CONFERENCIA INTERNACIONAL

BIREDIAL-ISTEC

18-20 de octubre de 2023

MONTEVIDEO • URUGUAY

Taller de preservación digital en repositorios institucionales

Dra. Marisa R. De Giusti

19 de octubre de 2023

PREBI prebi.unlp.edu.ar SEDICI sedici.unlp.edu.ar

UNIVERSIDAD NACIONAL DE LA PLATA

CESGI Centro de Servicios en Gestión de Información

COMISIÓN DE INVESTIGACIONES CIENTÍFICAS

Temas a tratar cuando se habla de preservación en un RI

1) **Acciones globales en el repositorio**
   - **a)** Gestión de copias de seguridad (backups): definir esquema de rotación temporal (cuantas y por cuánto tiempo), depósitos offsite, validación de completitud de copias, previsión de escenarios de intrusión que puedan modificar o anular las copias.
   - **b)** Análisis y gestión de riesgos de infraestructura sobre servidores, equipos y software.
     - i) Los riesgos pueden ser:
       - (1) Caída de servicios por errores. Suele implicar caídas por periodos cortos de tiempo.
       - (2) pérdida de datos y aplicaciones por daño. Suele implicar caídas por periodos largos de tiempo, sólo cuando puede recuperarse.
     - ii) Las fallas pueden ser:
       - (1) no intencionales: fallas de hardware, firmware, software de base de datos, filesystem, problemas ambientales,
       - (2) intencionales: hacking, ransomware, atentado, robo
   - **c)** Tener un plan de contingencia ante situaciones anómalas ¿quién/qué/cómo hace si
     - i) se quemó el repositorio
     - ii) se rompió un disco
     - iii) se encriptaron los servidores
   - **d)** Contar con una política y, sobre todo con un plan de preservación que oriente qué se va a guardar y durante cuánto tiempo
   - **e)** Gestión de recursos económicos y humanos
   - **f)** Contar con todos los permisos para la transformación en el tiempo de los objetos digitales
   - **g)** Seguimiento de grado del cumplimiento por parte del repositorio de las actividades vinculadas al almacenamiento, integridad de los datos, control de la información, metadatos y contenidos (autoevaluación al estilo NDSA).

Temas a tratar cuando se habla de preservación en un RI

2) Tópicos vinculados al objeto digital en sí mismo

a) Chequeo de completitud de envío

b) Resguardar la copia de los archivos originales, por más que hayan sufrido transformaciones posteriores.

c) Perfilamiento a lo largo del ciclo de vida de los objetos digitales que están en el repositorio: vía herramientas como DROID o a través del agregado a cada ítem de un metadato de formato que luego se podría consultar

d) Selección de formatos elegidos para todas las actividades del repositorio incluida la de preservación pero también el formato de visualización y el de uso. Conversión en lote por ejemplo en el caso de PDF. Validación. En este punto es importante también tener definido el procedimiento de transformación entre formatos. Ej para pasar a pdf/A se usarán las herramientas h1 y h2, aplicando tales parámetros en h1 y luego tales otros en h2.

e) Control del proceso de conversión (herramientas utilizadas, parámetros aplicados para cada una de ellas).

f) Acciones repetidas a realizar sobre el objeto digital:

- i) asignación de identificador persistente,
- ii) cálculo de checksum para comprobar la no alteración del archivo,
- iii) mejora (ej agregar OCR) o transformación de formato,
- iv) chequeo de virus…

g) Control de cambios (asociado a tener metadatos que sirvan para el control de cambios)

Temas a tratar cuando se habla de preservación en un RI

3) **En relación a los metadatos:**

a) licencias (de difusión y de uso),

b) identificadores persistentes: guardar todos los que se tengan tanto de las obras como de sus autores y organizaciones. Ej doi, handle, isbn, issn, ror, scopusAuthorId, pubmedId, arxivId, orcid, etc. Justificación: Los identificadores permiten vincular existencias de obras en la web. En caso que alguna se pierda, corrompa o no pueda ser usado por algún motivo, es muy probable que se pueda recuperar a partir de existencias en otros espacios relacionados.

c) metadatos:

i) Completitud de metadatos de acuerdo a las directrices del país solicitadas por un nodo nacional y si no lo hay, por OpenAire

ii) Agregado de metadatos de preservación

iii) Extracción automática de metadatos técnicos por ejemplo metadato de formato e ingesta en el repositorio

iv) cambio de sintaxis a nivel de provenance para pensar agente y evento al modo PREMIS (cambio mayor)

Temas a tratar cuando se habla de preservación en un RI

Cuando se incorpora la etapa de digitalización previo a la ingesta del material en el repositorio, es necesario atender a las siguientes etapas y acciones:

1. Recepción, análisis y evaluación del material a digitalizar (estado, tamaño y mucho más)
    a.- Características físicas del material (dimensiones, encuadernación y estado de conservación).
    b.- Cuidado del material para evitar daños o contaminaciones cruzadas.
2. Carga de materiales en el sistema de gestión de tareas o flujo de trabajo (Redmine en nuestro caso)
3. Elección de metodología de escaneo
4. Captura de imágenes
    a.- Calibración de escáneres (resolución, gamma, brillo y contraste).
    b.- Calibración de cámara de fotos (ISO, velocidad de obturador, apertura de diafragma, balance de blancos).
    c.- Calibración de la fuente de iluminación (Índice de Reproducción Cromática, CRI).
    d.- Calibración de dispositivos de reproducción (brillo, contraste y gamma).
    e.- Uso de cartas de calibración para el color y la resolución de imagen.
5. Edición de imágenes
    a.- Criterio estético: equilibrio entre forma y contenido.
    b.- Eficiencia en el uso del espacio de almacenamiento.
6. Guardado y validación de archivos para preservación y difusión: carpetas.
    a.- Nombrado de archivos y jerarquía de carpetas.
    b.- Reconocimiento óptico de caracteres (OCR).
    c.- Compilación de las imágenes con su texto en uno o dos ficheros PDF/A para preservación y difusión.
    d.- Compresión de ficheros con los proyectos (tar para máster, zip para proyectos de OCR).

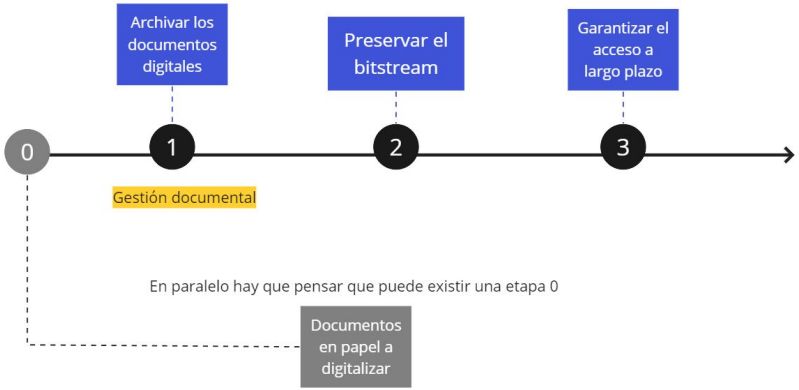
Temas a tratar cuando se habla de preservación en un RI

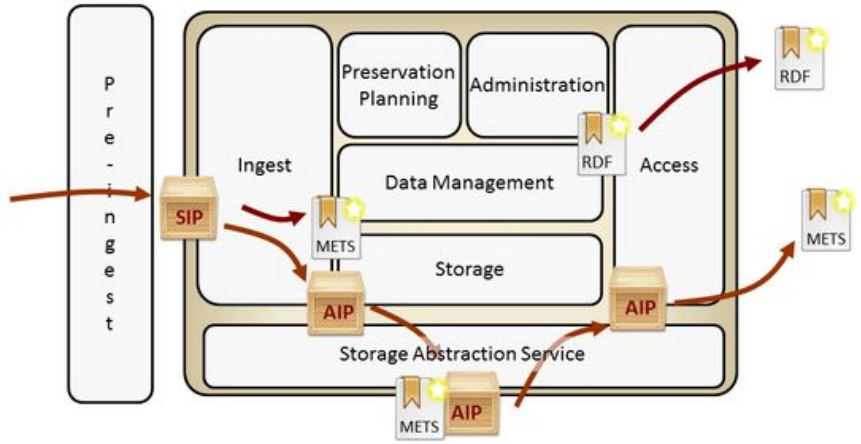
Las acciones y procedimientos en el repositorio deben contar con una contraparte documental, por ejemplo:

1. Documentación de procesos de backups: ¿Cómo se realizan los backups? ¿Qué herramientas se utilizan? ¿Cuál es el procedimiento?
2. Definición escrita sobre qué hacer ante cada potencial situación de amenaza sobre los datos
3. Manuales de carga y procedimientos de carga
4. Documentación del proceso de digitalización de documentos
5. Documentación de las tareas de mantenimiento y desarrollo
6. Inventario del equipamiento del repositorio
7. Plan para el control de obsolescencia de los equipos y su eventual reemplazo

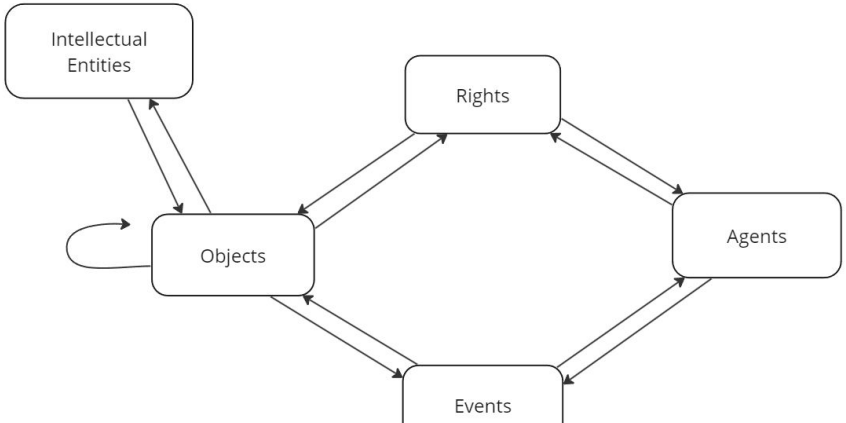
http://sedici.unlp.edu.ar/handle/10915/101101

Preservación y digitalización de documentos



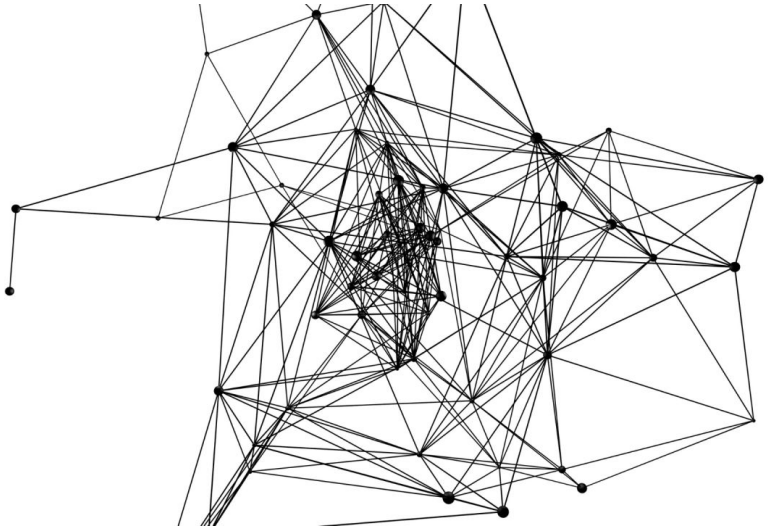


Documentos



Preservación digital

Advertencia: la experiencia deja en claro que la preservación es cosa compleja. Depende de cuestiones técnicas, depende de tener infraestructura y recursos humanos formados, pero también depende de cómo hacemos lo que hacemos.



Objetivos

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

Introducción

- En la actualidad, los recursos que se generan como resultado de los conocimientos de las personas y de sus expresiones “nacen”, cada vez más, en formas digitales, sean de carácter cultural, educativo, o engloben información de diferentes áreas del saber, ya sean de naturaleza técnica, artística o administrativa. Los productos de origen digital pueden no contar con un respaldo físico, por ejemplo en papel.
- Muchos de estos recursos son valiosos y constituyen un verdadero patrimonio a conservar a futuro para la sociedad.
  Además del acceso abierto al material de investigación, la preservación digital es una motivación importante para crear RIs y para asegurar que los materiales de investigación digitales estén disponibles y sean accesibles a largo plazo.

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.

Entonces 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 integridad y autenticidad del recurso.

Problemas en la preservación de OD

1. La propia naturaleza de los objetos digitales los hace efímeros.

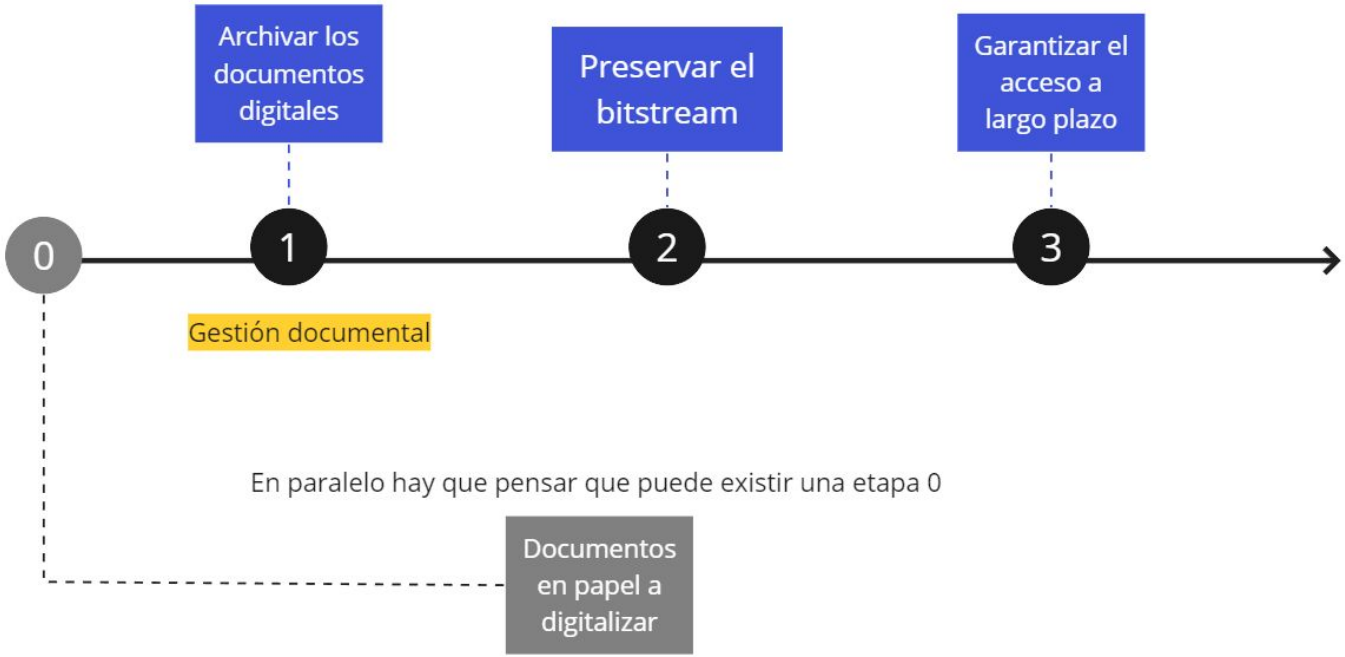
2. La obsolescencia de los medios informáticos: dado que los OD siempre están mediados por la tecnología que cambia constantemente; una inadecuada vigilancia o falta de transformaciones puede dejarlos inaccesibles. La incompatibilidad entre sistemas nuevos y antiguos sumado a que los formatos, medios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

Preservación digital

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

(**Keefer; Gallart, 2007).**

Etapas en la preservación



La preservación supone que:

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

Preservación digital

La preservación digital se define como el conjunto de prácticas de naturaleza política, estratégica y acciones concretas, destinadas a asegurar la preservación, el acceso y la legibilidad de los objetos digitales a largo plazo. Siempre hay que guiarse por una política y un plan para la permanencia de los contenidos.

Noción de preservación de UNESCO

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

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

Objeto digital

Acciones en su ciclo de vida para mantener el acceso

"UNA METODOLOGÍA DE EVALUACIÓN DE REPOSITORIOS DIGITALES PARA ASEGURAR LA PRESERVACIÓN EN EL TIEMPO Y EL ACCESO A LOS CONTENIDOS"
Autora: Ing. Marisa R. De Giusti
Directora: Dra. Silvia Gordillo

Preservación de los contenidos de un RI

Criterios nuevos para los recursos digitales:

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

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 del contenido de los RI

¿Qué materiales hay en los RI?

- resultados académicos y de investigación (tesis, artículos, presentaciones en congresos, etcétera);
- objetos de docencia y aprendizaje;
- datos;
- imágenes, audio, video, materiales multimediales;
- materiales digitalizados;
- material administrativo;
- y todo lo que la institución considere pertinente.

Preservación del contenido de los RI

¿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.

Preservación del contenido de los RI

Selección de recursos para su preservación

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

¿Qué atributos se quieren preservar?

datos, funcionalidad

apariencia, esencia

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

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. Lo mismo con PREMIS.

Preservación de contenido

- Hay una muy importante necesidad de preservar el contenido digital en el tiempo, con el objetivo de conservarlo accesible frente a riesgos como: incendios, inundaciones, robos, problemas de hardware (rotura de discos, etc.) y cambios tecnológicos constantes.
- *Es un proceso continuo*
- Además de lo técnico, los esfuerzos de preservación incluyen retos legales, económicos e institucionales.

Obsolescencia

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

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

Cornell University Library creó la "Cámara de los horrores" http://dpworkshop.org/dpm-eng/oldmedia/chamber.html

Preservación de contenido. “Obsolescencia digital”

Mantener tecnologías obsoletas requiere conservar

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

Suelen ser opciones muy difíciles de mantener y muy costosas.

Preservación de contenido. Estrategias

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

- Migración
- Adhesión a estándares internacionales
- Emulación
- Encapsulamiento
- Metadatos de preservación
- Políticas de backup
- Algo muy importante está vinculado al MODO de trabajo para asegurar la trazabilidad

Preservación de contenido. Metadatos de preservación

Los metadatos de preservación sirven para registrar información (además de la descriptiva) relativa a la evolución de los recursos en el tiempo para tener explícitas las acciones de preservación aplicadas, incluyendo información sobre formatos, usos, actividades de preservación realizadas, responsables de dichas actividades, etcétera.

Varias iniciativas:

- PREMIS: PREservation Metadata: Implementation Strategies
- 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

Preservación de contenido. Migración continua

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

- Es una de las opciones de mayor uso
- Asegura el acceso en todo momento (los datos son siempre accesibles mediante una tecnología actual)
- Requiere transformación de los datos originales
- Decisiones sobre qué se desea preservar
    - Más adelante se verá de qué habla la Norma ISO 14721

Un avance: estándares que sirven

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.

Preservación de contenido. Elección de formatos de archivo

- **Elegir formatos comunes al campo disciplinar al que se está trabajando:** Para asegurar la interoperabilidad y la reutilización de los datos, es fundamental elegir un formato relacionado a la disciplina científica en el que se desarrolla el estudio.
- Tener en cuenta el **tiempo** que se espera conservar los datos: Es uno de los factores más relevantes al momento de realizar una elección correcta de formatos de conservación. Cuanto mayor sea el periodo de tiempo que se desea conservar los datos, mayor será la necesidad de seleccionar formatos abiertos, estandarizados y bien documentados y considerar los medios de almacenamiento confiables y de calidad.

Preservación de contenido. Elección de formatos de archivo

- **La conversión de archivos puede provocar la pérdida de datos:** Para evitar la pérdida de información al momento de convertir los archivos, es importante considerar formatos de multiplataforma común que respondan a estándares específicos. Si la conversión a un formato de datos abiertos puede generar alguna pérdida de los datos, se puede considerar guardar los datos tanto en el formato propietario como en un formato abierto y revisar siempre el estado completo de los datos antes y después de hacer la conversión, ya que, pueden ocurrir errores en al momento de utilizar el software

- **Verificar los requisitos del repositorio de datos:** Muchas revistas, archivos y repositorios requieren de formatos específicos al momento de cargar la información. Es importante tomar en cuenta esto desde el comienzo del proyecto, para poder elegir el mejor formato de archivo al momento de recopilar, procesar y compartir los datos, junto con la conversión en las diferentes fases de la investigación.

