Curso de extensão de Programa de Preservação Digital

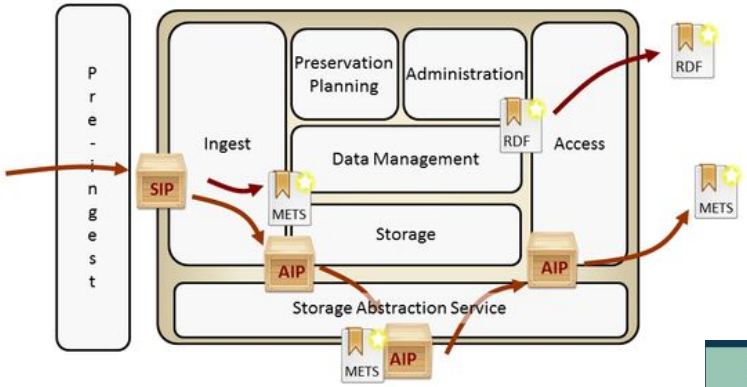
Curso de extensión del Programa de Preservación Digital

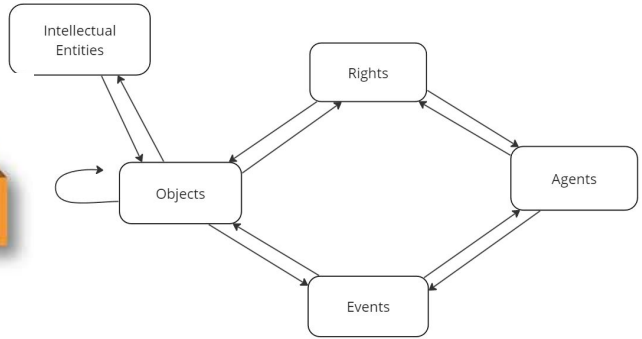
Clase 4: Preservación en repositorios

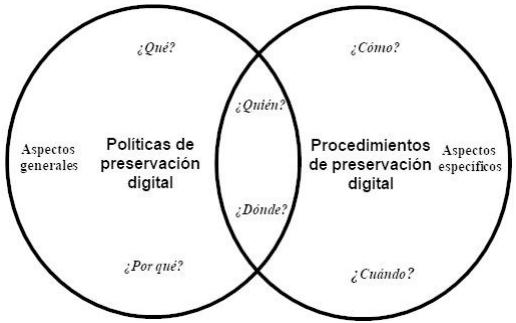
Prof. Dra. Marisa De Giusti
Universidad Nacional de La Plata e ISTEC

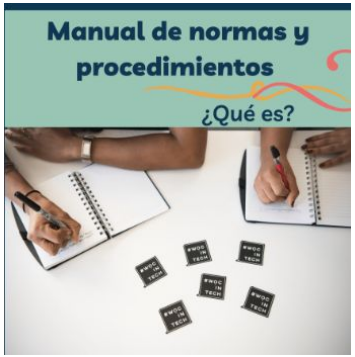
26 de setiembre de 2024

Algunos protagonistas de la preservación digital



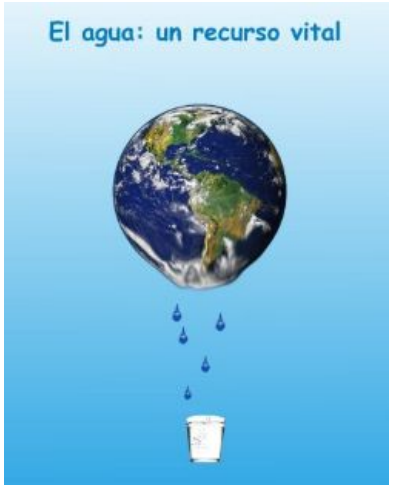






Características de la información que produce conocimiento

- Debe ser suficiente para la toma de decisiones
- No debe ser demasiada porque no puede gestionarse
- Contaminada no sirve
- Desactualizada no sirve
- Si no se usa, se inutiliza, desaparece
- Si no se conoce la fuente no es confiable
- Si no se puede acceder, es inútil



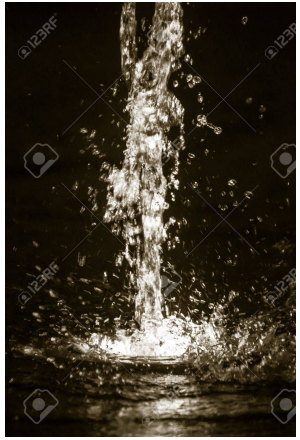
La información digital tiene los mismos objetivos pero características distintas de producción, organización, distribución, acceso y preservación.

- La información digital puede disociarse de su soporte: mejor diseminación pero mayor fragilidad.

Información digital

- Se representa por medio de una secuencia de bits
- Precisa de un formato para su representación
- Se registra en un objeto digital
- Se almacena en un soporte digital
- Necesita de hardware para registrarla, gestionarla así como dar acceso.
- Necesita de un software para crearla e interpretarla.
- Precisa de mecanismos para buscarla.

Información digital



- Precisa de cambios en el tiempo para mantenerla y por eso hay que asegurar su autenticidad.
- Puede estar representada en objetos digitales distintos.
- Puede estar almacenada en soportes digitales diferentes.
- Posibilita la compartición electrónica: para muchos y muy rápido.
- Puede almacenarse en grandes cantidades en un soporte que ocupa poco espacio físico.

La preservación de los contenidos

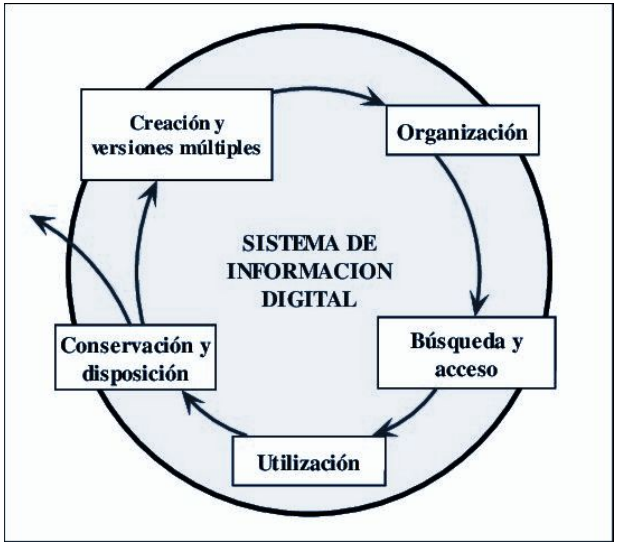
En los documentos en papel se habla de “negligencia benigna”: el olvido de un manuscrito en un arcón, puede que lo preserve. En los digitales, no existe negligencia benigna: un disco olvidado 5 años… no sirve.

En el caso de un documento digital se debe tener en cuenta:

- No a la negligencia benigna.
- No a la preservación basada en las condiciones ambientales.
- No se conserva para cualquier usuario futuro sino para una comunidad designada: el conjunto de los consumidores que tienen que entender la información almacenada.
- No necesariamente se conserva la integridad externa del documento sino las propiedades significativas.
- Se debe asegurar la autenticidad del recurso.

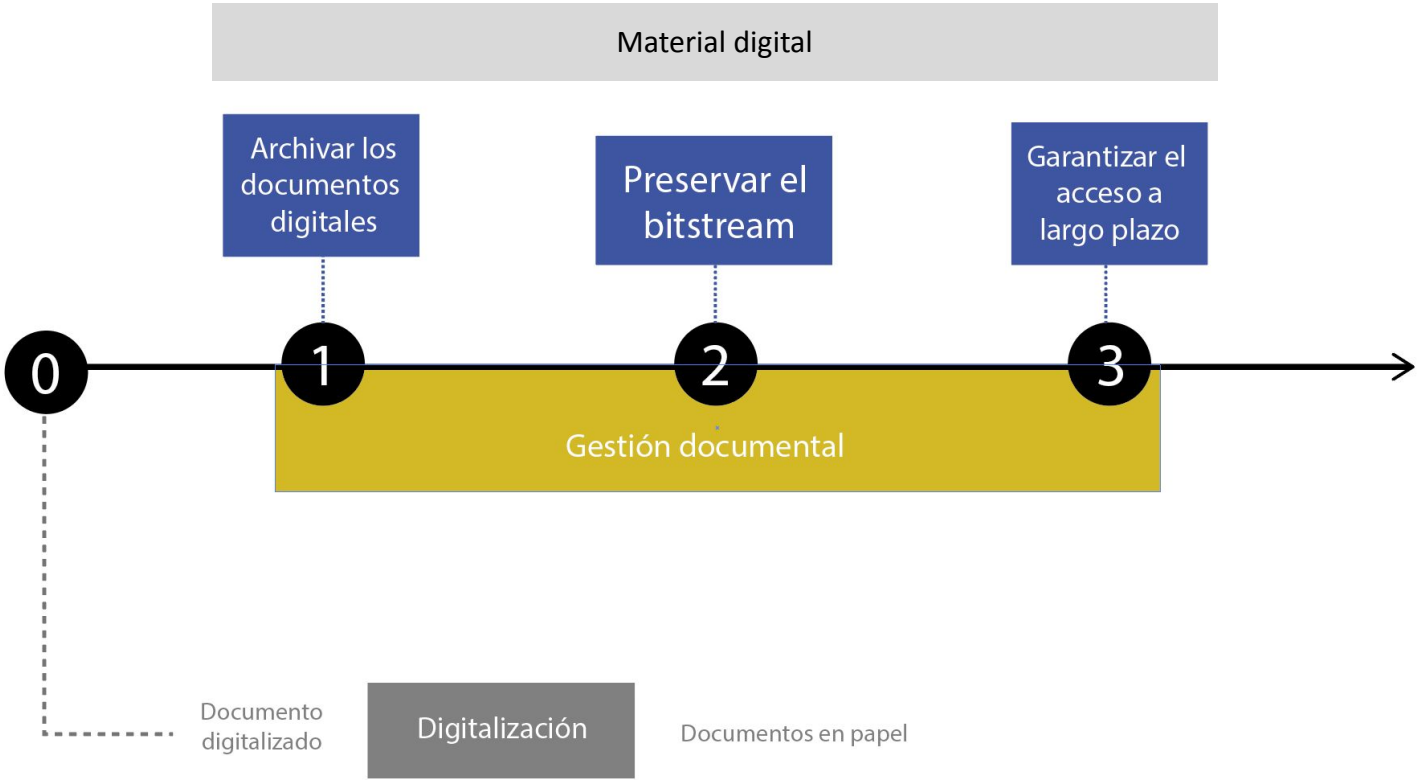
Objeto digital: información almacenada en un medio digital

Está claro que hay que realizar acciones en todo su ciclo de vida para mantener el acceso



De Giusti, M. R. (2014). *Una metodología de evaluación de repositorios digitales para asegurar la preservación en el tiempo y el acceso a los contenidos* (Doctoral dissertation, Universidad Nacional de La Plata). Disponible en: https://sedici.unlp.edu.ar/handle/10915/43157

En el repositorio: objetos nacidos digitales o digitalizados



SEDICI

UNIVERSIDAD NACIONAL DE LA PLATA

Buscar material

Busque entre los **161365** recursos disponibles en el repositorio

Tipo de documento	Formato de acceso
Artículo (58102)	PDF/PDF-A
Audio (528)	MP3
Boletín (593)	PDF/PDF-A
Capítulo de libro (3402)	PDF/PDF-A
Clase (243)	PDF/PDF-A
Comunicación (6302)	PDF/PDF-A
Conjunto de datos (96)	
Contribucion a revista (4998)	PDF/PDF-A
Convenio (66)	PDF/PDF-A
Discurso (87)	MP3
Documento de trabajo (814)	PDF/PDF-A
Documento institucional (1015)	PDF/PDF-A
Edición de revista (1406)	PDF/PDF-A
Edición de revista (1)	PDF/PDF-A
Entrevista (291)	MP3
Imagen fija (71)	JPEG
Informe (256)	PDF/PDF-A
Informe de proyecto (9)	PDF/PDF-A
Instrumento científico (17)	JPEG, MPEG-4, enlaces
Instrumento musical (162)	JPEG, enlaces
Libro (2576)	PDF/PDF-A
Material complementario (467)	PDF/PDF-A
Música (15)	MP3 y en caso de partituras PDF
Objeto de aprendizaje (41)	ZIPs y enlaces externos
Objeto de conferencia (48885)	PDF/PDF-A
Objeto físico (310)	JPEG+ descripción en PDF, enlace
Ordenanza (108)	PDF/PDF-A
Otros (85)	
Placa espectrográfica (253)	FITs, PNGs
Plano (20)	PDF
Preprint (1442)	PDF/PDF-A

Tipo de documento	Formato de acceso
Programa (198)	Programas de cátedra. PDF/PDF-A
Proyecto de extensión (1058)	PDF/PDF-A
Proyecto de investigación (25)	PDF/PDF-A
Publicación seriada (298)	PDF/PDF-A
Reporte (175)	PDF/PDF-A
Reporte técnico (378)	PDF/PDF-A
Resolución (933)	PDF/PDF-A
Resumen (5046)	PDF/PDF-A
Revisión (4409)	PDF/PDF-A
Tesis de doctorado (6143)	PDF/PDF-A
Tesis de grado (5127)	PDF/PDF-A
Tesis de maestría (2048)	PDF/PDF-A
Testimonio (122)	MP3
Trabajo de especialización (2063)	PDF/PDF-A
Trabajo práctico (310)	PDF/PDF-A
Video (306)	MPEG-4 y enlaces a Youtube

Desafíos para la preservación en los repositorios

- Gran cantidad y variedad de objetos digitales.
- Frágiles: pueden ser modificados, borrados por negligencia, error o acción consciente. La fragilidad tiene implicancias en la autenticidad e integridad.
- La obsolescencia de los medios informáticos: dado que los OD siempre están mediados por la tecnología que cambia constantemente (fueron creados para un dado software y hardware); 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.
- Cuestiones legales, derechos.
- Mecanismos de búsqueda y recuperación.
- Recursos humanos y tecnológicos
- Cultura de cambio.

A pesar de todo las instituciones buscan preservar

- Por razones patrimoniales
- Por memoria
- Por aspectos legales
- Para brindar un servicio a los usuarios
- Para que la información pueda reutilizarse
- Para no repetir…

BIBLIOTECA NACIONAL DE ESPAÑA

BNE

https://www.bne.es/es/blog/blog-bne/que-es-y-que-no-es-la-preservacion-digital

La preservación en el repositorio supone que:

1. Los datos se mantendrán en el repositorio sin sufrir daños, sin perderse o sin ser alterados de forma malintencionada/o no.
2. Los datos podrán ser localizados y entregados al usuario.
3. Los datos podrán ser interpretados y comprendidos por el usuario.

1, 2 y 3 serán realizables a largo plazo.

Preservación de la información: aspectos a considerar

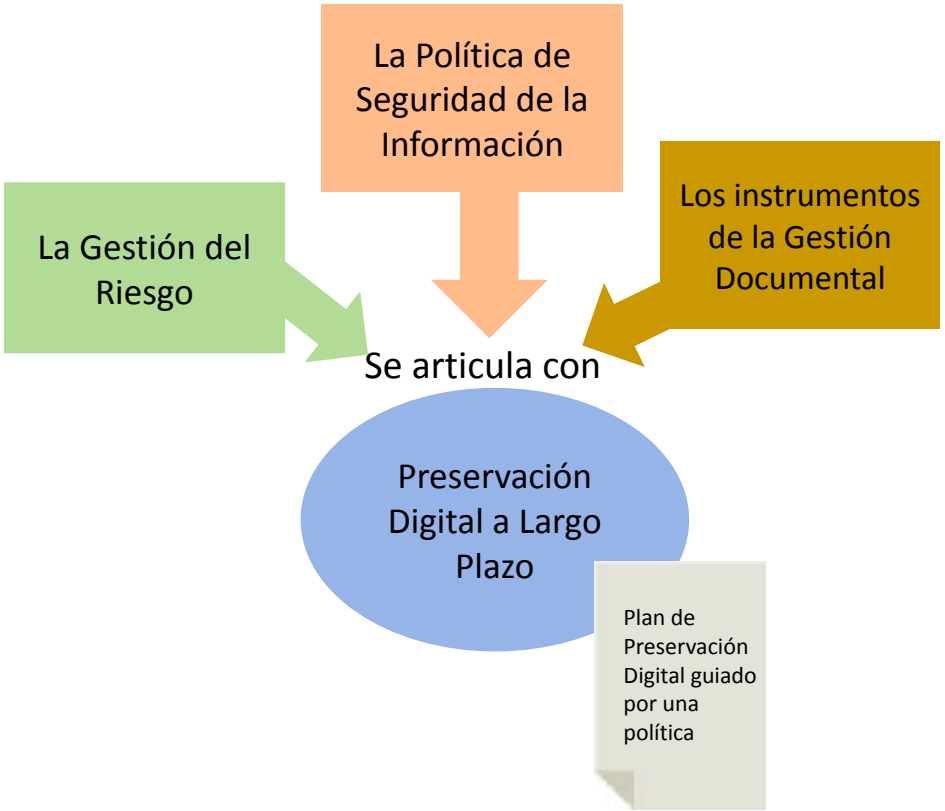
- que la institución tenga pleno derecho a manipular los datos para asegurar su acceso en entornos informáticos del futuro;
- que el recurso sea de un formato legible actualmente y previsiblemente en el futuro;
- que el recurso esté en un soporte gestionable para su transferencia y/o almacenamiento;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

Preservación del contenido en repositorios

La clave: PLANIFICAR

¿Qué materiales se tienen que preservar a largo plazo? Se atiende a 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.



