20 años 1999-2019

FACULTAD DE INFORMATICA

UNIVERSIDAD NACIONAL DE LA PLATA

Bibliotecas y Repositorios Digitales. Tecnología y aplicaciones 2022

Prof. a cargo: Dra. Marisa R. De Giusti

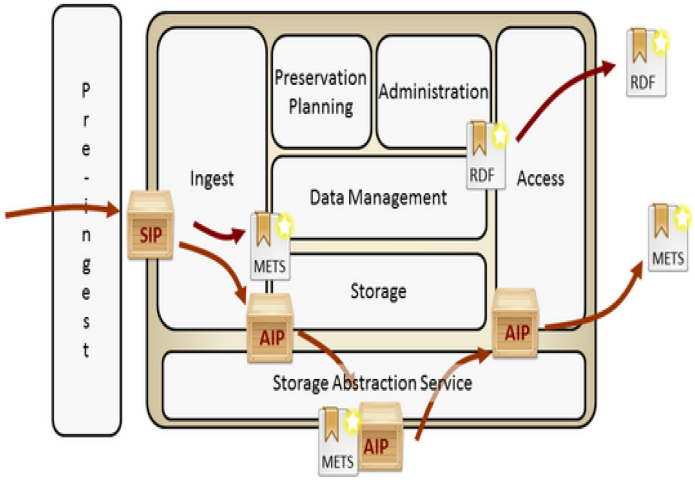
PREBI prebi.unlp.edu.ar SEDICI sedici.unlp.edu.ar CIC DIGITAL

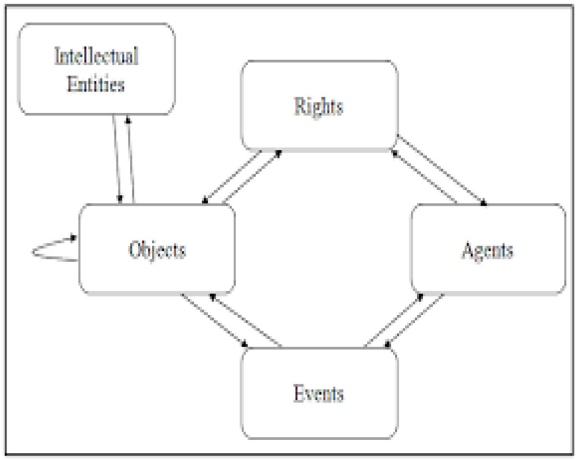
Repositorio Institucional
Comisión de Investigaciones Científicas

Este material ha sido elaborado en conjunto con el personal del repositorio SEDICI y del repositorio CIC Digital

Clase5

Preservación y digitalización de documentos





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.

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, la negligencia benigna no sirve: un disco olvidado 5 años… no sirve.

- 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

1. Archivar los documentos digitales

   gestión documental

2. Preservar el *bitstream*

3. Garantizar el acceso a largo plazo

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.

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

¿Qué materiales se tienen que preservar a largo plazo?

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

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

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.

Un avance: estándares

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

Problemas en la preservación: software

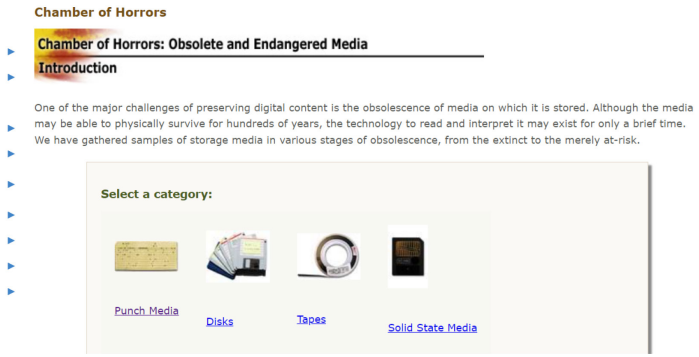
- Muchos problemas en lo relativo a la preservación derivan de una configuración deficiente del software que soporta el repositorio. Es necesario revisar las facilidades del software que soporta el repositorio en comparación con el modelo de preservación OAIS y realizar las personalizaciones necesarias 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

Preservación de contenido. Migración continua

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

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

Preservación de contenido. Adhesión a estándares internacionales

Es una estrategia que busca apoyarse en la afirmación de que los estándares internacionales son relativamente estables en el tiempo.

- En la actualidad, los estándares evolucionan casi tan rápido como las tecnologías
- Es una estrategia que debería usarse en combinación con otras
- https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/formatos

Preservación de contenido. Emulación

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

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

Preservación de contenido. Encapsulamiento

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

Como elementos a encapsular podemos tener:

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

Preservación de contenido. Metadatos de preservación.

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

Varias iniciativas:
- PREMIS: PREservation Metadata: Implementation Strategies
- OAIS: Open Archival Information System

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) como se hablara en clases precedentes.
- También es necesario elegir los formatos de acuerdo a los criterios que se mencionaron: uso de una gran comunidad, apertura, licenciamiento libre...

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, DNG, GIF, BMP, SVG
- Datos tabulares: CSV, TXT
- Texto: XML, PDF / A, HTML, JSON, TXT, RTF
- Archivo web: WARC

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

Selección de formatos: generalidades

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

Selección de formatos: generalidades

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

Selección de formatos: generalidades

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

Sobre PDF/A

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

Sobre PDF/A

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

Otros elementos de la compatibilidad con PDF/A

El contenido de audio y video está prohibido.

Java script y enlaces a archivos ejecutables están prohibidos.

Todas las fuentes deben estar embebidas, y también deben ser legalmente embebibles para renderización ilimitada y universal. Esto significa para un usuario poder abrir el documento y que los caracteres se muestren de manera correcta (de aquí a X años) aunque no tenga esa tipografía en su computadora.

Los espacios de colores deben ser especificados de una manera independiente del dispositivo.

Se prohíbe la encriptación.

El uso de metadatos basados en estándares se mantiene.

Otros elementos de la compatibilidad con PDF/A

Las referencias a contenidos externos están prohibidas.

La compresión de imágenes LZW y JPEG2000 están prohibidas en PDF/A1, pero JPEG 2000 se permite en PDF/A2.

Capas y objetos transparentes están prohibidos en PDF/A1 pero no en PDF/A2.

Firmas digitales provisionales se permiten en PDF/A2.

Los archivos embebidos están prohibidos en PDF/A1, pero PDF/A2 permite embeber archivos PDF/A.

PDF/A3 permite embeber cualquier formato como XML, CSV, CAD, archivos de Word, planilla de cálculo, otros PDF/A, etc. como objetos archivados completos.

Niveles de cumplimiento

PDF/A posee dos niveles de cumplimiento:

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

PDF/b aplica integridad visual.

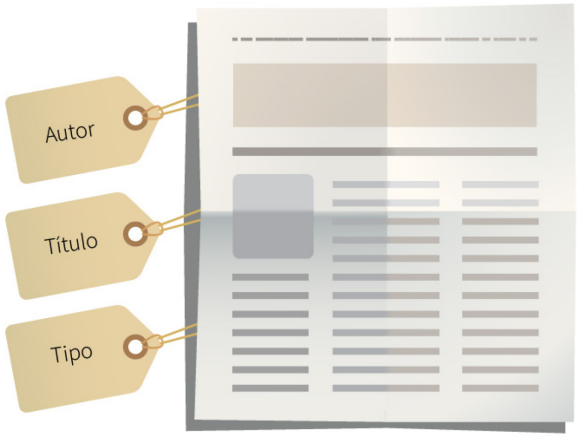
Formatos. ¿Cómo conocer lo que tiene un RI?

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…

Bibliotecas y repositorios digitales

Capítulo 8:

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

Estándares

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

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.

Referencia:
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).

http://www.oais.info/

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

Conceptos en OAIS

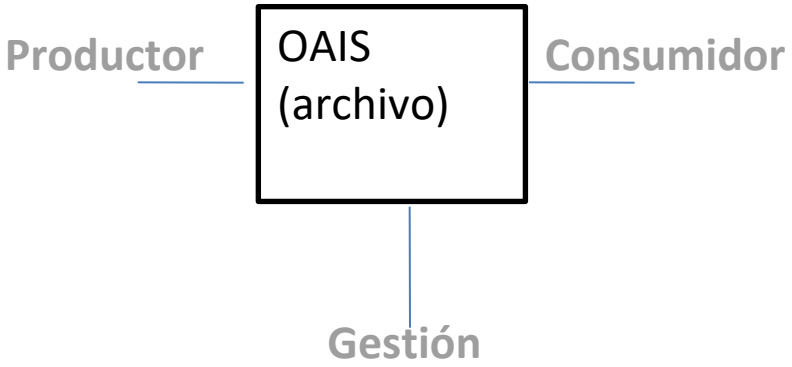
El propósito de esta sección es motivar y describir varios conceptos clave, de alto nivel del OAIS. Un punto de vista más completo y una modelización formal de estos conceptos, se da en la sección 4.

Medioambiente OAIS

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

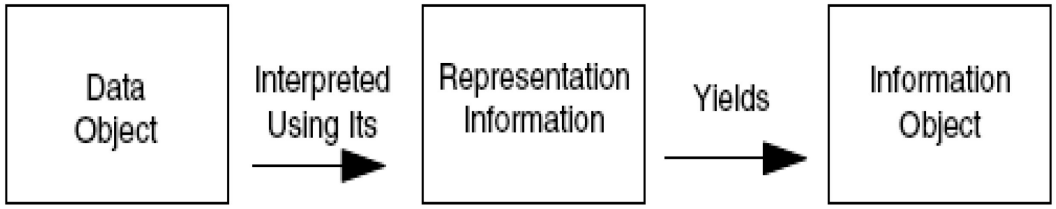
Actores en el modelo

- Producer-Consumer-Management



Conceptos en OAIS

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



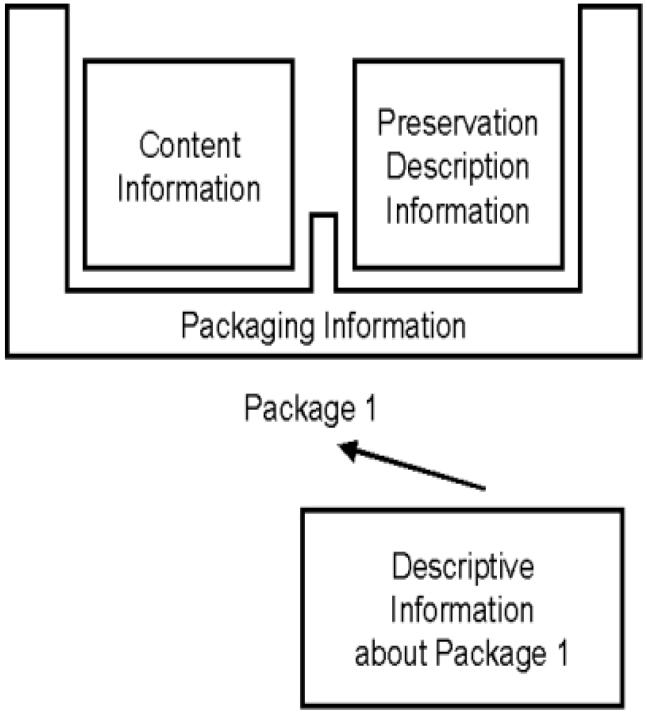
Conceptos en OAIS

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

Conceptos en OAIS

- La unidad de intercambio entre un OAIS y su medioambiente es el paquete de información –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

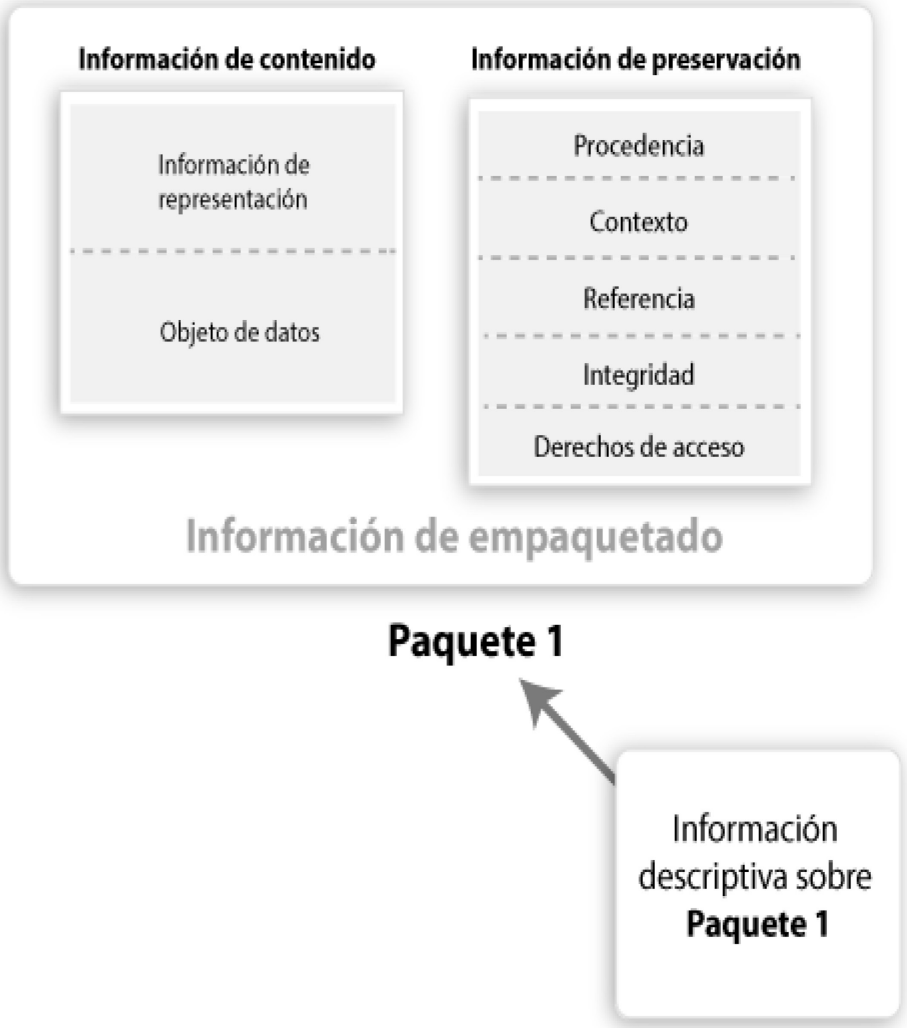


ISO 14721: Fig 2-3: Paquete de información: conceptos y relaciones

Conceptos en OAIS

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

Estructura del Paquete de Información



El paquete de información (IP)