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

Preservación de contenido. Elección de formatos de archivo. Bibliografía

ANDS. (2016b, diciembre). ANDS Guide: File Formats. Australian National Data Services. Recuperado a partir de http://www.ands.org.au/__data/assets/pdf_file/0003/731775/File-Formats.pdf

Library of Congress. (s. f.-a). Format Description Categories - Sustainability of Digital Formats [Webpage]. Recuperado 22 de junio de 2017, a partir de https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml

MIT Libraries. (s. f.). File formats for long-term access [Blog]. Recuperado 22 de junio de 2017, a partir de https://libraries.mit.edu/data-management/store/formats/

Open Knowledge International. (s. f.). Formato de Archivos [Webpage]. Recuperado 22 de junio de 2017, a partir de http://opendatahandbook.org/guide/es/appendices/file-formats/

The University of Edinburgh. (2014). File formats and transformation - MANTRA Research Data Management Training [Web page - Free online course]. Recuperado 23 de junio de 2017, a partir de http://mantra.edina.ac.uk/fileformatandtransformation/

UK Data Service. (s. f.-b). File formats & software [Webpage]. Recuperado 23 de junio de 2017, a partir de https://www.ukdataservice.ac.uk/manage-data/format/file-formats.aspx

Grupo de la OPF

International Comparison of Recommended File Formats

Comparación accesible en: Comparación de formatos por el grupo de la OPF

Video: https://www.youtube.com/watch?v=Iik4YIC0-1k&ab_channel=OpenPreservationFoundation

Recomendaciones de formatos

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

Recomendaciones de formatos

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, JPEG 2000, PDF, PNG, 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, el PNG (Portable Network Graphics), cuya compresión experimenta apenas pérdidas en la resolución y además es muy ligero y el JPEG.

Selección de formatos: algunos especiales

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

Sobre PDF/A

Porque los RI tienen mucho texto

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

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 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.

Sobre PDF/A

PDF se creó como un formato que podía ofrecer interoperabilidad a través de diferentes software, ordenadores y plataformas. PDF/A amplia la idea en el tiempo: asegura que los documentos PDF seguirán pudiéndose abrir en el futuro.

PDF/A es un formato PDF destinado al archivado, a la conservación a largo plazo y al intercambio de documentos electrónicos. La aspecto visual de los documentos electrónicos se mantiene en el tiempo, independientemente de las herramientas y sistemas que se hayan utilizado para su producción, almacenamiento y reproducción. Las fuentes de documentos pueden ser de papel, correos electrónicos, documentos PDF «normales», páginas web y muchos más. PDF/A realiza una captura digital instantánea y fiable de cualquier documento que sigue permitiendo realizar búsquedas y sigue siendo plenamente procesable.

Texto extraído de: https://pdf.abbyy.com/es/learning-center/pdf-standards/

Estándares PDF

Lo que es importante entender sobre el estándar PDF y sus diferentes especificaciones es que las versiones siguientes no eliminan las anteriores. Cada nueva versión amplía las capacidades de formato, pero eso no significa que todas deban usarse al crear documentos PDF, ni que los documentos creados según las especificaciones anteriores se vuelvan obsoletos con la introducción de PDF 2.0. De hecho, todavía no circulan demasiados documentos PDF 2.0 y aún son menos los que utilizan las últimas funciones del formato, mientras que la mayoría de documentos, incluso los creados ahora, son PDF 1.7 o incluso anteriores. La razón es simple y básicamente responde al propósito y filosofía del formato PDF: si una especificación anterior es suficiente para representar bien el contenido de un documento, es mejor que se use para ofrecer la máxima compatibilidad con diferentes software de PDF.

Texto extraído de: https://pdf.abbyy.com/es/learning-center/pdf-standards/

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.

Para aprender también puede verse: https://pdf.abbyy.com/es/learning-center/pdf-standards/

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	
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).

Criterios de uso de PDF/A en SEDICI y CIC Digital

- En el área de Digitalización de PREBI-SEDICI se usa PDF/A-1 y 2 en todos sus subniveles. Se privilegia el uso de PDF/A-2 porque permite mayor compresión de imágenes y transparencias y capas de imágenes.
- Se evita PDF/A-3, porque permite cualquier tipo de formato embebido, lo que presenta un problema de preservación y seguridad.
- Ghostscript y las herramientas *open source* basadas en él, como OCRmyPDF, permiten generar un PDF/A-1; 2 ó 3, pero siempre en el subnivel b.
- ABBYY FineReader permite elegir entre los estándares 1; 2 y 3 todos los subniveles: dependiendo de qué tan bien el OCR comprendió la estructura del documento digitalizado, se elige entre el a y el b. El u se usa cuando hay caracteres distintos del alfabeto latino español, como æ u œ.
- 3H hace un análisis del PDF y determina y aplica el estándar más adecuado entre PDF/A-1; 2 ó 3, en todos sus subniveles.

	PDF/A-1	PDF/A-2
Fuentes embebidas	✅	✅
Incrustación de otros archivos	x	✅
Metadatos estandarizados	✅	✅
Inclusión de perfiles de color	✅	✅
Transparencias	x	✅
Compresión ZIP	✅	✅
Compresión JPEG 2000	x	✅

PDF/UA

«UA» significa accesibilidad universal («Universal Accessibility» en inglés), y PDF/UA es una especificación que define cómo hacer que las tecnologías de ayuda (software especiales o incluso dispositivos) puedan leer un documento PDF, de manera que un ordenador pueda leer en voz alta el contenido de tal documento a cualquier persona que dependa de estas tecnologías. Debido a que los documentos PDF se han vuelto muy frecuentes en nuestras vidas, en especial en ámbitos como los servicios públicos, la banca, los servicios básicos, el empleo, la medicina, la educación y muchos otros, asegurar una accesibilidad fácil y por igual a los mismos es realmente crucial.

Texto extraído de: https://pdf.abbyy.com/es/learning-center/pdf-standards/

PDF/UA

Un documento PDF/UA tiene una estructura lógica definida con claridad y corrección, y correctamente descrita. Al usar esta descripción de estructura, la tecnología de ayuda sabrá y podrá decir cuál es el encabezado del documento, en qué orden leer los párrafos y las columnas del texto, cuáles son las listas, dónde están las imágenes y qué muestran, saltarse la lectura de los encabezados y pies de página con numeración que se repiten, etc.

FineReader PDF puede tanto convertir documentos PDF existentes de cualquier tipo en PDF/UA como crear documentos PDF/UA a partir de archivos de otros formatos como DOCX, XLSX, PPTX, RTF, archivos de imagen y otros. Esto es posible gracias a la tecnología OCR de ABBYY, que es capaz de analizar la estructura de cualquier documento independientemente de su formato.

Texto extraído de: https://pdf.abbyy.com/es/learning-center/pdf-standards/

Preservación de contenido

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

Y despues... la vigilancia continua

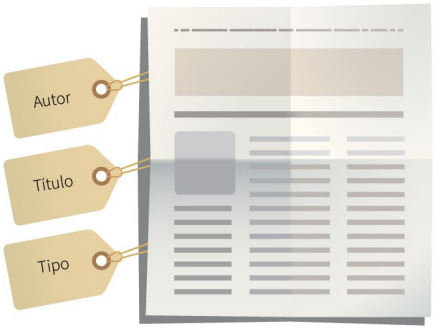
Formatos. ¿Cómo conocer lo que tiene un RI?

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

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

¿Qué acciones se proponen?

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

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…

https://docs.google.com/document/d/1euH8mxYM1OlVzOTzBva_qHq-F3_hre-I/edit

Modelo Funcional ISO 14721 entre muchas cosas

La preservación en el repositorio institucional: el Modelo OAIS ISO 14721. Comparación con las facilidades que ofrece DSPACE.

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.

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

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).

OAIS-DPC Online: https://wiki.dpconline.org/index.php?title=OAIS_Structure

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.

El modelo de Referencia OAIS

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

Sección 1

Justificación del Modelo de referencia

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

Antecedentes

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

Funciones del Modelo de referencia

- Las dos funciones principales del modelo son **conservar** la información y **garantizar el acceso** a la misma.

- El modelo funcional OAIS, que se propone lograr estos objetivos amplios, en cierta medida, define la arquitectura aproximada de cualquier tipo de sistema de software diseñado para cumplir con esta norma y con todo tipo de flujos de trabajo asociados con el repositorio.

Propósito y campo de aplicación

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

Propósito y campo de aplicación

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

Sección 2

- **2 OAIS CONCEPTS (2-1)**
  - 2.1 OAIS ENVIRONMENT (2-2)
  - 2.2 OAIS INFORMATION (2-3)
    - 2.2.1 INFORMATION DEFINITION
    - 2.2.2 INFORMATION PACKAGE DEFINITION
    - 2.2.3 INFORMATION PACKAGE VARIANTS
  - 2.3 OAIS HIGH-LEVEL EXTERNAL INTERACTIONS (2-8)
    - 2.3.1 MANAGEMENT INTERACTION
    - 2.3.2 PRODUCER INTERACTION
    - 2.3.3 CONSUMER INTERACTION

Conceptos en OAIS

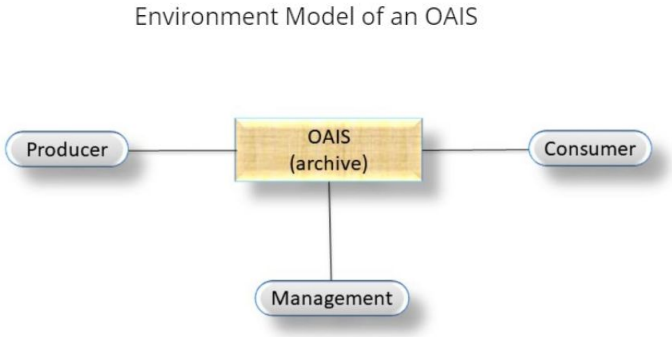
El propósito de esta sección es motivar y describir varios conceptos clave, de alto nivel del OAIS. Un punto de vista más completo y una modelización formal de estos conceptos, se da en la Sección 4.

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

Definición de información

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

Conceptos en OAIS
Definición de información

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

Conceptos en OAIS
Paquete de información

- La unidad de intercambio entre un OAIS y su medioambiente es el paquete de información –IP.
- Un IP contiene 2 tipos de información:
  De contenido
- De descripción de preservación (PDI)
- La información de contenido y la PDI pueden verse como encapsuladas e identificables por medio de la información de empaquetado.
- El paquete resultante es recuperable en virtud de la información descriptiva: DI.

Conceptos en OAIS

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**.

Conceptos en OAIS

Paquete de información

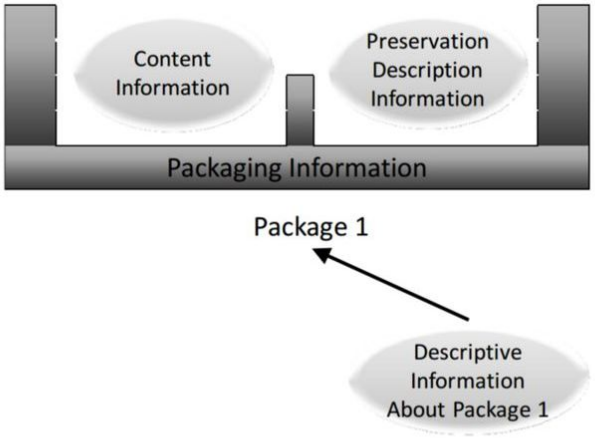


Figure 2-3: Information Package Concepts and Relationships

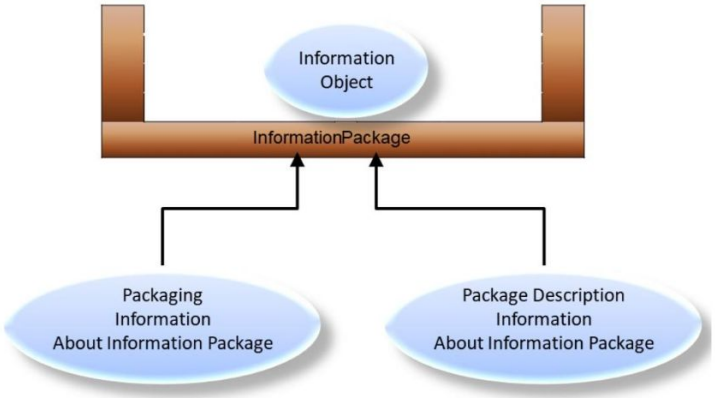


Figure 2-3: Information Package Concepts and Relationships

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

Conceptos en OAIS: IP en la nueva versión

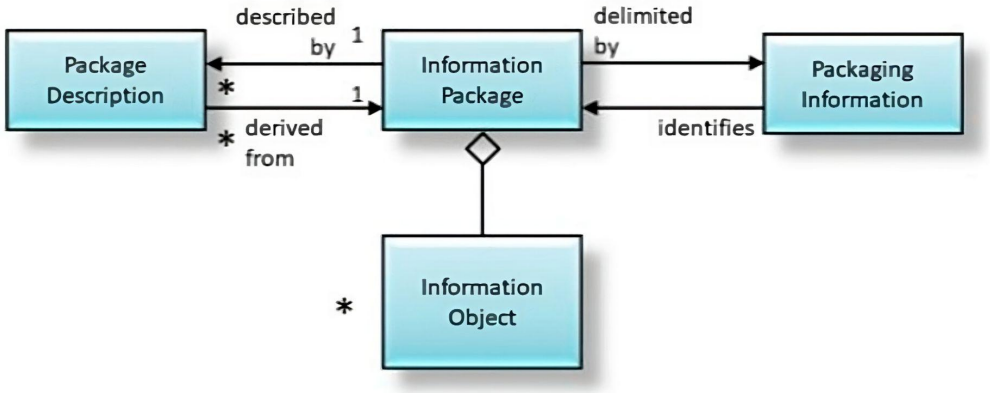


Figure V-2 Updated Information Package

Estructura del Paquete de Información

Conceptos en OAIS

Información de empaquetado y descriptiva

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

Elementos de la PDI

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

Conceptos en OAIS

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

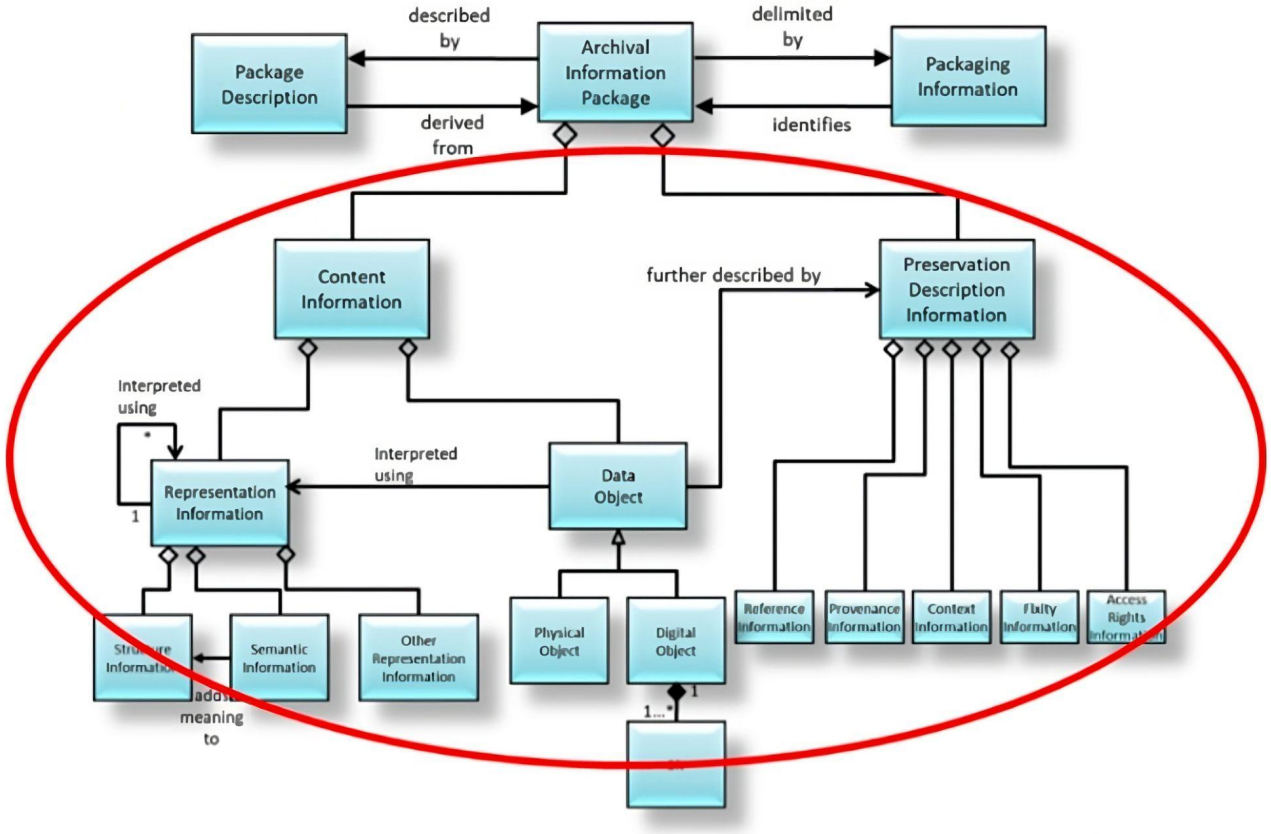
Clases de IP según su función

Submission Information Package (**SIP**): es el paquete que proviene del productor y se va a incorporar al OAIS. Suele contener menos información que el AIP.

Clases de IP según su función

Archival Information Package (**AIP**): contiene, como mínimo, suficiente información de un objeto como para garantizar la preservación a largo plazo. Busca mantener la mayor calidad posible de información descriptiva de preservación y de representación de los objetos representados o contenidos.

AIP


described by
Package Description
derived from
Archival Information Package
delimited by
identifies
Packaging Information
Content Information
further described by
Preservation Description Information
Interpreted using
1
Representation Information
Interpreted using
Data Object
Structure Information
Semantic Information
Other Representation Information
meaning to
Physical Object
Digital Object
1
1...*
Reference Information
Provenance Information
Context Information
Fixity Information
Access Rights

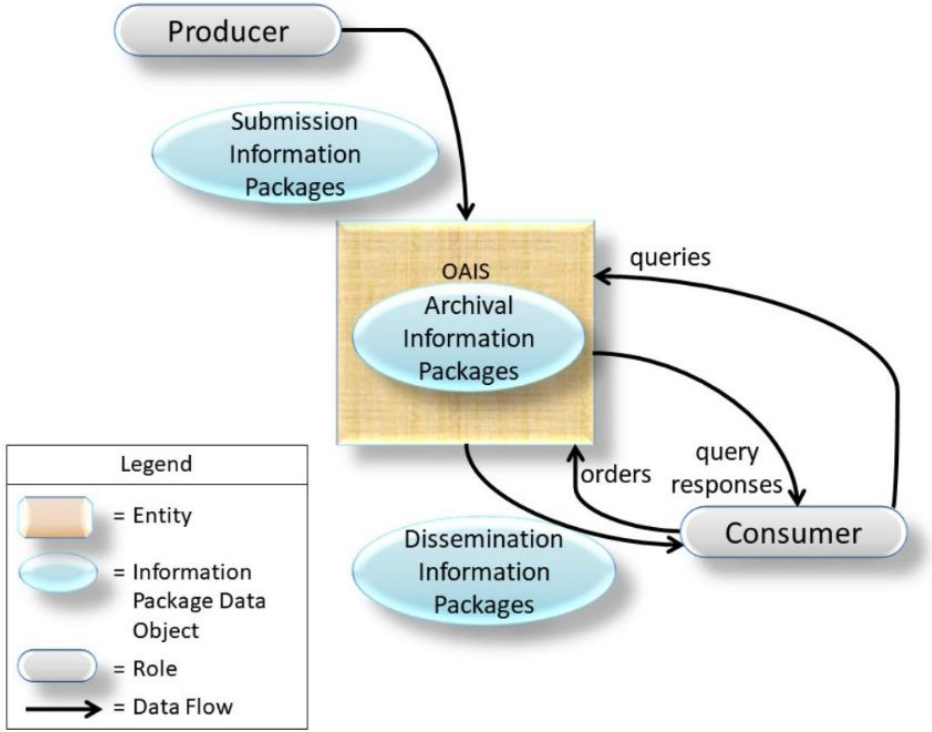
Clases de IP según su función

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

OAIS interacciones externas de alto nivel

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

OAIS interacciones externas



Visión de alto nivel de las interacciones en un entorno OAIS

- Interacción de la gestión
  - financiación, utilización de recursos, pagos, resolución de conflictos.