¿Por dónde comenzar en PD?

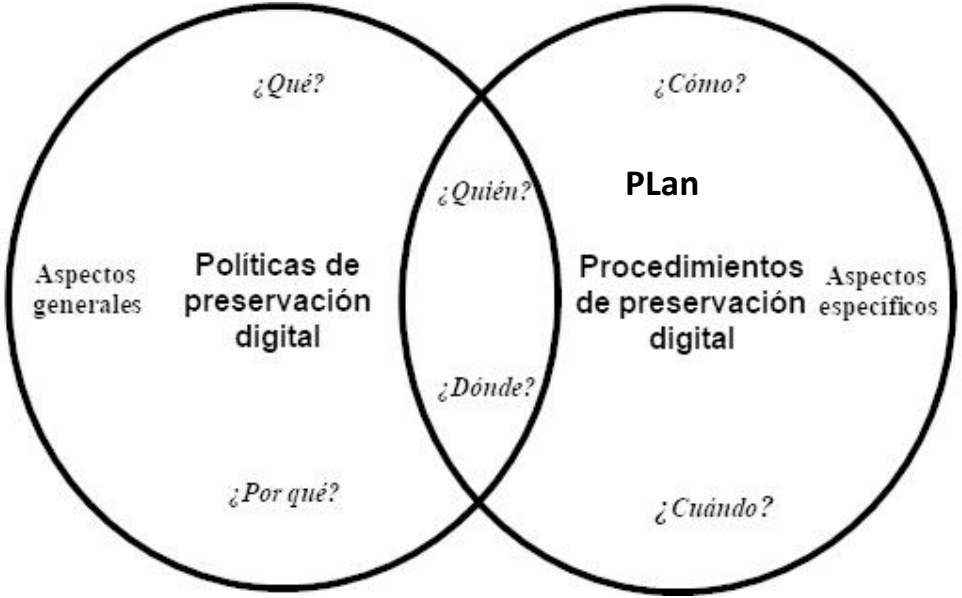
- Pensar desde la institución
  - ¿Por qué se debe preservar?
  - ¿Qué se debe preservar?
  - ¿Qué beneficios acarreará a la institución?
  - ¿Qué personal deberá involucrarse?
  - ¿Qué acciones se deberán llevar adelante?
  - ¿Qué plazos?
  - ¿Con qué recursos?

¿Por dónde comenzar en PD?

Elementos de la PD

1. Objetivos de la institución
2. Equipo interdisciplinario
3. Responsabilidades/ Responsables
4. Recursos económicos
5. Aspectos legales
6. Selección de materiales
7. Modelos, patrones, recomendaciones
8. Infraestructura tecnológica
9. Repositorios institucionales
10. Estrategias de preservación
11. Auditoría y certificación

Operativo y documental



Tener un plan de preservación

- El plan debe exponer los motivos, principios y sobre qué contenidos va a centrarse para garantizar la conservación, el acceso y la comprensión a largo plazo de esos fondos.
- Debe identificar necesidades y prioridades y aportar un cronograma pormenorizado que muestre la distribución de las tareas en el periodo de vigencia del plan.
  - El primer paso es identificar los contenidos en digital (qué prima en el repositorio).
  - En qué formatos y versiones.
  - Diseñar los procedimientos de prevención de riesgos y desastres.
  - Diseñar los procedimientos que indiquen las acciones sobre los ODs según tipología y su frecuencia, determinando claramente los responsables.
  - Hacer un cronograma y marcar el estado de las acciones.

Plan de preservación: estructura

1. Situación actual del repositorio: análisis de activos.
2. Definición del plan de preservación:
    a. Ámbito
    b. Objetivos: por ejemplo uso de estándares internacionales, metadatos…
    c. Procedimientos
3. Estrategias:
    a. Modelo de gestión
        i. Recursos humanos
        ii. Financiación
4. Diseño: dimensión de la colección, requisitos del sistema de almacenamiento, digitalización, metodología.
5. Ejecución y difusión.

Planes de preservación:
https://repositorio.uloyola.es/bitstream/handle/20.500.12412/2164/PlanPreservacionBiblioteca.pdf?sequence=1&isAllowed=y
https://drive.google.com/drive/u/0/folders/1NMu3lhXEKdtZR1cIfXM_yAoaxsvBfeRu

Elementos técnicos destacados

Buscan garantizar:

- autenticidad de los objetos digitales
- identificación unívoca
- formatos válidos
- descripción de objetos digitales por metadatos
- uso de modelos y estándares
- infraestructura tecnológica necesaria
- aplicación de estrategias de preservación
- garantía de acceso a objetos digitales preservados
- Seguimiento

Autenticidad de los objetos digitales

Fixity o fijeza, se refiere a que los archivos mantengan su estructura original a nivel de bit, sin alteración alguna. Si las palabras en un archivo de texto se cambian, o se cambia la organización del texto, el archivo cambia a nivel de bit. Por consiguiente, pierde su fijeza, ya que su integridad a nivel de bit se ha perdido. Asegurar la fijeza es esencial en términos de proteger la confiabilidad de un objeto digital.

Un plan de preservación digital, además de atender el almacenamiento de los objetos digitales y el acceso a los mismos, debe velar por que se mantenga la fijeza de dichos objetos.

Los "checksums" se utilizan a menudo para confirmar la fijeza de los archivos. Los "checksums" se generan aplicando un conjunto de operaciones matemáticas al archivo. Esto produce una secuencia de caracteres, la cual constituye el "checksum" de ese archivo. Si el archivo se altera de alguna manera, el "checksum" que produce también va a cambiar. Así es que los "checksums" ayudan a detectar alteraciones a los objetos digitales.

Identificadores persistentes

Los identificadores persistentes son una referencia imperecedera a un documento, archivo, página web u otro objeto. La función principal de los identificadores persistentes es crear enlaces fiables y permanentes entre diferentes elementos dentro del ecosistema global de investigación.

Existen múltiples identificadores persistentes para una diversidad de objetos:

- Persistent URL (PURL)
- Digital Object identifier (DOI), Handle, Archival Resource Key (ARK)
- ORCID (Open Research and Contributor ID), Research ID, INSI (Identificador Estándar Internacional de Nombres)
- ROR (Research Organization Registry)

Técnico y político

Identificadores persistentes: el habitual en repositorios Dspace

¿Qué es el Handle?

- Es un sistema abierto que permite la asignación de identificadores persistentes a los objetos digitales de Internet (artículos, revistas, imágenes, etc.), es decir, es una URL que no varía aunque la página cambie de ubicación.
- Cada *handle* desarrollado por la **CNRI (Corporation for National Research Initiatives)** se estructura en dos bloques: por ejemplo en el http://hdl.handle.net/10230/22749
  - el **prefijo 10230** identifica al productor (universidad, editorial, revista, etc.). en este caso la UPF
  - el **sufijo 22749** identifica a cada uno de los documentos o obras ditigales (artículos, libros, capítulos, etc.)

Recomendaciones de formatos: accesibilidad

Aunque la definición de los formatos para preservación puede variar de institución a institución, se recomienda que estos sean:

- No propietarios
- Estándares abiertos y documentados
- Utilizados comúnmente dentro de la comunidad de investigación
- Transmitidos mediante formas de representación estándar (ASCII, Unicode)
- No encriptados
- Sin compresión

https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/formatos

Preservación de contenido. Elección de formatos de archivo. Bibliografía

MIT Libraries. (s. f.). File formats for long-term access [Blog]. https://libraries.mit.edu/data-management/store/formats/

Open Knowledge International. (s. f.). Formato de Archivos [Webpage]. Recuperado 26 de setiembre de 2024, a partir de http://opendatahandbook.org/guide/es/appendices/file-formats/ Repaso de formatos en general.

https://library.duke.edu/using/policies/recommended-file-formats-digital-preservation. Recuperado 26 de setiembre de 2024.

Recomendaciones de formatos para REUSO

Formatos de archivo FAIR

- Contenedores: TAR, GZIP, ZIP
- Bases de datos: XML, CSV, JSON
- Geoespacial: SHP, DBF, GeoTIFF, NetCDF
- Video: MPEG, AVI, MXF, MKV
- Sonido: WAVE, AIFF, MP3, MXF FLAC
- Estadísticas: DTA, POR, SAS, SAV
- Imágenes: TIFF, JPEG2000, PDF, DNG, GIF, BMP, SVG
- Datos tabulares: CSV, TXT
- Texto: XML, PDF / A, HTML, JSON, TXT, RTF
- Archivo web: WARC

https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/formatos

Selección de formatos: algunos especiales

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: algunos especiales

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 y el JPEG2000.

Selección de formatos: algunos especiales

En cuanto a los Formatos mixtos (contenedores) se recomienda 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 particularmente el PDF/A

Sobre PDF/A

PDF/A es un estándar para codificar documentos en un formato que es portable entre sistemas y ampliamente usado para distribución y archivado de documentos.

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, si se usa o no 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.

Estándares de PDF/A

Con el tiempo surgieron nuevos estándares para el PDF/A, que no implican –por definición– la obsolescencia de las versiones anteriores, sino la ampliación de las posibilidades de archivo.

Estándar **PDF/A**	Subnivel	Norma ISO	PUID	Versión PDF
1	a	ISO 19005-1:2005	fmt/95	PDF Reference third edition pdf 1.4 fmt/18
	b		fmt/354	
2	a	ISO 19005-2:2011	fmt/476	ISO 32000-1 pdf 1.7 fmt/276
	b		fmt/477	
	u		fmt/478	
3	a	ISO 19005-3:2012	fmt/479	ISO 32000-1 pdf 1.7 fmt/276
	b		fmt/480	
	u		fmt/481	
4	e	ISO 19005-4:2020		ISO 32000-2 pdf 2 fmt/1129
	f			

	PDF	PDF/A-1	PDF/A-2	PDF/A-3	PDF/A-4
Enlaces a recursos externos	✅	x	x	x	x
Fuentes embebidas	x	✅	✅	✅	✅
Código ejecutable	✅	x	x	x	✅
Cifrado	✅	x	x	x	x
Audio	✅	x	x	x	x
Video	✅	x	x	x	x
Incrustación de otros archivos	✅	x	✅*	✅**	✅
Metadatos estandarizados	x	✅	✅	✅	✅
Inclusión de perfiles de color	x	✅	✅	✅	✅
Transparencias	✅	x	✅	✅	✅
Objetos 3D interactivos	✅	x	x	x	✅

Diferencias entre PDF y PDF/A (1, 2, 3 y 4)

NOTAS:
* Solo archivos PDF/A-1 o PDF/A-2.
** Permite la inclusión de archivos que no sean a su vez PDF/A-1 o PDF/A-2, aunque les impone ciertas restricciones (definidas en el estándar ISO 32000-1).

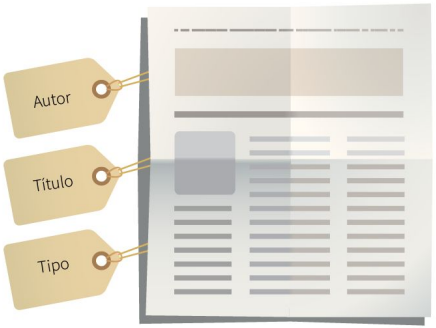
Siguiendo con los repositorios. Formatos.
¿Cómo conocer lo que tiene un RI?

Perfilamiento automatizado de los objetos del repositorio Realizar el perfil con DROID que contrasta con el registro PRONOM y brinda un reporte.

El punto precedente se considera importante a la hora de pensar la preservación y la 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		2



Información descriptiva
(DI)

De Giusti, Marisa R. (2014). Tesis doctoral: “UNA METODOLOGÍA DE EVALUACIÓN DE REPOSITORIOS DIGITALES PARA ASEGURAR LA PRESERVACIÓN EN EL TIEMPO Y EL ACCESO A LOS CONTENIDOS”. Disponible en: http://hdl.handle.net/10915/43157

Los metadatos

De acuerdo con la norma UNE-ISO 23081-1: 2008 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”

Datos sobre datos

Los metadatos incluyen una amplia información que se puede utilizar para identificar, autenticar y contextualizar los objetos, las personas, los procesos de negocio, la regulación y sus relaciones.

Los metadatos

A nivel sintáctico definen los elementos, su orden y su formato de etiquetado o codificado y a nivel semántico ofrecen recomendaciones de uso de los elementos y de vocabularios especializados, a la vez que las reglas de contenido indican qué datos se registran en cada elemento y de qué modo. **El trabajo de normalización de la información, es de vital importancia en este sentido.**

Agencia Nacional de Investigación y Desarrollo

acceso abierto

Lanzamiento de Directrices de Metadatos y Mecanismos de Interoperabilidad para repositorios: Impulsando la ciencia abierta en Chile

Marzo 7, 2024

- **Metadatos descriptivos,** describen el recurso y son útiles para su recuperación y contextualización: título, autor, colaboradores, versión, resumen, fuente de publicación, vínculos, editor, año de publicación, etc…
- **Metadatos administrativos,** brindan la información necesaria para la gestión del recurso
  - metadatos técnicos, necesarios para decodificar y procesar el recurso
  - metadatos de preservación, necesarios para la gestión y migración en el largo plazo
  - metadatos de derechos, referidos a la propiedad intelectual
- **Metadatos estructurales,** describen las relaciones que existen entre las partes de un conjunto de datos. Por ejemplo, un esquema que representa las relaciones entre tablas de una base de datos.

https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/metadatos

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.

Preservación de contenido. Metadatos de preservación

Buscan registrar información relativa a la evolución de los recursos en el tiempo según las acciones de preservación aplicadas, incluyendo información sobre formatos, usos, actividades de preservación realizadas, responsables de dichas actividades en el tiempo, etc.

- Agregar metadatos técnicos. Los más importantes vinculados a los formatos pueden extraerse con algunas herramientas e incorporarse en el flujo de trabajo. ver: Herramientas para modificar y crear metadatos de una gran variedad de archivos: http://sedici.unlp.edu.ar/handle/10915/139859

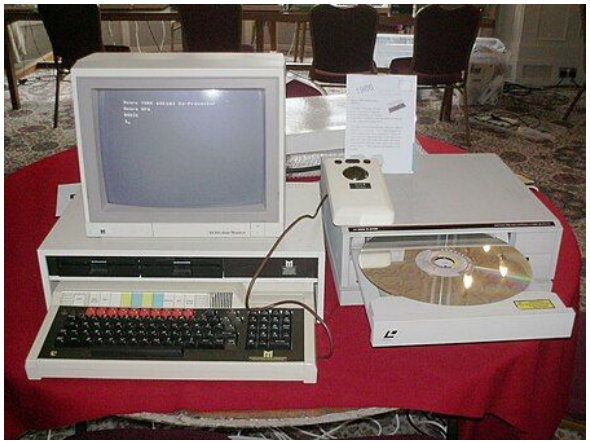
Un avance: estándares

El estándar 14721 (OAIS), los metadatos PREMIS y las directrices para la preservación, en conjunto con el esquema METS, constituyen el marco ideal para la gestión de un repositorio, para asegurar su interoperabilidad y dar preservación a sus contenidos.

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 muy particulares, pero no en la mayoría.



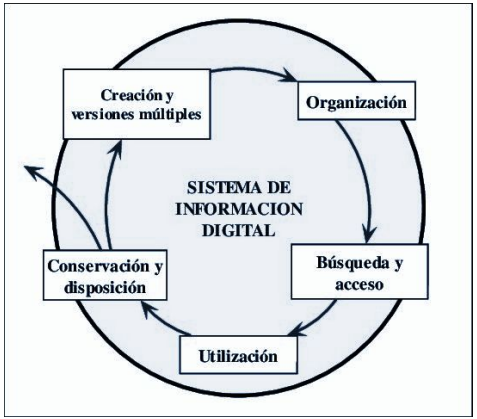
Una máquina de Domesday Project (1986) con su Laser disc modificado.

https://es.wikibrief.org/wiki/Digital_obsolescence

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



Es importante el MODO de trabajo para asegurar la trazabilidad. Saber quién hace qué.

