Creación de procedimientos de contingencia ante desastres.

Auditoría interna/externa de seguridad.

Plan de preservación...



Plan de preservación



Plan de preservación: cuestiones a tener en cuenta

¿Qué materiales hay en los RI?

- resultados de la investigación (tesis, e-prints);
- objetos de docencia y aprendizaje;
- datos no elaborados;
- fondos digitalizados;
- material administrativo.

¿Qué materiales se tienen que preservar a largo plazo?

Criterios tradicionales para tomar la decisión sobre la preservación a largo plazo, principalmente los factores de: valor, pertinencia, uso

- Otros condicionantes: misión, disponibilidad de recursos humanos, económicos, materiales, obligaciones legales o contractuales.

Plan de preservación: cuestiones a tener en cuenta

Criterios nuevos para los recursos digitales:

- que la institución tenga pleno derecho a manipular los datos para asegurar su acceso en entornos informáticos del futuro;
- que el recurso sea de un formato legible actualmente y previsiblemente en el futuro;
- que el recurso esté en un soporte gestionable para su transferencia y/o almacenamiento;
- que el recurso disponga de documentación, incluyendo los metadatos.

Plan de preservación: cuestiones a tener en cuenta

Selección de recursos para su preservación

¿Qué formatos? ¿qué versiones? ¿qué material adicional incluir?

¿Qué atributos se quieren preservar?

datos, funcionalidad

apariencia, esencia

La decisión dependerá de la misión institucional, las necesidades de la comunidad de usuarios, la capacidad técnica/ tecnológica institucional y los recursos disponibles.





¡Muchas gracias!

Por consultas: marisa.degiusti@sedici.unlp.edu.ar

Nuestros sitios

<http://sedici.unlp.edu.ar>

<http://digital.cic.gba.gob.ar/>

<http://cesgi.cic.gba.gob.ar/>

<http://prebi.unlp.edu.ar>

<http://www.istec.org/liblink/>

<http://revistas.unlp.edu.ar/cientificas/>

<http://revistas.unlp.edu.ar>

<http://congresos.unlp.edu.ar>

<http://ibros.unlp.edu.ar>

La presentación está disponible en la colección de SEDICI <http://sedici.unlp.edu.ar/handle/10915/25295>