- Interacción del productor
  - los acuerdos de ingesta. Acuerdo por los SIPs que va a mandar, tiempo (acuerdo por data submission session)

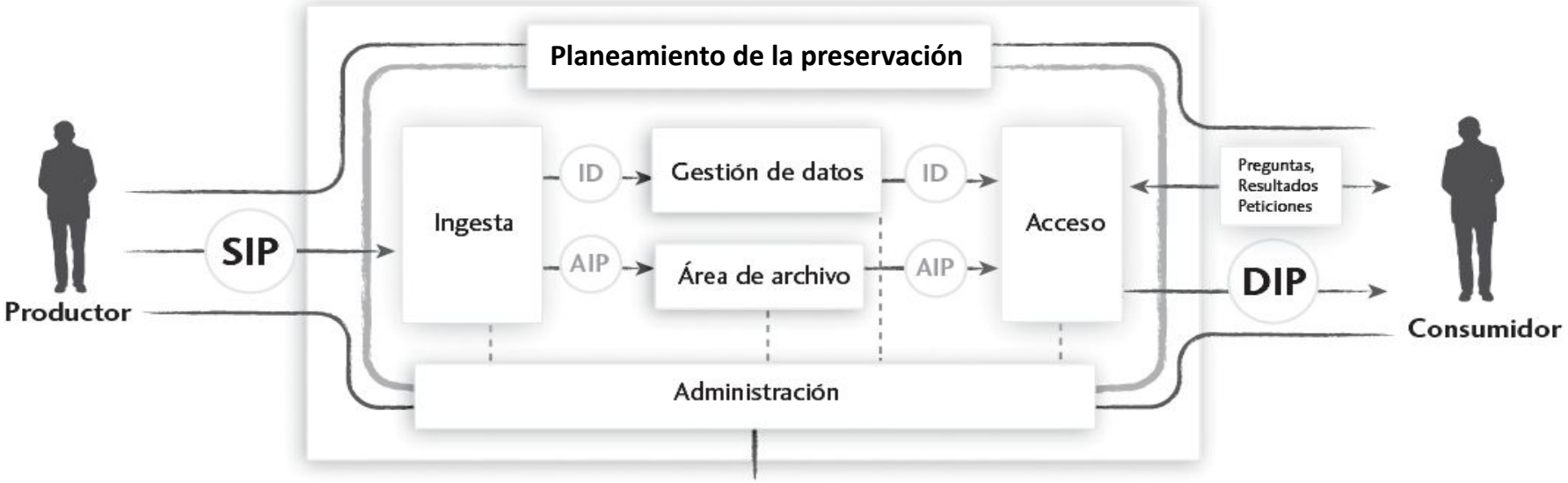
- Interacción de los consumidores
  - Ayudas, descubrimiento de información, ordenamiento de la información. (Data dissemination session).

Sección 4

Open Archival Information System

Modelo Funcional

Sección 4.1



Este modelo abstracto es como una especie de lista de requerimientos que se deberían cumplir para asegurar la preservación y el acceso a largo plazo.

OAIS Modelo funcional

Seis entidades funcionales e interfaces relacionadas:

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

Modelo OAIS

El 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.

Modelo OAIS

Los datos relacionados con los documentos y el repositorio mismo se mantienen organizados a través de la entidad *data management*. Luego hay una entidad *administration* dedicada a la administración adjunta a la gestión (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.

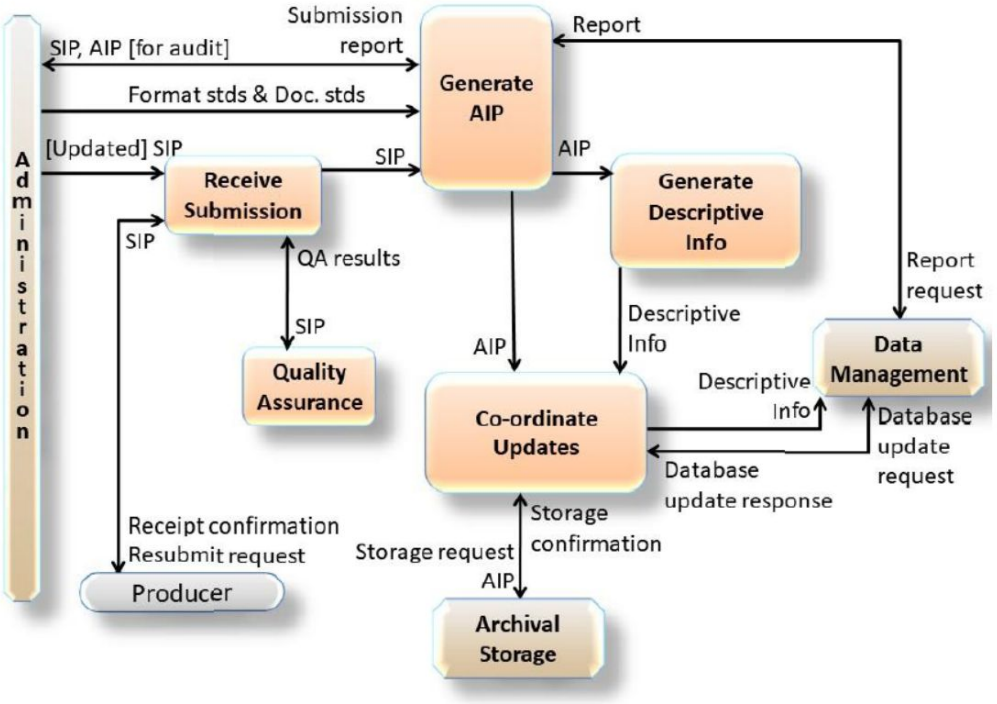
Modelo OAIS

El módulo de *planificación de la preservación* desarrolla estrategias y normas de conservación, monitorea las últimas novedades y avances en el campo, y monitorea los cambios en la comunidad designada, para que toda la información nueva que se solicite, se pueda adjuntar a los AIP correspondientes.

Ingesta/Ingest/presentación

Provee los servicios y funciones para aceptar el paquete de información presentado (SIP) por parte de los Productores (o a partir de elementos internos bajo control de la administración) y preparar los contenidos para almacenaje y gestión dentro del archivo.

Funciones de Ingest/Ingesta/Ingreso



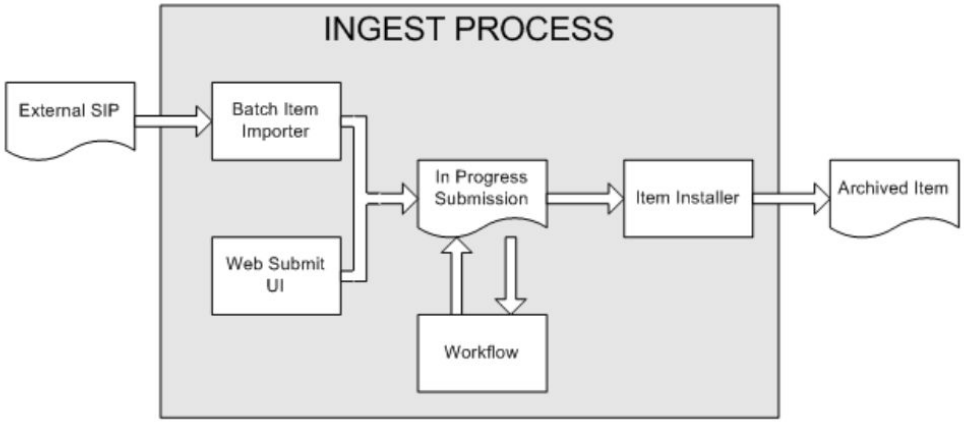
Entidad **Ingest**

- Provee los servicios y funciones para aceptar un SIP por parte de los Productores o bajo el control de la Administración: puede ser una transferencia de custodia (e incluir la licencia adecuada) e incluso puede requerir del agregado de permisos especiales para el acceso.
- Prepara los contenidos para almacenamiento y gestión dentro del archivo.
- Realiza el aseguramiento de calidad/validación de los SIPs.
- Genera el AIP que cumple con los estándares de formato de datos y documentos.
- Extrae la información descriptiva y la envía al data management.
- Coordina las actualizaciones en el archival storage y en el data management de la base de datos.

Entidad **Ingest**

- La evidencia de autenticidad puede tomar muchas formas. La evidencia está diseñada para respaldar la afirmación de que el objeto es lo que se supone que es. La evidencia inicial es proporcionada por el Productor como parte de la PDI en la presentación/envío y esta evidencia es mantenida, actualizada y/o incrementada por el archivo/repositorio a lo largo del tiempo.
    - Es posible que sea necesario cambiar el objeto digital de alguna manera para que siga siendo comprensible de forma independiente para la comunidad designada. Es importante que estos cambios se documenten como parte de la Información de procedencia (provenance information) del objeto para que el objeto pueda rastrearse hasta el objeto original enviado al Archivo por el Productor. También es importante que cualquier cambio en el objeto no cambie la información del contenido hasta el punto de que ya no transmita la información prevista del objeto original. Un método para proporcionar evidencia que respalde la afirmación de la autenticidad del objeto modificado es el uso de descripciones de propiedades de la información.

Procedimiento INGEST en DSPACE



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

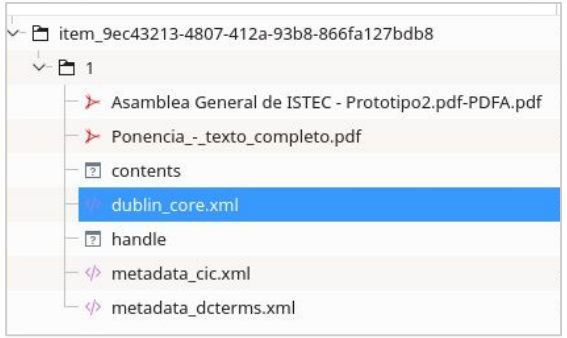
Sería muy importante saber agentes y eventos que se anota o no se anotan según de dónde provenga el OD.

Entidad Ingest en DSpace

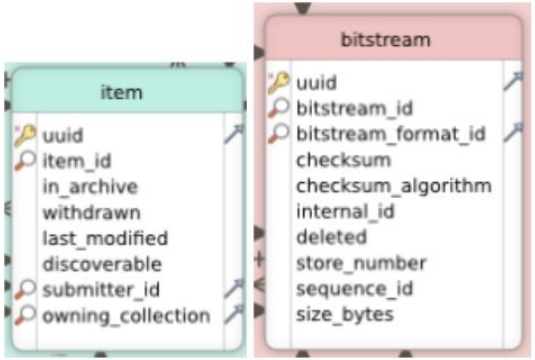
- El proceso de carga tradicional o Submission process, permite configurar para cada colección y para cada tipología documental qué metadatos se requieren de forma opcional u obligatoria, y cuál es la secuencia de etapas que debe cumplir el ítem antes de ingresar al repositorio.
- La cosecha se realiza 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

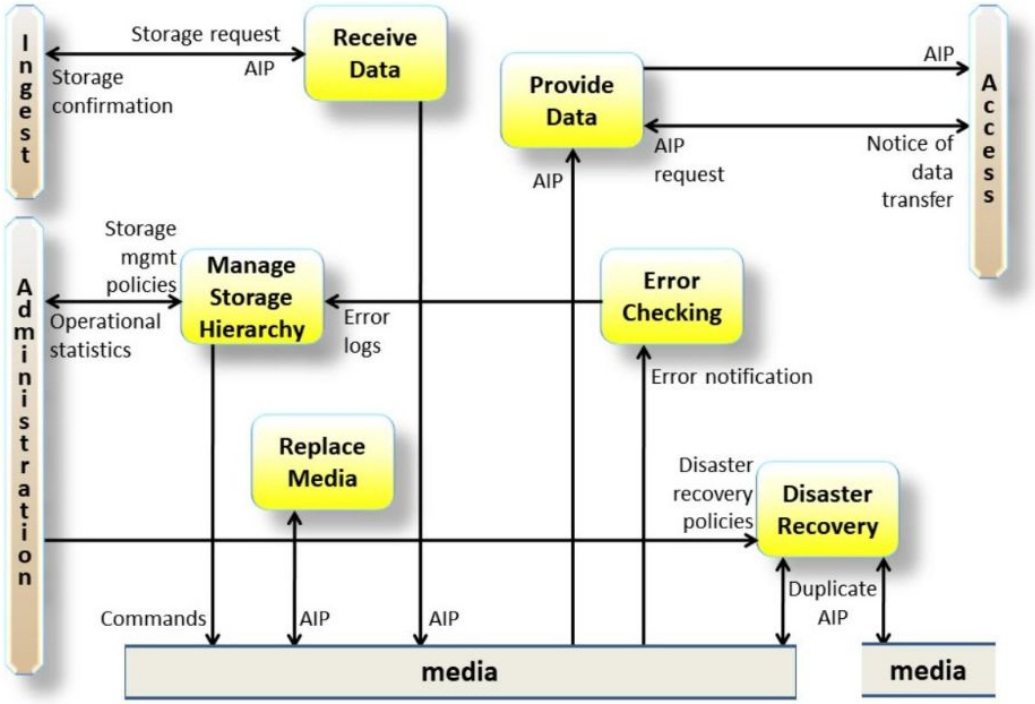
Aspectos mejorables

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

Es necesario contar con políticas de contenidos adecuadas, controles sobre lo que se sube y antivirus que hagan control periódico (ej 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.

Functions of 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

La función **Recibir datos** recibe una solicitud de almacenamiento y un AIP de Ingest y mueve el AIP al almacenamiento permanente dentro del Archivo. Es posible que la solicitud de transferencia deba indicar la frecuencia prevista de utilización de los Objetos de datos que componen el AIP para permitir que se seleccionen los dispositivos o medios de almacenamiento apropiados para almacenar los AIP. Esta función seleccionará el tipo de medio, preparará los dispositivos o volúmenes y realizará la transferencia física a los volúmenes de Archival Storage. Una vez completada la transferencia, esta función envía un mensaje de confirmación de almacenamiento a Ingest, incluida la identificación de almacenamiento de los AIP.

Entidad Archival Storage

La función **Administrar jerarquía de almacenamiento** posiciona, a través de comandos, el contenido de las AIP en los medios apropiados según las políticas de administración de almacenamiento, las estadísticas operativas o las instrucciones de Ingest a través de la solicitud de almacenamiento. También se ajustará a cualquier nivel especial de servicio requerido para la AIP, o cualquier medida de seguridad especial que se requiera, y asegura el nivel apropiado de protección para la AIP. Estos incluyen el almacenamiento en línea, fuera de línea o casi en línea, la tasa de rendimiento requerida, la tasa de error de bits máxima permitida o los procedimientos especiales de manejo o copia de seguridad. Supervisa los registros de errores para garantizar que los AIP no se dañen durante las transferencias. Esta función también proporciona estadísticas operativas a la administración que resumen el inventario de medios disponibles, la capacidad de almacenamiento disponible en los distintos niveles de la jerarquía de almacenamiento y las estadísticas de uso.

Entidad Archival Storage

La función **Reemplazar medios** brinda la capacidad de reproducir los AIP a lo largo del tiempo. Dentro de la función Reemplazar medios, la Información de contenido y la Información de descripción de preservación (PDI) no deben modificarse. Sin embargo, los datos que constituyen la Información del Empaquetado pueden cambiarse *siempre y cuando continúen realizando la misma función y exista una implementación sencilla que no provoque la pérdida de información*. La estrategia de migración debe seleccionar un medio de almacenamiento, teniendo en cuenta las tasas de errores esperadas y reales encontradas en varios tipos de medios, su rendimiento y sus costos de propiedad. Si se han incluido atributos dependientes de los medios (p. ej., tamaños de bloque de cinta, información de volumen de CD-ROM) como parte de la información de contenido, se debe encontrar una manera de preservar esta información al migrar a medios de mayor capacidad con diferentes arquitecturas de almacenamiento. Anticipándose a la terminología de 5.1.3, esta función puede realizar 'Renovación', 'Replicación' y 'Reempaquetado' que es sencillo. Un ejemplo de tal 'Reempaquetado' es la migración a nuevos medios bajo un nuevo sistema operativo y sistema de archivos, donde la información de contenido y la PDI son independientes de los sistemas de archivos. Sin embargo, el 'Reempaquetado' complejo y toda la 'Transformación' se realizan bajo la supervisión de la Administración mediante la función de Actualización de información de archivo para garantizar la preservación de la información. (Consultar 5.1.3 para obtener una descripción detallada de los problemas de migración).

Entidad Archival Storage

La función de **Comprobación de errores** proporciona una seguridad estadísticamente aceptable de que ningún componente del AIP se corrompe en el almacenamiento de archivos o durante cualquier transferencia de datos de almacenamiento de archivos interno. Esta función requiere que todo el hardware y el software dentro del archivo proporcionen una notificación de posibles errores y que estos errores se enruten a los registros de errores estándar que son verificados por el personal. La información de fijeza de PDI proporciona cierta seguridad de que la información de contenido no ha sido alterada a medida que se mueve y se accede a la AIP. Se necesita información similar para proteger la PDI. También se puede utilizar un mecanismo estándar para rastrear y verificar la validez de todos los Objetos de datos dentro del Archivo. Por ejemplo, se podrían mantener CRC para cada archivo de datos individual. También podría proporcionarse un nivel superior de servicio, como la codificación Reed-Solomon para admitir la detección y corrección de errores combinados. Los procedimientos de la instalación de almacenamiento deben prever la verificación aleatoria de la integridad de los objetos de datos utilizando CRC o algún otro mecanismo de verificación de errores.

Entidad Archival Storage

La función de **Recuperación ante Desastres** proporciona un mecanismo para duplicar los contenidos digitales de la colección del Archivo y, por ejemplo, almacenar el duplicado en una instalación físicamente separada. Esta función normalmente se logra copiando el contenido del Archivo en alguna forma de medio de almacenamiento extraíble (por ejemplo, cinta lineal digital, CD-ROM), pero también se puede realizar a través del transporte de hardware o transferencias de datos de red. Los detalles de las políticas de recuperación ante desastres los especifica la Administración.

Entidad Archival Storage

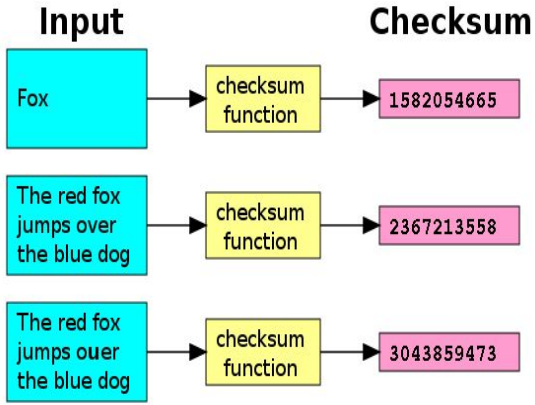
La función **Proporcionar datos** proporciona copias de los AIP almacenadas para ACCESS. Esta función recibe una solicitud de AIP que identifica las AIP solicitadas y las proporciona en el tipo de medio solicitado o las transfiere a un área de almacenamiento temporal. Esta función también envía un aviso de transferencia de datos a Access al completar un pedido.