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

Elementos de la PDI

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

Conceptos en OAIS

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

Clases de IP según su función

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

Clases de IP según su función

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

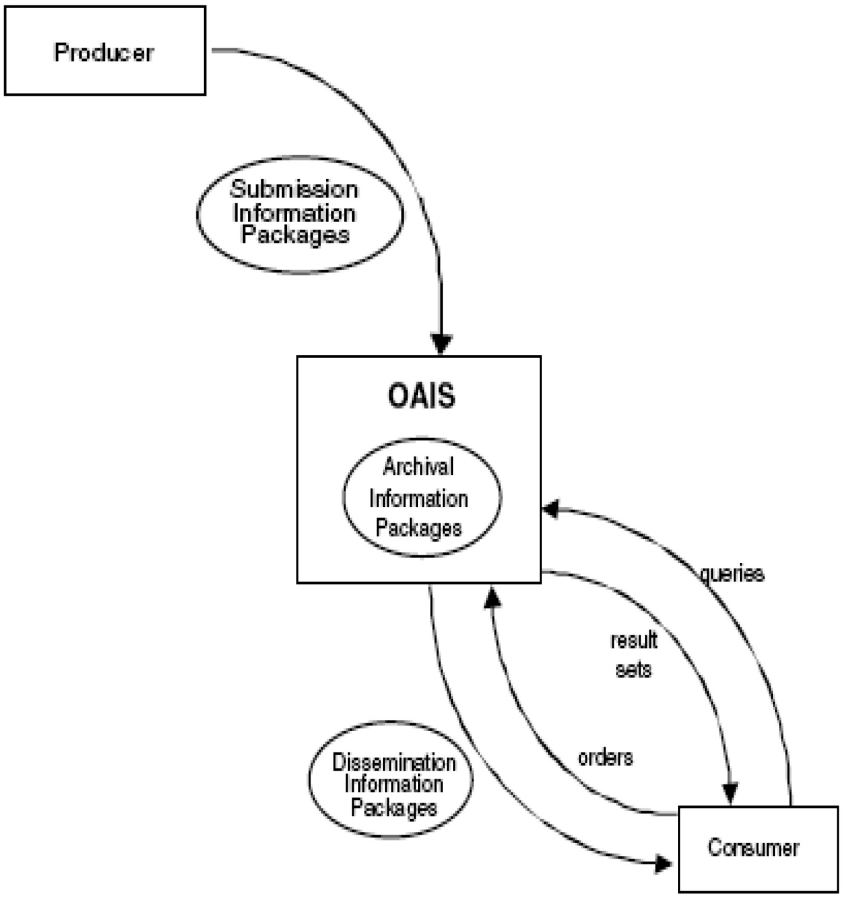
Clases de IP según su función

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

OAIS interacciones externas de alto nivel

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

OAIS interacciones externas



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

OAIS

Modelo Funcional

Sección 4.1


Planeamiento de la preservación
Productor
SIP
Ingesta
ID
AIP
Gestión de datos
Área de archivo
ID
AIP
Acceso
Preguntas,
Resultados
Peticiones
DIP
Consumidor
Administración
Gestión
REFERENCIA
Actores
Entidades

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.

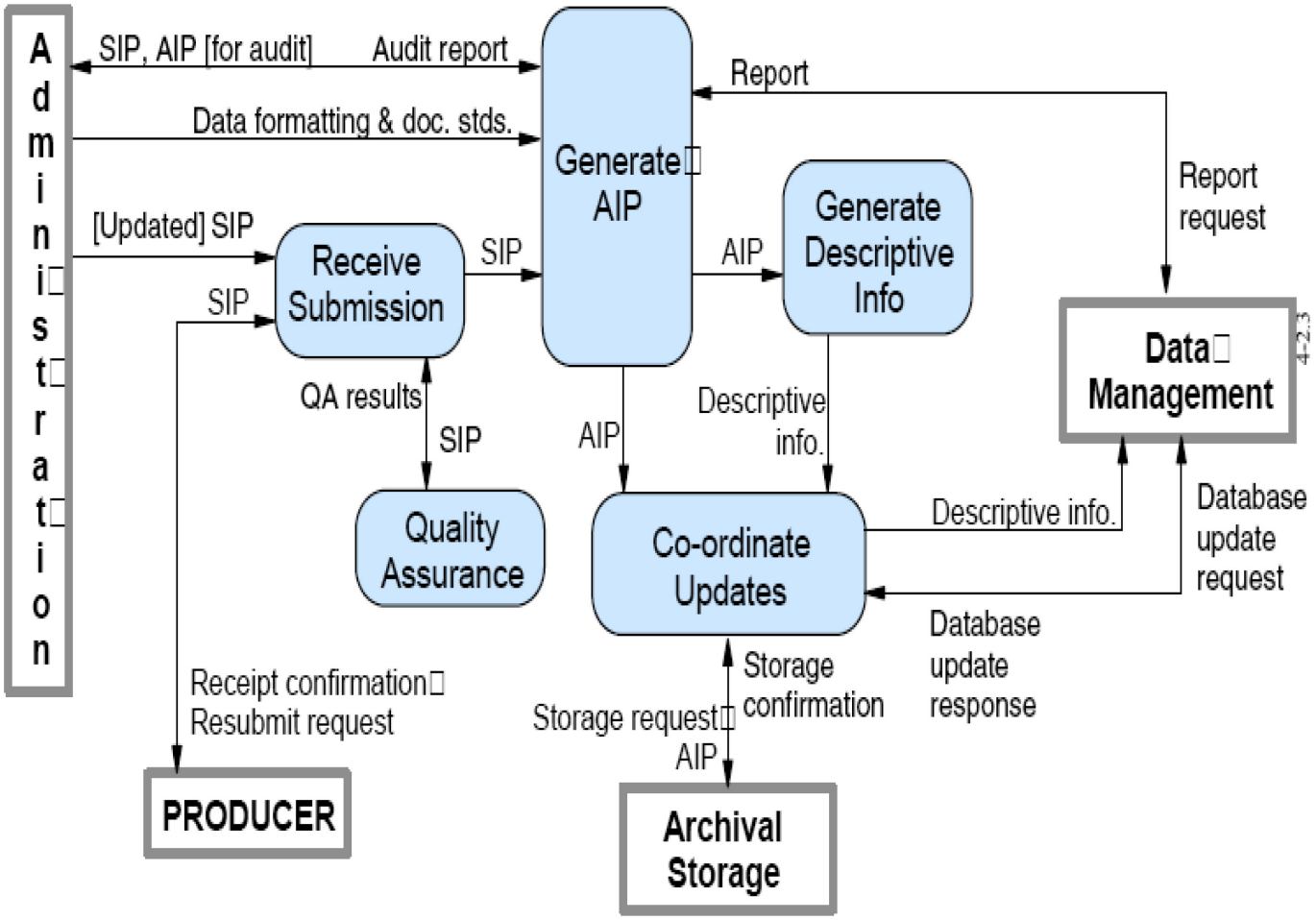
Modelo OAIS

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.

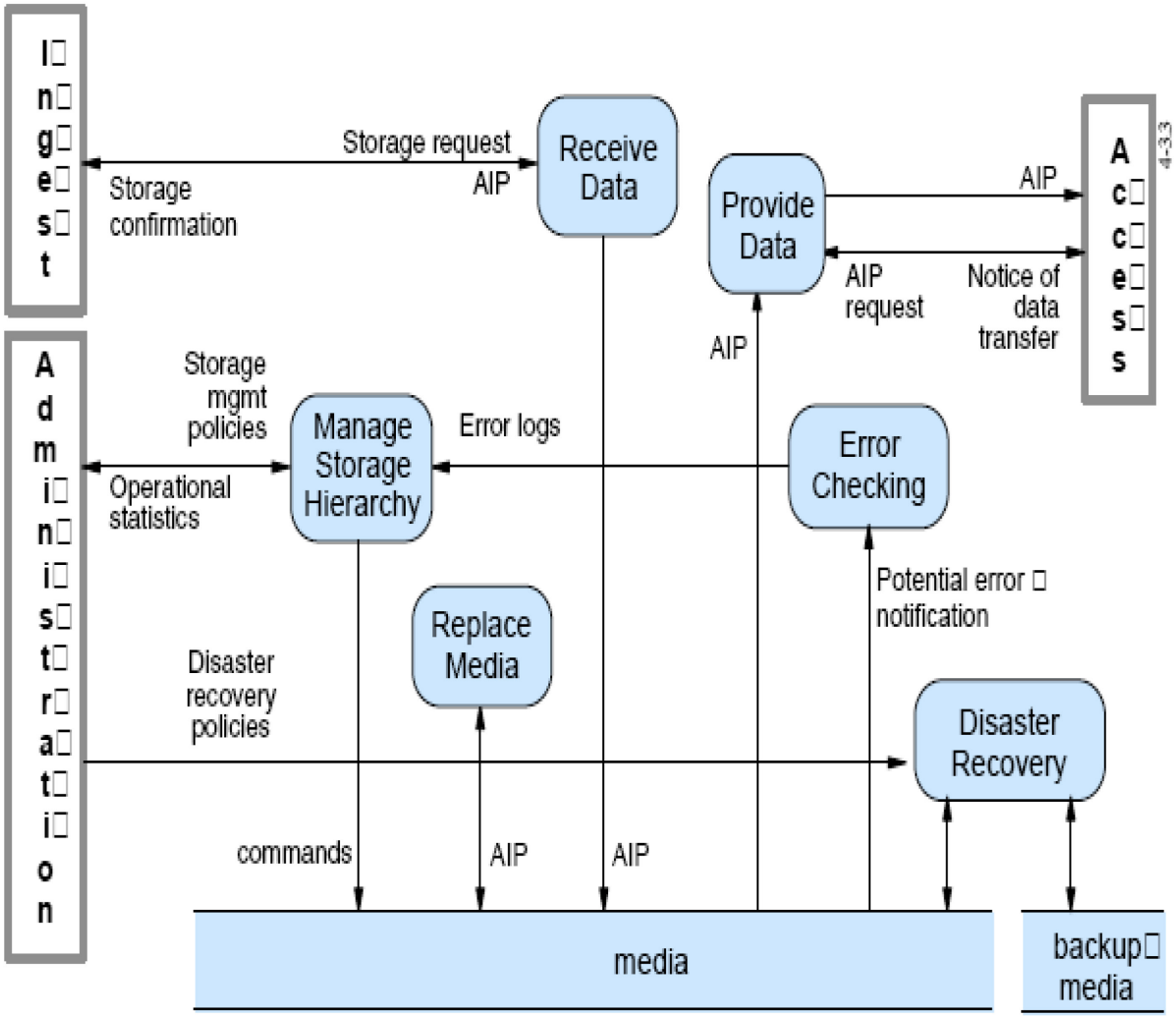
Functions of Ingest



Entidad OAIS Ingest

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

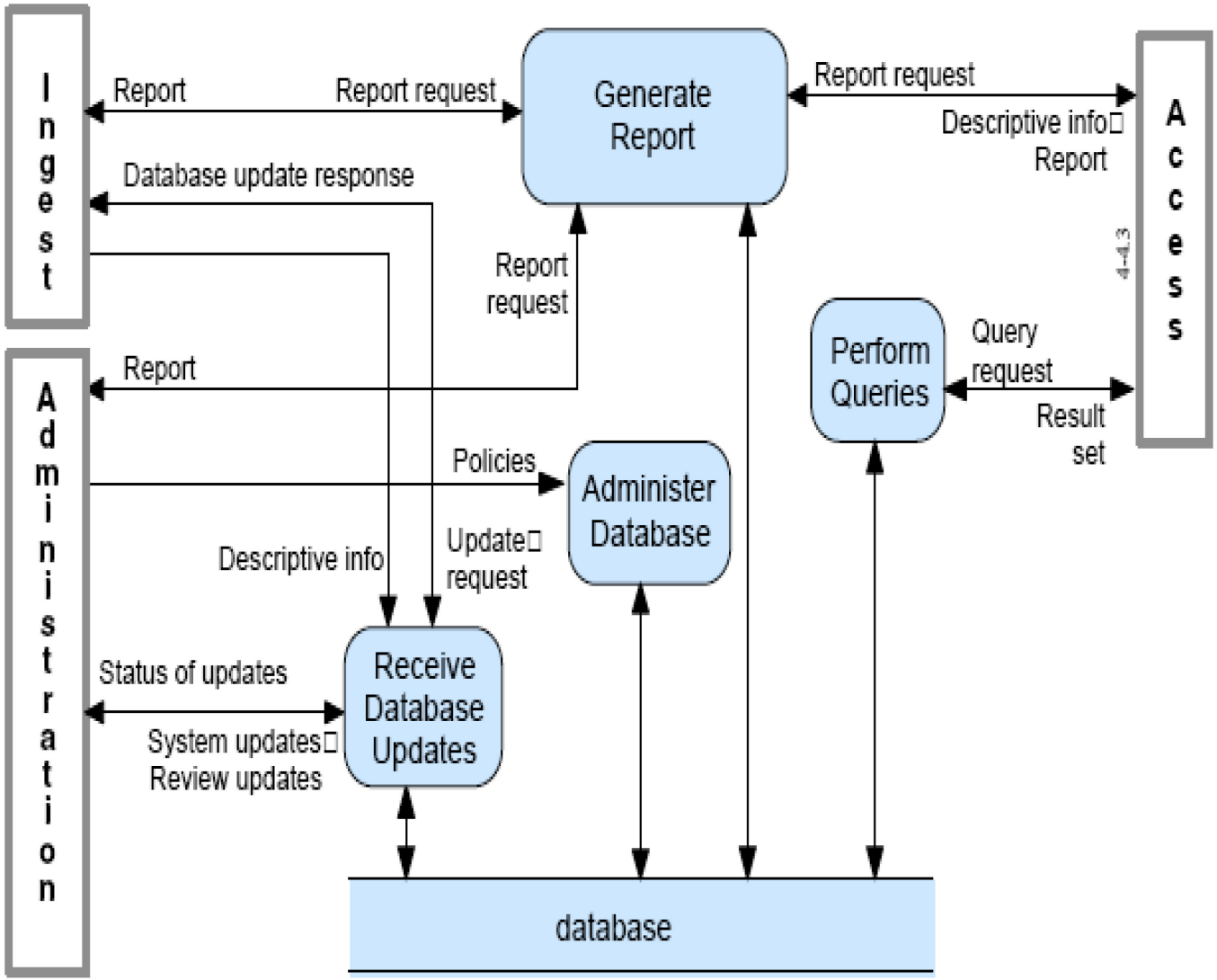
Functions of Archival Storage



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 una lugar físico separado. Provee copia de los AIPs almacenados a la entidad *access*.