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…

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.

Vamos a centrarnos en las funciones que propone la norma ISO 14721 y mostrar qué tiene y qué no un repositorio.

De Giusti, M. R., Lira, A. J., Villarreal, G. L., & Texier, J. D. (2012). Las actividades y el planeamiento de la preservación en un repositorio institucional. In *BIREDIAL-Conferencia Internacional Acceso Abierto, Comunicación Científica y Preservación Digital*.http://sedici.unlp.edu.ar/handle/10915/26045

Estándares


ESTANDARES - ESTANDARES – ESTANDARES – ESTANDARES – ESTANDARES

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).

SECCIONES

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.

https://wiki.dpconline.org/index.php?title=OAIS_Structure

Giaretta, D., Garrett, J., Conrad, M., Zierau, E., Longstreth, T., Hughes, J. S., ... & Engel, F. (2019). OAIS Version 3 Draft Updates. In Proceedings of the 16th International Conference on Preservation of Digital Objects.

https://scholar.archive.org/work/dzbkqoaxjrcxxbzyqey6ggos5e/access/wayback/https://services.phaidra.univie.ac.at/api/object/o:1079787/diss/Content/download

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.

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.

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 funcionalidad requerida del 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

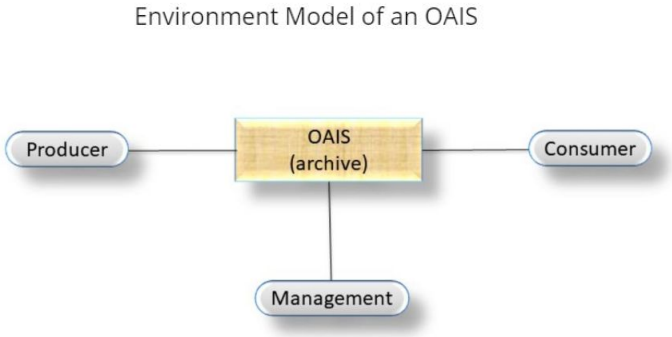
- Estandariza las relaciones y los componentes de un sistema de archivos. Es un framework que sirve para entender mejor de qué se habla.
- Establece un vocabulario común.
- Ofrece un marco consensuado internacional para la definición de entidades, procesos y funciones de los archivos de datos.
- Facilita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

Conceptos en OAIS

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



Conceptos en OAIS

- La unidad de intercambio entre un OAIS y su medioambiente es el IP.
- Un IP contiene 2 tipos de información:

  De contenido y 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.

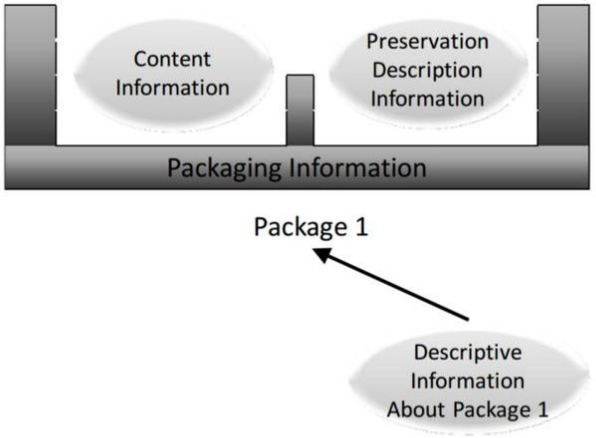


Figure 2-3: Information Package Concepts and Relationships

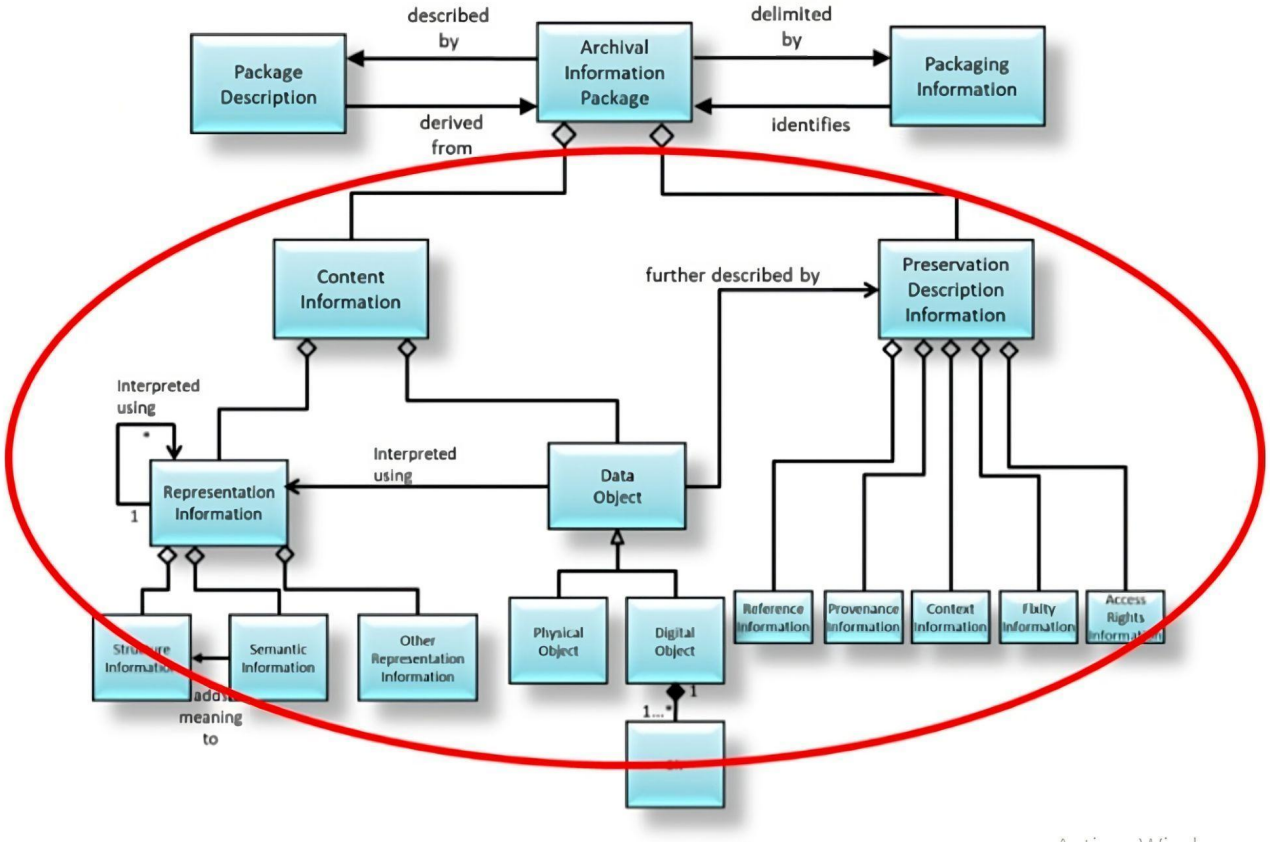
El paquete de información (IP)

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

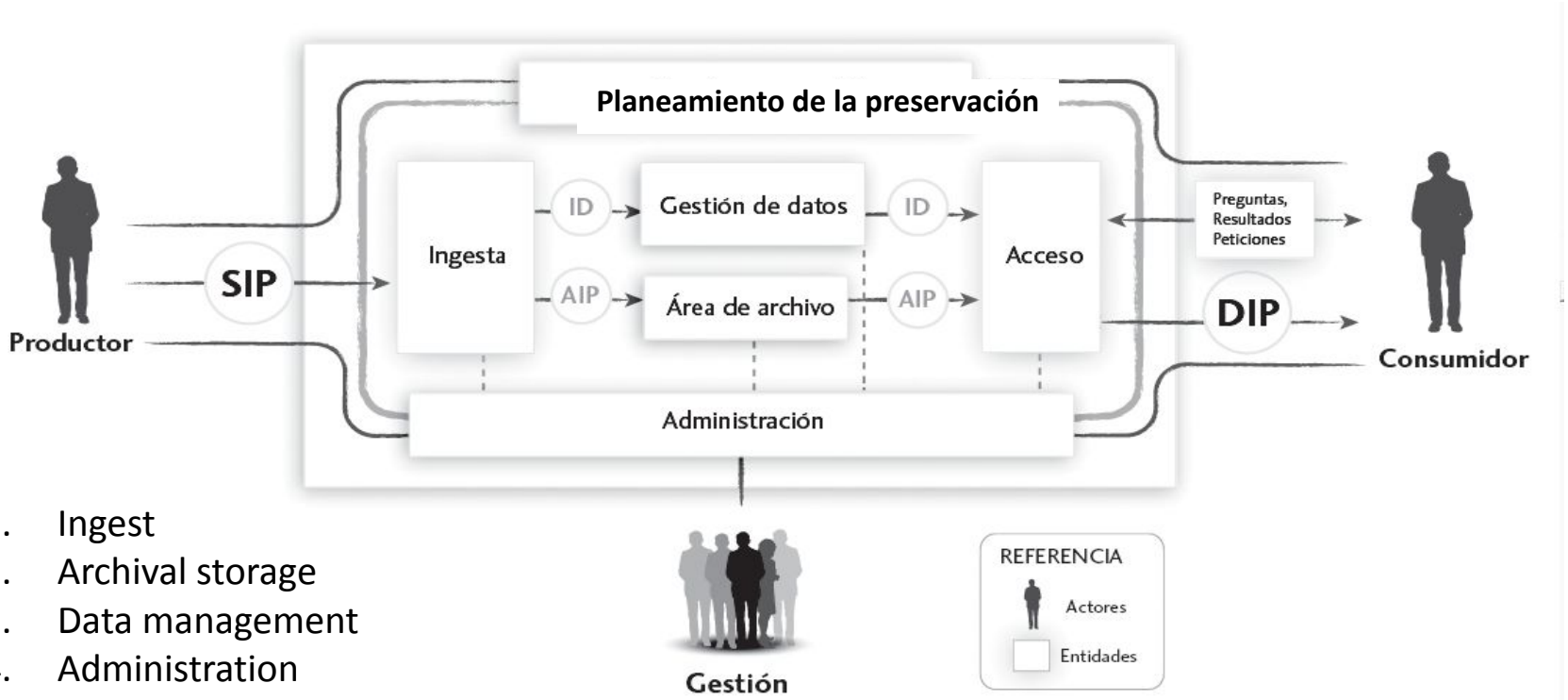
La **procedencia**, más allá de describir la fuente, incluye los procesos que se han realizado sobre la información: la historia del objeto, cambios, versiones y responsables. El **contexto** muestra las relaciones con otras fuentes de información o contenidos. La **referencia** provee una identificación 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.

AIP



Giaretta, D., Garrett, J., Conrad, M., Zierau, E., Longstreth, T., Hughes, J. S., ... & Engel, F. (2019). Oais versiOn 3 draftupdates. In *Proceedings of the 16th International Conference on Preservation of Digital Objects*.

Modelo Funcional: Sección 4.1: 6 entidades



1. Ingest
2. Archival storage
3. Data management
4. Administration
5. Preservation Planning
6. Access

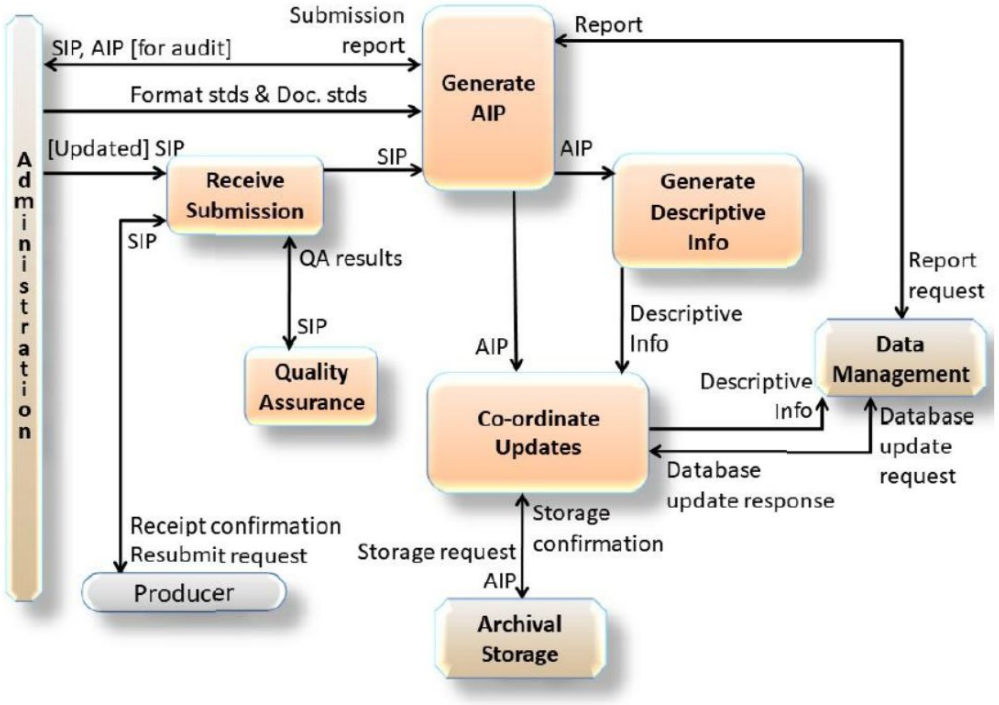
Modelo OAIS

El proceso puede iniciarse cuando el productor suministra el recurso (paquete de entrada) llamado SIP a través del ingest, que luego se convierte en AIP terminando en la entidad archival storage. El flujo puede continuar cuando 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 se mantienen organizados a través de la entidad *data management*. La entidad *administration* son los administradores y responsable del repositorio y esta entidad se relaciona con las secciones de ingesta, *gestión de datos, almacenamiento de archivos* y *planificación de la preservación*. Esto permite una gestión estructural y ayuda a mantener los AIP a lo largo del tiempo.

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.

Funciones de la entidad Ingest/Ingesta/Ingreso



Entidad OAIS Ingest

- **Descripción:** Provee los servicios y funciones para aceptar un SIP por parte de los Productores o bajo el control de la Administración.
- Prepara los contenidos para almacenamiento y gestión en el 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.

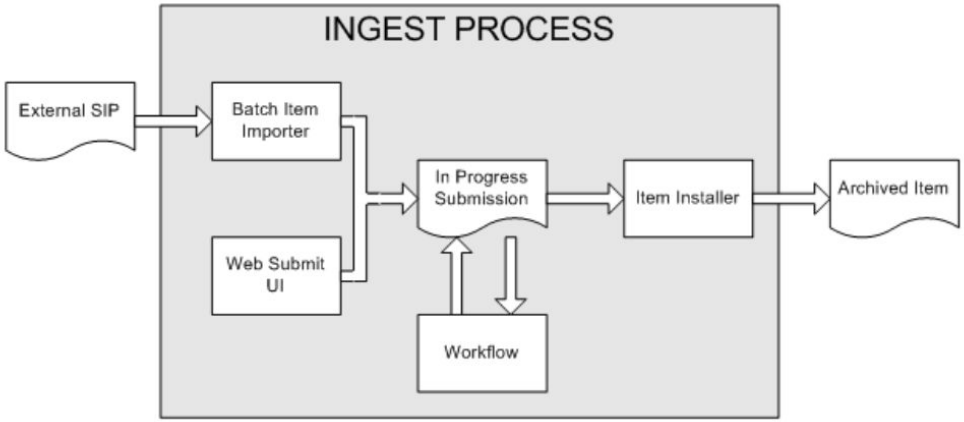
¿Qué hacer cuando se ingesta un archivo?

Los archivos que llegan de los autores o los sistemas informáticos deberían pasar una serie de controles:

- Control de procedencia e integridad
- Control de virus
- Control de formatos

Después de pasados los controles se deberían extraer los metadatos técnicos, realizar el checksum, agregar el identificador persistente, las licencias, la procedencia y el contexto. Cada vez que se realiza un cambio en el archivo, recalcular el checksum.

Procedimiento INGEST en DSPACE



La ingesta puede realizarse por importación en masa, cosecha, depósito SWORD o proceso de carga tradicional.