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.

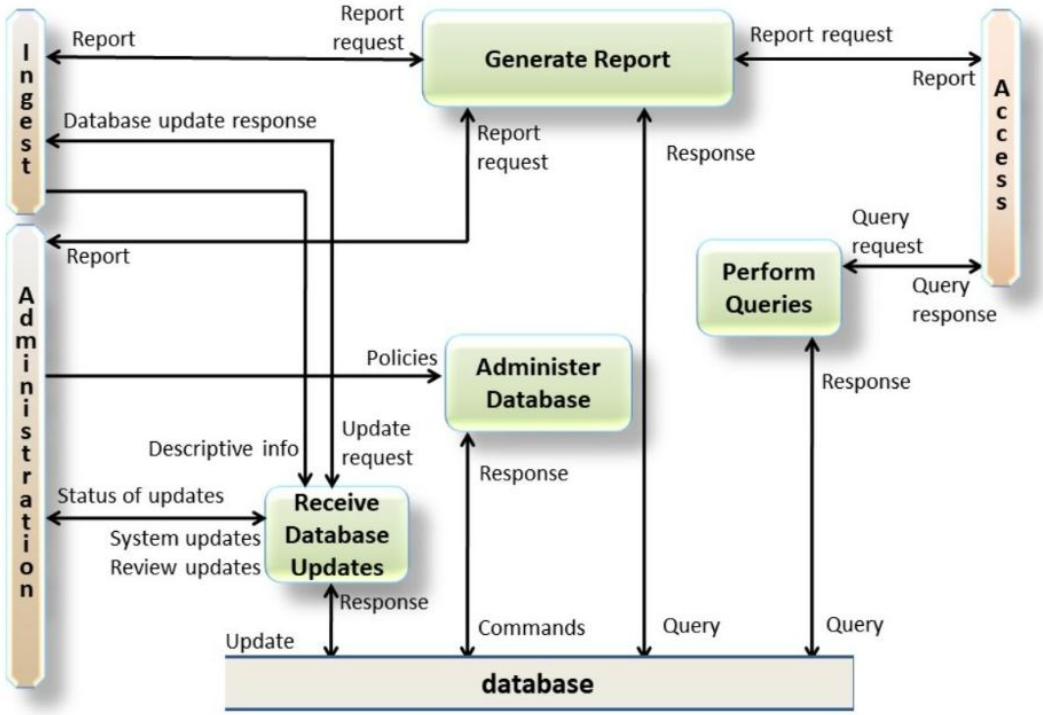
Entidad Archival Storage en DSpace

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)

Functions of 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.
- 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:

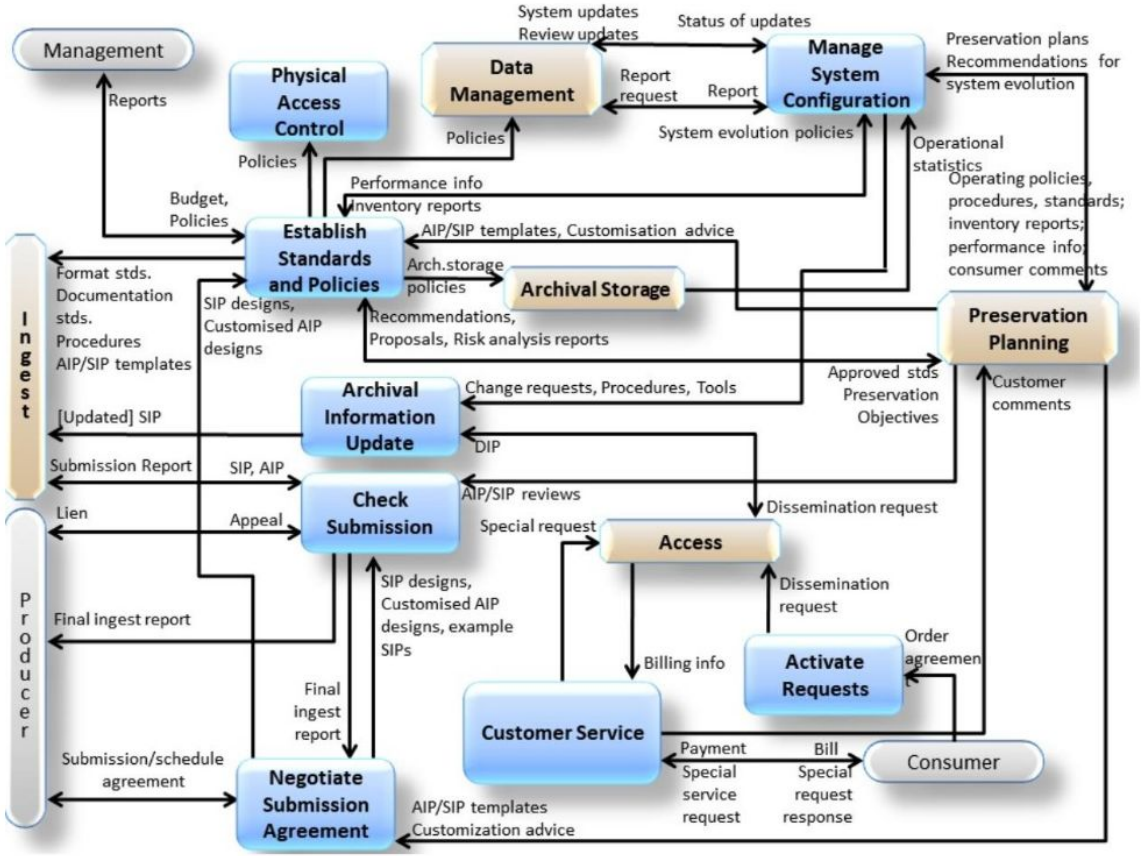
- de sólo lectura o modificación de elementos, para control y transformación respectivamente.
- ejecutadas automática o manualmente
- de ejecución única o periódica.
- aplicadas sobre una colección o sobre todo el repositorio

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

http://sedici.unlp.edu.ar/handle/10915/139884

Functions of Administration



Entidad OAIS Administration

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

Solicita la información necesaria sobre los archivos y negocia los acuerdos con los Productores.

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

Entidad OAIS administration

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

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

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 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

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.

Panel de control

Información Java | Configuración de Dspace | Alertas del Sistema | Recolectando | **Current Activity**

PARAR el registro de la actividad de usuarios anónimos.

EMPEZAR el registro de la actividad de bots.

Actividad actual (máximo 250 páginas)

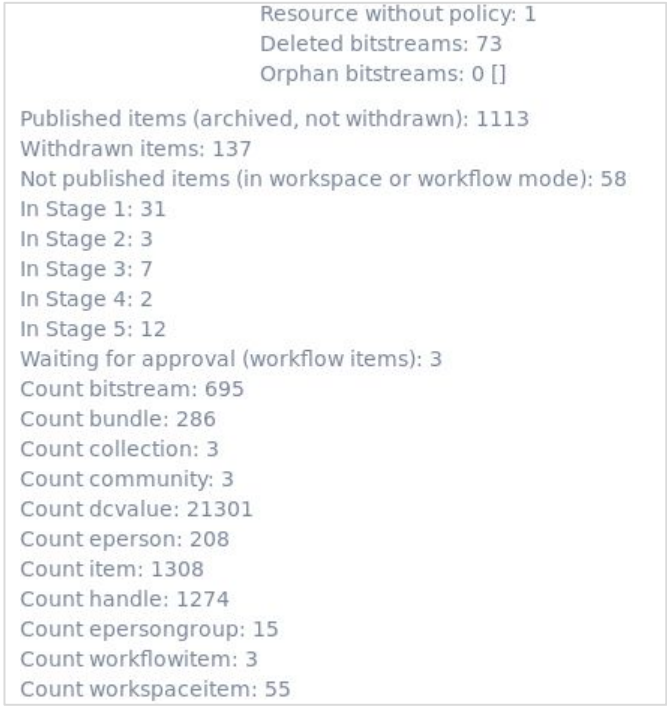
Marca horaria	Usuario	Dirección IP	Página URL	Navegador
0 s	Facundo Gabriel Adorno	163.10.34.221	/admin/panel	Firefox
2 s	Analía Pinto	163.10.34.197	/handle/10915/50/submit /continue	Chrome
5 s	Analía Pinto	163.10.34.197	/handle/10915/50/submit /continue	Chrome
7 s	Coordinación de egreso FBA UNLP	163.10.39.130	/handle/10915/50/submit /continue	Chrome
15 s	Analía Pinto	163.10.34.197	/handle/10915/61601	Chrome

Entidad Administration en DSpace: Reportes

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

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

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

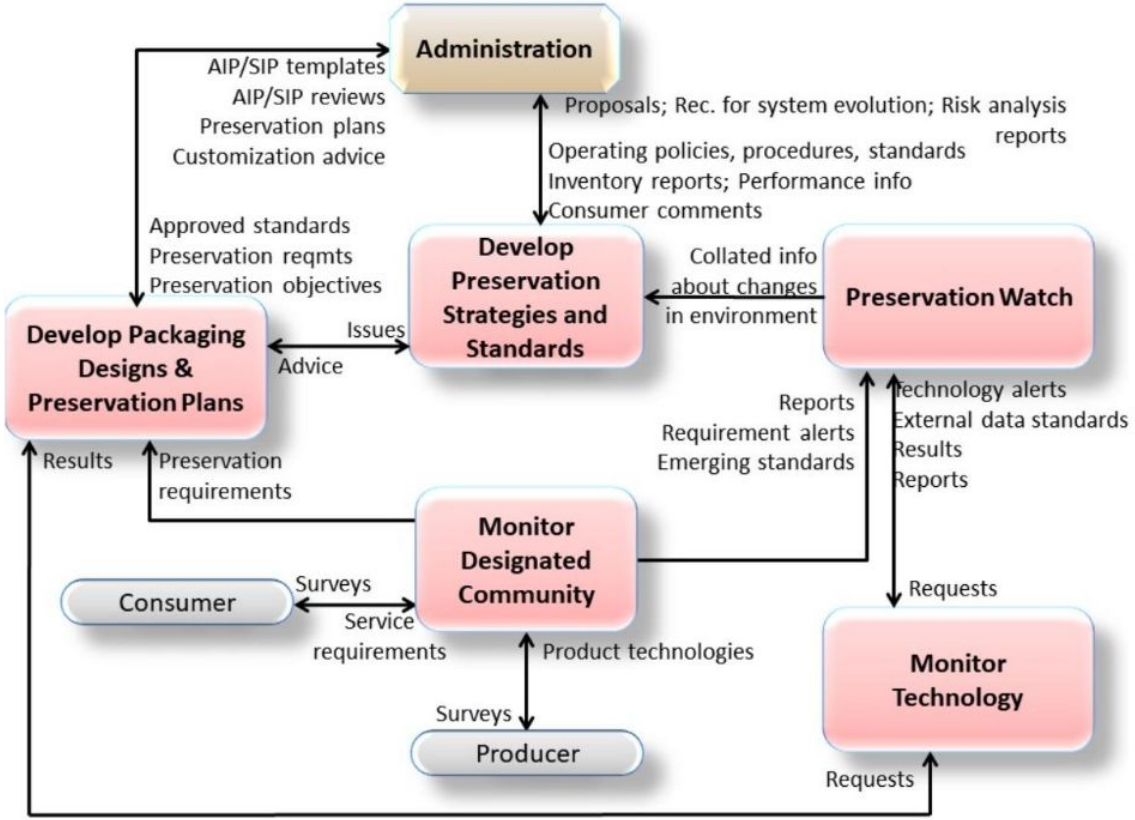


Entidad Administration en DSpace: Performance

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...)

Functions of 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.

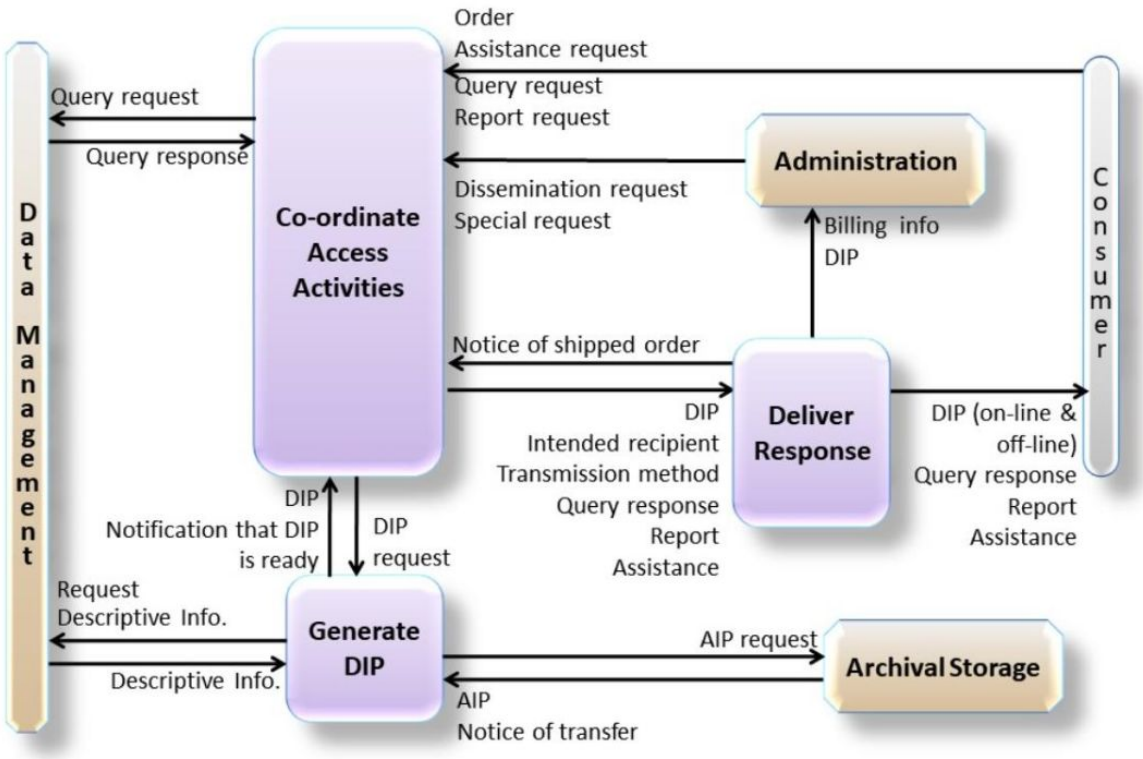
Entidad Preservation Planning en DSpace

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.

Functions of 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:

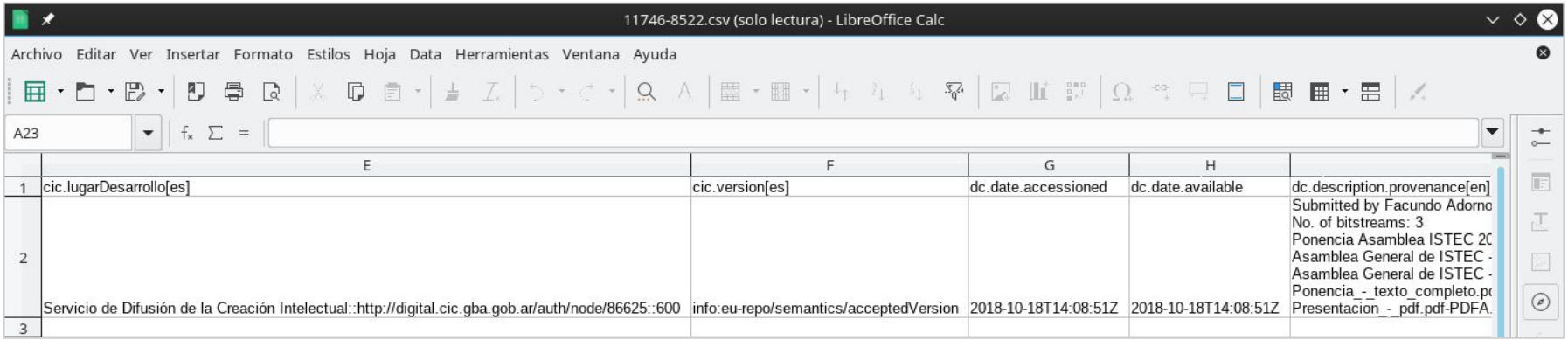
- dc.description.provenance presenta información sobre las acciones realizadas sobre el item para su depósito en el repositorio.
- dc.rights que contiene información de sobre la licencia de uso.
- handle identificador persistente del item
- checksum de bitstreams embebidos en el metadato provenance

Entidad Access en DSpace

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

Cada columna presenta el siguiente formato

schema.element.[qualifier].[content_language]

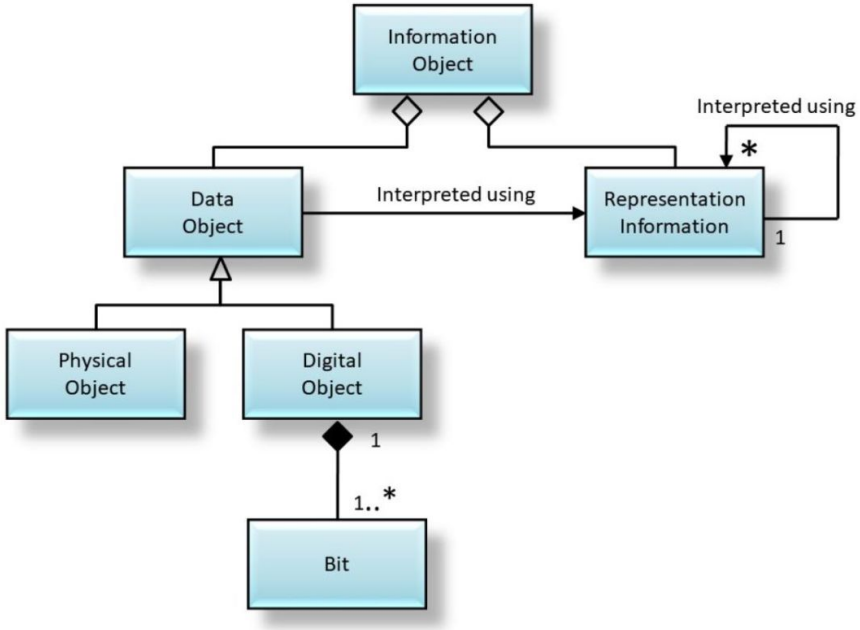


OAIS

Modelo de Información

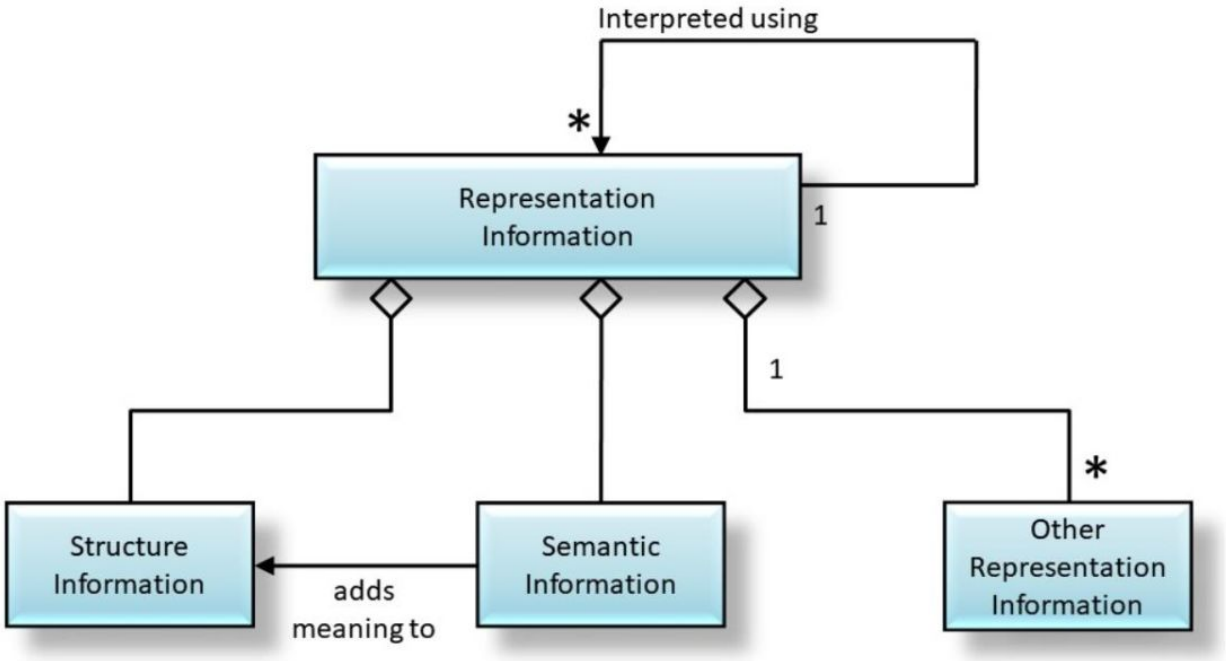
Sección 4.2

OAIS Objeto de información

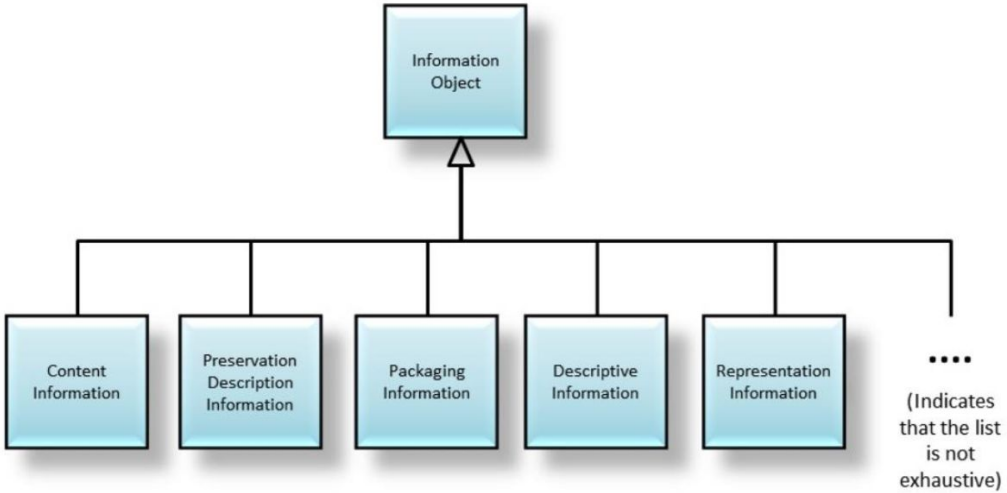


- El **Objeto de Información** está compuesto de un Objeto de Datos, que puede ser físico o digital, e Información de Representación que permite la interpretación completa de los datos.