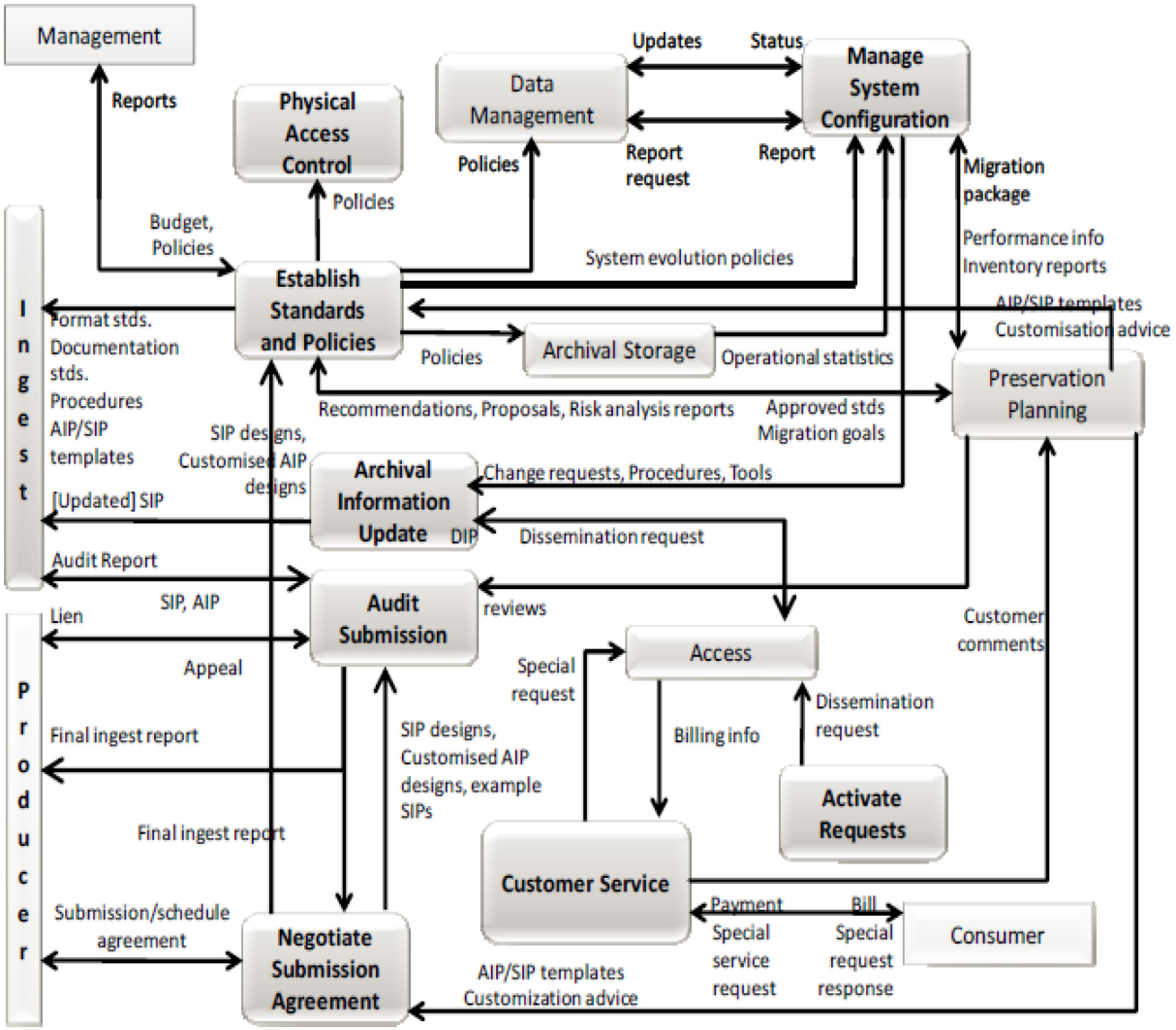
Functions of Data Management



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*.

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.

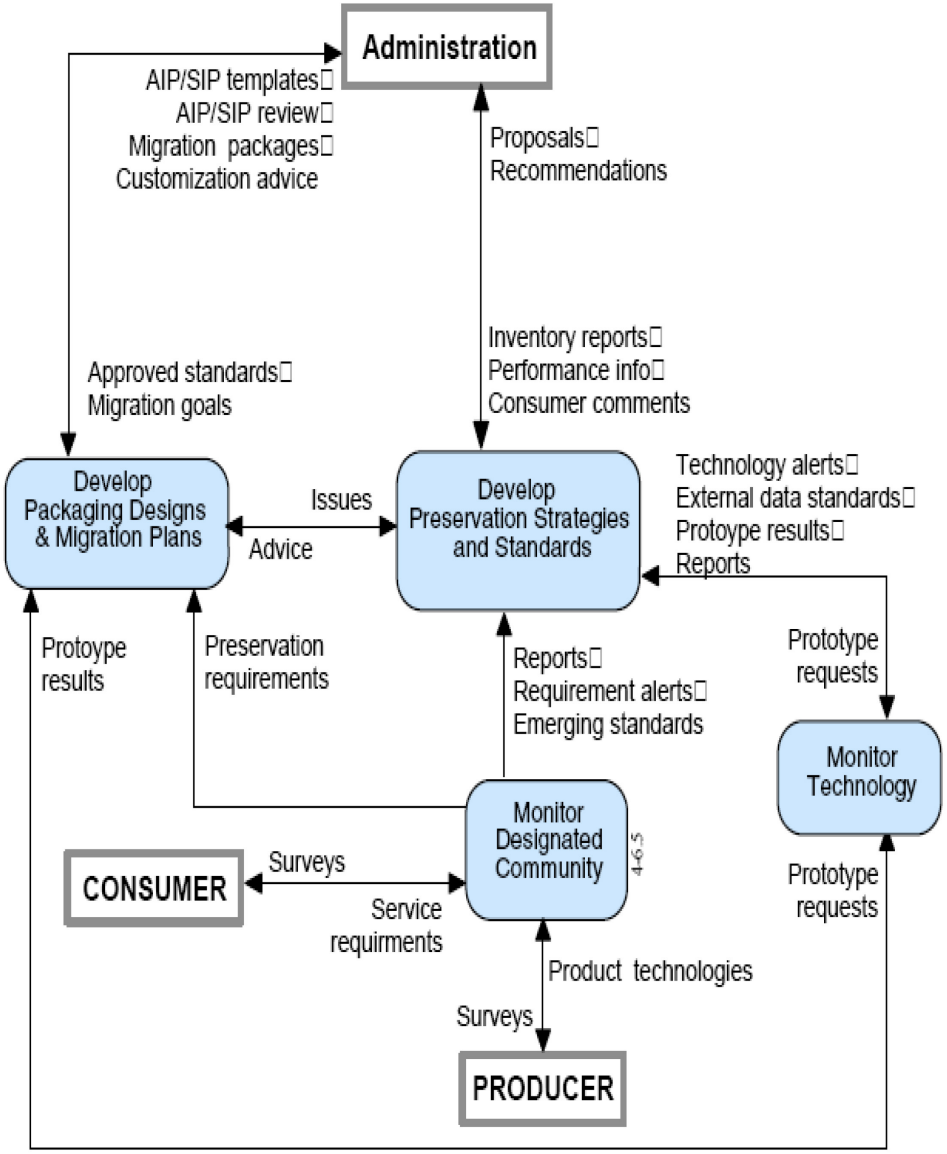
Entidad OAIS administration

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.

Functions of Preservation Planning



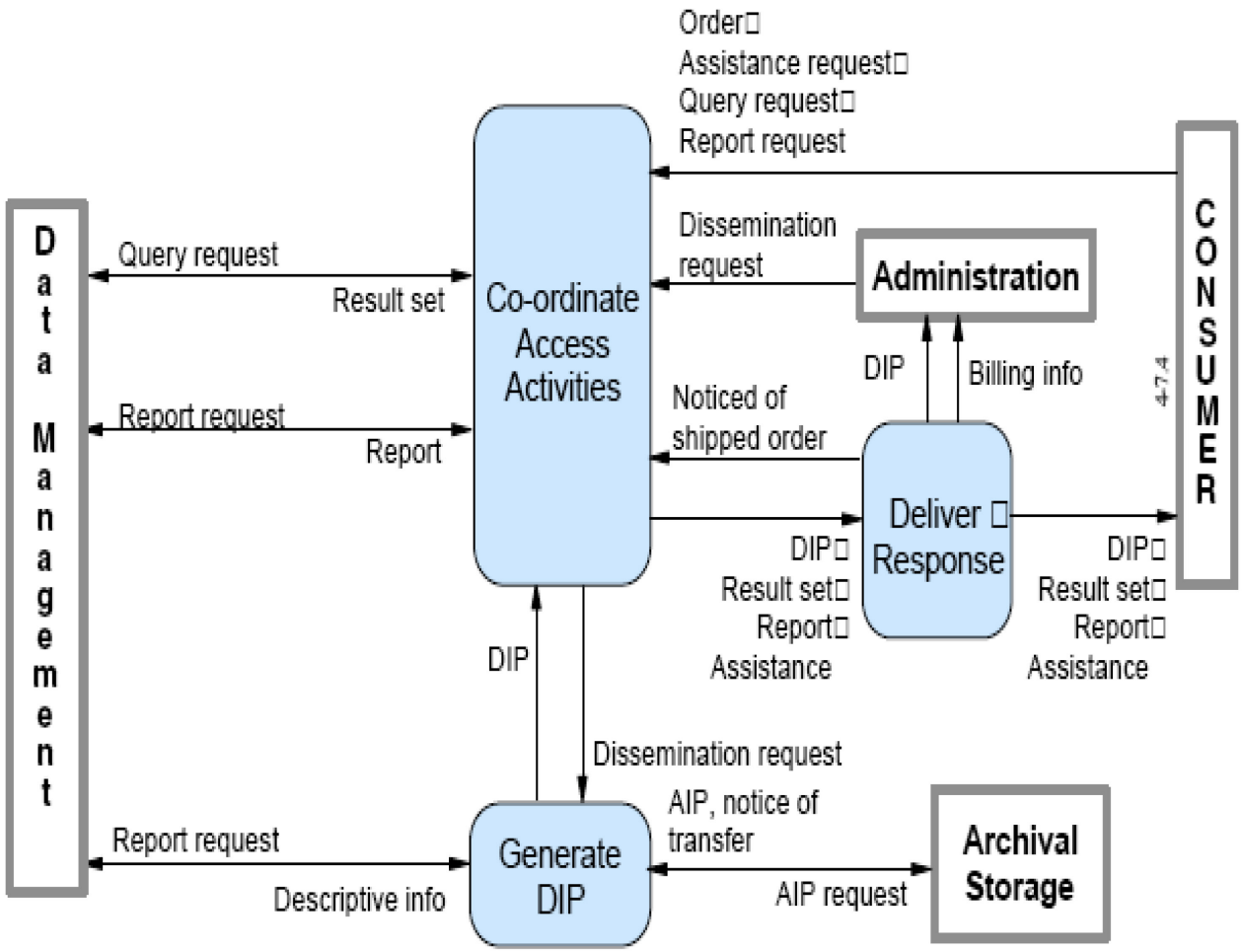
Entidad OAIS Preservation Planning

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.

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.

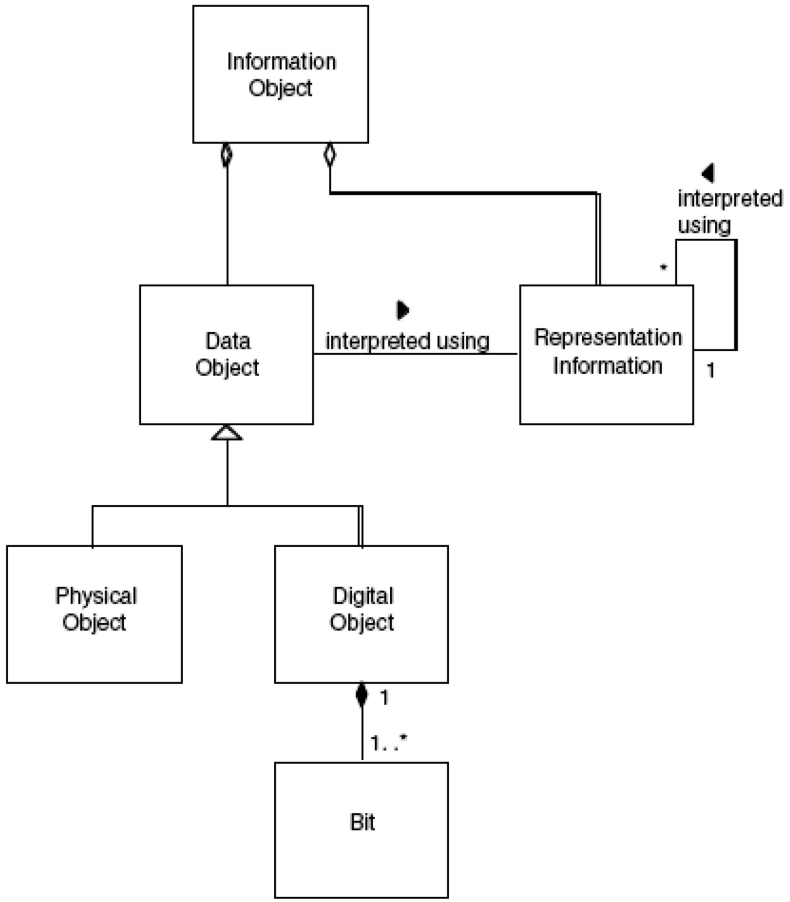
Entrega las respuestas en línea y fuera de línea de los consumidores.