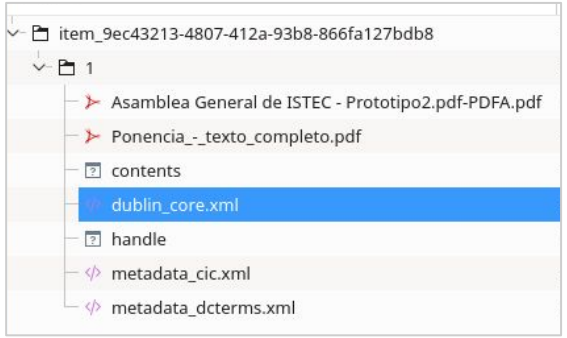
Sería muy importante saber agentes y eventos qué se anota o no se anota según de dónde provenga el OD. Importación masiva:
https://sedici.unlp.edu.ar/handle/10915/165890
Autoarchivo:
http://sedici.unlp.edu.ar/handle/10915/153550

Entidad Ingest en DSpace

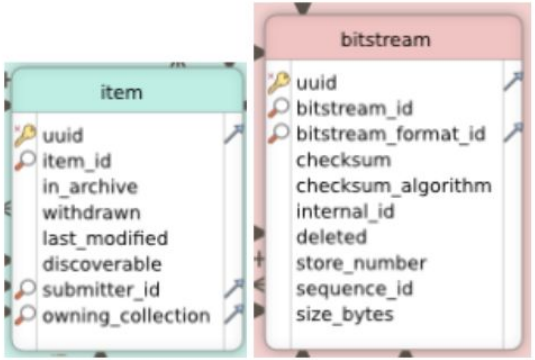
- El proceso de carga tradicional o Submission process, permite configurar para cada colección y para cada tipología documental qué metadatos se requieren de forma opcional u obligatoria, y cuál es la secuencia de etapas que debe cumplir el ítem antes de ingresar al repositorio.
- La cosecha se realiza 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 indice, metadatos y binarios,
  - METSSIPImporter en formato METS o,
  - en cualquier otro formato,
    si se desarrolla un packager plugin ad-hoc.



Entidad Ingest en DSpace



Evidencia de autenticidad: lo que se tiene como "auditoría" es

- submitter (usuario de dspace que sube el ítem),
- checksum de cada bitstream
- nombre original del bitstream

Cuando se hace el install también se guardan los metadatos:

- dc.description.provenance (resumen de auditoría en forma textual),
- dc.date.accessioned (fecha de disponibilización del ítem) y
- dc.identifier.uri entre los metadatos del item.

dc.date.accessioned	2014-05-15T13:35:30Z
dc.identifier.uri	http://sedici.unlp.edu.ar/handle/10915/35446
dc.description.provenance	Step: SeDiCILevelReview - action:editaction Approved for entry into archive by Carlos Nusch(carlosnusch@prebi.unlp.edu.ar) on 2014-05-15T13:35:30Z (GMT)
dc.description.provenance	Made available in DSpace on 2014-05-15T13:35:30Z (GMT). No. of bitstreams: 1 ortiz-rodriguez-mayra.pdf: 108844 bytes, checksum: 4a87033c07f711b89f43e9915908bd2a (MD5)

Entidad Ingest en DSpace

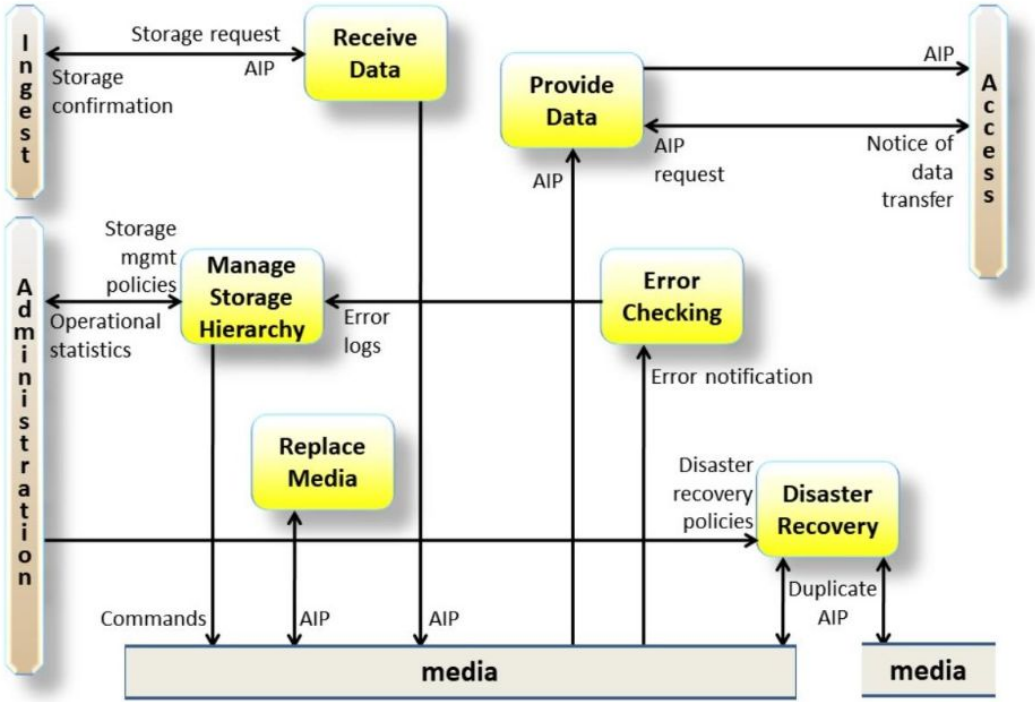
Aspectos mejorables

- Seguridad: un usuario malintencionado podría subir código malicioso
  a) ejecutable por un responsable del repositorio
  b) ejecutable por el público al usar los archivos por ejemplo de código fuente.

Es necesario contar con políticas de contenidos adecuadas, controles sobre lo que se sube y antivirus que hagan control periódico (ejemplo: tarea de curation antivirus CLAMAV)

- metadatos técnicos
  - se deberían extraer automáticamente al ingestar un recurso con un componente binario (un bitstream) para que luego sean usados por las demás entidades del sistema.

Funciones de de la entidad Archival Storage



Entidad OAIS Archival Storage

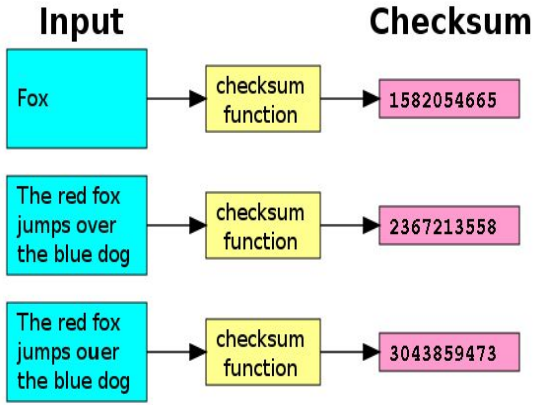
- **Descripción:** Provee los servicios y funciones para el almacenamiento, mantenimiento y recuperación de los AIPs.
- Recibe el AIP de la entidad Ingest y lo almacena. Gestiona las jerarquías de almacenamiento. Configura niveles especiales de servicio, seguridad y protección (por ejemplo backups). 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 un mecanismo de duplicación de los contenidos en un lugar físico separado. 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 (bitstreams, archivos) 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 ⇒ 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. HABILITAR.



- 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.

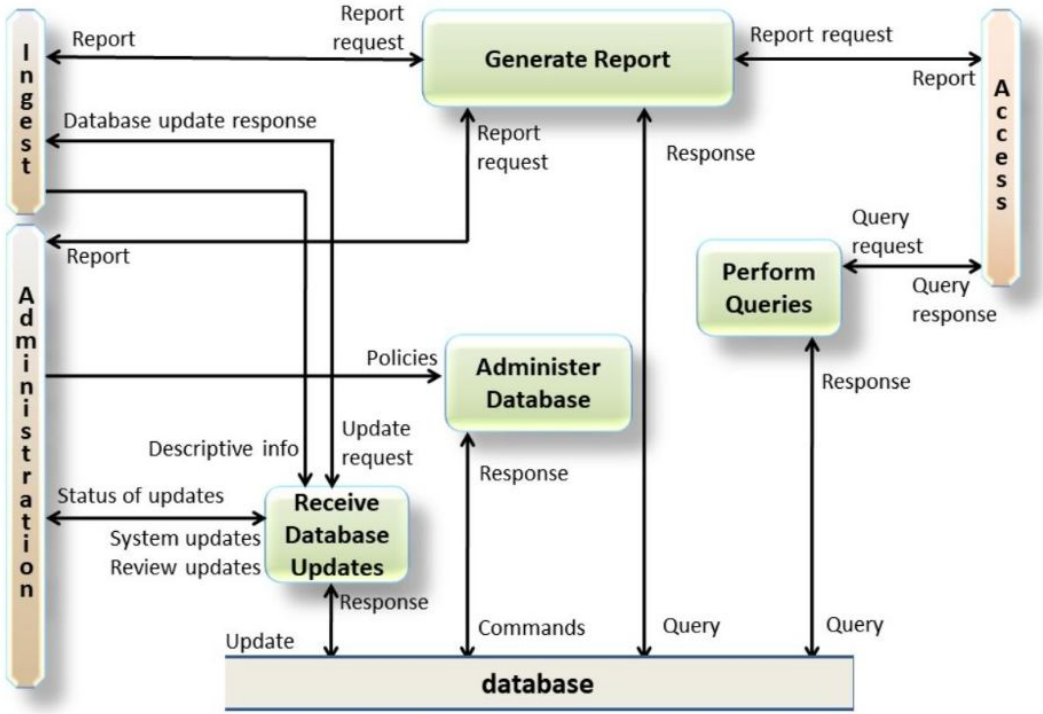
Entidad Archival Storage en DSpace

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 corrección 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)

Funciones de la entidad Data Management



Entidad OAIS Data Management

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

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 (*worfklow*). 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.
    - https://wiki.lyrasis.org/display/DSDOC8x/Discovery
- 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.

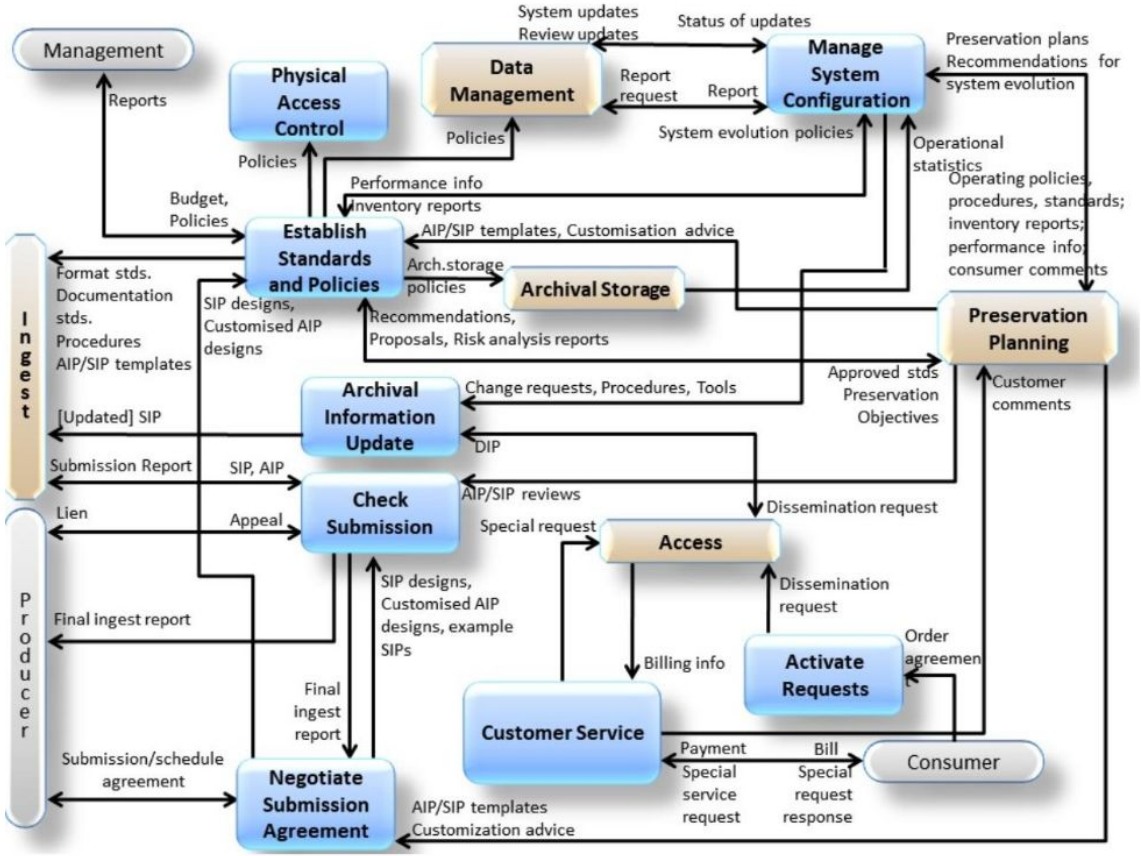
Las tareas de curation pueden ser:

- 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

Funciones de la entidad 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 diseminación 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:

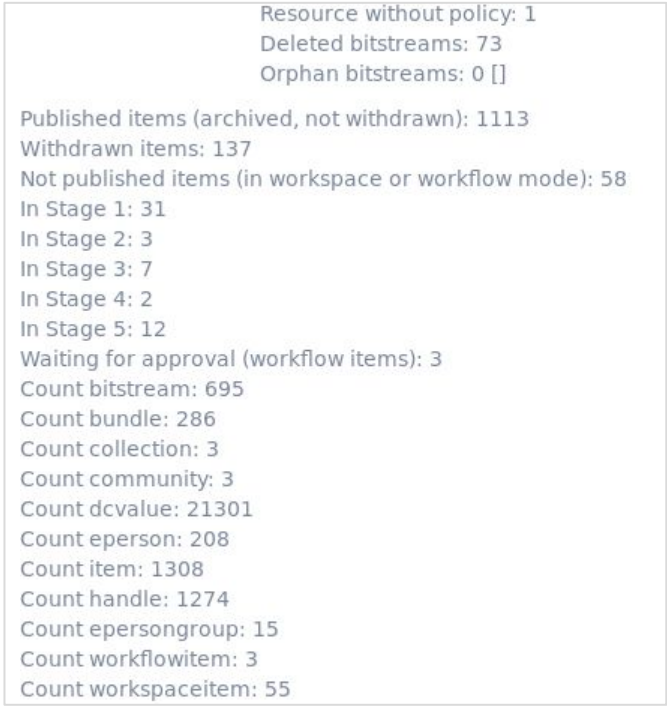
- Version de JVM utilizada
- Memoria disponible
- Uso de la cache Java
- Parámetros de configuración activos
- Peticiones realizadas al repositorio
- etc.

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 HealthCheck, que debe activarse mediante tareas programadas del sistema (cronjobs).

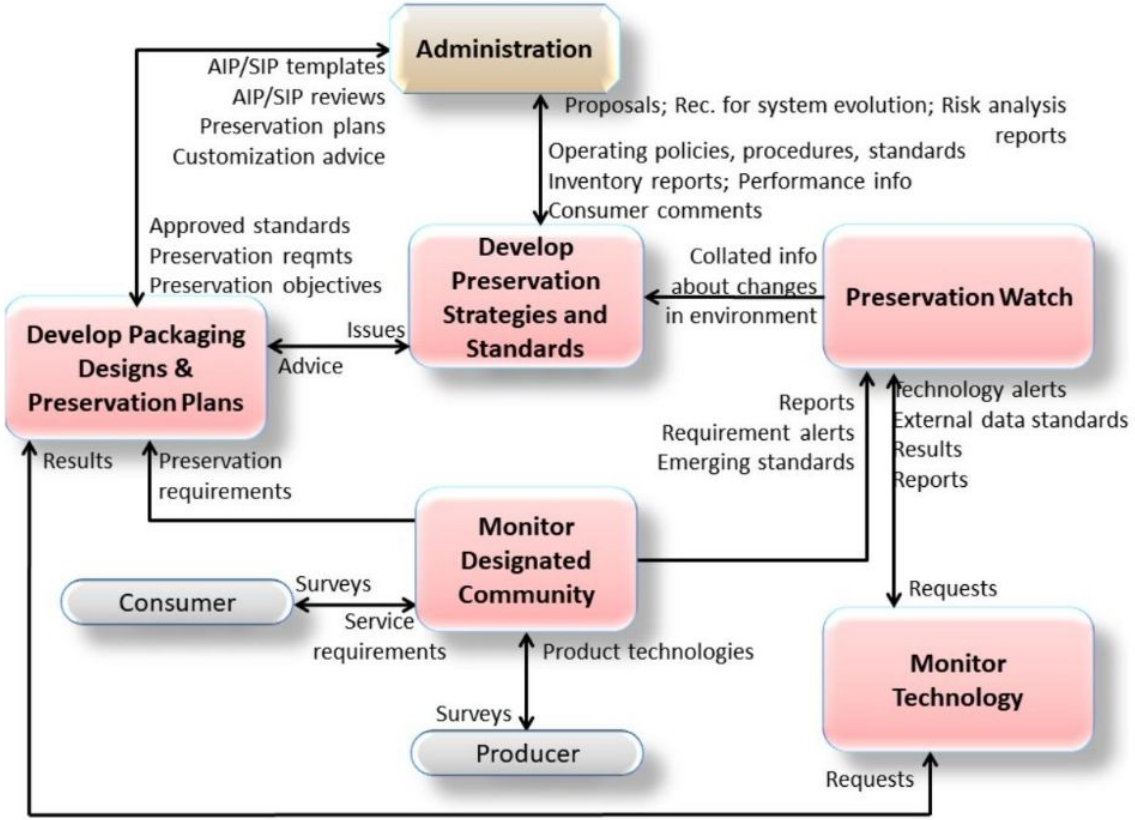


Entidad Administration en DSpace: Performance

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...)