Representation Information Object



Tipos de objetos de información

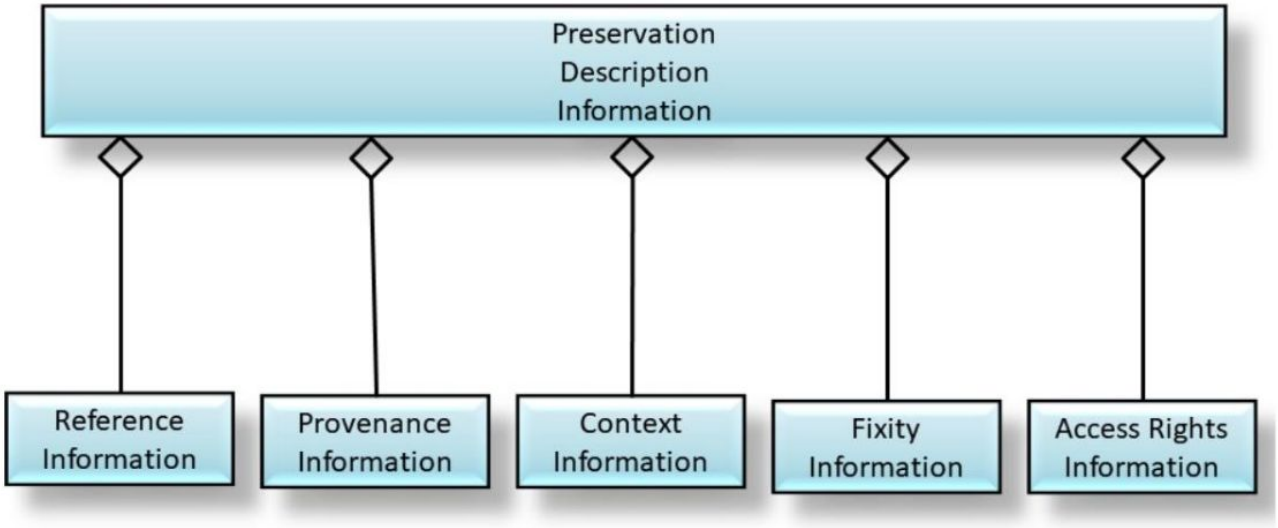


Los objetos de información se clasifican por su contenido y función como : objetos de información de contenido, de descripción de la preservación, de empaquetado y de información descriptiva.

Información de contenido

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

Información descriptiva de preservación (PDI)



Información descriptiva de preservación

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

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

Información descriptiva de preservación

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

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

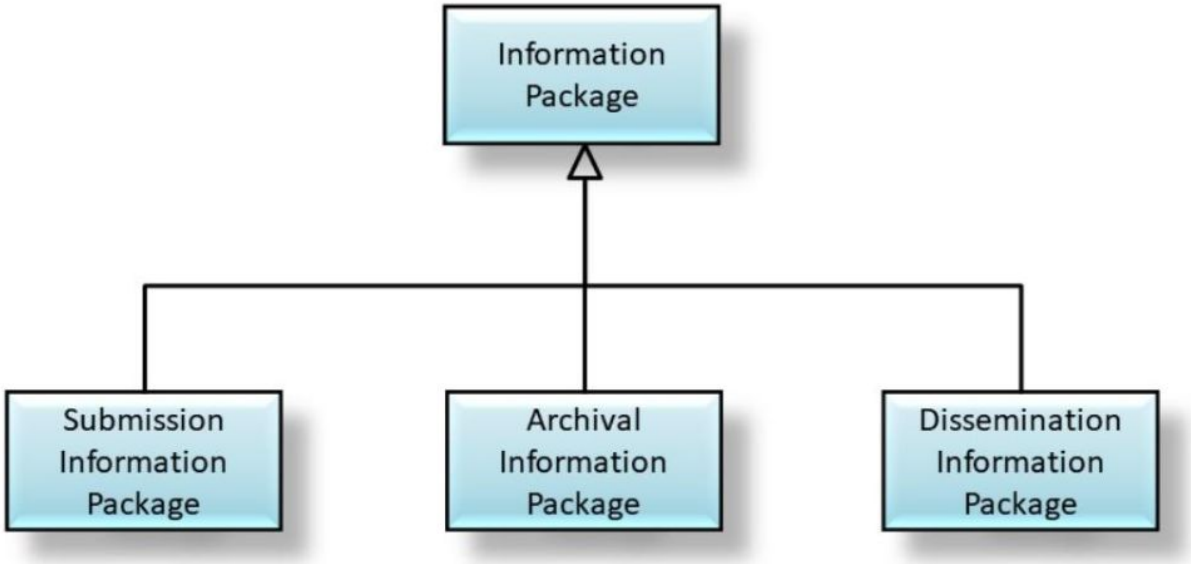
Información de sobre derechos de acceso: proporciona los permisos de uso de la información de contenido.

Paquetes de información en OAIS

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

Paquetes de información en OAIS

Tipos de paquetes de información



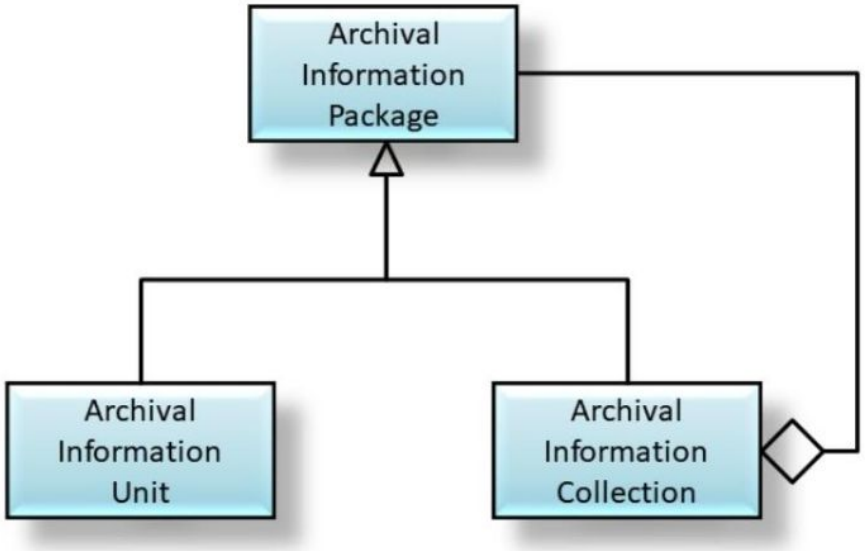
SIP

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

AIP

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

Tipos de AIPs



DIP

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

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

Preservación en el repositorio

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

Metadatos

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

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

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

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

Metadatos de preservación

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

Metadatos en Dspace

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

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

Referir al texto sobre qué metadatos agregar para conocer de dónde provienen los recursos que provienen de operaciones distintas: de procesos de digitalización, de ingesta masiva…

PREMIS

PREMIS es un grupo de trabajo internacional patrocinado por Online Computer Library Center (**OCLC**) y Research Libraries Group (**RLG**) que, como su nombre lo indica, se enfoca en estrategias de implementación de metadatos de preservación en Archivos Digitales.

En 2008, este grupo elaboró el Diccionario de Datos PREMIS para Metadatos de Preservación, el cual define los metadatos de preservación como "*la información que utiliza un repositorio para dar soporte al proceso de preservación digital".*

Diccionario de datos PREMIS

El diccionario define un conjunto de *unidades semánticas*, propiedades, e información que la mayoría de los repositorios necesita conocer de sus entidades para asegurar la preservación.

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

Modelo de Datos PREMIS

Las entidades que este modelo define se denominan:

- Entidades intelectuales
- Objetos
- Derechos
- Agentes
- Eventos

PREMIS

Entidad intelectual: conjunto coherente de contenido que se describe como una unidad: por ejemplo, un libro, un mapa, una fotografía, una publicación periódica,… etc. Una entidad intelectual puede incluir otras entidades intelectuales: por ejemplo, un sitio web, puede incluir una página web, una página web puede incluir una fotografía.
Una entidad intelectual puede tener una o más representaciones…

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

PREMIS

Evento: Acción que incluye al menos un Objeto Digital y/o un agente conocido en el repositorio de preservación
Agente: Actor (humano, máquina o software) asociado con uno o más eventos asociados a un objeto digital
Derechos: Afirmación de uno o más derechos o permisos que pertenecen a un objeto digital y/o a un agente

Entidad intelectual

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

Objetos

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

Eventos

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

Agentes

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

Derechos

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

Bibliografía: METS Y Metadatos Orientados A La Preservación Digital: PREMIS. Biblioteca Nacional de España:

http://www.bne.es/export/sites/BNWEB1/webdocs/Inicio/Perfiles/Bibliotecarios/bibliografia-oposiciones/21._METS_PREMIS.pdf

DIGITALIZACIÓN

Introducción a la Digitalización

La preservación digital se define como el conjunto de prácticas de naturaleza política, estratégica y acciones concretas, destinadas a asegurar la preservación, el acceso y la legibilidad de los objetos digitales a largo plazo.

Una estrategia de preservación es la de adoptar estándares y directrices internacionales, de modo tal que nos apoyemos en su relativa estabilidad en el tiempo. 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).

Introducción a la Digitalización

Según las guias FADGI, los formatos de archivos recomendados para preservación son: **TIFF, JPEG2000 y PDF/A.**

Dentro del repositorio SEDICI utilizamos el formato TIFF para el guardado de archivos "**maestros"** y PDF/A para archivos de "**preservación y difusión**".

PDF/A (Portable Document Format) es uno de los mejores formatos para preservar documentos electrónicos. Se utiliza la versión PDF/A (Archive) para archivar documentos con fines de preservación, pues contiene todos los elementos necesarios para reproducir el contenido tal como se generó, independientemente del programa con que se creó.

TIFF (Tagged Image File Format) es un formato de imágenes muy usado y de estándar abierto. Los archivos pueden utilizar compresión sin pérdidas y es utilizado para la creación de archivos maestros de imagen.

Circuito de digitalización en SEDICI

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:

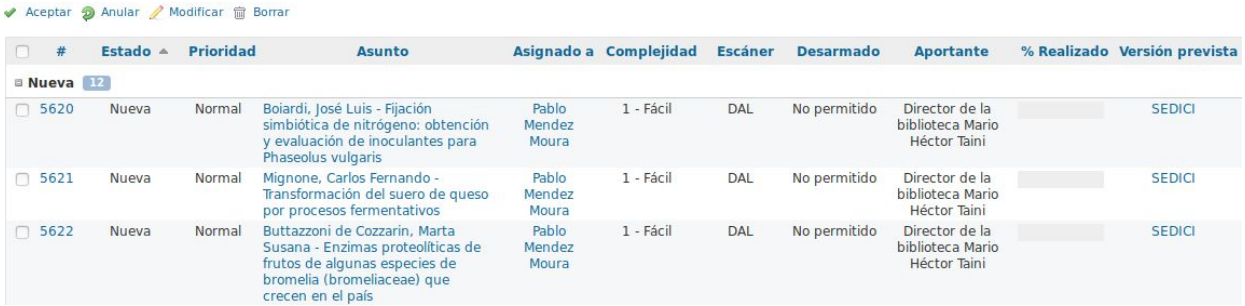
- Estado general de conservación
- Dimensiones
- Manifestaciones de la obra
- Tipos de encuadernación
- Importancia histórica, educativa, institucional: Dependiendo de la utilidad y el interés del material, los procesos de edición de imagen tienen mayor o menor automatización y revisión posterior. El 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)

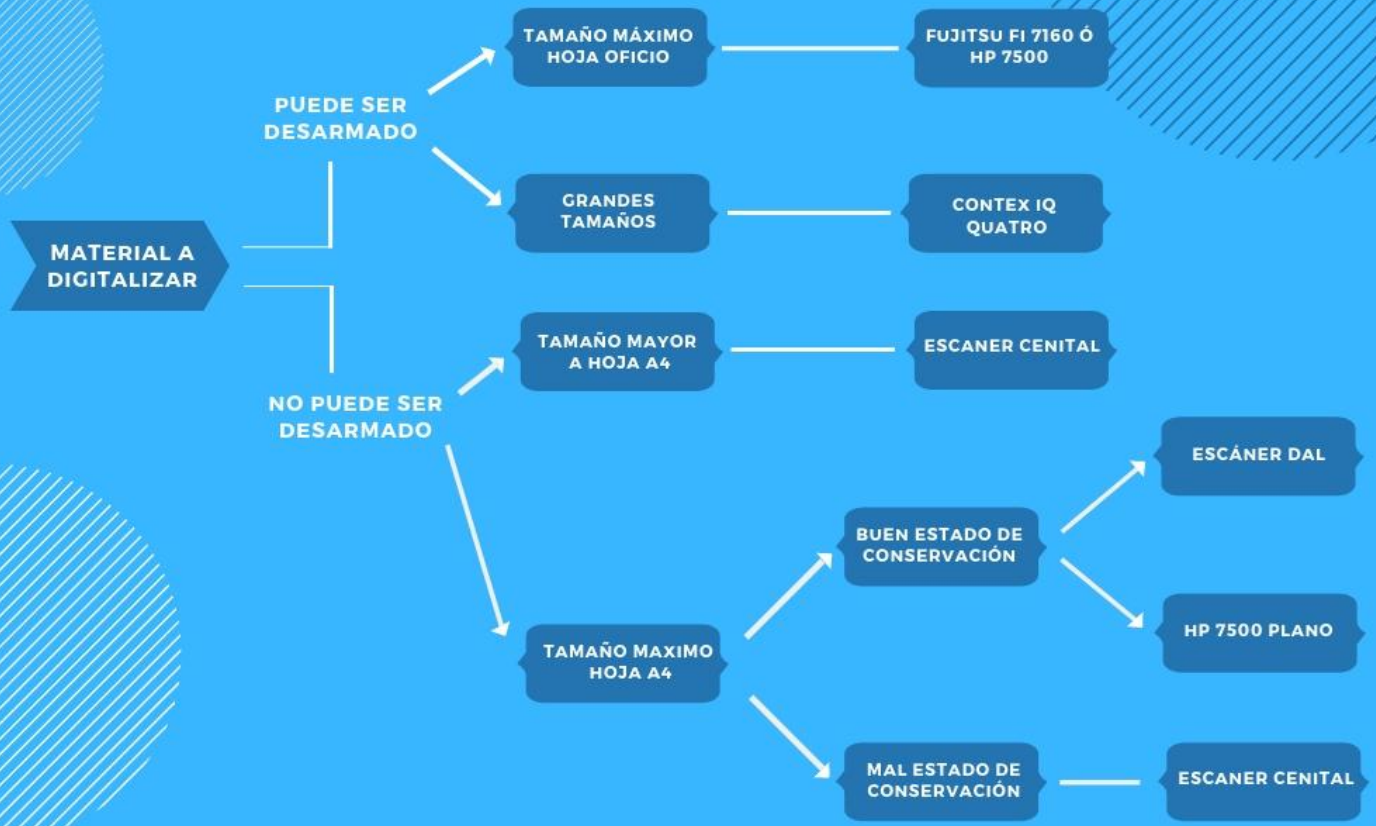
A medida que el trabajo pasa por sus distintas etapas, cada una de ellas queda asentada en el sistema hasta que el proceso finaliza.



#	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

FUJITSU FI7160

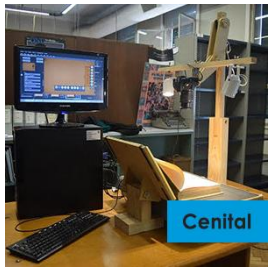
HP 7500

de gran formato

Contex IQ Quattro

de libros

Archivista



Cenital

Escáneres automáticos

Este tipo de escáner cuenta con alimentación automática, lo que permite un mayor flujo de material y una mayor velocidad de procesamiento. Además del alimentador automático el modelo HP 7500 trae una cama plana para digitalizar hojas sueltas, que por distintas razones (por ejemplo, friabilidad del papel), no pueden ser procesadas a través del alimentador automático.

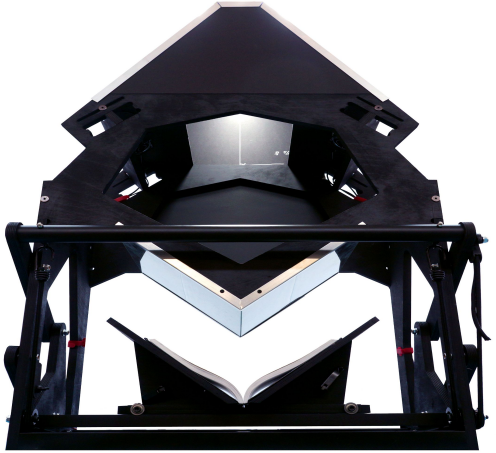
Escáner de gran formato

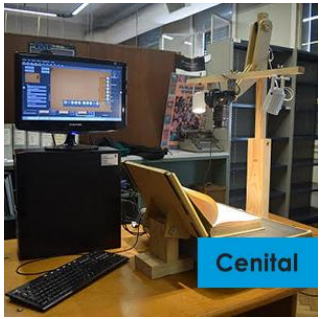
Permite digitalizar mapas, planos, dibujos arquitectónicos, posters, y otros tipos de obras de hasta 44 pulgadas. Por sus características (se trata de un escáner de rodillo compuesto de varios escáneres concatenados), es necesario realizar una calibración periódica y utilizar una protección de mylar para la captura del documento.