OAIS

Modelo de Información

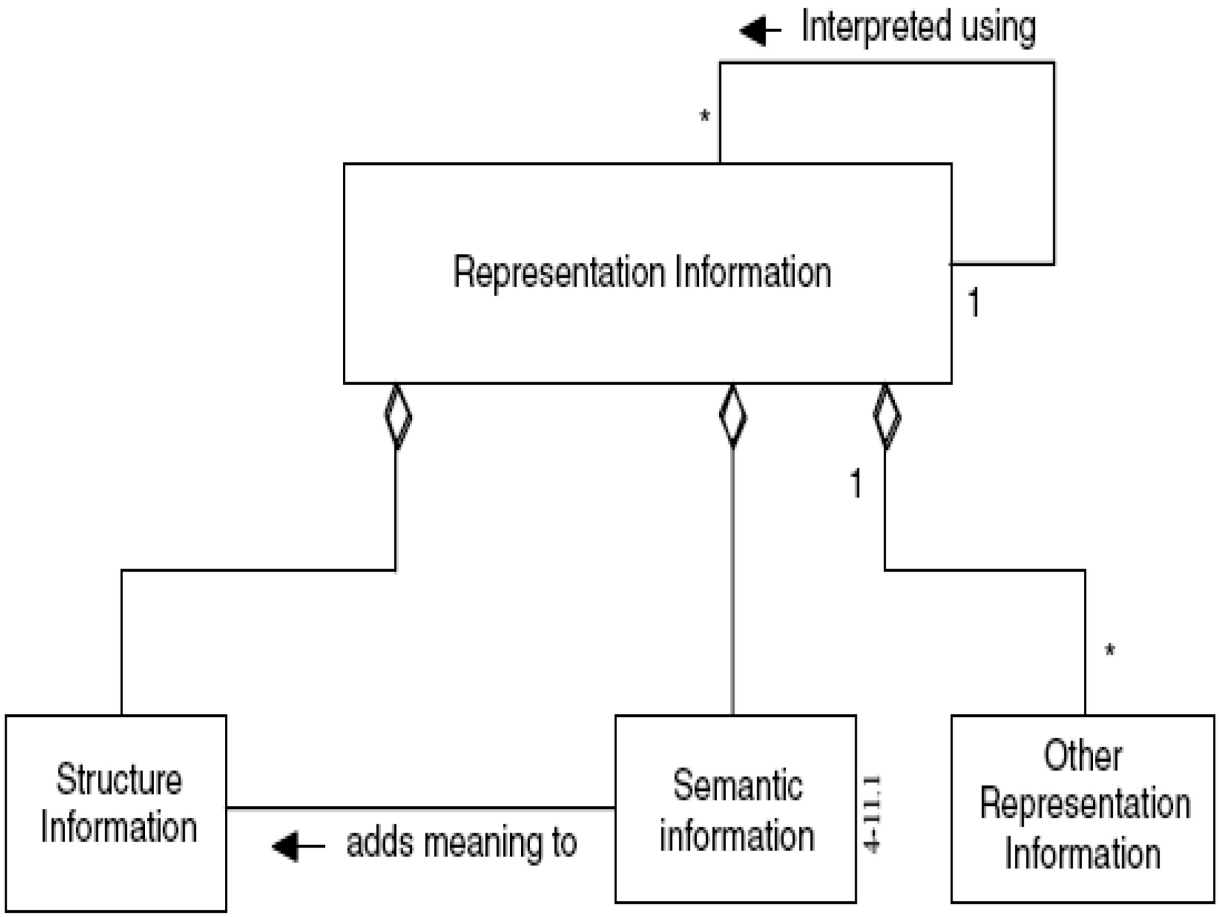
Sección 4.2

OAIS Objeto de información

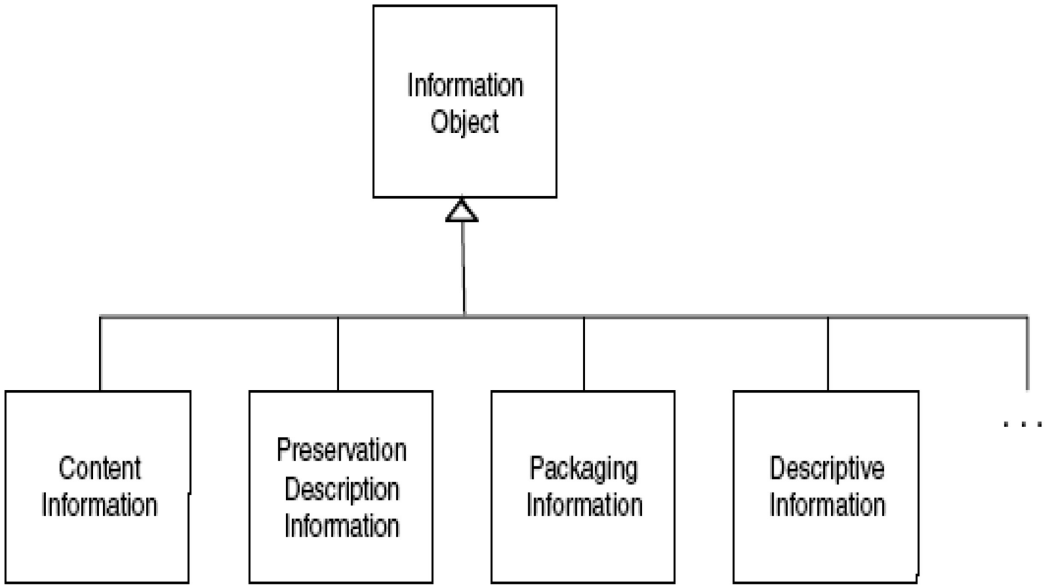


- El **Objeto de Información** está compuesto de un Objeto de Datos, que puede ser físico o digital, e Información de Representación que permite la interpretación completa de los datos.

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)

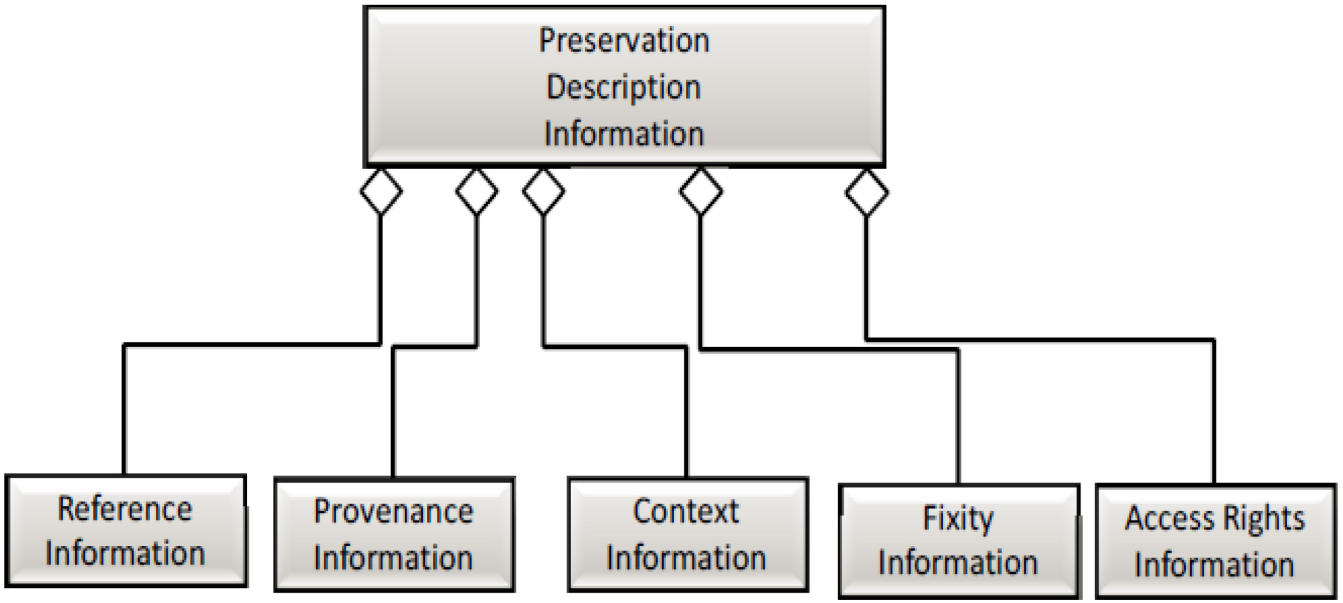


Figure 4-16: Preservation Description Information

Información descriptiva de preservación

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

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

Información descriptiva de preservación

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

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

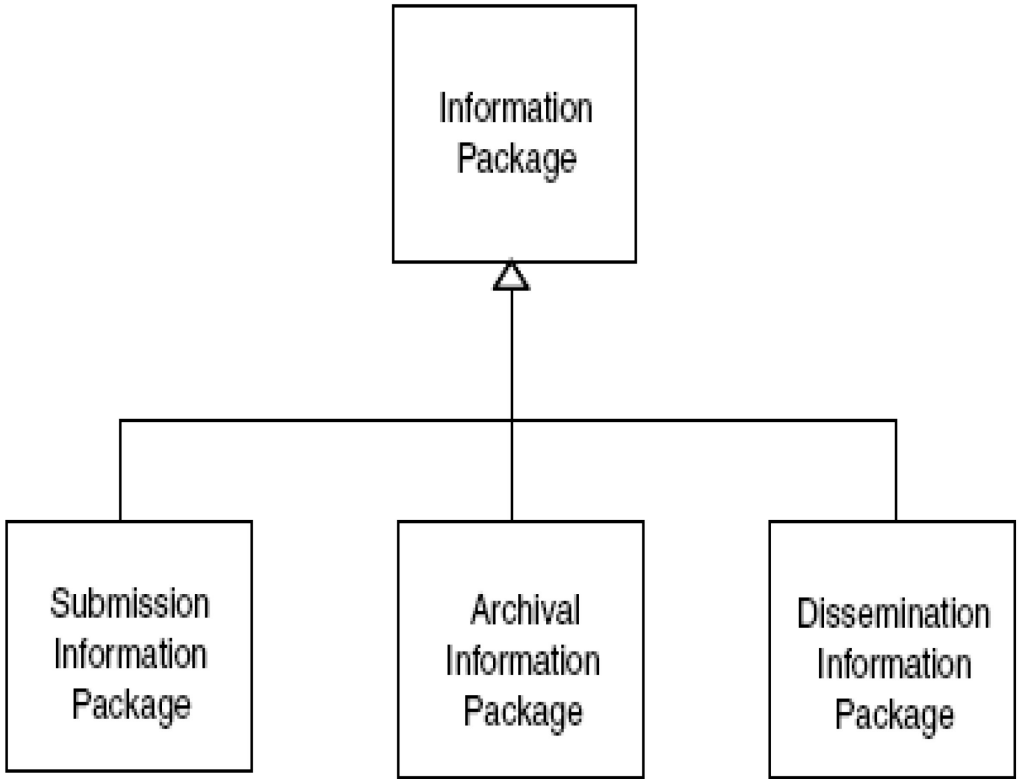
Información de sobre derechos de acceso: proporciona los permisos de uso de la información de contenido.

Paquetes de información en OAIS

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

Paquetes de información en OAIS

Tipos de paquetes de información



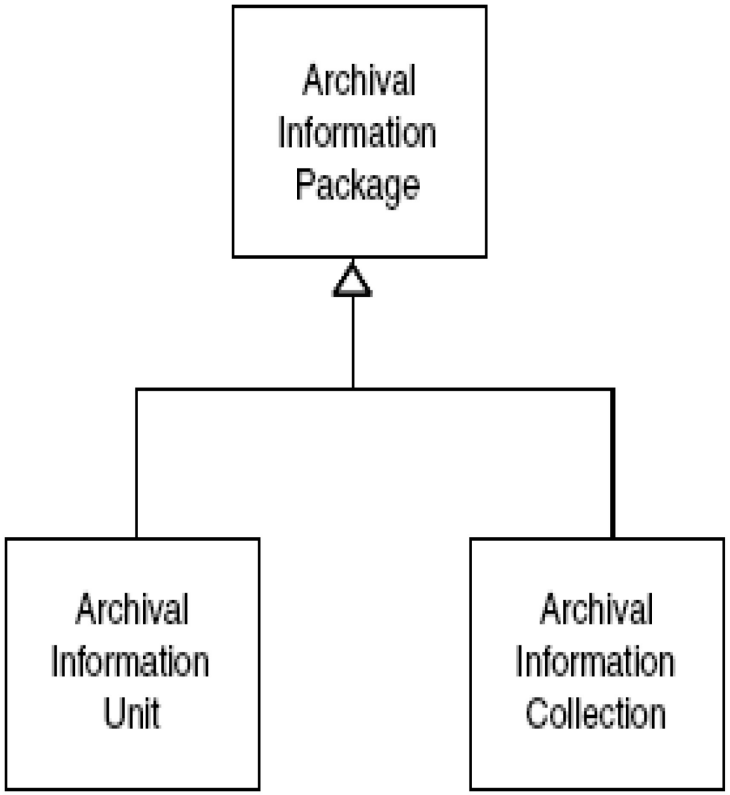
SIP

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

AIP

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

Tipos de AIPs



DIP

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Participantes

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

Sección 5: Perspectivas sobre preservación

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

Saliendo de la 14721

Aproximaciones a la preservación

Existen numerosas estrategias para asegurar la preservación de la información:

- 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.

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 internacionales, es decir, apoyarse en la afirmación de que los estándares internacionales son relativamente estables en el tiempo.