Funciones de la entidad Preservation Planning



Entidad OAIS 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.

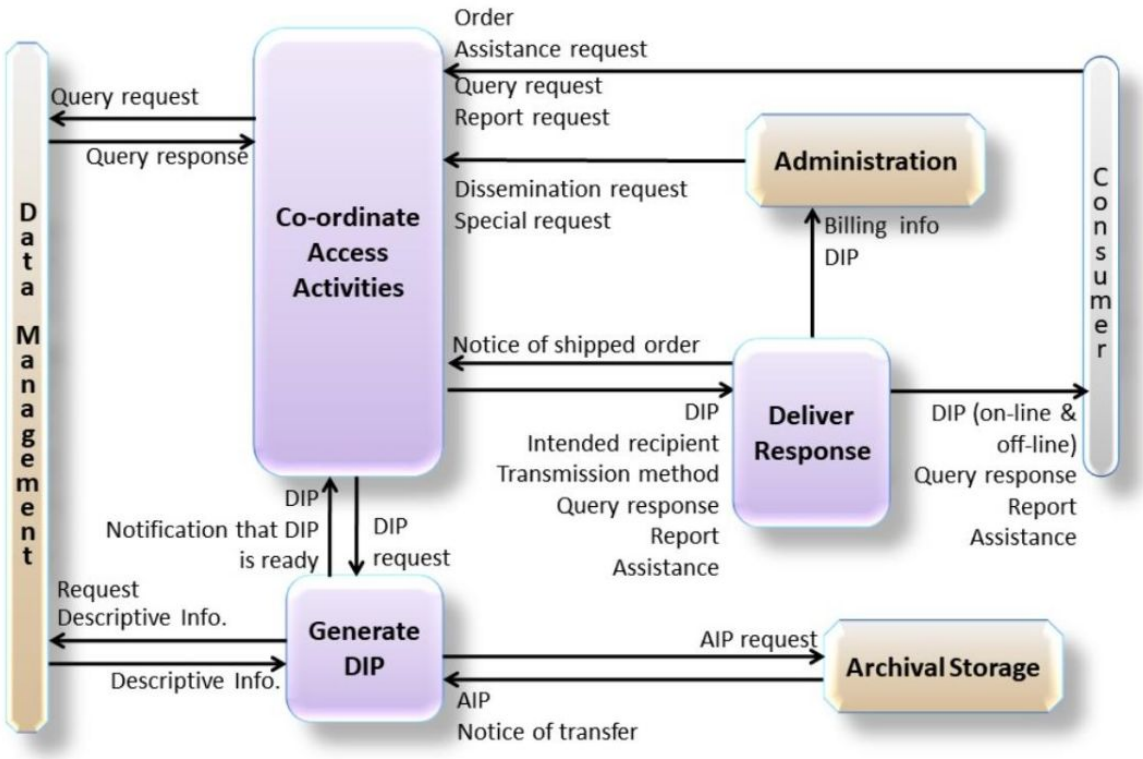
Entidad Preservation Planning en DSpace

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.
- El guardado de un PUID

Funciones de la entidad 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*.

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:

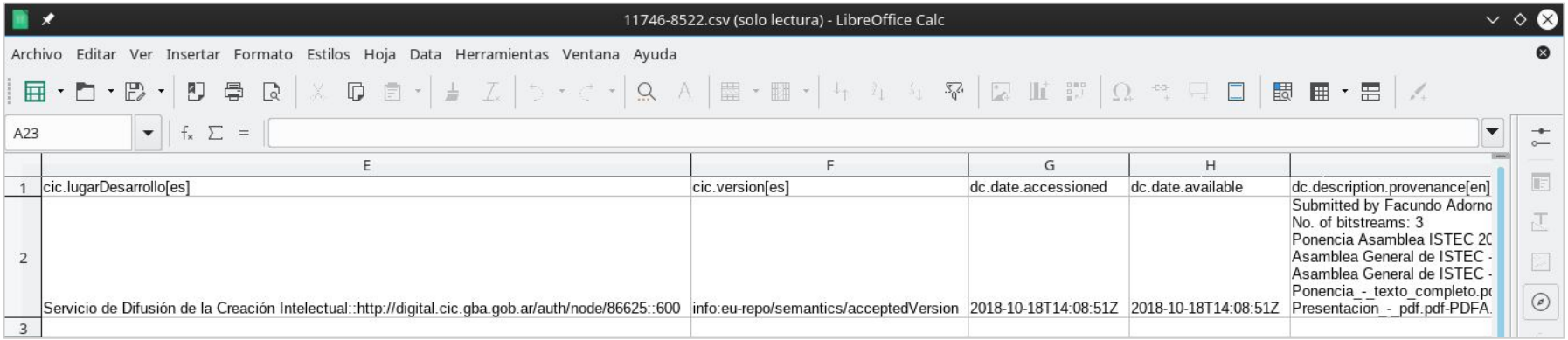
- dc.description.provenance presenta información sobre las acciones realizadas sobre el item para su depósito en el repositorio.
- dc.rights que contiene información de sobre la licencia de uso.
- handle identificador persistente del item
- checksum de bitstreams embebidos en el metadato provenance

Entidad Access en DSpace

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

Cada columna presenta el siguiente formato

schema.element.[qualifier].[content_language]



Saliendo de la 14721

Aproximaciones a la preservación

Existen numerosas estrategias para asegurar la preservación de la información:

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

The National Archives

Search our website

About us | Education | Records | Information management | Archives sector

You are here: Home > Information management > Digital preservation > PRONOM

The technical registry
PRONOM

Welcome : About Add an entry
Search Help Information resources

Welcome to PRONOM

PRONOM changes and DROID signature file release notes.

DROID signature files.

Find out more about our plans to make PRONOM's data available in a linked open data format on The National Archives Labs.

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 ›

https://www.nationalarchives.gov.uk/PRONOM/

The technical registry
PRONOM

Search Results

Help : report on simple search

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

You searched for: "pdfa"

Save as... XML | CSV Print

page 1

Acrobat PDF/A - Portable Document Format (1b)

The Portable Document Format/ Archive is a format designed for long term preservation by Adobe Systems. PDF/A is a simplified version of PDF 1.4, with all of the features from PDF 1.4 that would impede long term preservation removed. Removed features include Javascript, Audio/Visual content, LZW compression and encryption. A major principle of PDF/A is that it is self contained and not reliant on externalities thus all font and colour information is encoded into the file. PDF/A files are larger than other types of PDF files due to the need for embedded information. PDF/A supports two levels of compliance PDF/A1-a (Accessible) and PDF/A1-b (Basic). PDF/A1-a is fully ISO 19005-1:2005 (PDF/A-1) compliant whereas PDF/A1-b is less stringent and not compliant. PDF/A1-a requires tagged PDF and Unicode whereas PDF/A1-b does not. The signature identification proposed for PDF/A 1b has been tested against a number of PDF/A and PDF files for verification. However, PDF itself is a complex format, and while the PDF/A 1b signature contained in PRONOM will identify many PDF/A 1b files, DROID can only identify what a file asserts itself to be. As there are a number of ways in which PDF/A 1b files can violate the PDF/A 1b conformance, which can not be spotted until the file is parsed, it is recommended that PDF files are tested against validation software when attempting to confirm identification. Further information concerning PDF/A violations, and a test suite of PDF items which fail to conform to PDF/A standards, can be found at http://www.pdfa.org/2011/08/isartor-test-suite/

The technical registry
PRONOM

Search Results

Help : report on file format

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

You searched for: "xml"

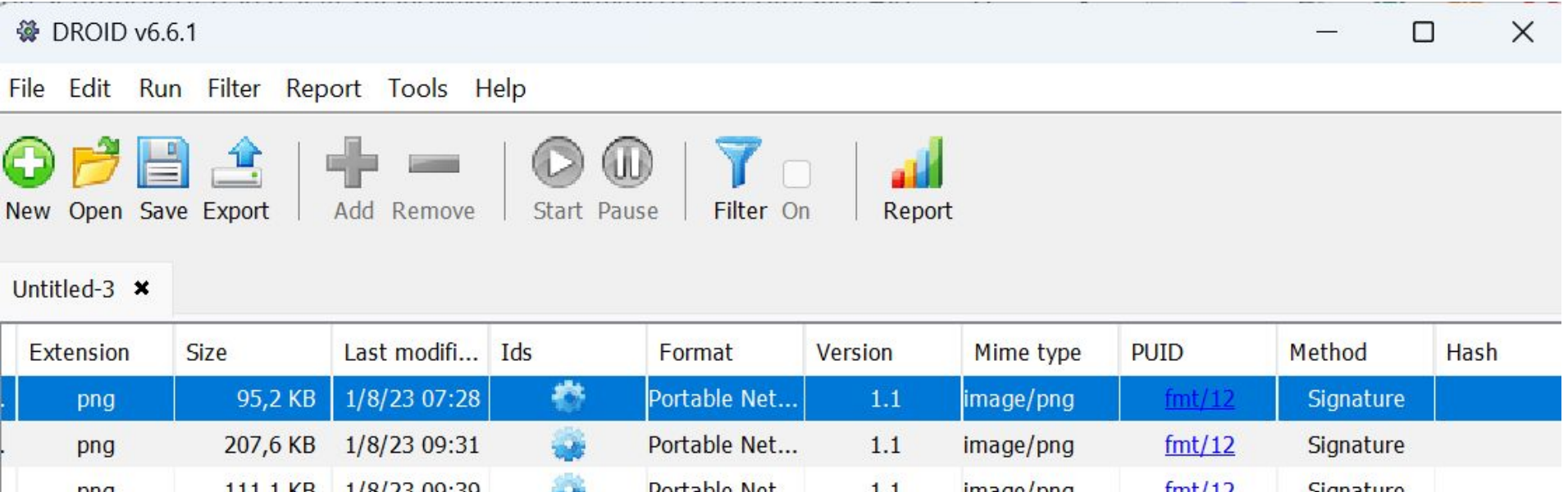
Save as... XML | CSV Print

page 1 2 3

PRONOM Unique ID	Format Name	Format Version	Extension	Format Risk
fmt/1946	Draw.io Diagram (XML) File		drawio xml	
fmt/120	DROID File Collection File Format	1.0	xml	
fmt/121	DROID Signature File Format	1.0	xml	
fmt/1729	Esri Shapefile Geospatial Metadata File		xml	
fmt/101	Extensible Markup Language	1.0	xml	
fmt/1776	Extensible Markup Language	1.1	xml	

Perfilar el repositorio

La herramienta DROID (Digital record object identification service) que usa los perfiles de formato 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.



Metadatos de preservación

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.

Metadatos de preservación

- 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.

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".*

PREMIS define un subconjunto de los metadatos de preservación

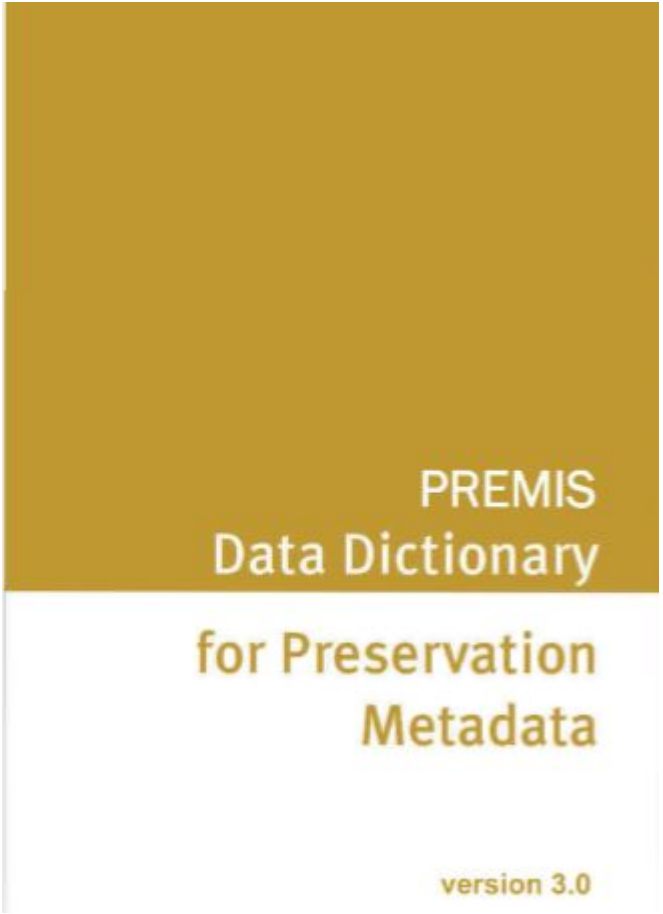
Las personas encargadas de diseñar aplicaciones de software para repositorios de preservación deben utilizar PREMIS como una guía sobre la información que debe obtener y registrar la aplicación.

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.

https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

https://www.loc.gov/standards/premis/UnderstandingPREMIS_espanol.pdf



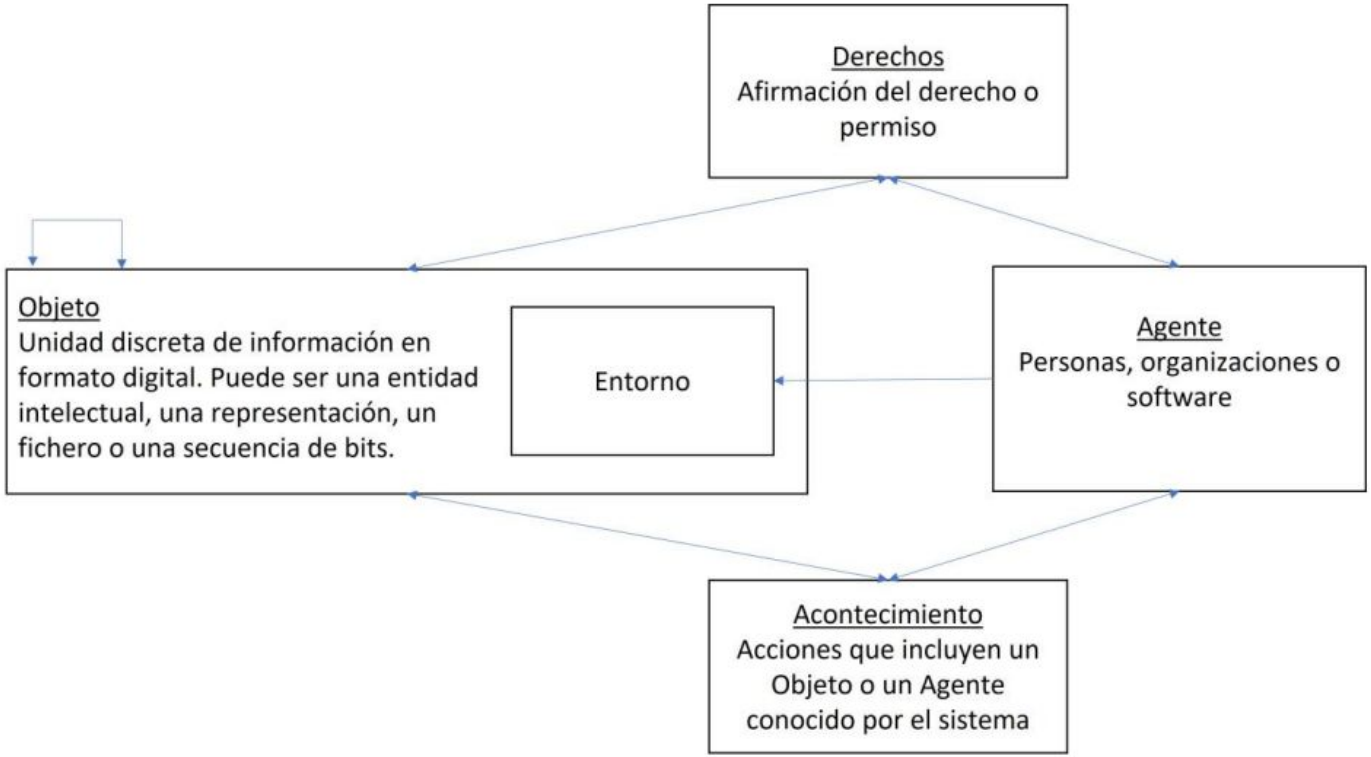


Gráfico 2: Modelo de datos PREMIS, versión 3

Imagen extraída de: https://www.loc.gov/standards/premis/understandingPREMIS_spanish_2021.pdf

Objetos

Los ***Objetos*** son unidades discretas de información en forma digital, que se clasifican en cuatro tipos: ***archivo (file), representación (representation) y cadenas de bits (bitstream) y entidad intelectual.*** 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*** (p.e. un sitio web). Los objetos *cadenas de bits* Por ejemplo, si tenemos un fichero en formato AVI (audio-vídeo intercalado) quizás se quiera diferenciar la secuencia de bits de audio de la secuencia de bits de vídeo, y describirlas como objetos-secuencia de bits independientes.