Escáneres de libros

- **Archivista 2014**
  Este escáner fue fabricado íntegramente en SEDICI bajo las pautas propuestas por http://diybookscanner.org. Cuenta con dos cámaras Nikon reflex D5300 y es controlado por el software gratuito y de código abierto DigiCamControl http://digicamcontrol.com/
- **Cenital**
  Cuenta con una cámara Nikon D5600 y permite digitalizar materiales con dificultades en la manipulación y encuadernaciones frágiles.

Ambos cuentan con lámparas de luz cálida cuyo CRI es de 90.

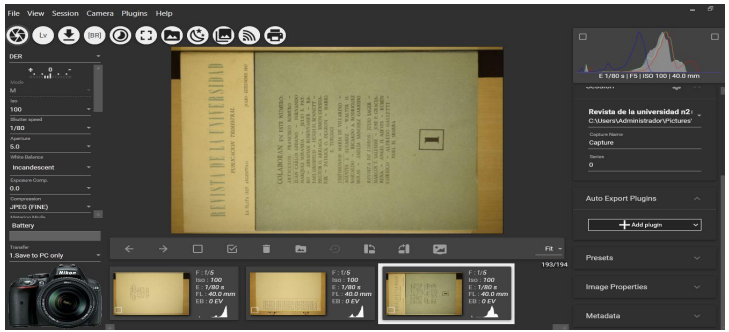




4) Captura de imágenes

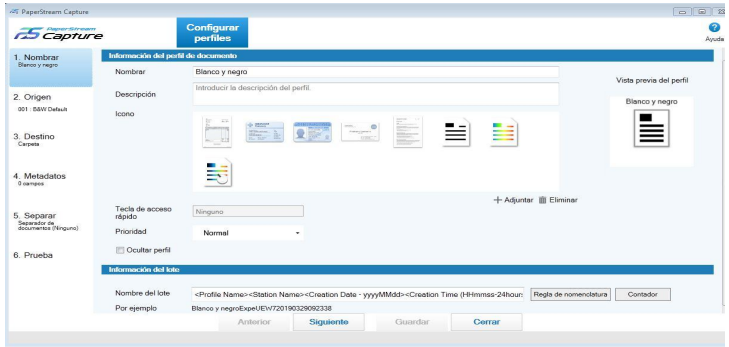
Captura con **“digiCamControl”**

Este software permite la configuración y control completo de las cámaras que se utilizan tanto en el escáner Archivista como en el cenital.



Captura con “**Paperstream**”

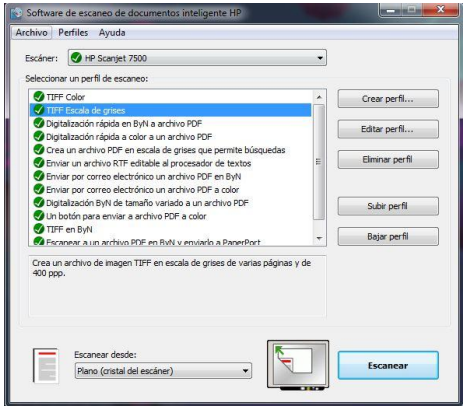
Este es el software utilizado para la captura con el escáner Fujitsu FI 7160.



4) Captura de imágenes

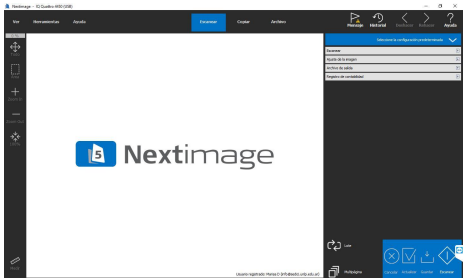
Captura con **“Software de escaneo de documentos inteligente” de HP**

Este programa es utilizado para controlar el escáner HP 7500



Captura con **“NextImage”**

Este software es utilizado para controlar el escáner de formato grande Contex IQ Quattro.

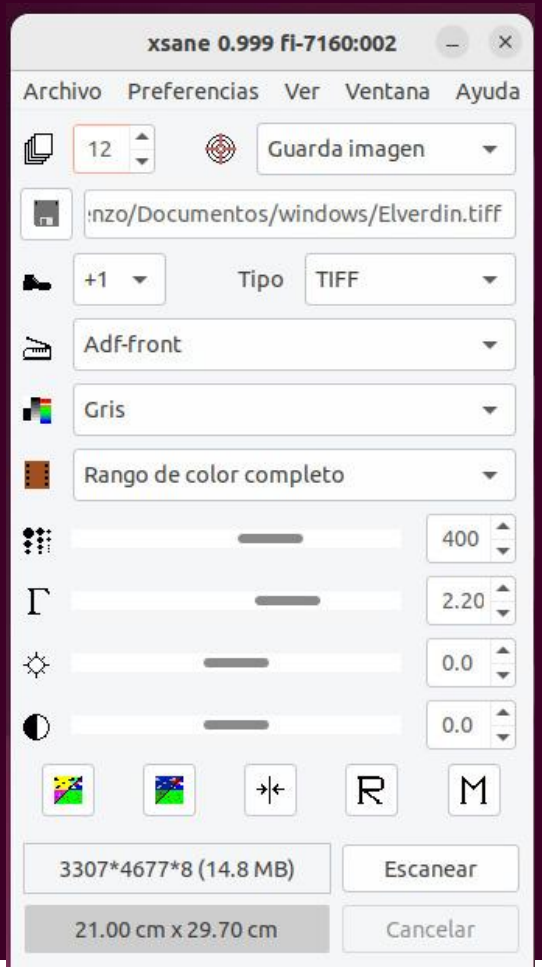


4) Captura de imágenes

Captura con **“xsane (Scanner Access Now Easy)”**

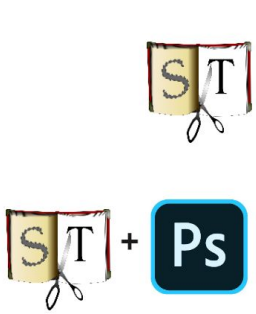
Este programa permite controlar los escáneres que cuentan con drivers compatibles con Linux.

Permite el control de las variables de captura (resolución, gamma, brillo y contraste), la selección del formato de salida y hasta ofrece la posibilidad de realizar un OCR de lo capturado.



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:

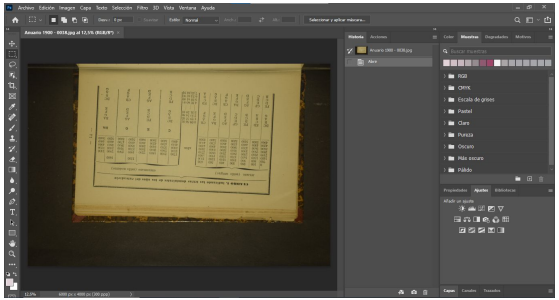


1. Rotar páginas
2. Enderezar las imágenes
3. Ajustar márgenes
4. Eliminar manchas, puntos indeseados
5. Normalizar color
6. Mejorar contraste entre texto y fondo

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

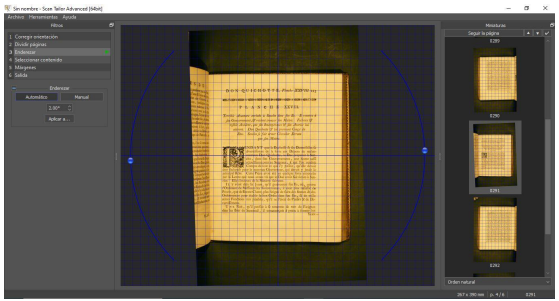
Edición con Photoshop

Este programa es uno de los más potentes del mercado para la edición de imágenes, y es utilizado en casos que presentan muchas dificultades en la visualización o legibilidad.



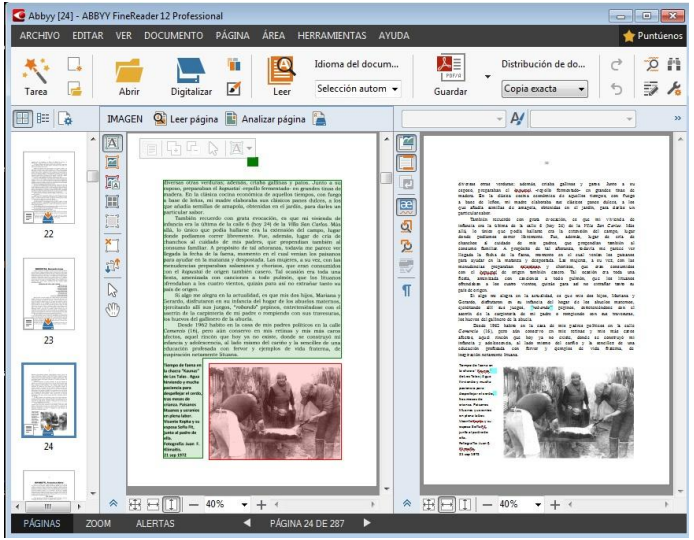
Edición con ScanTailor Advanced

Scantailor es una herramienta gratuita de código abierto que permite corregir o modificar las imágenes capturadas. Soporta los siguientes formatos de entrada: *.tif, *.tiff, *.png, *.jpg, *.jpeg y genera archivos con formato tiff de salida (uno por cada página).



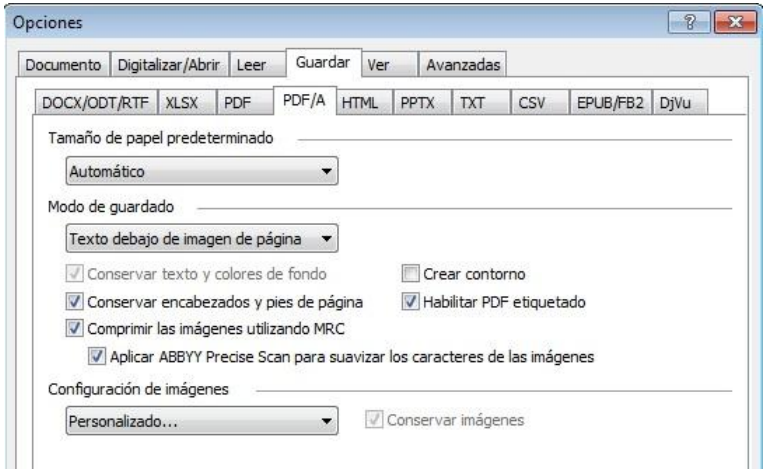
6) OCR y compilación en PDF/A: ABBYYFineReader 15

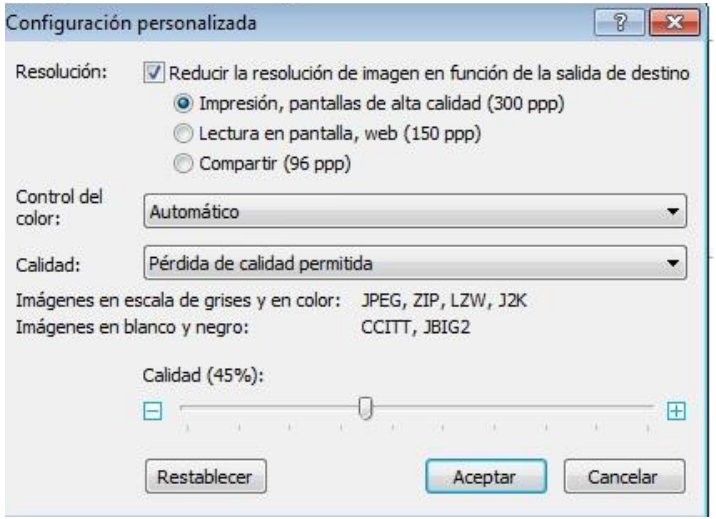
Luego de editar las imágenes se realiza el OCR con Abbyy FineReader 15. En esta etapa del proceso se selecciona el contenido según sea texto, imagen o cuadro. Luego se revisa el resultado del OCR y se generan los archivos PDF/A.



Compresión de pdf

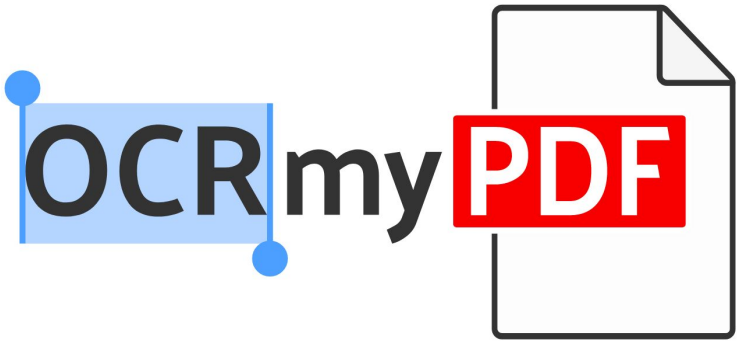
Por último, en el momento del guardado, el programa nos permite modificar la compresión para obtener documentos más pequeños, que pueden ir desde compresiones sin pérdida a compresiones con pérdida de calidad.





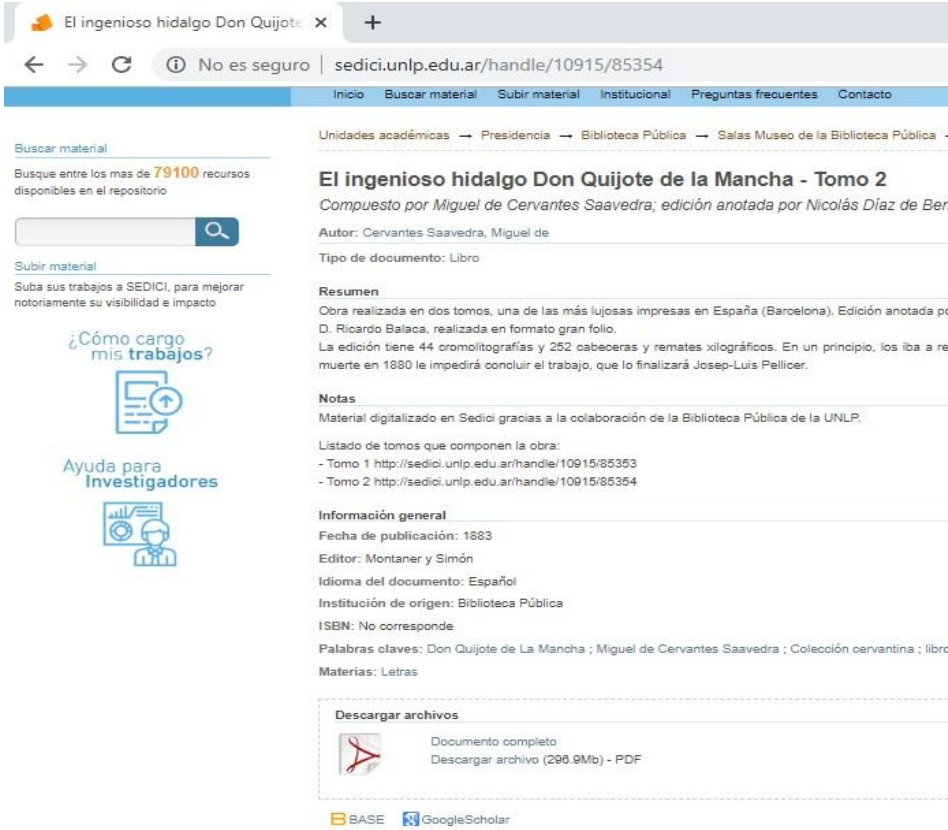
6) OCR y compilación en PDF/A: ocrmypdf

- OCRmyPDF es un software libre desarrollado por James R. Barlow.
- Se utiliza mediante línea de comandos.
- Combina otros software libres: unpaper (edición de imagen), tesseract (OCR) y Ghostscript (manipulación de PDF).
- Los parámetros mínimos son el archivo de entrada y de salida: ocrmypdf 'input_pdf_or_image' 'output_pdf'
- Puede también cambiarse el idioma de OCR (-l).
- https://ocrmypdf.readthedocs.io



6) Guardado de archivos para preservación y difusión

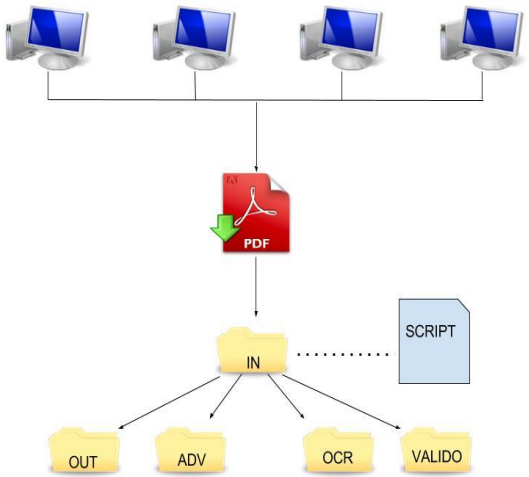
- Se generan dos archivos PDF/A de cada obra: uno de alta calidad de imagen, que se destina a la preservación digital y otro comprimido que se utiliza para difusión en el repositorio.
- Para el caso de algunos libros editados por la UNLP, también se crean libros electrónicos en formato .epub y .mobi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



6) Conversión por lote: 3-HEIGHT

Este software posee una arquitectura cliente servidor, que permite convertir por lotes archivos de distintos formatos a PDF/A. Además también es utilizado para verificar si los archivos PDF/A cumplen con la norma.