Según las guias "Technical Guidelines for Digitizing Cultural Heritage Materials" (FADGI), los formatos de archivos utilizados 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 /A (Archival) 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**

Clasificación del material según relevancia

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

1) Recepción, análisis y evaluación del material a digitalizar

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

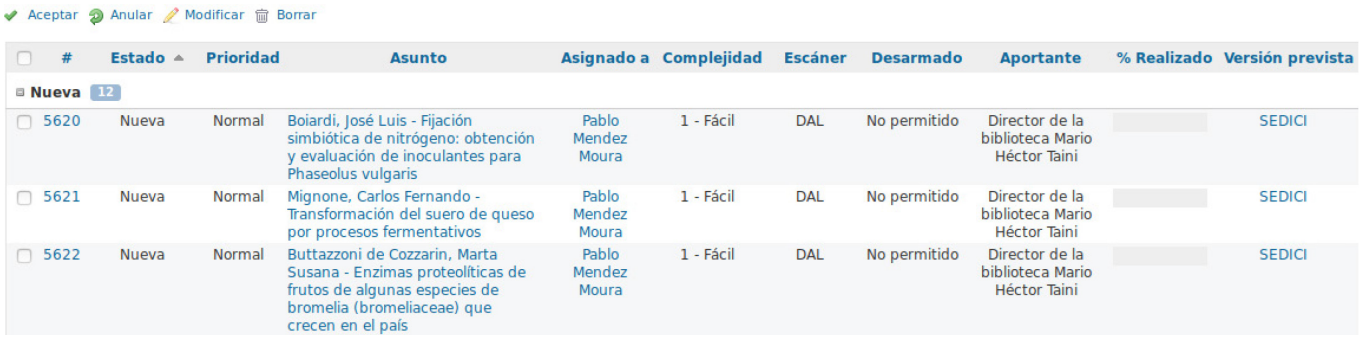
- Estado general de conservación
- Dimensiones
- Formatos
- Tipos de encuadernación
- Importancia histórica, educativa, institucional

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 todos los datos propios del material (Autor, Título etc)

A medida que las obras van pasando por distintas etapas, también se verá reflejado 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

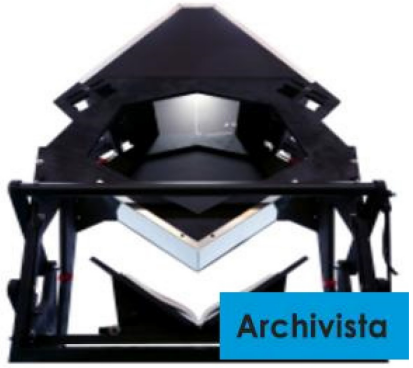
3) Elección de metodología de escaneo

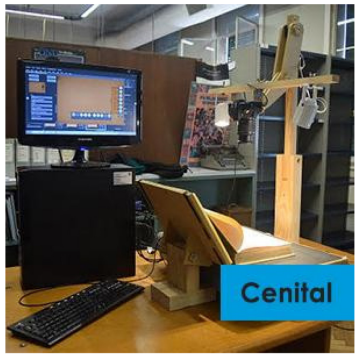
Tipos de escáneres utilizados

automáticos

de gran formato

de libros





Escáneres automáticos

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

Escáner de gran formato

Permite digitalizar hojas de hasta 44 pulgadas, por ejemplo: mapas, planos, dibujos arquitectónicos, posters, etcétera.

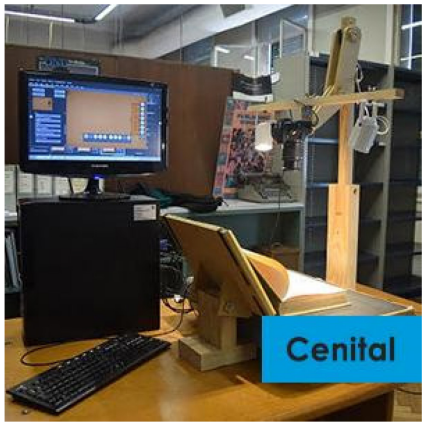
Escáneres de libros

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

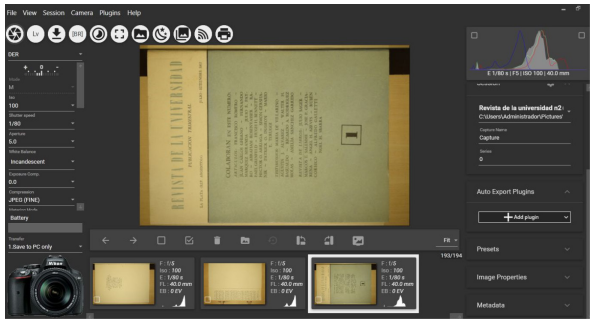
Fue adaptado en SEDICI para digitalizar materiales con dificultades en la manipulación y encuadernaciones frágiles. Por ejemplo las Joyas de la colección Cervantina



4) Captura de imágenes

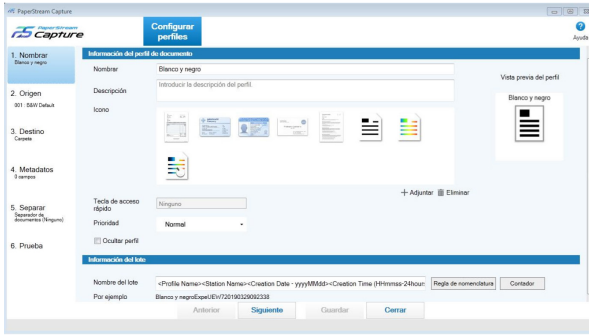
Captura con **“digiCamControl”**

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

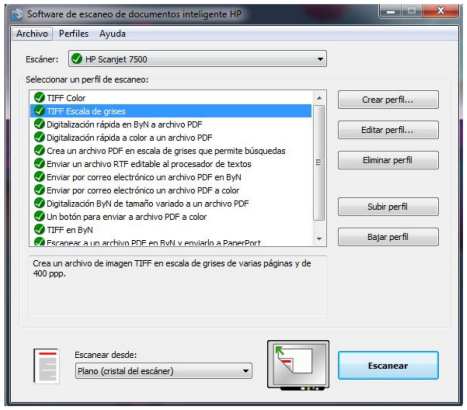


Captura con “**Paperstream**”

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



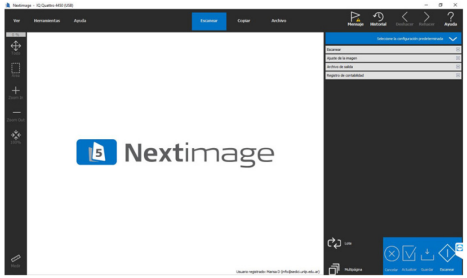
Captura con **“Software de escaneo de documentos inteligente” de HP**



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.



5) Edición de imagen

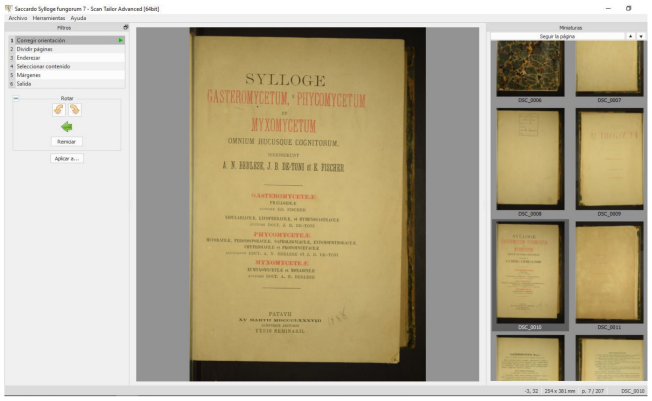
Luego de obtener las imágenes escaneadas, se editan en Photoshop o ScanTailor. De esta manera se puede corregir la orientación, dividir y alinear las páginas, seleccionar el contenido, limpiar los márgenes, eliminar las manchas y modificar el contraste.

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.

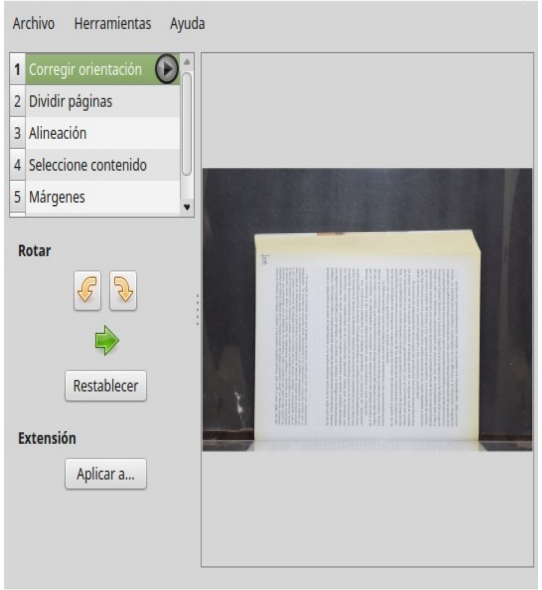
Edición con ScanTailor Advanced

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).

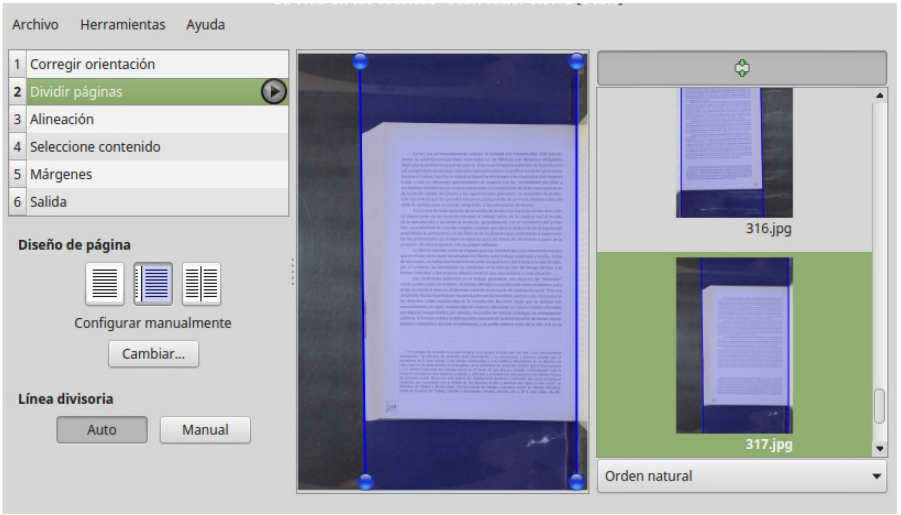


Algunas de las funciones principales de ScanTailor son las de:

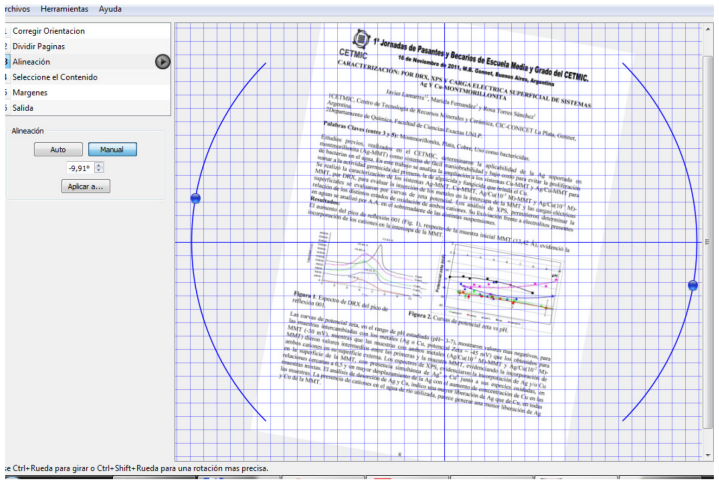
Rotar páginas

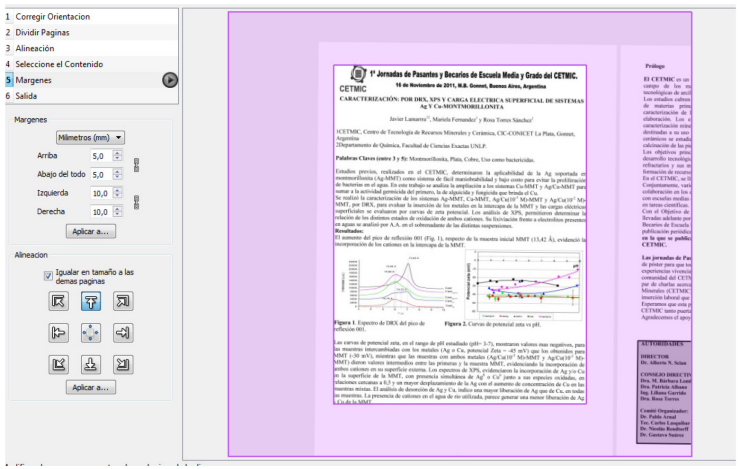


Dividir páginas

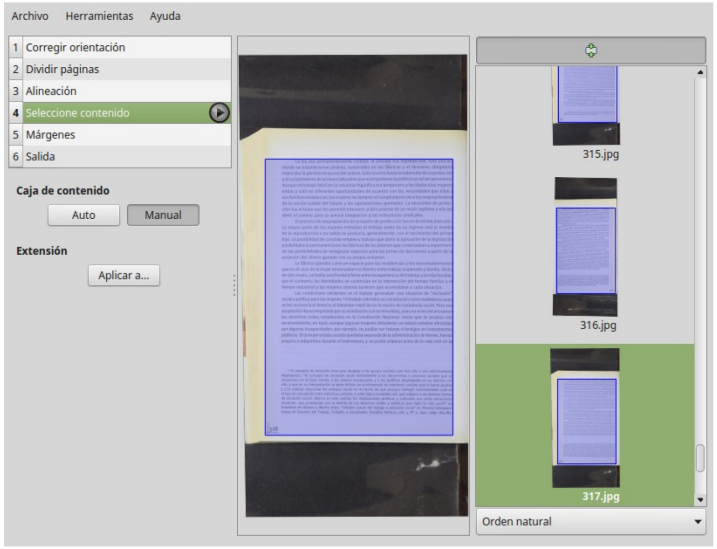


También es posible corregir la orientación de las páginas, seleccionar los márgenes, eliminar manchas y por último ajustar el color y contraste de las páginas.