3.1. Entidad Objeto

Los *Objetos* son lo que realmente se almacena y gestiona en un repositorio de preservación. La mayor parte de PREMIS se dedica a describir objetos digitales. Entre la información que se puede registrar se incluye:

- el identificador único del objeto (tipo y valor)
- información de fijeza, como la suma de verificación (compendio del mensaje) 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 está almacenado
- información sobre la firma digital
- relación con otros objetos y otros tipos de entidades

Entidad entorno

Para registrar el entorno de un objeto se definen varias unidades semánticas, es decir, qué hardware y software son necesarios para su reproducción y qué dependencias existen de otros objetos. En las versiones 1 y 2, esa información forma parte de la descripción del objeto. En la versión 3, el entorno, como un tipo especial de objeto-entidad intelectual, puede vincularse desde los ficheros, representaciones y secuencias de bits que lo utilizan.

3.2. Acontecimientos

La *entidad acontecimiento* agrega información sobre acciones que afectan a los objetos del repositorio. Disponer de un registro preciso y fiable de los acontecimientos resulta imprescindible para mantener la procedencia digital de un objeto, lo que, a su vez, es importante para demostrar la autenticidad del objeto.

La información que se puede registrar sobre los acontecimientos incluye:

- el identificador único del acontecimiento (tipo y valor)
- el tipo de acontecimiento (creación, ingesta, migración, etc.)
- le fecha y la hora en las que tuvo lugar el acontecimiento
- una descripción detallada del acontecimiento
- un 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

3.3. Agentes

Los *Agentes* son actores que tienen una función en los acontecimientos, en las declaraciones de derechos (véase el apartado 3.4, Derechos), y en los objetos-entorno. Los agentes pueden ser personas, organizaciones, aplicaciones de software, o hardware. PREMIS solo define un número mínimo de unidades semánticas necesarias para identificar a los agentes, puesto que existen varios estándares externos que se pueden utilizar para registrar información más detallada. Un repositorio puede elegir entre utilizar un estándar independiente para registrar información adicional sobre los agentes, y utilizar el identificador del agente para señalar la información registrada externamente.

El Diccionario de Datos incluye:

- un identificador único del agente (tipo y valor)
- el nombre del agente
- la designación del tipo de agente (persona, organización, software)
- la versión del agente (software o hardware)
- una nota general sobre el agente
- acontecimientos asociados con el agente
- declaración de derechos asociados con el agente
- objetos-entorno asociados con el agente

3.4. Derechos

La información que se puede registrar en una declaración de derechos incluye:

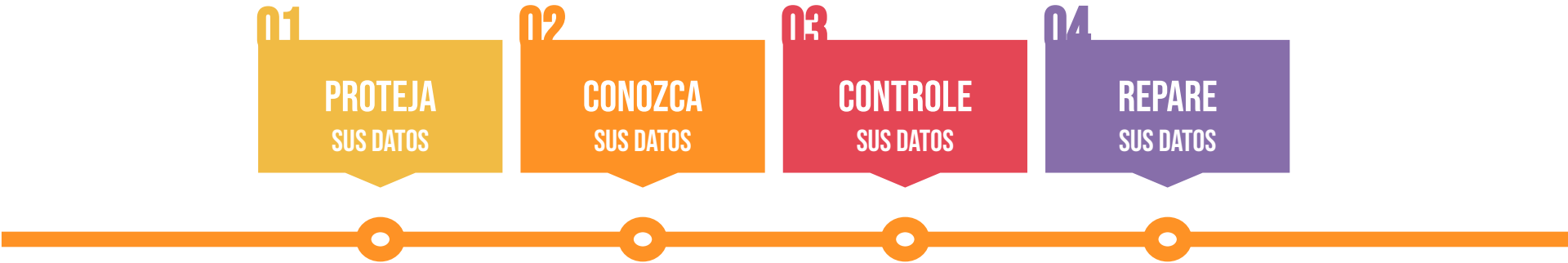
- un identificador único de la declaración de derechos (tipo y valor)
- si la base de la reclamación del derecho es el copyright, la licencia, la ley, u otra (p. ej. la política de la institución)
- información más detallada sobre el estado del copyright, las condiciones de la licencia, o la ley, según su aplicabilidad
- la acción o acciones que permita la declaración de derechos
- cualquier restricción sobre la acción o acciones
- el plazo del otorgamiento o la restricción de derechos, o el período de vigencia de la declaración
- el objeto u objetos a los que resulta aplicable la declaración
- los agentes implicados en la declaración de derechos y sus funciones

Autoevaluación: NDSA

- **Aspectos/ Áreas:**

1. Sobre el almacenamiento: copias y localización
2. Sobre la no alteración de los archivos y la integridad de los datos.
3. Sobre la seguridad de la información: quién ha hecho qué con los contenidos.
4. Metadatos.
5. Formatos.

- **Niveles de cumplimiento de cada área (complejidad creciente)**



Matriz de auto-evaluación NDSA

Área Funcional	Nivel			
	Nivel 1 - (Conocer su contenido)	Nivel 2 - (Proteger su contenido)	Nivel 3 - (Controlar su contenido)	Nivel 4 - (Mantener su contenido)
Almacenamiento	Tener dos copias completas en ubicaciones separadas Documentar todos los medios de almacenamiento donde este almacenado el contenido Poner el contenido en soportes de almacenamiento estables	Tener tres copias completas con al menos una copia en una ubicación geográfica distinta Documentar el almacenamiento y medios de almacenamiento, indicando los recursos y las dependencias que estos requieren para funcionar	Tener al menos una copia en una ubicación geográfica con amenaza de desastre diferente a las otras copias Tener al menos una copia en un medio de almacenamiento de diferente tipo Rastrear la obsolescencia del almacenamiento y los medios	Tener al menos tres copias en ubicaciones geográficas distintas, cada una con una amenaza de desastre diferente Maximizar la diversificación del almacenamiento para evitar puntos únicos de falla Tener un plan y realizar acciones para abordar la obsolescencia del hardware, software y medios de almacenamiento
Integridad	Verificar que la información de integridad se ha proporcionado con el contenido Generar información de integridad si esta no ha sido proporcionada con el contenido Se verifica virus en todo el contenido; se aisla el contenido en cuarentena según sea necesario	Verificar la información de integridad al mover o copiar contenido Usar bloqueadores de escritura cuando se trabaja con medios originales Hacer una copia de seguridad de la información de integridad y almacenar una copia en una ubicación separada del contenido	Verificar la información de integridad del contenido en intervalos fijos Documentar los procesos y resultados de verificación de información de integridad Realizar una auditoría de la información de integridad bajo demanda	Verificar la información de integridad en respuesta a eventos o actividades específicas Reemplazar o reparar el contenido dañado según sea necesario
Control	Se determinan los agentes humanos y de software qué deben estar autorizados para leer, escribir, mover y eliminar contenido	Documentar a los agentes humanos y de software autorizados para leer, escribir, mover y eliminar contenido y aplicar estos	Mantener los registros (logs) y se identifican a los agentes humanos y de software que realizaron acciones sobre el contenido.	Se realizan revisiones periódicas de acciones / registros (logs) de acceso
Metadatos	Crear un inventario de contenido, documentando también la ubicación de almacenamiento actual de estos Hacer una copia de respaldo del inventario y se almacena al menos una copia por separado	Almacenar suficientes metadatos para saber cuál es el contenido (esto podría incluir alguna combinación de aspectos administrativos, técnicos, descriptivos, de preservación y estructurales)	Determinar qué estándares de metadatos aplicar Encuentra y completa los vacíos en sus metadatos para cumplir con esos estándares	Registrar las acciones de preservación asociadas con el contenido y cuándo ocurren esas acciones Implementa los estándares de metadatos elegidos
Contenido	Documentar los formatos de archivo y otras características de contenido esenciales, incluido cómo y cuándo fueron identificados	Verificar los formatos de archivo y otras características de contenido esenciales Establecer relaciones con los creadores de contenido para fomentar la elección sostenible de archivos	Monitorear la obsolescencia y los cambios en las tecnologías de las que depende el contenido	Realizar migraciones, normalizaciones, emulación y actividades similares que garanticen el acceso al contenido

Acciones iniciales: Revisión de prácticas y procedimientos

- Copias: de qué y en dónde.
- Documentos escritos: cuáles.
- Formatos utilizados.
- Autorizaciones, tipos de usuarios y permisos.
- Control de agentes y cambios sobre los bitstreams: metadatos.
- Metadatos usados: administrativos + técnicos, de preservación (PDI).

Sobre el almacenamiento

Definir el requerimiento de copias de seguridad en lo que respecta a cantidad, rotación y ubicación de las mismas.

- Se debe definir qué resguardar, cómo, dónde y por cuánto tiempo.
- Copias en localizaciones de fallas distintas
- Copias en la nube en cumplimiento con legislación nacional
- Requerimiento de encriptación.

Sobre el almacenamiento: ejemplos prácticos

Nivel 1 (más básico): tener dos copias completas de la información por separado y en servidores propios (no en discos de manera aislada).

Nivel 2: tres copias y una en una localización geográfica diferente.

Nivel 3: tres copias y una en una localización geográfica diferente con previsión de desastres distinta. Controlar el proceso de obsolescencia del almacenamiento y el soporte.

Nivel 4: como mínimo tres copias en tres localizaciones geográficas diferentes, cada una con previsión de desastres distinta. (Alguna puede ser en la nube). Disponer de un plan integral para mantener datos y metadatos accesibles en los sistemas y soportes.

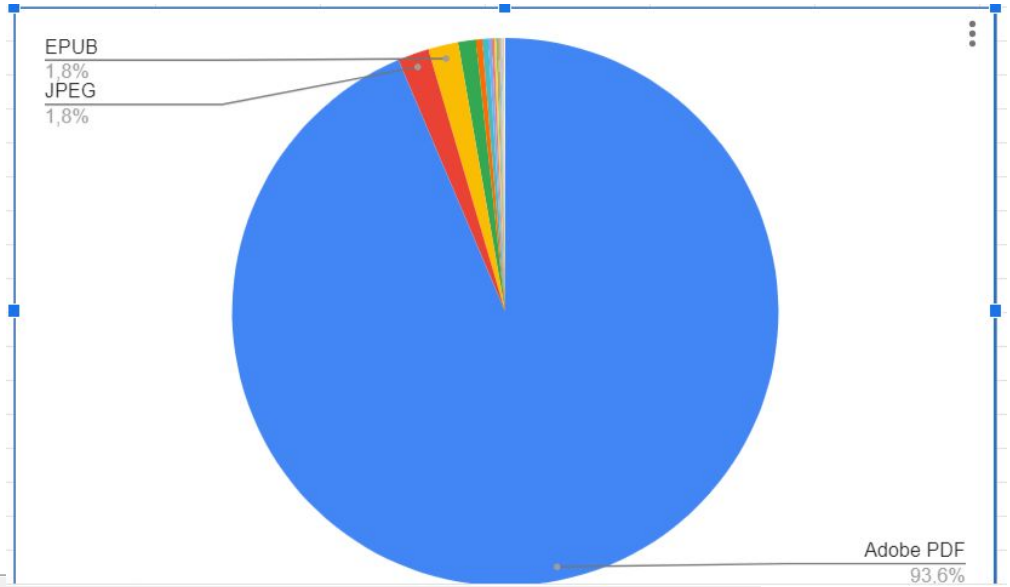
- En paralelo con la acción documentar el sistema elegido (en el nivel que sea).
- Escribir un plan.

Acciones iniciales: Revisión de contenidos y formatos

- Realizar un perfilamiento básico de los contenidos del repositorio, entendido en el estricto sentido de saber las grandes tipologías o algo como “súperclases!”: texto, video, audio y datos.
  - En cada caso saber “qué aspectos de la información se deben preservar”.
- En función de las cantidades de cada qué, establecer un orden de prioridad que puede ser un plan de acción que comience por lo que hay menos (que se resuelve rápido) o por lo que hay más. Observar que esto es una definición política-administrativa-de recursos.
- Después hacer un perfilamiento detallado para cada formato (DROID por ejemplo).

SEDICI

- Texto
- Audio
- Video
- Imágenes
- Datasets
- Otros (Binarios, comprimidos, software, etc)



DROID v6.5

File Edit Run Filter Report Tools Help

New Open Save Export Add Remove Start Pause Filter On Report

Untitled-1 Untitled-2

Resource	Extension	Size	Last modified	Ids	Format	Version	Mime type	PUID	Method	Hash
C:\Users\mari...	zip	114,2 MB	30/6/22 09:23		ZIP Format		application/zip	x-fmt/263	Signature	
C:\Users\mari...	pdf	1,3 MB	30/6/22 08:06		Acrobat PDF 1.6 - Por...	1.6	application/pdf	fmt/20	Signature	
C:\Users\mari...	pdf	122,7 KB	23/2/22 09:08		Acrobat PDF 1.6 - Por...	1.6	application/pdf	fmt/20	Signature	
C:\Users\mari...	pdf	5,5 MB	11/4/22 07:53		Acrobat PDF 1.4 - Por...	1.4	application/pdf	fmt/18	Signature	
C:\Users\mari...	pdf	9 MB	23/2/22 11:52		Acrobat PDF 1.6 - Por...	1.6	application/pdf	fmt/20	Signature	
C:\Users\mari...	docx	55,9 KB	4/9/21 10:45		Microsoft Word for W...	2007 onwards	application/vnd.open...	fmt/412	Container	
C:\Users\Publi...	lnk	2 KB	16/4/21 09:08		Microsoft Windows S...			x-fmt/428	Signature	
C:\Users\mari...	docx	21 KB	10/3/22 18:05		Microsoft Word for W...	2007 onwards	application/vnd.open...	fmt/412	Container	
C:\Users\mari...	pdf	2,7 MB	13/9/22 09:21		Acrobat PDF 1.4 - Por...	1.4	application/pdf	fmt/18	Signature	
C:\Users\mari...	pdf	67 KB	26/3/22 12:57		Acrobat PDF 1.7 - Por...	1.7	application/pdf	fmt/276	Signature	

Acciones de preservación - Documentación

- **Documentación de procesos de backups**
  - ¿Cómo se realizan los backups? ¿Qué herramientas se utilizan? ¿Cuál es el procedimiento?
  - Definir qué hacer ante cada potencial situación de amenaza sobre los datos
- **Manuales de carga**
  - Formato correcto de cada metadato
  - Consideraciones a la hora de seleccionar autores, materias, etc.
- **Documentación del proceso de digitalización de documentos**
- **Documentación de las tareas de mantenimiento y desarrollo**
  - Se suele usar un sistema de tickets (trac)
- **Inventario del equipamiento del repositorio**
  - Plan para el control de obsolescencia de los equipos y su eventual reemplazo

Acciones de preservación - Documentación: resguardo de datos

- **Escribir un plan de riesgos sobre los datos:** ciberseguridad, fallas de hardware, errores humanos, desastres naturales, conflictos bélicos/atentados, etcétera.
  - ISO 16363.