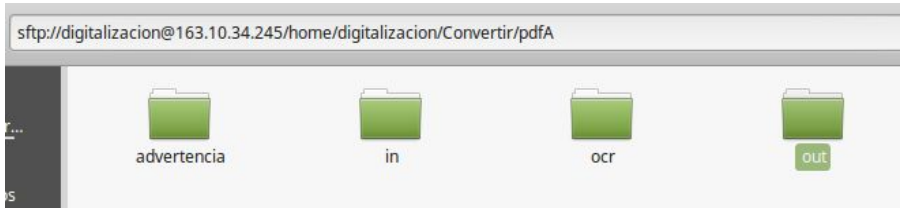
- Deteccion de archivos
- Análisis
- Conversión
- Verificación



6) Conversión por lote: 3-HEIGHT

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

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



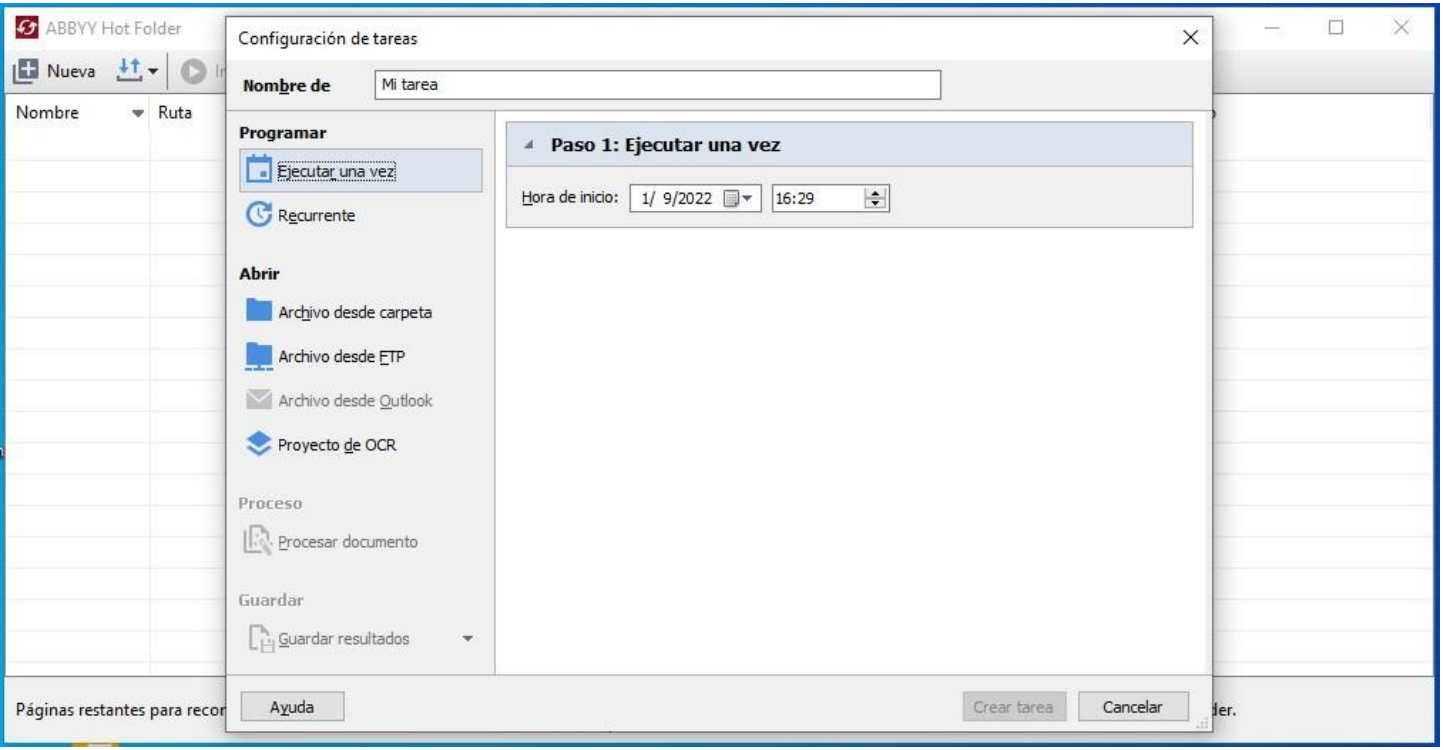
6) Conversión por lote: 3-HEIGHT

3-HEIGHT analiza el PDF y elige en qué versión va a convertirlo. Si la conversión sale bien, en la carpeta out tendremos los siguientes archivos:

El archivo con la fecha 13-20-16.txt presenta el log de la ejecución del script..
El archivo pdf original.
El archivo convertido con la terminación: -PDFA.pdf
El último archivo txt da más detalles de la conversión del archivo original

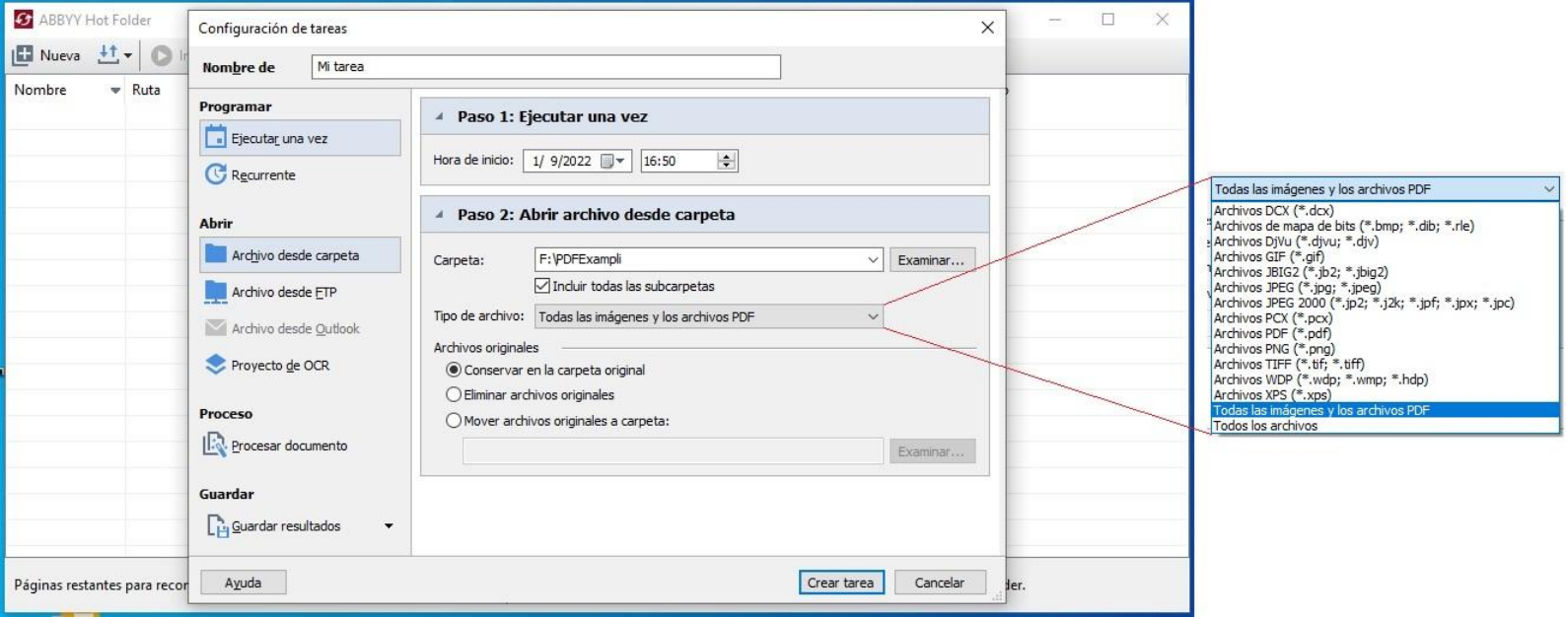
6) Conversión por lote: AbbyHotFolder15

Para crear y convertir en lote los PDF/A, seleccionamos en el inicio de Windows la herramienta Hot Folder de ABBY 15.

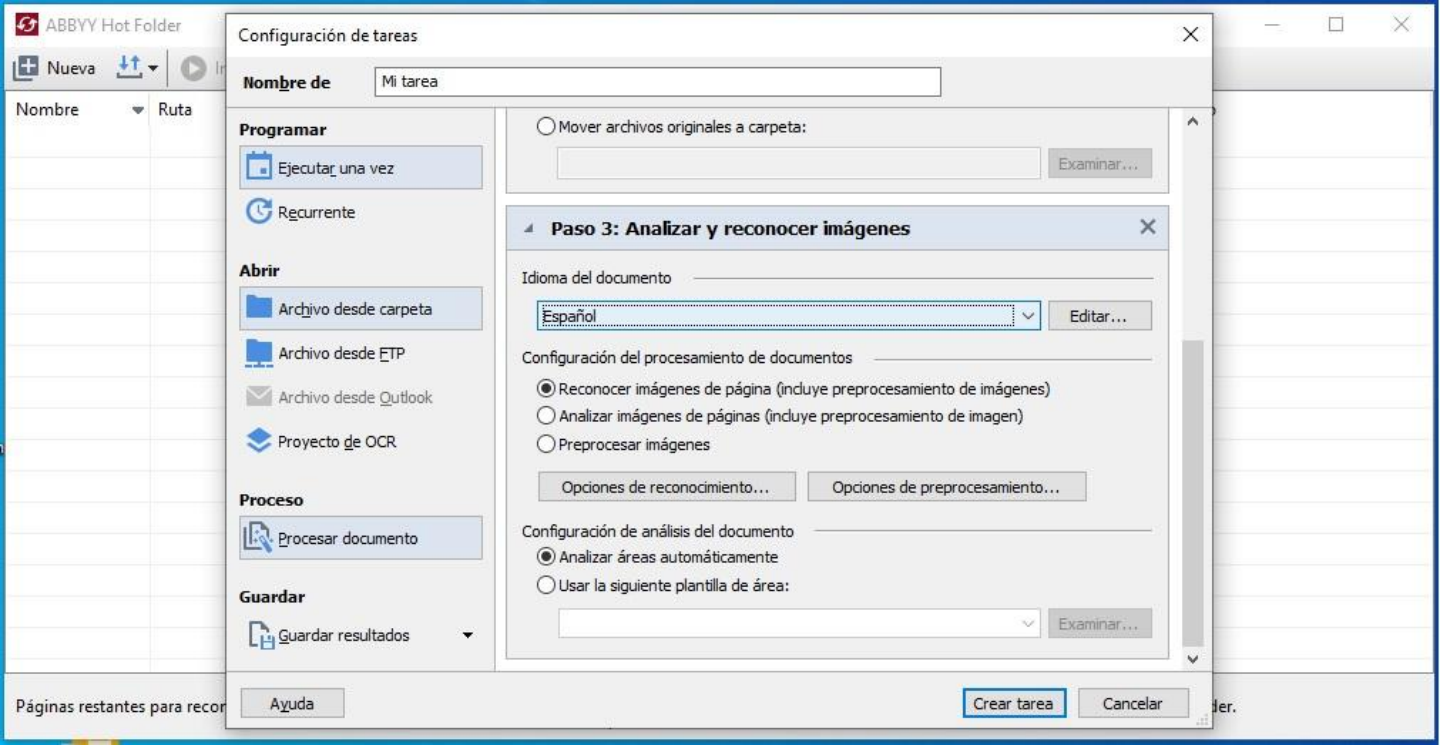


Paso 1:

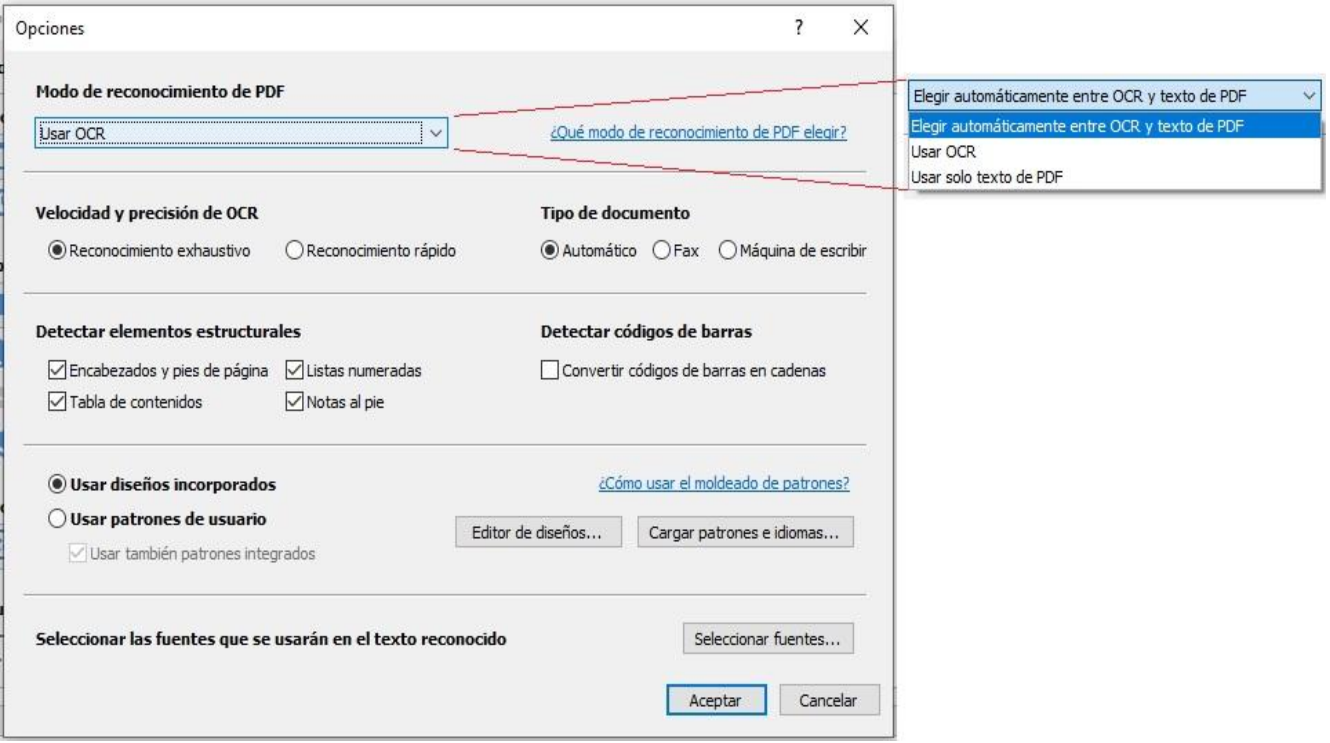
Programamos la tarea…



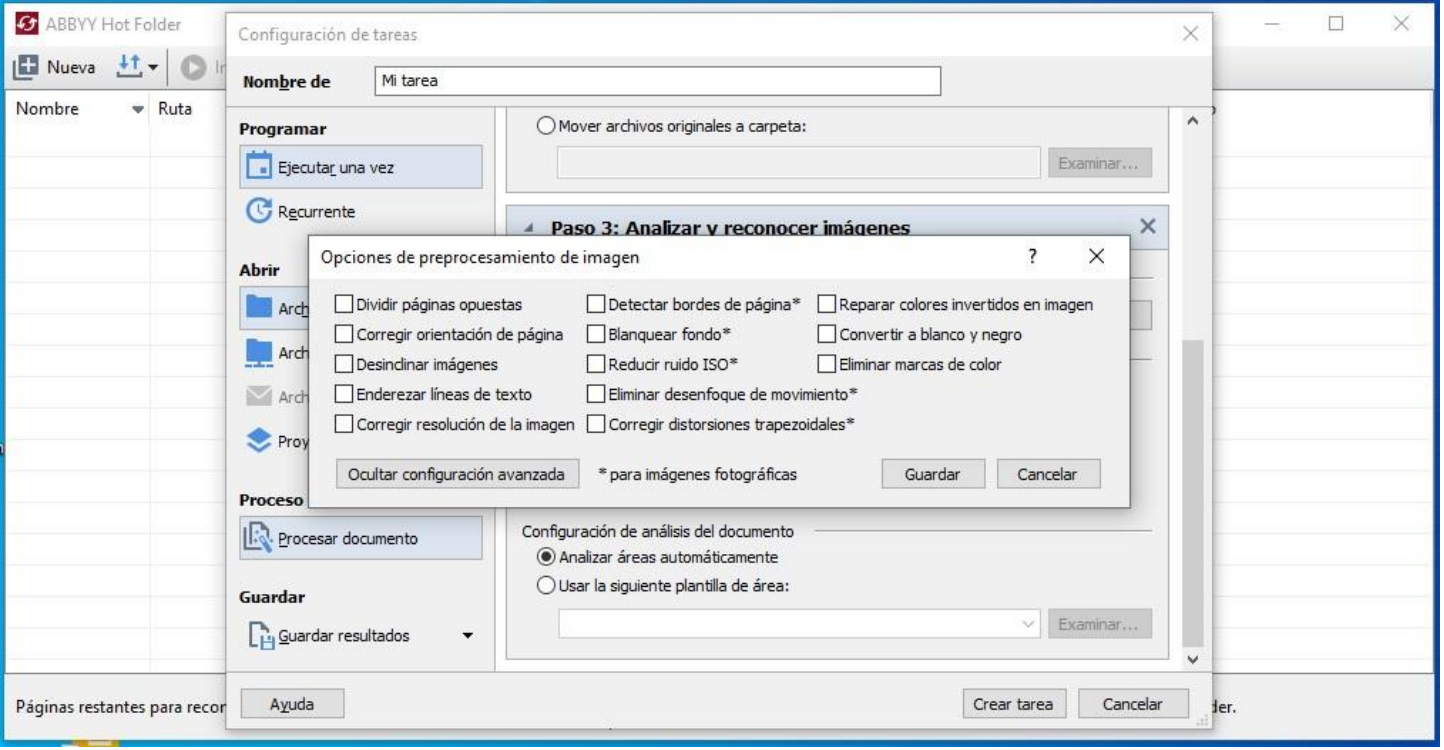
Paso 2: Definimos el origen…



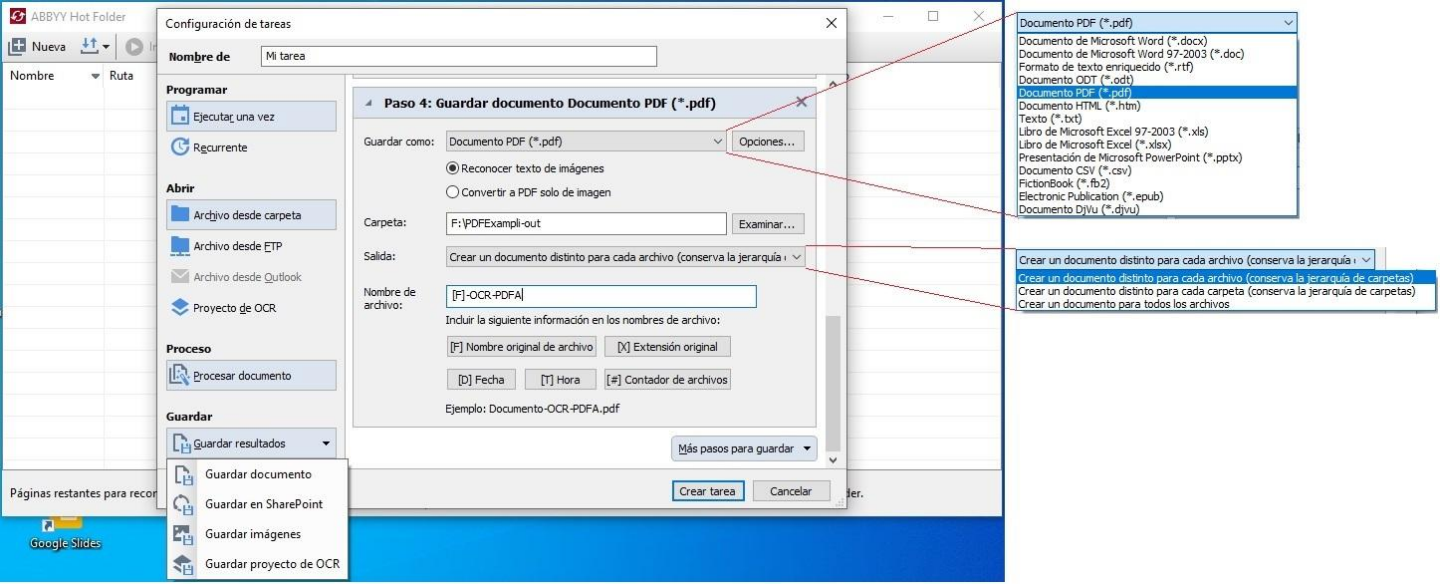
Paso 3: Definimos el proceso…



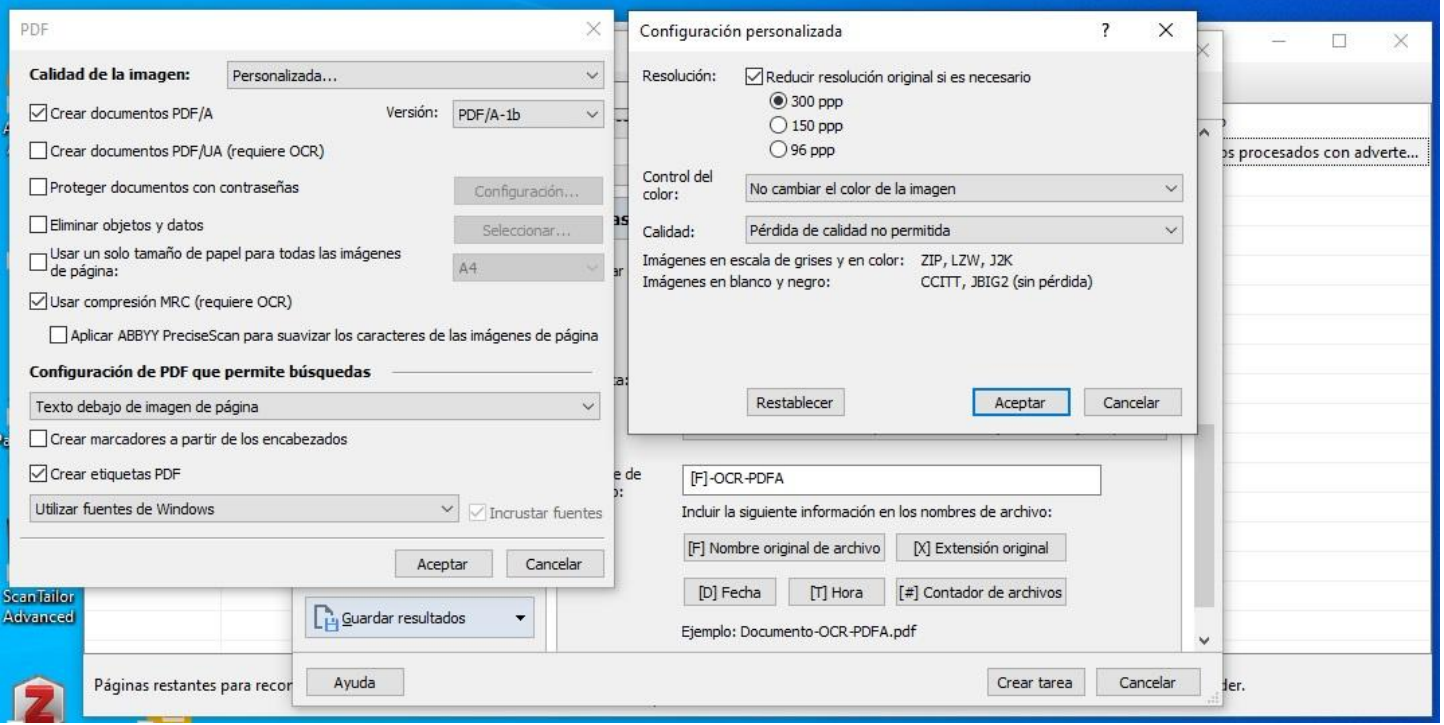
Paso 3: Opciones de OCR



Paso 3: Opciones de preprocesamiento

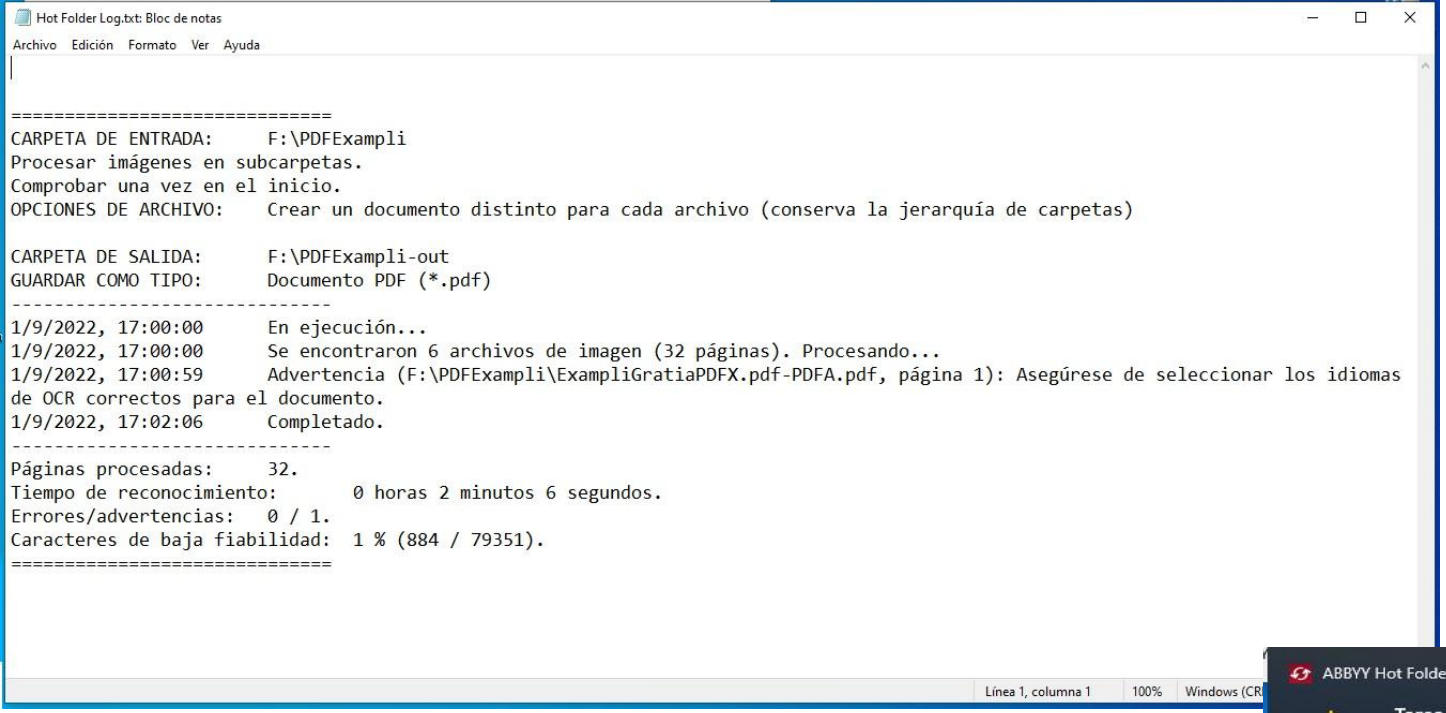


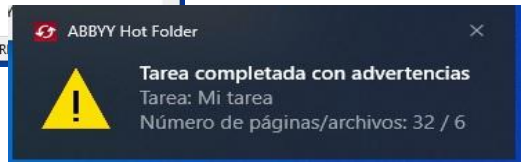
Paso 4: Definimos la salida…



Paso 4:

Opciones de salida: Calidad de la imagen y estándares (PDF/A-1;2 y 3)



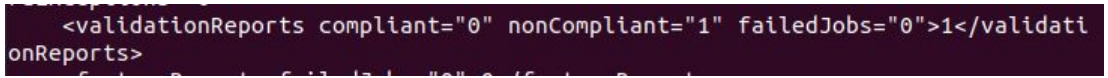


Tarea completada e informe (.txt, en carpeta de salida)

6) Validación de PDF/A

Una vez obtenido el fichero PDF, es necesario validarlo para comprobar si efectivamente cumple con el estándar PDF/A, puesto que puede haberse guardado en el mismo el metadato técnico que permite a los lectores reconocerlo como tal, pero aún así persistir algunos problemas tales como glifos no definidos, espacios de color dependientes de dispositivos, interpolaciones de píxeles en la imagen que ocasionan errores.

Contamos con dos programas de validación de PDF/A: AcrobatDC y VeraPDF.

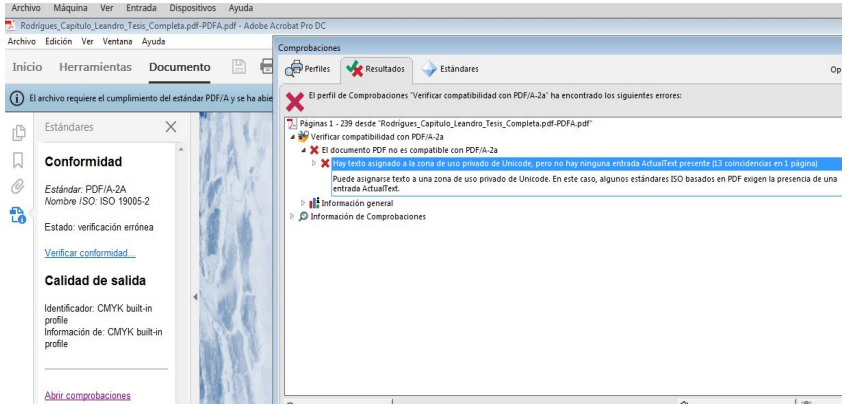


Ver: http://sedici.unlp.edu.ar/handle/10915/139212

6) Validación de PDF/A: Acrobat DC

Acrobat DC cuenta con un validador de estándares de PDF. Si la validación es errónea, podemos arreglarlo desde el mismo Acrobat: el propio programa analiza y aplica los cambios necesarios para convertirlo correctamente al estándar PDF/A que seleccionemos (aplica espacios de color, incrusta fuentes, elimina caracteres no definidos, entre otros). Es conveniente, sin embargo, reconocer el error desde el validador y volver al proyecto OCR original para resolverlo.

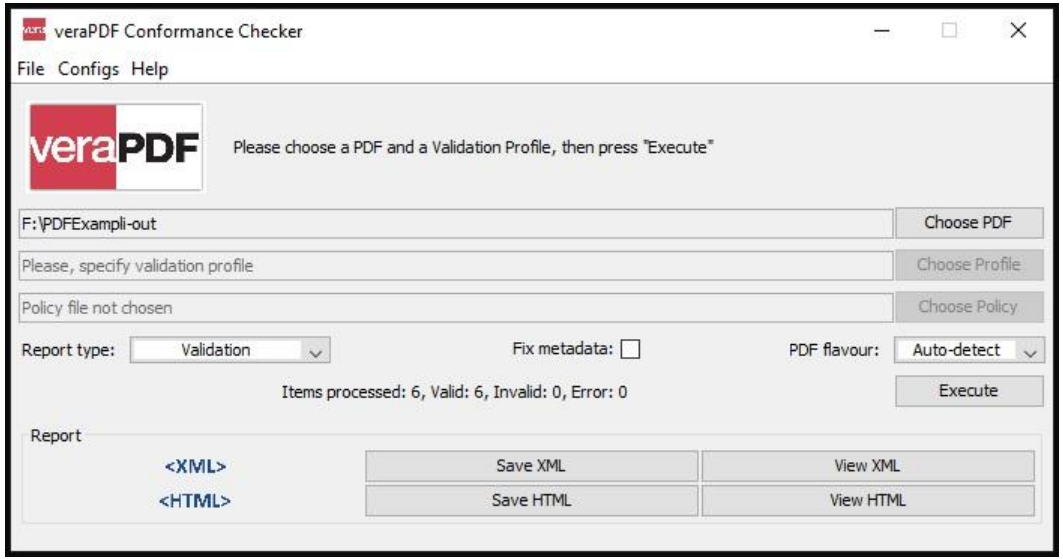
Adobe Acrobat DC



6) Validación de PDF/A: VeraPDF

- VeraPDF es un software libre, desarrollado por Open Preservation Foundation, que contiene todos los estándares de PDF/A y sus niveles de conformidad necesarios para la preservación.
- El programa analiza y produce un informe de validación que muestra qué estándares se verifican y si los documentos PDF seleccionados los cumplen.
- En forma más detallada es posible ver sus características (metadatos, fuentes incrustadas, espacios de color, etc.).
- Se puede usar por medio de la interfaz como se muestra a continuación o por líneas de comando de la forma: **verapdf "ruta del archivo.pdf"**
- La opción **-x** nos muestra las características del PDF.
- Y una forma de guardar el informe generado en .xml es agregando: **> "ruta del archivo\nombre del archivo.xml"**
- A modo de ejemplo: **verapdf -x "ruta del archivo.pdf" > "ruta del archivo\nombre del archivo.xml"**

6) Validación de PDF/A: VeraPDF (GUI)



6) Validación de PDF/A: VeraPDF

veraPDF

Validation Report

Build Information

Version:	1.20.3
Parser:	PDFBox
Build Date:	2022-05-19T09:10:00-03:00

Batch Summary

Processing time	Total Jobs	Failed to Parse	Encrypted	Compliant	Not Compliant
00:00:04.612	6	0	0	6	0

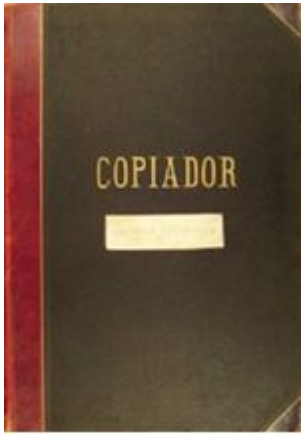
6) Validación de PDF/A: VeraPDF

Job Summary

File Name	Validation Profile	Compliance	Passed Rules	Failed Rules	Passed Checks	Failed Checks	Duration
F:\PDFExampli-out\ExampliGratiaPDF-OCR-PDFA.pdf	PDF/A-1B validation profile	Passed	101	0	63768	0	00:00:02.305
F:\PDFExampli-out\ExampliGratiaPDFX.pdf-PDFA-OCR-PDFA.pdf	PDF/A-1B validation profile	Passed	101	0	246905	0	00:00:01.715
F:\PDFExampli-out\ExampliGratiaTIFF-001-OCR-PDFA.pdf	PDF/A-1B validation profile	Passed	101	0	3229	0	00:00:00.040
F:\PDFExampli-out\ExampliGratiaTIFF-002-OCR-PDFA.pdf	PDF/A-1B validation profile	Passed	101	0	9858	0	00:00:00.061
F:\PDFExampli-out\ExampliGratiaTIFF-003-OCR-PDFA.pdf	PDF/A-1B validation profile	Passed	101	0	8176	0	00:00:00.053
F:\PDFExampli-out\ExampliGratiaTIFF-004-OCR-PDFA.pdf	PDF/A-1B validation profile	Passed	101	0	7384	0	00:00:00.050

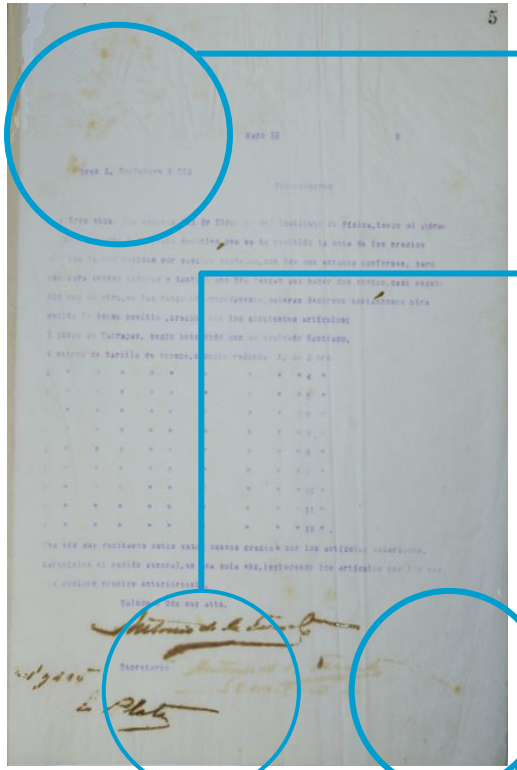
Ejemplo de caso de proceso completo de digitalización:

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



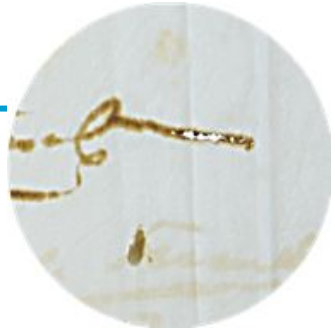
SEDICI y el **Museo de Física de la Facultad de Ciencias Exactas** de la **UNLP** destinaron personal para la digitalización de un documento archivístico: el libro *Copiador – Facultad de Ciencias Físicas, Matemáticas y Astronómicas (1918-1925).* Se siguieron los estándares internacionales para la digitalización (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

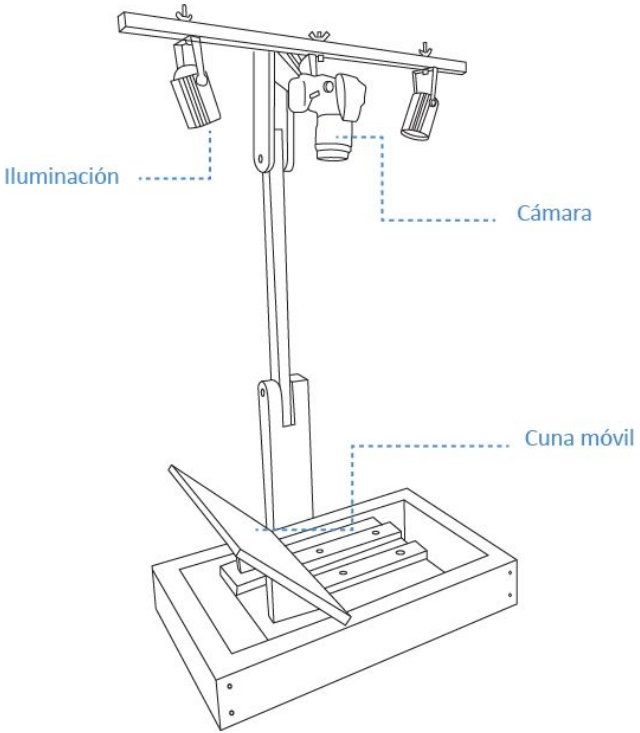
Dobleces y desprendimientos



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

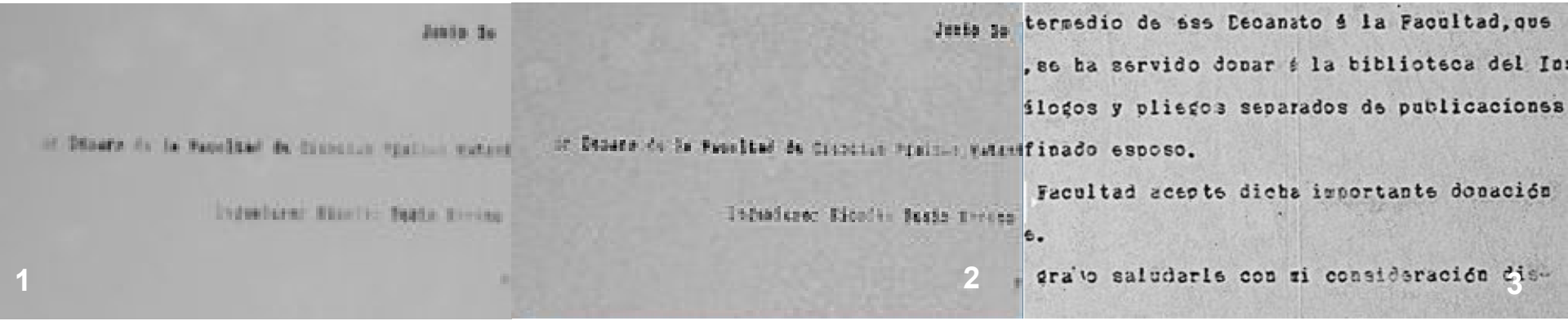
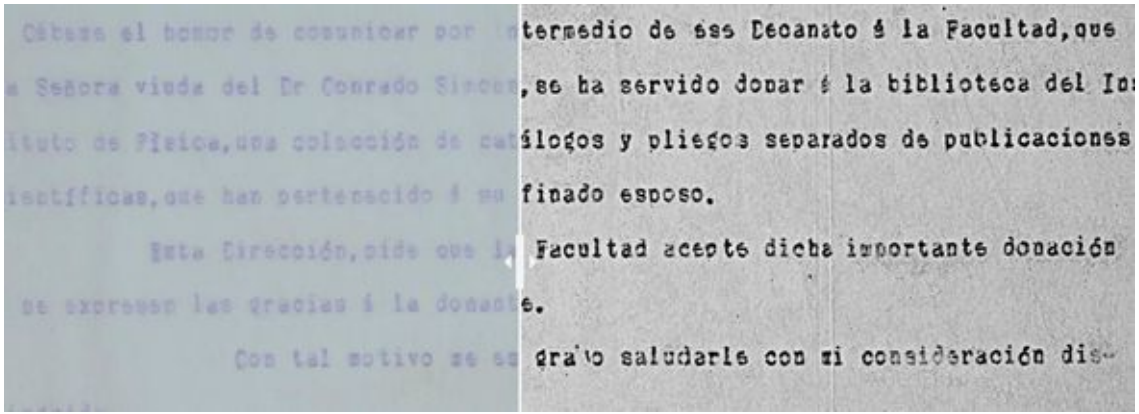


Imagen original e imagen mejorada lista para reconocimiento de texto (abbyy FineReader)



Advertencia: la experiencia deja en claro que la preservación es cosa compleja. Depende de cuestiones técnicas, depende de tener infraestructura y recursos humanos formados, pero también depende de cómo hacemos lo que hacemos.

CONFERENCIA INTERNACIONAL

BIREDIAL-ISTEC

18-20 de octubre de 2023 MONTEVIDEO • URUGUAY

'23

Nuestros sitios

http://sedici.unlp.edu.ar

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

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

http://prebi.unlp.edu.ar

Este material está disponible en la colección de **SEDICI** http://-

Marisa R. De Giusti marisa.degiusti@sedici.unlp.edu.ar

PREBI prebi.unlp.edu.ar SEDICI sedici.unlp.edu.ar

UNIVERSIDAD NACIONAL DE LA PLATA

CESGI Centro de Servicios en Gestión de Información cesgi.cic.gba.gob.ar

COMISIÓN DE INVESTIGACIONES CIENTÍFICAS