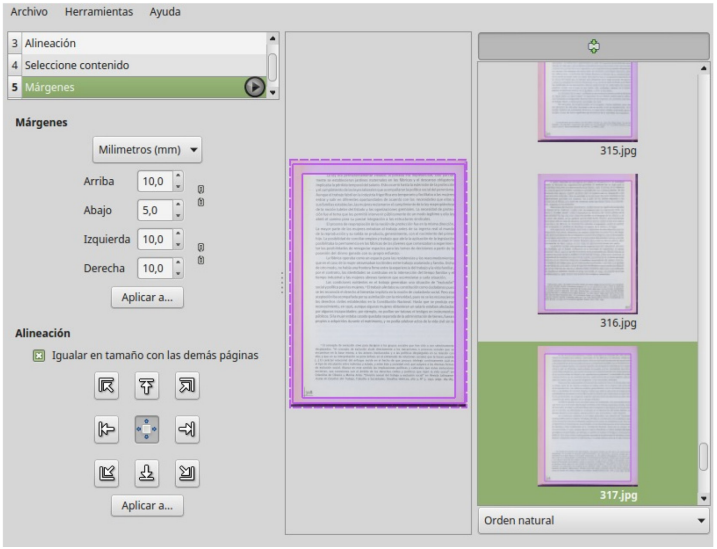




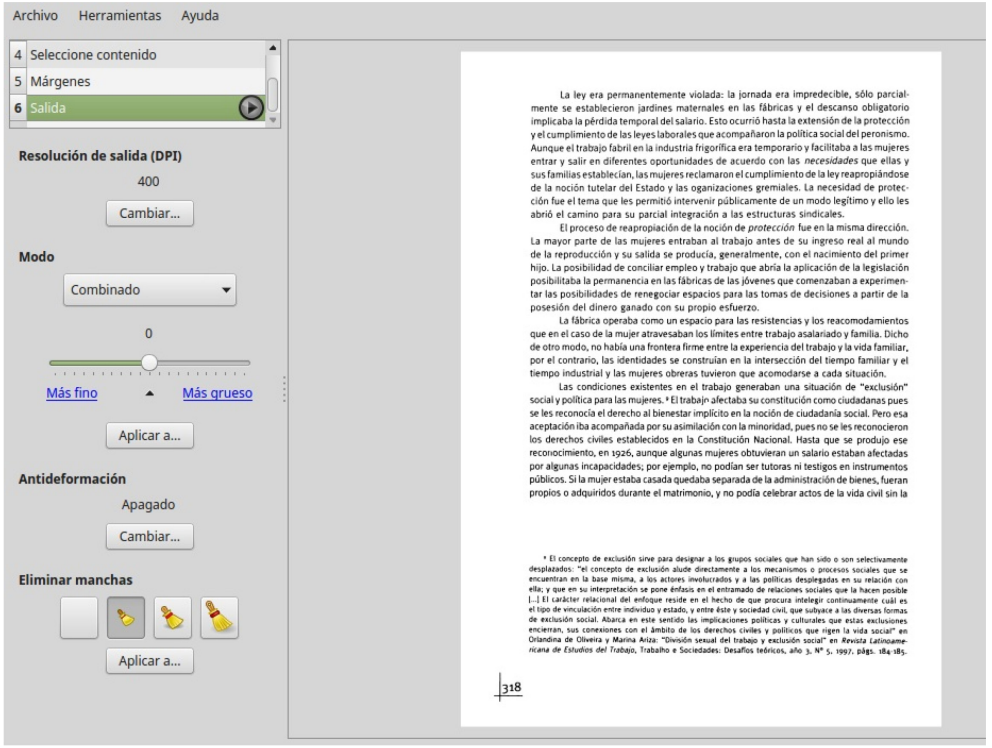
Seleccionar contenido



Agregar márgenes



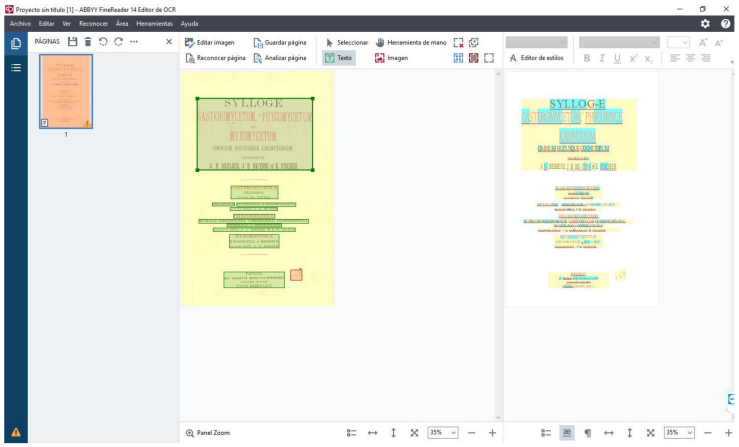
Eliminar manchas



6) Guardado de archivos para preservación y difusión

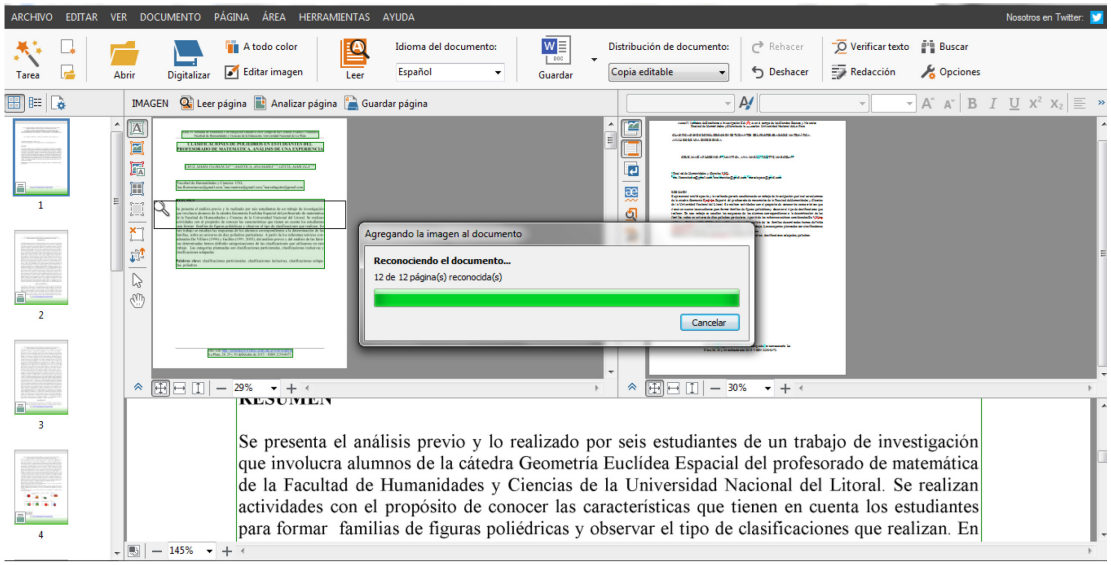
Reconocimiento de caracteres (OCR) y guardado en formato PDFA con Abbyy FineReader

Este software permite realizar un reconocimiento óptico de caracteres, posee un editor de texto donde se corrige manualmente las palabras que contienen errores y por último los archivos se guardan en formato PDFA

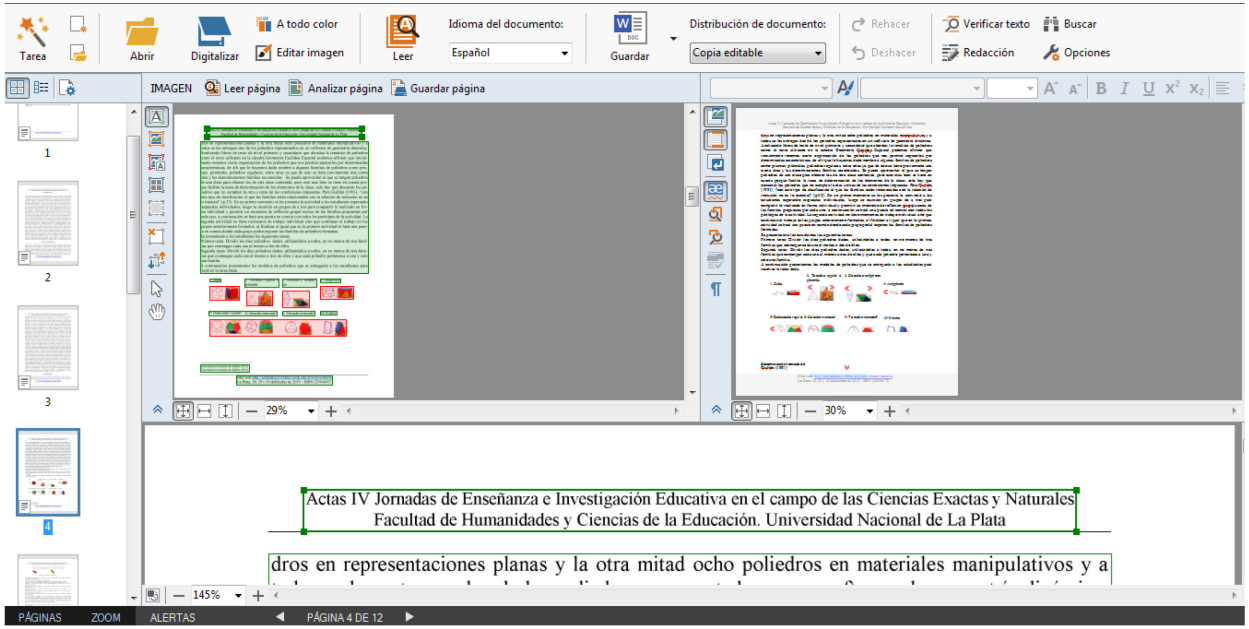


Generación de OCR con ABBYY FineReader

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

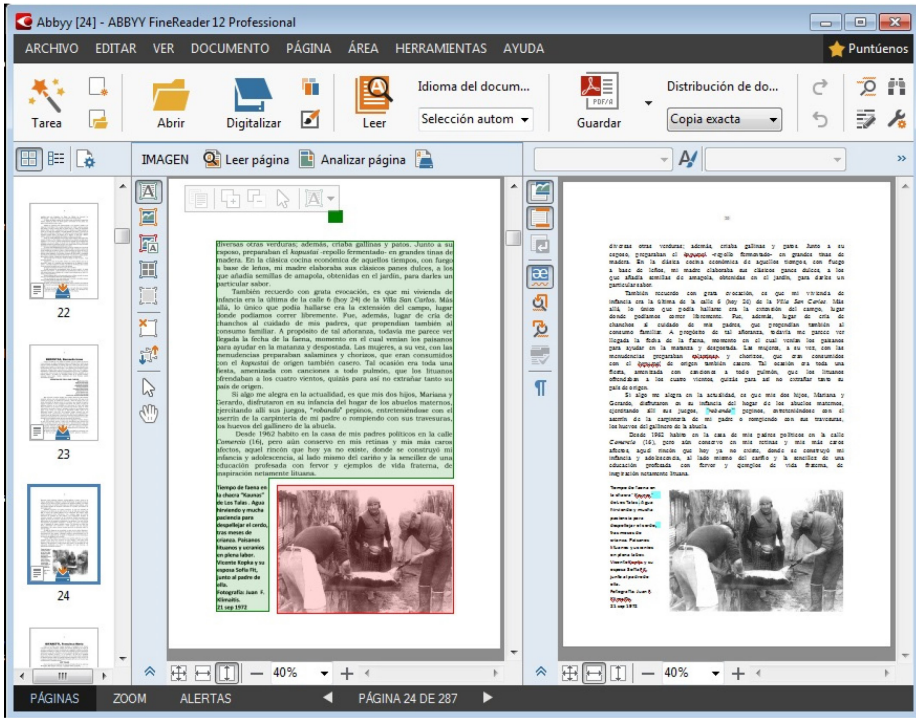


Este software permite seleccionar imágenes e indicar qué parte de la página debe ser reconocido y qué no. Posee un motor de reconocimiento óptico de caracteres muy potente además también permite la corrección manual.



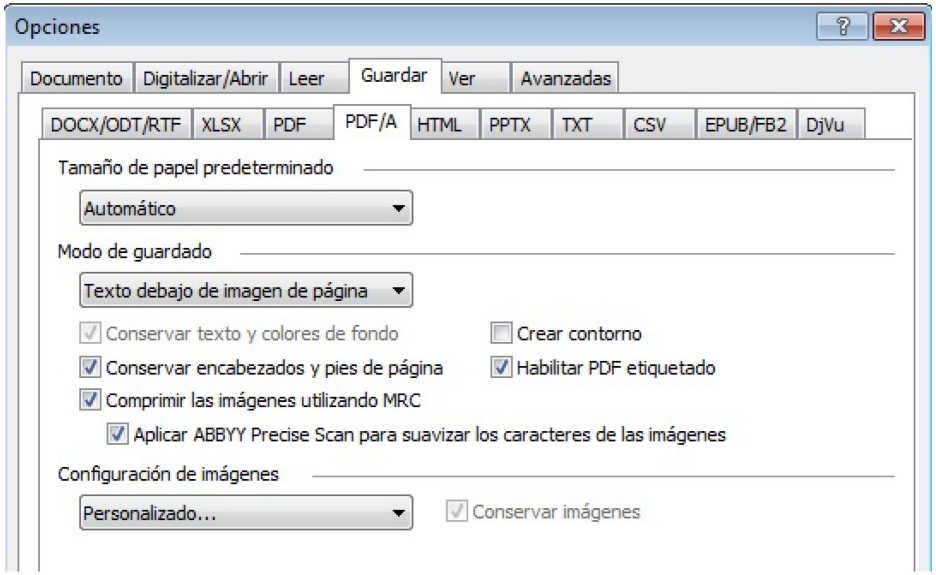
Reconocimiento de caracteres (OCR)

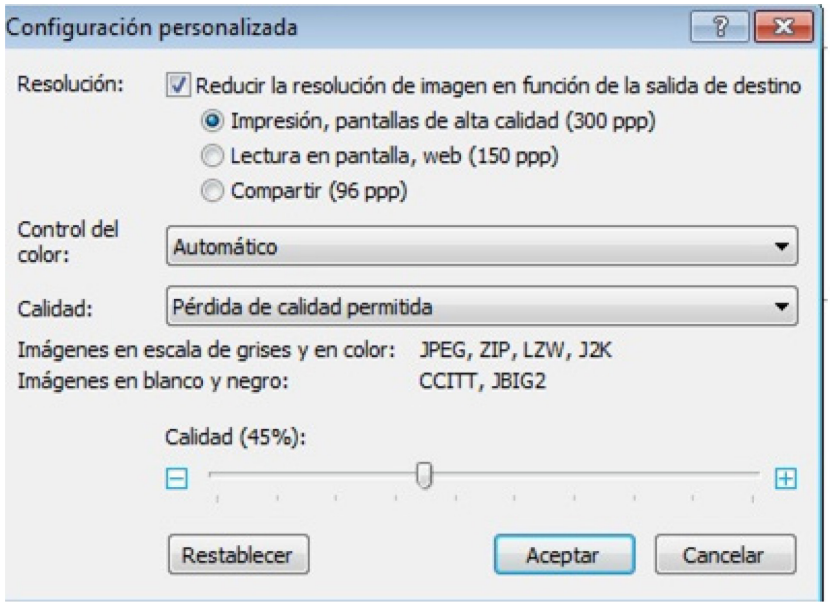
Luego de editar las imágenes se realiza el OCR con el Abbyy FineReader. En esta etapa del proceso se selecciona el contenido según sea texto, imagen o cuadro. Luego se revisa el resultado del OCR y se generan los archivos PDF/A.