Acciones de preservación - Integridad

- Análisis de checksum sobre el contenido original
- Checksum sobre los backups
- Automatización de estos procesos - tareas de curation

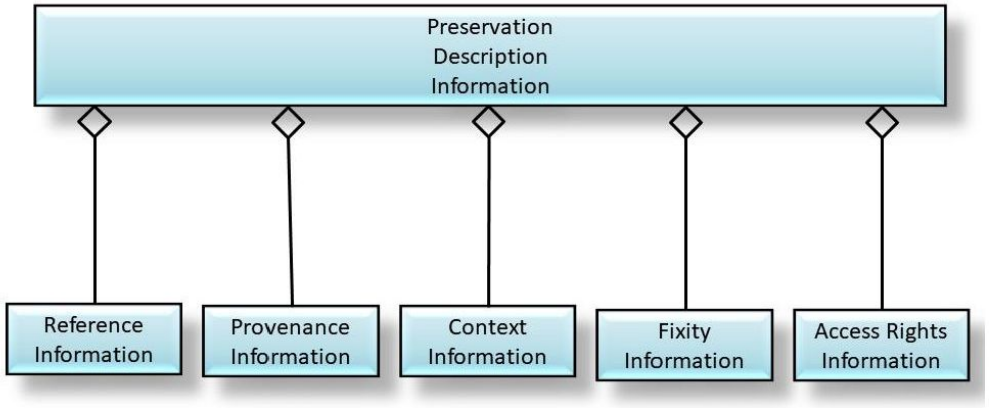


Figure 4-17: Preservation Description Information *Management Council of the Consultative Committee for Space Data Systems (CCSDS). (2019). OAIS final v3 draft with changes wrt OAISv2 20190924-rl.docx.* P. 4-41

Acciones de preservación - Control de cambios

- Dejar registrado en metadatos quien modifico un archivo o los metadatos de un ítem (provenance).
- Definir distintos permisos sobre los metadatos y objetos digitales para los distintos tipos de usuarios (resuelto en DSpace con permisos y grupos).
- Tener documentados los permisos de los usuarios sobre los archivos
- Tener versiones de los objetos digitales ante una modificación

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

Acciones de preservación - Metadatos

- Separar el almacenamiento de los metadatos del objeto digital
- Almacenar metadatos de administración y de preservación (ej PREMIS)
- Uso de identificadores persistentes
    - Para el ítem y para el OD
- Formatos estándares
- Uso de vocabularios controlados
- Uso de metadatos técnicos (como los que arroja exiftool)

Acciones de preservación - Control de contenido

- Perfilamiento de archivos para revisión de formatos por problemas de obsolescencia. (Raramente se hace).
- Actualización de versiones entre formatos.
- Transformación/migración de un formato a otro.
- Análisis de virus.
- Definir un listado de los formatos aceptados.

Conclusiones de la experiencia realizada

- Alcances de NDSA y otras posibilidades
  - NDSA es interesante porque permite la autoevaluación y porque es relativamente simple. De lo mencionado precedentemente en cuanto acciones necesarias para asegurar la PD, todas las descritas por este estándar son de naturaleza técnica.
  - Hay acciones que exceden a los objetos digitales y la infraestructura, por ejemplo acciones que hacen a lo organizacional, que no cubre NDSA, básicamente ahí debe escalarse a una norma como la ISO 16363 que además permite certificar el repositorio pero es bastante compleja.

DIGITALIZACIÓN

Introducción a la Digitalización

- Una estrategia de preservación es adoptar estándares y directrices internacionales. En SEDICI, se utilizan como guía las directrices:
  - "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).
- A partir de estas normas y la propia experiencia, se desarrolló un procedimiento de trabajo presentado en un manual publicado en 2021 (http://sedici.unlp.edu.ar/handle/10915/101101).

Introducción a la Digitalización

- Según las guías FADGI (se aplican al caso de imágenes fijas), los formatos de archivos recomendados para preservación son: **TIFF, JPEG2000 y PDF/A.**
  - **TIFF (Tagged Image File Format)** es un formato de imágenes de estándar abierto que permite compresión sin pérdida. Es utilizado para la creación de archivos maestros de imagen.
  - **PDF/A (Portable Document Format/Archive)** es un estándar de PDF que exige que todos los elementos necesarios para reproducir el contenido tal como se generó estén embebidos en el archivo, independientemente del programa con que se creó.
- Dentro del repositorio SEDICI utilizamos el formato TIFF para el guardado de archivos "**maestros"** y PDF/A para archivos de "**preservación y difusión**".

Circuito de digitalización en el repositorio SEDICI (UNLP)

1. Recepción, análisis y evaluación del material a digitalizar
2. Carga de materiales en el sistema de gestión (Redmine)
3. Elección de metodología de escaneo
4. Captura de imágenes
5. Edición de imágenes
6. Guardado de archivos para preservación y difusión

1) Recepción, análisis y evaluación del material a digitalizar

Todas las obras antes de ingresar al flujo de trabajo son evaluadas teniendo en cuenta estos criterios:

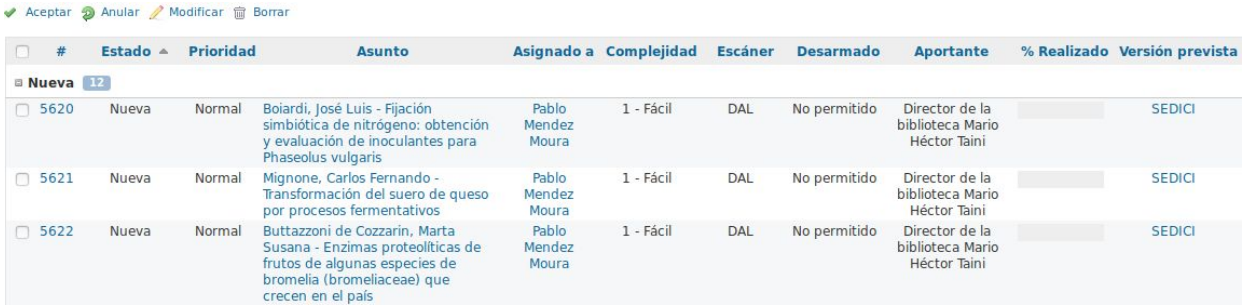
- Estado general de conservación
- Dimensiones
- Formatos
- Tipos de encuadernación
- Importancia histórica, educativa, institucional: Dependiendo de la utilidad y el interés del material, los procesos de edición de imagen tienen mayor o menor automatización y revisión posterior. El material de alta 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 portada e índice y una revisión general del resto. En cambio, los materiales de digitalización rápida requieren una revisión general y un proceso casi totalmente automatizado.

2) Carga de materiales en el sistema de gestión (Redmine)

Luego de tener en claro todas las particularidades de cada caso se:

- asigna el estado de conservación del material
- selecciona el escáner apropiado de acuerdo al formato
- asigna una persona responsable
- determina la complejidad
- agregan los datos propios del documento (Autor, Título etc)

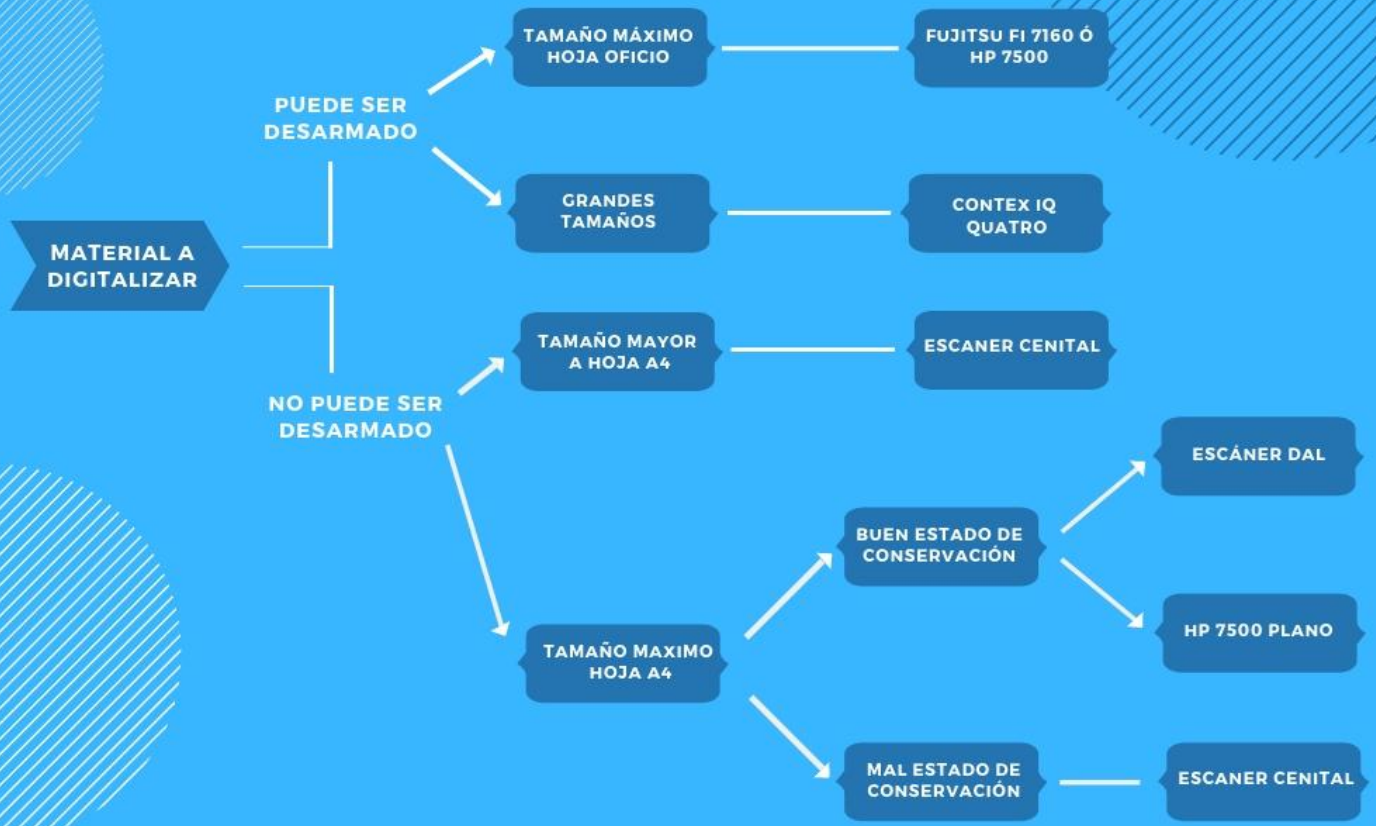
A medida que el trabajo pasa por sus distintas etapas, cada una de ellas queda asentada en el sistema hasta que el proceso finaliza.



#	Estado	Prioridad	Asunto	Asignado a	Complejidad	Escáner	Desarmado	Aportante	% Realizado	Versión prevista
Nueva 12										
5620	Nueva	Normal	Boiardi, José Luis - Fijación simbiótica de nitrógeno: obtención y evaluación de inoculantes para Phaseolus vulgaris	Pablo Mendez Moura	1 - Fácil	DAL	No permitido	Director de la biblioteca Mario Héctor Taini		SEDICI
5621	Nueva	Normal	Mignone, Carlos Fernando - Transformación del suero de queso por procesos fermentativos	Pablo Mendez Moura	1 - Fácil	DAL	No permitido	Director de la biblioteca Mario Héctor Taini		SEDICI
5622	Nueva	Normal	Buttazzoni de Cozzarin, Marta Susana - Enzimas proteolíticas de frutos de algunas especies de bromelia (bromeliaceae) que crecen en el país	Pablo Mendez Moura	1 - Fácil	DAL	No permitido	Director de la biblioteca Mario Héctor Taini		SEDICI

3) Elección de metodología de escaneo

SELECCIÓN DE ESCANER

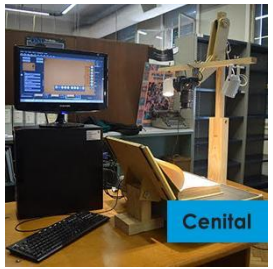


Tipos de escáneres utilizados

automáticos

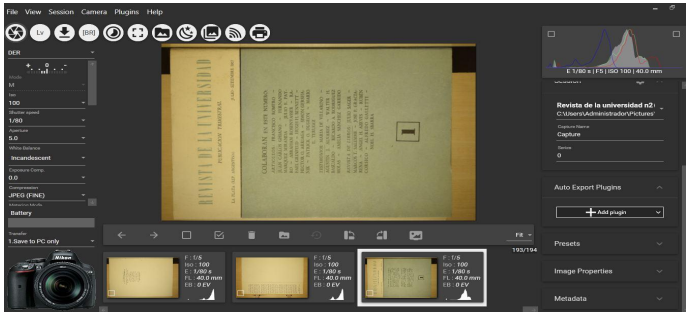
de gran formato

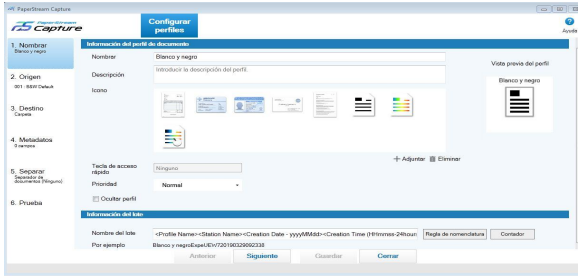
de libros

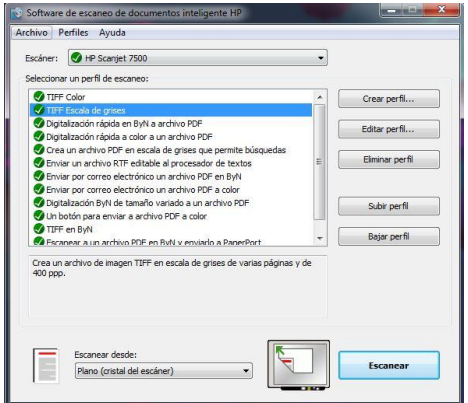


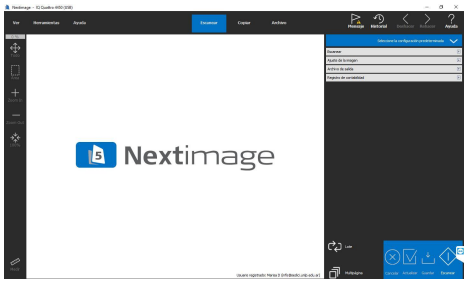
4) Captura de imágenes

- Captura con **“digiCamControl”**
- Captura con “**Paperstream”**
  - Utilizado para la captura con el escáner Fujitsu FI 7160.
- Captura con **“Software de escaneo de documentos inteligente” de HP**
- Captura con **“NextImage”**



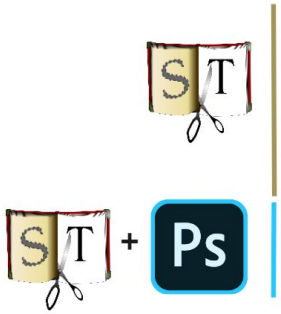






5) Edición de imagen

Para la edición y mejoramiento de imagen se toman los archivos generados en la etapa de captura, y se procesan las imágenes para:

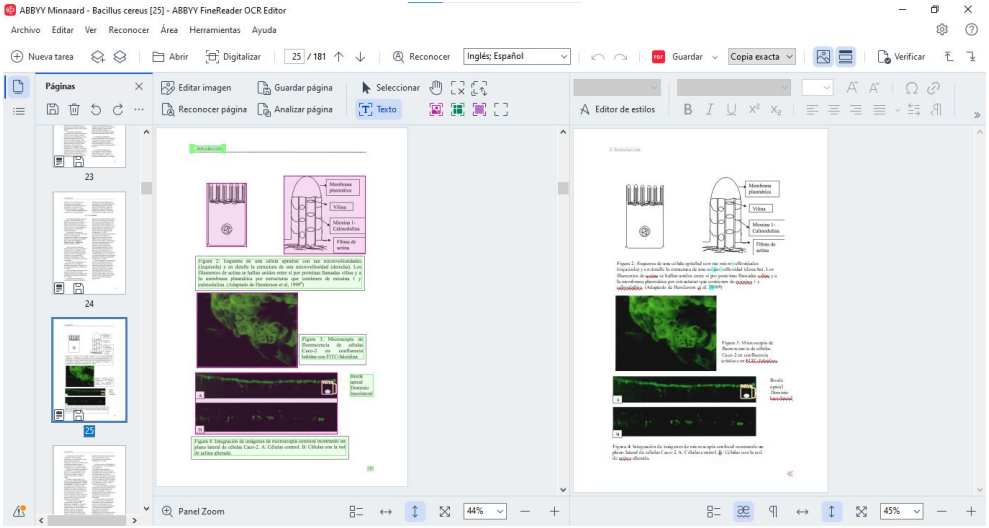


1. Rotar páginas
2. Enderezar las imágenes
3. Ajustar márgenes
4. Eliminar manchas, puntos indeseados
5. Normalizar color
6. Mejorar contraste entre texto y fondo

Para realizar estas acciones se utilizan los productos **Scantailor Advanced** y **Adobe Photoshop**

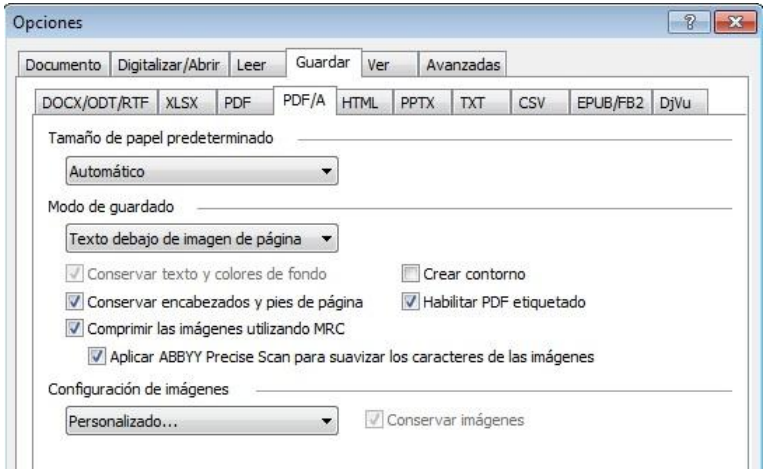
6) OCR y compilación en PDF/A: ABBYYFineReader 16

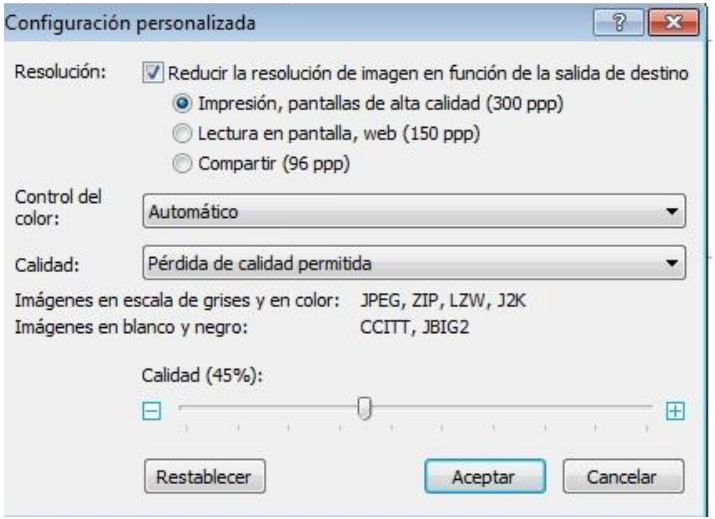
Luego de editar las imágenes se realiza el OCR con Abbyy FineReader 16. En esta etapa del proceso se selecciona el contenido según sea texto, imagen o cuadro. Luego se revisa el resultado del OCR y se generan los archivos PDF/A.