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.

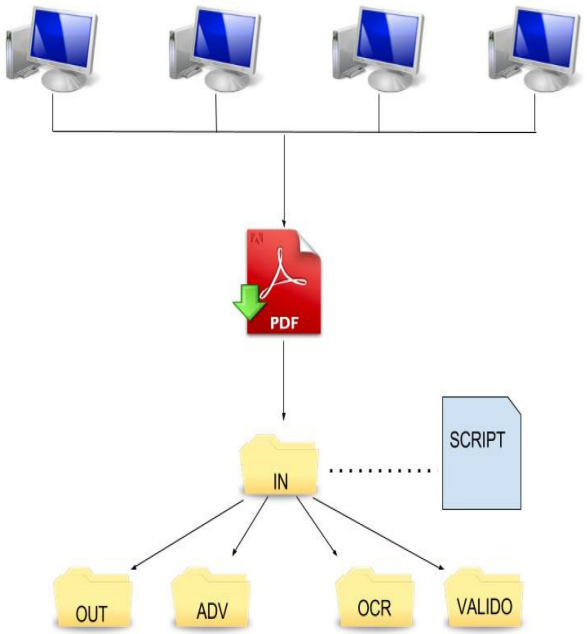




3-HEIGHT

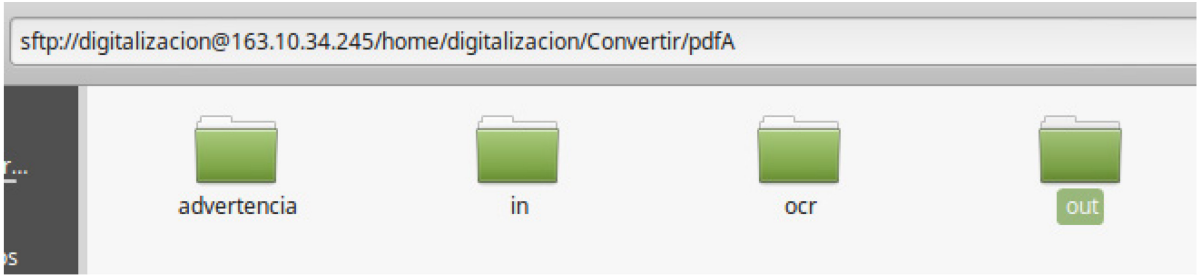
Este software posee una arquitectura cliente servidor, que permite convertir por lotes archivos de distintos formatos a pdf/a. Además también es utilizado para verificar si los archivos pdf/a cumplen con la norma.

- Deteccion de archivos
- Análisis
- Conversión
- Verificación



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:

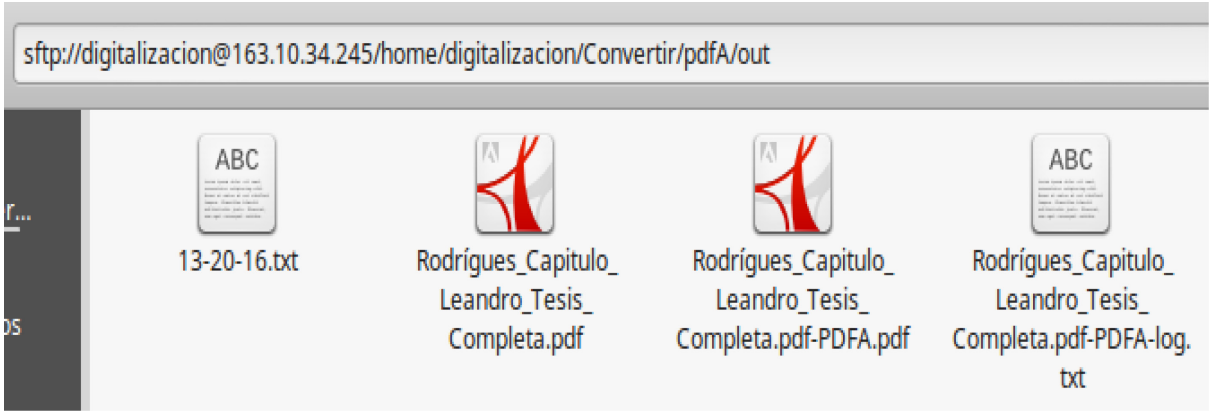
- Una carpeta “in” para ingresar los archivos a procesar
- Una Carpeta “out” donde se depositarán los archivos resultantes.
- Y dos carpetas destinadas a diferentes tipos de errores llamadas “advertencia” y “ocr”



Generación de PDF/A con 3-HEIGHT

El 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:

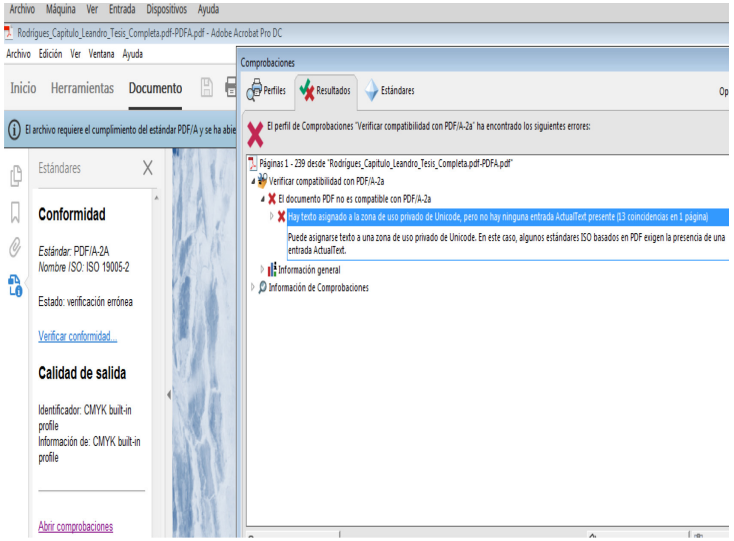
El archivo con la fecha 13-20-16.txt presenta el log de la ejecución del script..
El archivo pdf original.
El archivo convertido con la terminación: -PDFA.pdf
El último archivo txt da más detalles de la conversión del archivo original



Validación de PDF/A - Acrobat DC

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

Adobe Acrobat DC



Generación de PDF/A

Un formato de preservación para documentos de texto es el estándar PDF/A, descrito en las normas ISO 19005-(1-2-3).
Este formato está basado en el estándar PDF 1.4, al que le incorpora algunos requerimientos adicionales, por ejemplo:

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

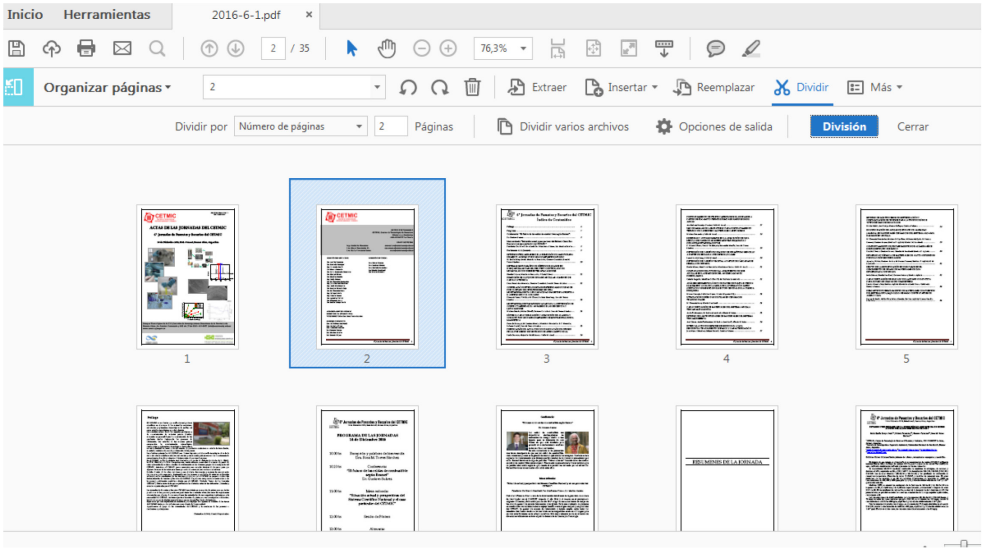
Uno de los propósitos de los requerimientos del estándar PDF/A es de proveer soporte para personas con capacidades diferentes, por ejemplo, incorporando la información requerida y necesaria para aplicaciones que hagan el pasaje de texto a voz.

Acrobat DC - Tratamiento y mejoras de archivos PDF

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

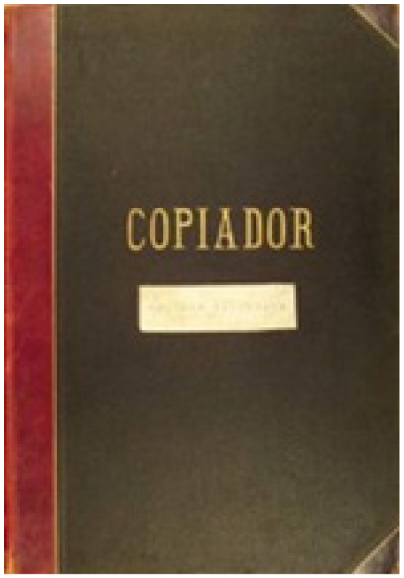
Acrobat DC - División de Archivos

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



Ejemplo de caso de proceso completo de digitalización:

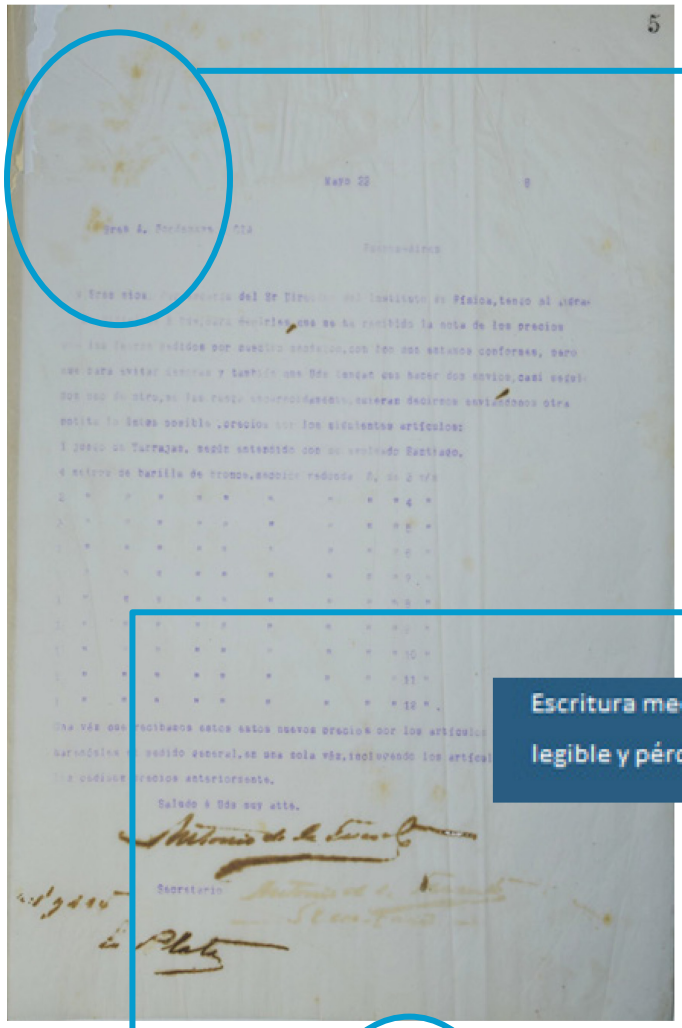
LIBRO COPIADOR - FACULTAD DE CS. FÍSICAS, MATEMÁTICAS Y ASTRONÓMICAS (1918-1925)



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**.

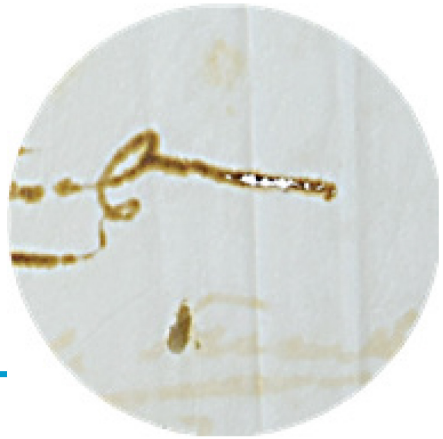
Estado de conservación

- Papel amarronado, débil, friable con roturas y desprendimientos.
- Escritura manuscrita con tinta difundida en el papel y transferida a los siguientes, con pérdida de nitidez e imagen doble.
- Escritura mecanográfica poco legible.
- Encuadernación con especiales requerimientos de manipulación.



Dobleces y desprendimientos

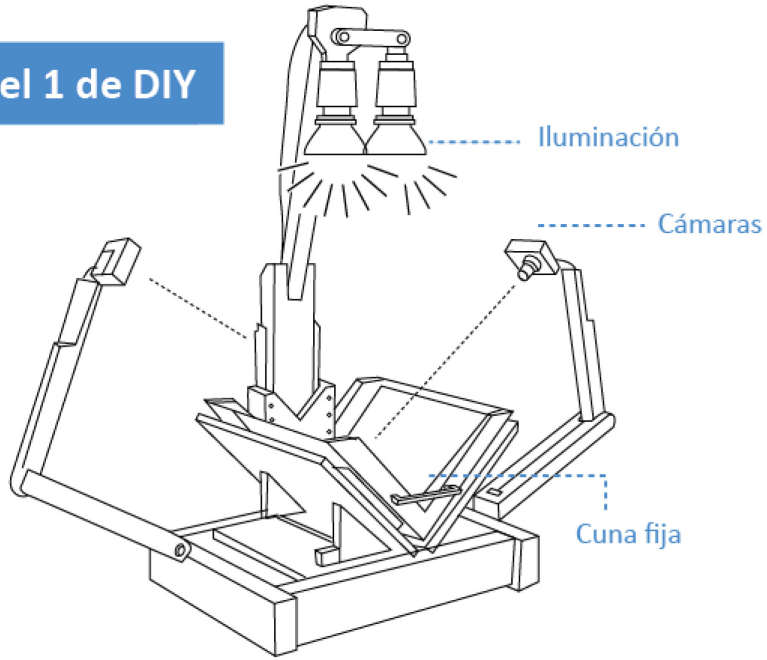
Escritura mecanografiada poco legible y pérdida de nitidez



Tinta difundida en el papel y transferida a los consecutivos.

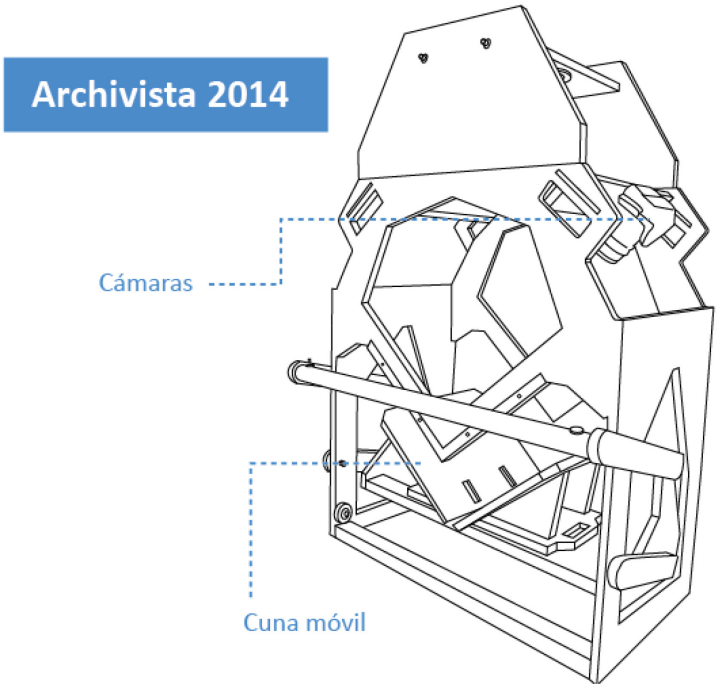
Escáneres y software descartados

Model 1 de DIY



Los escáneres **DAL** utilizados para el escaneo de libros en buen estado no pudieron utilizarse porque el giro de las páginas hacía que el papel pudiera quebrarse.

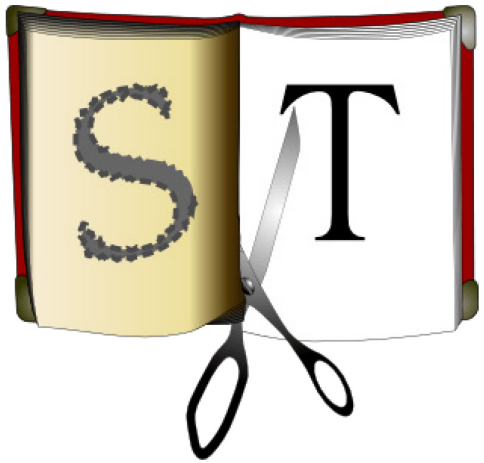
Escáneres y software descartados



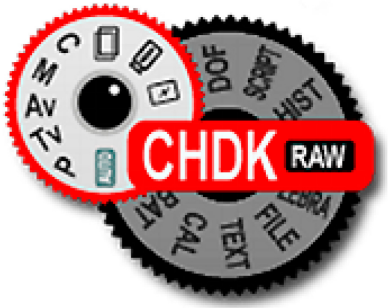
Modelo Archivista 2014

Presentaba muchas mejoras pero no pudo emplearse por el tamaño y estado de deterioro del Libro Copiador.

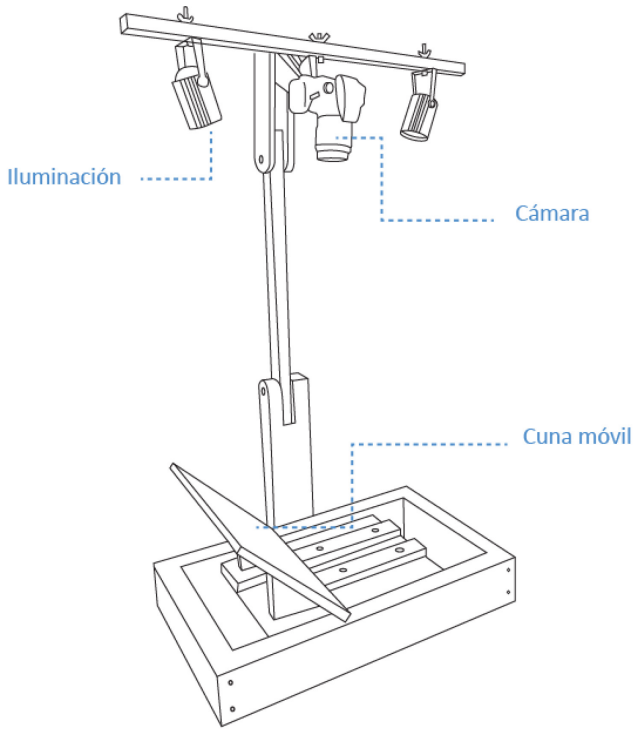
Escáneres y software descartados



- Los softwares utilizados para la captura de imágenes y posprocesamiento debieron adaptarse y transformarse dejando de lado el firmware modificado (**CHDK**) y el **Scan Tailor.**
- El **Scan Tailor** no tenía funciones avanzadas de ajuste de colores y nitidez.
- Firmware modificado CHDK: requería la configuración *in situ* de cada cámara por separado y el uso de una tarjeta de memoria para mover las imágenes a la computadora.



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.

Software de captura

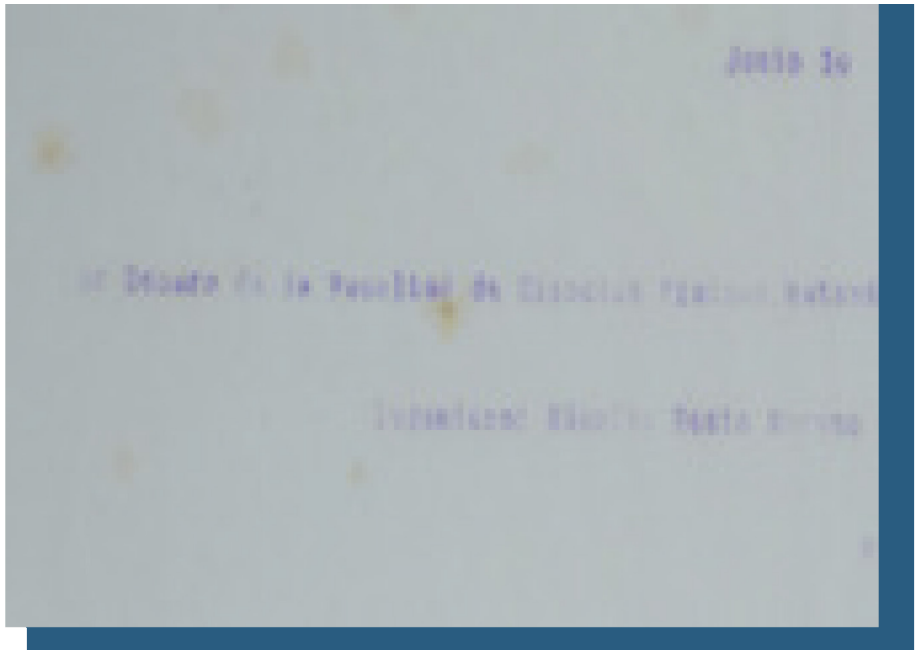
El **digiCamControl 2.0.72.0** fue un programa rápido, confiable y versátil para la captura de imágenes y permitió el manejo de las cámaras directamente desde la computadora.

Software de edición

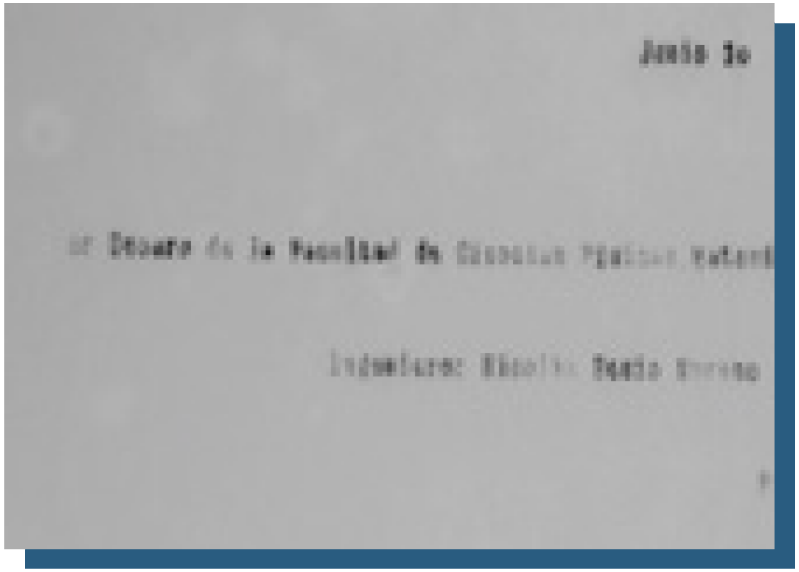
Photoshop CS6 permitió mayor libertad en la manipulación de las imágenes destinadas al reconocimiento de texto. Se aplicaron filtros y se automatizó el procedimiento estándar para todas las imágenes

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

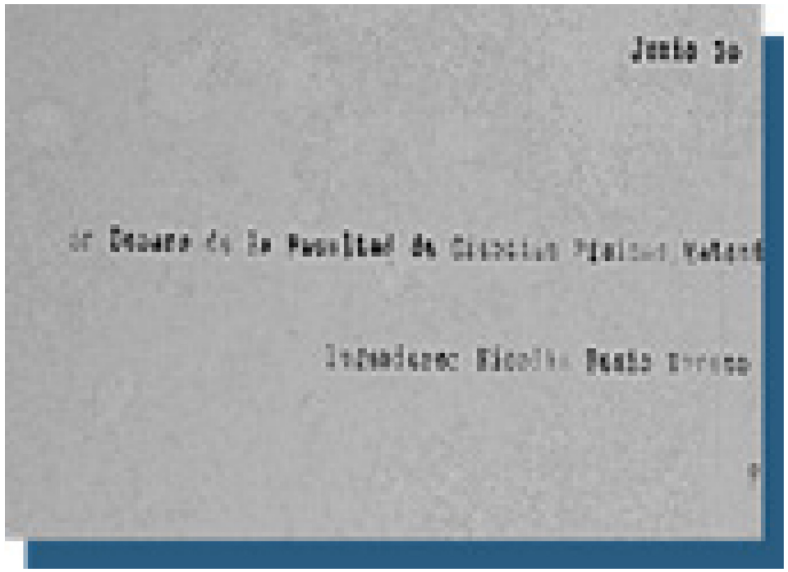
Se utilizaron dos filtros para mejorar la imagen capturada con el fin de hacer el OCR del documento



- **Desaturación por color (black and white filter):** este filtro desatura los colores por separado. Esto permite seleccionar las tonalidades que representan manchas, suciedades y atenuarlos hasta que la superficie se vea homogénea.

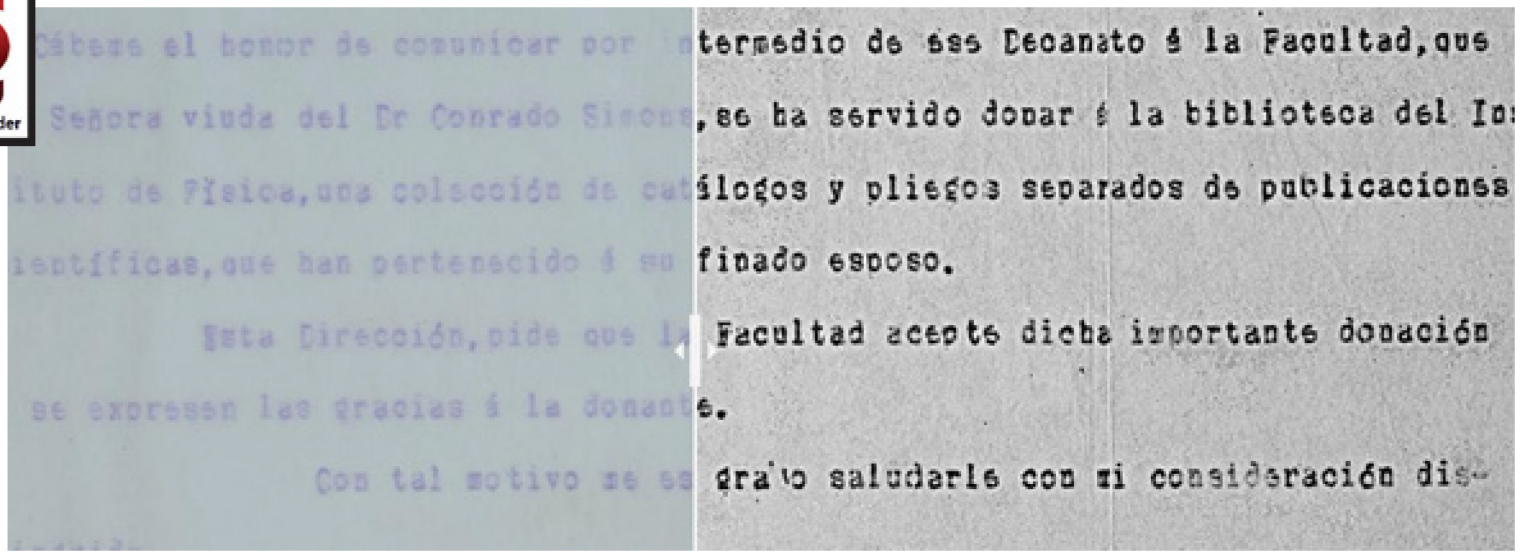


- **Enfocar (smart sharpen)** para acentuar el borde de la tipografía en la imágen y mejorar el contraste con el fondo.



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)



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

Nuestros sitios

http://sedici.unlp.edu.ar

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

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

http://prebi.unlp.edu.ar

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

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

http://revistas.unlp.edu.ar

http://congresos.unlp.edu.ar

http://ibros.unlp.edu.ar