6) Compresión de pdf

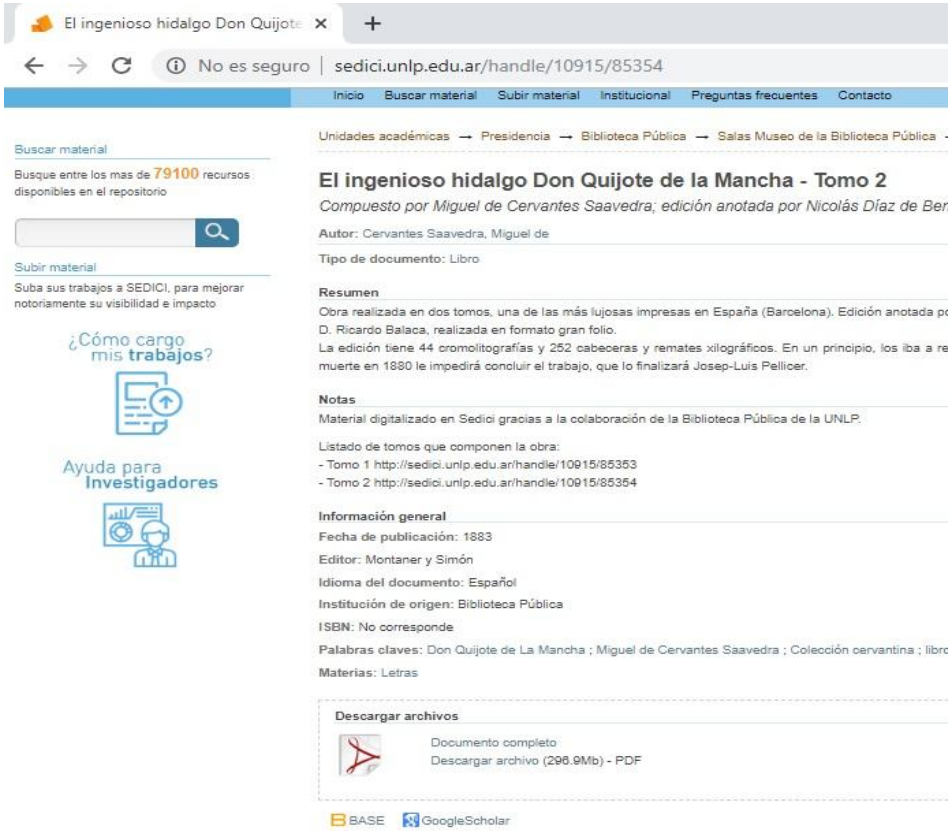
Por último, en el momento del guardado, el programa nos permite modificar la compresión para obtener documentos más pequeños, que pueden ir desde compresiones sin pérdida a compresiones con pérdida de calidad.





6) Guardado de archivos para preservación y difusión

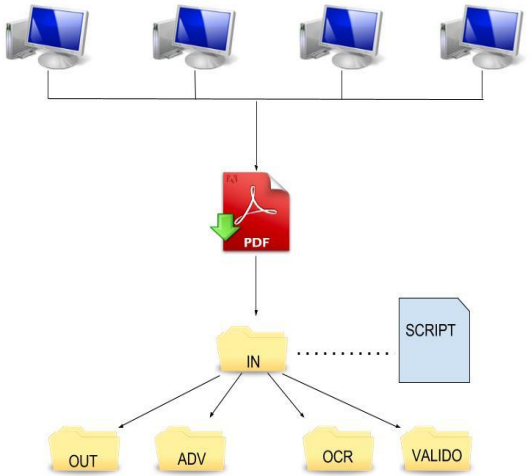
- Se generan dos archivos PDF/A de cada obra: uno de alta calidad de imagen, que se destina a la preservación digital y otro comprimido que se utiliza para difusión en el repositorio.
- Para el caso de algunos libros editados por la UNLP, también se crean libros electrónicos en formato .epub y .mobi para difusión.



6) Conversión por lote: 3-HEIGHT

- Este software, desarrollado por pdf-tools, permite la conversión de archivos PDF a cualquiera de los subestándares de PDF/A-1;2 y 3.
- En el repositorio se ha desarrollado un script escrito en bash que posee una arquitectura cliente servidor y convierte lotes de archivos PDF a PDF/A mediante 3- Height.

- Deteccion de archivos
- Análisis
- Conversión
- Verificación

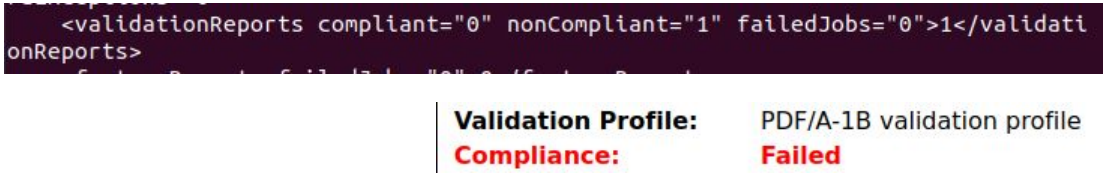


6) Validación de PDF/A

Adobe Acrobat DC

veraPDF

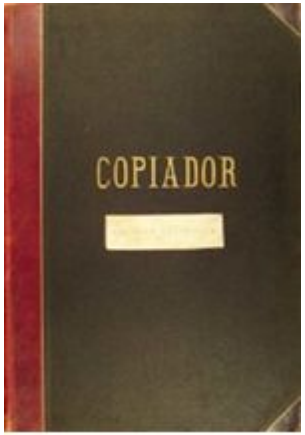
- Es necesario validar que un PDF/A cumpla efectivamente con el estándar.
  - “Un campo de metadatos es obligatorio: el identificador PDF/A. Este identificador se escribe normalmente de forma automática en el campo pertinente en su forma correcta por el convertidor PDF/A que se utiliza para crear el documento en cuestión.” (PDF Association, 2010, p.50)
  - Sin embargo, pueden persistir problemas tales como glifos no definidos, espacios de color dependientes de dispositivos, interpolaciones de píxeles en la imagen que ocasionan errores.



Ver: http://sedici.unlp.edu.ar/handle/10915/139212

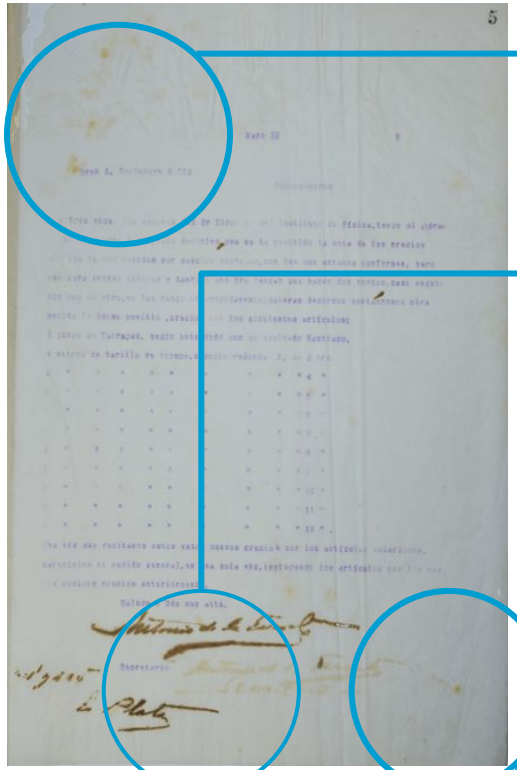
Ejemplo de caso de proceso completo de digitalización:

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



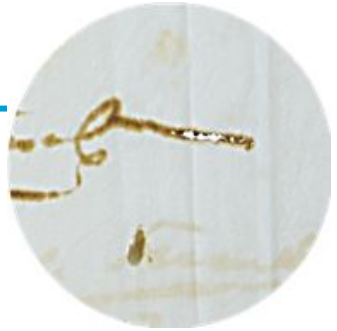
SEDICI y el **Museo de Física de la Facultad de Ciencias Exactas** de la **UNLP** destinaron personal para la digitalización de un documento archivístico: el libro *Copiador – Facultad de Ciencias Físicas, Matemáticas y Astronómicas (1918-1925).* Se siguieron los estándares internacionales para la digitalización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

SEDICI
SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL



Escritura mecanografiada poco legible y pérdida de nitidez

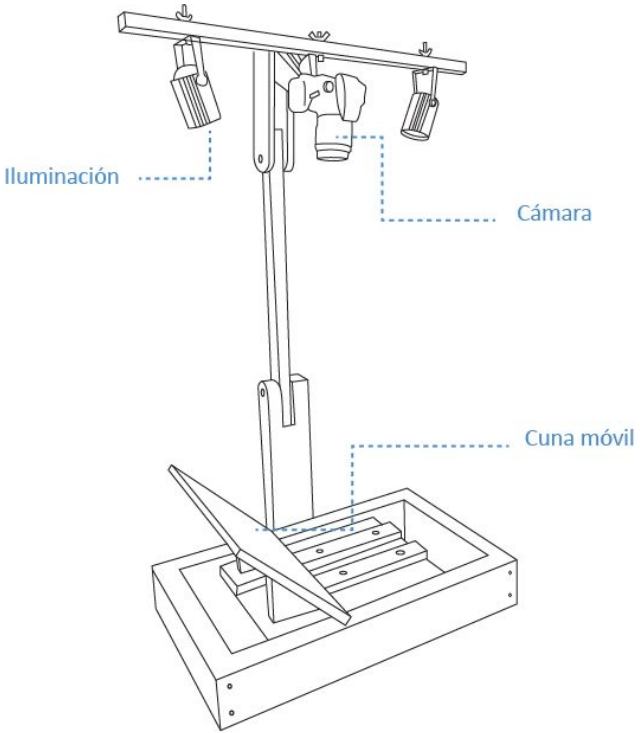
Dobleces y desprendimientos



Tinta difundida en el papel y transferida a los consecutivos.

Estado de conservación

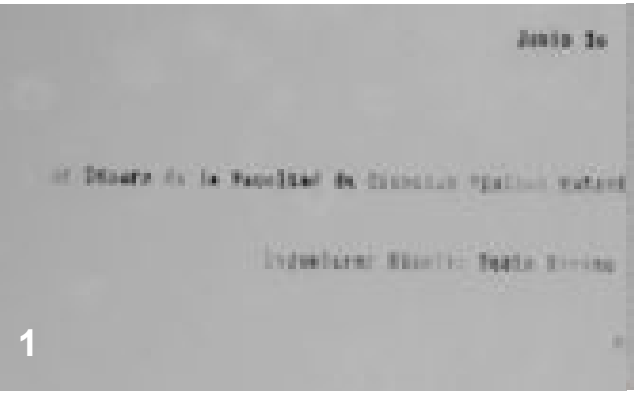
Escáner elegido: cenital



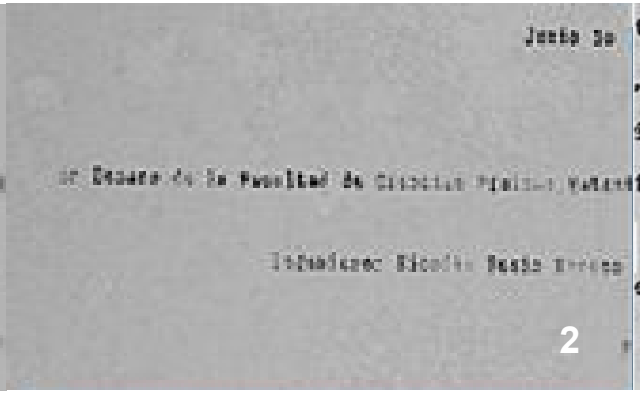
Se optó por un sistema de escaneo **rediseñado a partir del Model 1** de DIY, con una cámara cenital apuntando hacia el libro, junto con dos luces LED dicroicas de luz cálida cuya temperatura no daña el material.

Post-procesos de ajuste de imagen y enfoque (Photoshop)

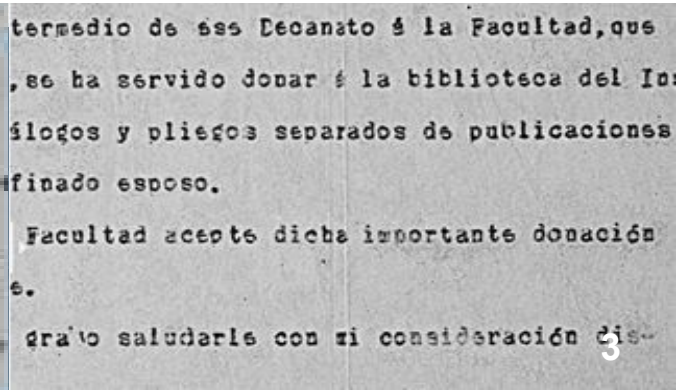
1. **Desaturación por color (black and white filter):** este filtro desatura los colores por separado. Esto permite seleccionar las tonalidades que representan manchas, suciedades y atenuarlas hasta que la superficie se vea homogénea.
2. **Enfocar (smart sharpen)** para acentuar el borde de la tipografía en la imágen y mejorar el contraste con el fondo.
3. Imagen mejorada lista para OCR.
4. El proceso completo se automatizó completamente por medio de las funciones Actions y Droplet de Photoshop



1

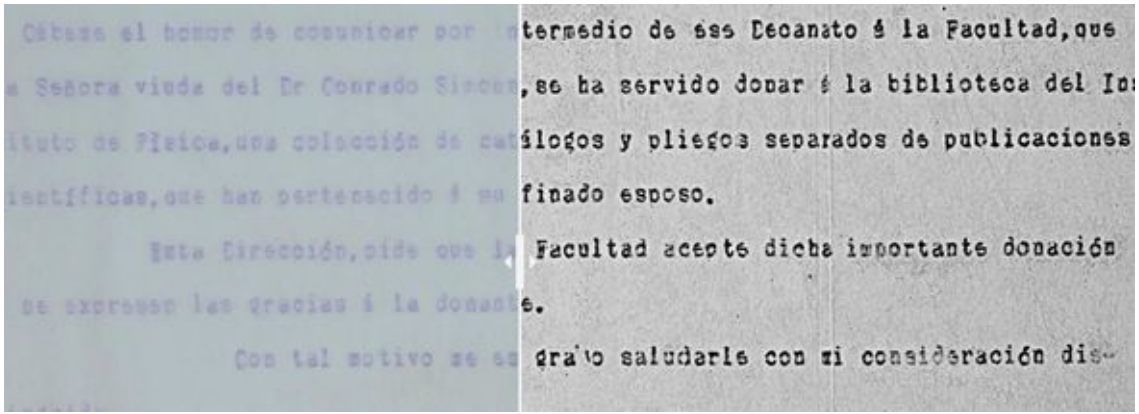


2



3

Imagen original e imagen mejorada lista para reconocimiento de texto (abbyy FineReader)



Por consultas: mdegiusti@gmail.com, marisa.degiusti@sedici.unlp.edu.ar